Deep learning methods for designing proteins scaffolding functional sites

Jue Wang[a,b,†], Sidney Lisanza[a,b,c,†], David Juergens[a,b,g,†], Doug Tischer[a,b,†], Ivan Anishchenko[a,b], Minkyung Baek[a,b], Joseph L. Watson[a,b], Jung Ho Chun[a,b,c], Lukas F. Milles[a,b], Justas Dauparas[a,b], Marc Expòsit[a,b,g], Wei Yang[a,b], Amijai Saragovi[a,b], Sergey Ovchinnikov[d,e,*], David Baker[a,b,f,*]

[a] Department of Biochemistry, University of Washington, Seattle, WA 98105, USA
[b] Institute for Protein Design, University of Washington, Seattle, WA 98105, USA
[c] Graduate program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA
[d] FAS Division of Science, Harvard University, Cambridge, MA 02138, USA
[e] John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138, USA
[f] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA
[g] Graduate program in Molecular Engineering, University of Washington, Seattle, WA 98105, USA
[†]These authors contributed equally to this work.

* To whom correspondence should be addressed. Email: dabaker@uw.edu, so@fas.harvard.edu

Abstract

Current approaches to *de novo* design of proteins harboring a desired binding or catalytic motif require pre-specification of an overall fold or secondary structure composition, and hence considerable trial and error can be required to identify protein structures capable of scaffolding an arbitrary functional site. Here we describe two complementary approaches to the general functional site design problem that employ the RosettaFold and AlphaFold neural networks which map input sequences to predicted structures. In the first “constrained hallucination” approach, we carry out gradient descent in sequence space to optimize a loss function which simultaneously rewards recapitulation of the desired functional site and the ideality of the surrounding scaffold, supplemented with problem-specific interaction terms, to design candidate immunogens presenting epitopes recognized by neutralizing antibodies, receptor traps for escape-resistant viral inhibition, metalloproteins and enzymes, and target binding proteins with designed interfaces expanding around known binding motifs. In the second “missing information recovery” approach, we start from the desired functional site and jointly fill in the missing sequence and structure information needed to complete the protein in a single forward pass through an updated RoseTTAFold trained to recover sequence from structure in addition to structure from sequence. We show that the two approaches have considerable synergy, and

AlphaFold2 structure prediction calculations suggest that the approaches can accurately generate proteins containing a very wide array of functional sites.

Main text

The biochemical functions of proteins are generally carried out by a small number of residues in a protein which 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 which produce the desired activity--this can be done using quantum chemistry calculations in the enzyme case (to identify ideal theozymes for catalyzing a desired reaction) (*1*–*3*) or fragment docking calculations in the protein binder case (*4*, *5*); alternatively functional sites can be extracted from native protein having the desired activity (*6*, *7*). In this paper, we focus on the second step: given a functional site description from any source, design an amino acid sequence which folds up to a three dimensional structure containing the site. Methods have been developed for functional site scaffolding for 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. Current methods also have the limitations that assumptions must be made about the secondary structure of the scaffold, and that the amino acid sequence must be generated in a subsequent sequence step, so there is no guarantee that the generated backbones are in fact designable (encodable by some amino acid sequence).

An ideal method for functional de novo protein design would 1) embed the functional site with minimal distortion in a designable scaffold protein; 2) 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 3) jointly generate backbone structure and amino acid sequence. We reasoned that the trRosetta neural network (*11*), which maps input

sequences to predicted structures, could be adapted for this purpose. Completely new proteins can be designed using trRosetta by starting from a random amino acid sequence, and carrying out Monte Carlo sampling in sequence space maximizing the probability that the sequence folds to some (unspecified) three dimensional structure (*12*). We refer to this process as "hallucination" as it produces solutions that the network considers ideal proteins but do not correspond to any actual natural protein (Fig. 1A); crystal and NMR 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*). We sought to extend this approach to scaffold functional sites using trRosetta by sampling in sequence space with a combination of the hallucination loss to favor folding to a unique structure, and a structure recapitulation loss to favor formation of the desired functional site (rather than the entire structure as in (*13*); Fig. 1B; Methods). While we succeeded in generating structures that had segments which closely recapitulated functional sites, Rosetta structure predictions suggested that the sequences poorly encoded the structures, and hence we used Rosetta design calculations to generate more optimal sequences (*14*). Several designs targeting PD-L1 generated by constrained hallucination with binding motifs derived from PD-1, followed by Rosetta design, were found to have binding affinities in the mid-nanomolar range (Fig. S1). While this experimental validation is encouraging, the requirement for sequence design using Rosetta is at odds with property (3) above-the joint design of sequence and structure.

We found following the development of RosettaFold (*15*) that using it, rather than trRosetta, to guide motif-constrained hallucination resulted in designed protein sequences that more strongly encoded their structures (Fig. S2), likely reflecting the better overall modeling of protein sequence-structure relationships evidenced by the superior structure prediction performance

(*15*). Constrained hallucination with RosettaFold has the further advantages that since 3D coordinates are explicitly modeled (trRosetta only generates residue-residue distances and orientations), motif 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 protein target (Fig. 1B, 1D).

In the following sections, we explore the use of the constrained RosettaFold hallucination method to design proteins containing a wide range of functionally diverse motifs (Fig. 2-4, Table S1). It is impractical to experimentally validate many designs for many different applications; we instead evaluate these designs using the AlphaFold (AF) protein structure prediction network (*16*) which has very high accuracy on *de novo* designed proteins (*17*). Although RoseTTAFold was inspired by AF, the two models were developed and trained independently, and hence AF predictions can be regarded as an orthogonal *in silico* test of whether RF designed sequences fold into the intended structures, analogous to traditional *ab initio* folding benchmarks (*13*, *18*). For almost all problems, we obtained designs that are closely recapitulated by AF with overall and motif RMSD typically <2 Å and <1 Å respectively with model confidence pLDDT > 80 (Table S2). While solving current challenges with protein design clearly requires making and characterizing proteins in the lab, this *in silico* AF test is well suited for testing performance of design methods on a wide range of problems, and is quite stringent, as discussed below.

Hallucinating immunogen candidates and receptor traps

We first applied the constrained hallucination method to the problem of antigen presentation for immunogen design, where the goal is to scaffold a native epitope recognized by a neutralizing antibody as accurately as possible (and thus elicit antibodies binding the target protein upon immunization). Additional interactions with the target antibody are undesirable because the goal is to elicit antibodies recognizing the original antigen, and hence we incorporate an additional repulsive term assessed on the complex 3D coordinates in the composite loss function to

penalize interactions with the antibody beyond those present in the epitope being scaffolded (Fig. 1D, S3). As a test case, we focused on respiratory syncytial virus, a leading cause of infant mortality whose F protein (RSV-F) contains antigenic epitopes for which structures with neutralizing antibodies have been determined (*7*, *9*, *10*). We sought to scaffold RSV-F site II, a contiguous helix-turn-helix motif that had previously been grafted successfully onto a 3-helix bundle architecture (*7*), as well as RSV-F site V, a helix-turn-strand motif that has not yet been scaffolded successfully (*19*). We were able to hallucinate designs for both epitopes with a variety of folds and motifs recapitulated to sub-angstrom Cα RMSD in the AF predicted structure of the designed sequence (Fig. 2A, Fig. S8, S11; for these and all designs below, full amino acid sequence and PDB files are in the SM, and comparisons of the design models to AF predictions, in Fig. S8-10--since they are virtually identical, to save space we show only one of these in the main text figures).

We next applied the hallucination method to the design of receptor traps, which neutralize viruses by mimicking their natural binding targets and thus are inherently robust against mutational escape. We again augmented the loss function with an explicit penalty on interactions beyond those present in the receptor to avoid opportunities for viral escape. As a test case, we scaffolded the interfacial helix of human angiotensin-converting enzyme 2 (hACE2) interacting with the receptor-binding domain (RBD) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein (*20*). The hallucinated hACE2 mimetics have a diverse set of helical topologies, and AF2 structure predictions recapitulate the binding interface with sub Å accuracy (Fig. 2B, S8, S10).

Hallucinating metal binding and enzyme active sites

We next explored the scaffolding of functional sites involved in metal-binding and catalysis. We designed scaffolds around a di-iron binding site, which is important in biological systems for iron storage (*21*) and also potentially harnessable for catalysis (*22*, *23*). The motif, composed of four

roughly parallel helical segments from *E. coli* bacterioferritin (cytochrome b1), was recapitulated with sub-angstrom RMSDs (Fig. 3A), in scaffolds with quite different helix connectivities than the parent (Fig. S9). For the calcium-binding EF-hand motif (*24*) composed of a 12 residue loop flanked by helices, the hallucination method readily generates a variety of scaffolds recapitulating either 1 or 2 EF-hand motifs within 0.5 Å RMSD of the calcium binding motif (Fig. 3C). When tasked with scaffolding one EF-hand motif, the method chooses to buttress the loop with a helix, avoiding the need for another long loop.

We next sought to hallucinate enzyme active sites. Carbonic anhydrase II, which catalyzes the interconversion of carbon dioxide and bicarbonate, enables CO_2 transport in humans (*25*), plays a key role in photosynthesis (*26*), and is emerging as a tool for CO_2 sequestration (*27*). The active site contains 3 Zn^{2+} coordinating histidines (PDB ID 5yui: His94,His96,His119) on two strands, and a hydrophobic loop containing Thr199 which sequesters and orients the CO_2. Despite the complexity of the irregular, discontinuous, 3 segment site, the method generated designs with sub angstrom motif RMSDs with correct His placement for Zn^{2+} coordination (Fig. 3E, S9); these are less than 100 residues, significantly smaller than the 261 residue long native protein.

To enable specification of sidechain geometry, we carried out iterative gradient descent using gradient information obtained by backpropagation through the AF neural network rather than RF, which currently does not explicitly model side chains (see Methods). As a test, we used the catalytic sidechain geometry of Δ^5-3-ketosteroid isomerase (1QJG: residues 14, 38, 99), which catalyzes the isomerization of Δ^5- to Δ^4-3-ketosteroid needed for synthesis of steroid hormones in mammals (*28*). In initial experiments, we were only able to obtain designs that fully recapitulated the catalytic sidechain geometry when optimization was over a multiple sequence alignment rather than a single sequence; the landscape may be too rugged with the high resolution sidechain-based loss in the single sequence case. To overcome this problem, we

developed a two-stage approach; with a first stage using both AF and trRosetta (to reduce the structure-prediction resolution and thus 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 two-stage approach led to multiple plausible solutions with predicted structures having a nearly exact match to the catalytic sidechain geometry (Fig. 3G, S9); however, we cannot use AF as an independent test of design accuracy in this case (given the very large number of model parameters, direct optimization against the output of a neural network has the potential to identify false optima, and hence independent *in silico* validation is important).

Hallucinating protein-protein interfaces

We next sought to design binding proteins which extend beyond an input binding motif to make additional favorable interactions with the target by explicitly including the sequence and structure of the target in the hallucination process (Figs S6, Methods). We designed binders of the anti-inflammatory cytokine interleukin 10 (IL-10) α-receptor that incorporate one of the two discontinuous binding sites in the domain-swapped IL10 dimer in a single chain; the resulting scaffolds recapitulate the IL10 binding region within 0.5A (Fig. 4A, S10). Starting from the complement cascade protein C3d which enhances immune responses to covalently attached antigens (*29*) we designed binders to complement receptor 2 (CR2) present on B-cell and dendritic cells (*30*). The designs are much smaller (<100 AAs) than native C3d (306 AAs), recapitulate the binding interface with sub angstrom accuracy (Fig. 4B, S6C).

As a test of building around beta strand motifs, we sought to design binders of the immune checkpoint protein CTLA-4 starting from B7-2, which binds CTLA-4 through four beta strands. Starting from a single five residue strand, hallucination in the presence of CTLA-4 generated designs having both alpha-beta and all beta topologies with novel binding modes and comparable interface contacts to native B7-2 (Fig. 4C, S10). As expected, designs hallucinated

in the presence of the target had considerably better Rosetta protein-protein interface metrics (*4*) (binding free energy, etc) than those designed without the receptor (Fig. S6).

Generalized protein function design by missing information recovery using RoseTTAFold

While quite 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 in sequence optimization. In the original training of RosettaFold for structure prediction a small fraction (15%) of tokens in the MSA are masked, and the network learns to recover this missing sequence information in addition to predicting structure. We reasoned that this ability to recover sequence information along with structural information could provide a second solution to the functional site scaffolding problem: given a functional site description, a forward pass through the network could potentially be used to complete, or “inpaint”, both protein sequence and structure (Fig. 1C; 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 (*31*) and completion of corrupted images using inpainting methods (*32*). As illustrated in Fig. 1E, a wide variety of protein structure prediction and design challenges can be similarly formulated as missing information recovery problems. We began from a RoseTTAFold model trained for structure prediction (*15*) and carried out further training on both fixed-backbone sequence design and fixed-sequence structure prediction tasks (Methods; Fig. S13; Algorithm S1). After training, the mean amino acid sequence recovery of the resulting model, denoted RF_{joint}, on a *de novo* protein test set was 33% (Fig. 5A; this is similar to Rosetta fixed backbone design performance), and there was also a slight increase in structure prediction accuracy (Fig. 5B). Thus, the model can both recover missing structure information given sequence and missing sequence information given structure.

We next considered design challenges where both sequence and structure information were missing for a portion of the protein. For smaller masked regions, the sequences and structures

recovered by RF_{joint} are close to those of the input native structure, and as the size of the masked regions increases the divergence of both sequence and structure increases as expected (Fig. S14). The extent of variation in the resulting designs can be controlled by the amount of input sequence and structure information provided (Fig. S18C). Since the calculations require a single forward pass (including recycling outputs back as input) through the network, only 1-10 seconds on an NVIDIA RTX2080 GPU (Methods) are required to generate both sequence and structure.

Encouraged by the excellent performance of RF_{joint} on simultaneous sequence and structure recovery despite being only trained on recovery of one or the other, we sought to improve this further by explicitly training on joint sequence/structure recovery tasks. Sequence and structure diversity is useful when designing proteins containing functional motifs, as subtle variations in the structure of the motif can drastically affect function (*33*), and hence we trained this new model to predict the sequence and structure of masked regions between two provided residue coordinates, in the absence of structural and sequence information of the residues flanking the two residue coordinates (to force the model to place structural elements based more on larger protein context than the local structure of the immediately connected chain segments). With this second model, which we call RF_{joint2}, the two residue coordinates can, at inference time, be varied, enabling the rapid generation of further sequence/structure diversity (Fig. 5D; a similar problem has been explored using Rosetta (*33*)). Of note, the degree of diversification in the inpainted region can be controlled by varying the distance by which the two residue coordinates are translated (Fig. 5D, left panel), while the structure of the templated (unmasked) protein remains remarkably stable.

We next explored the use of RF_{joint} and RF_{joint2} to generate complete protein structures around the functional sites described in Figs 2-4, and found that success depended on the size and context of the input functional regi

on. With the RF_{joint} model, we found that best results were obtained for the more minimalist functional sites by first building up extended versions using the constrained hallucination approach. Many alternative structure and sequence completions can then be generated by RF_{joint} in a network forward pass (Figure 6A, Figure S18). Almost all designs shown have sub-angstrom RMSD from the AF prediction to the native motif and <2 Å RMSD between design model and AF prediction (Fig. 6A, Fig. S19), and > 80 pLDDT. Diverse ensembles of such solutions to a specific design challenge can be very rapidly generated by varying the input sequence and structure information (Fig. S18). While RF_{joint} struggled to generate well-predicted proteins from native/minimalist motifs, we found that RF_{joint2} was able to generate complete and confidently-predicted (by AF2) protein models from smaller regions, such as a single EF hand motif (Fig. S18B). Further, RF_{joint2} could simultaneously scaffold two motifs while retaining good (<1 Å RMSD) alignment to both (Fig. 6B, top row). Remarkably, in some cases, RF_{joint2} was able to generate well-predicted scaffolds to complex, multi-chain motifs taken directly from a native crystal structure (Fig. 6B, middle and bottom row), as well as translationally symmetric proteins (Fig. S20), provided little more than the desired motif, in a single forward pass through the network.

Tests on the full range of challenges described here suggest that the two function design approaches are complementary: the constrained hallucination approach can build protein structures harboring minimalist functional sites but is quite compute and memory intensive since it requires a forward and backward pass (to generate gradient information to guide sequence optimization) through the neural network at each step of sequence optimization (Methods), while the missing information recovery method in most but not all cases requires extended functional site description but is much less compute intensive, and generally outperforms the hallucination method when more starting information is provided, as illustrated by the lower RMSDs on constrained regions (Fig. S15). This difference in performance can be understood by

considering the manifold in sequence-structure space corresponding to folded proteins; the space of all possible sequence-structure pairs is far larger than the set of sequence-structure pairs of folded proteins, and hence this manifold occupies a tiny fraction of the overall space. The missing information recovery approach can be viewed as projecting an incomplete or corrupted input sequence-structure pair onto the subset of this manifold (as represented by RosettaFold) containing the functional site--if insufficient starting information is provided, this projection is not necessarily well determined, but with sufficient information, it readily 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 as protein-like solutions as the inpainting approach-- on the other hand, since stochastic search can be initiated from any starting point, the hallucination approach can start from minimalist functional site descriptions, or, as in the fully unconstrained case (*12*), no sequence and structural information at all.

Evaluation of designs using AF2

New protein design methods have traditionally been evaluated by experimental testing, and for actual applications it is essential to make and characterize proteins in the lab. The high structure prediction accuracy of AF2 now enables evaluation of new design methodology in silico, which has the considerable advantage that a much wider variety of design challenges can be evaluated. In the work described here, AF2 was not used for any of the design calculations except for the sidechain active site design case of Fig. 3E, and hence provides an independent test of design accuracy. Both the backbone design challenge--generating a plausible protein backbone with a geometry capable of hosting a desired site, and the sequence design challenge--generating a sequence which strongly encodes this backbone, are quite formidable. For the backbone design problem, the very large set of structures predicted for naturally

occurring proteins using AF and recently made available (*34*) provides an excellent point of comparison: for the RSV-F site V immunogen design challenge described above, the frequency of non-homologous proteins in the AF proteomes database and the Protein Data Bank (PDB) (*35*) matching the functional site with equal or lower RMSDs than our designs was 3.9×10^{-6} (Fig. S17; Supplementary Text); similarly low frequencies of suitable natural scaffolds in the PDB were observed for other targets (Table S3). For the sequence design problem, the accuracy of native protein structure prediction based on single amino acid sequences provides a point of comparison; as shown in Fig. S16, our designs are predicted more confidently from sequence than the vast majority of native proteins with known crystal structures, and on par with structurally validated de novo designed proteins. This success in designing sequences confidently predicted to fold to structures harboring a wide range of functional sites derives in part from a key advance over 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.

Conclusions

The deep learning methods presented here are quite general, requiring no inputs other than the structure and sequence of the desired functional site, and unlike current non-deep-learning methods, do not require specification of the secondary structure or topology of the scaffold, and simultaneously generate both sequence and structure. Despite a recent surge of interest in using machine learning to design protein sequences (*36–43*), the design of protein structure is relatively underexplored, likely due to the difficulty of efficiently representing and learning structure (*44*). Generative adversarial networks (GANs) and variational autoencoders (VAEs) trained on specific fold families have been used to design biophysically plausible protein backbones (*45*, *46*), but not ones containing functional sites. RoseTTAFold and Alphafold have

been trained on the entire PDB, and thus generalize from a very wide range of known protein structures. Our “activation maximization” hallucination approach enables use of arbitrary loss functions tailored to specific problems without retraining for any sequence length. Complementary to this, the ability of our “missing information recovery” 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 combination of the two approaches is more powerful than either one alone, as ensembles of solutions to a given functional design problem can be generated very rapidly using the second approach starting from extended site descriptions identified in the first. The hallucination approach could, in theory, also be used to refine the more extensive designs generated by inpainting. The two approaches individually, and the combination of the two, should increase in power as more and more accurate protein structure, interface, and small molecule binding prediction networks are developed moving forward.

References

1. O. Khersonsky, A. M. Wollacott, L. Jiang, J. Dechancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, D. Baker, Kemp elimination catalysts by computational enzyme design. **453** (2008), doi:10.1038/nature06879.

2. L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard, D. Baker, De Novo Computational Design of Retro-Aldol Enzymes. *Science*. **319**, 1387–1391 (2008).
3. J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St. Clair, J. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael, D. Baker, Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science*. **329** (2010), doi:10.1126/science.1190239.
4. L. Cao, B. Coventry, I. Goreshnik, B. Huang, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, S. Halabiya, B. Hammerson, W. Yang, S. Benard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, “Robust de novo design of protein binding proteins from target structural information alone” (2021), p. 2021.09.04.459002, , doi:10.1101/2021.09.04.459002.
5. A. A. Chevalier, D. Silva, G. J. Rocklin, R. Derrick, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, G. Yao, C. D. Bahl, S. Miyashita, I. Goreshnik, T. James, M. Bryan, D. A.

Fernández-velasco, L. Stewart, M. Dong, X. Huang, Massively parallel de novo protein design for targeted therapeutics. *Nat. Publ. Group* (2017), doi:10.1038/nature23912.
6. E. Procko, G. Y. Berguig, B. W. Shen, Y. Song, S. Frayo, A. J. Convertine, D. Margineantu, G. Booth, B. E. Correia, Y. Cheng, W. R. Schief, D. M. Hockenbery, O. W. Press, B. L. Stoddard, P. S. Stayton, D. Baker, A Computationally Designed Inhibitor of an Epstein-Barr Viral Bcl-2 Protein Induces Apoptosis in Infected Cells. *Cell*. **157**, 1644–1656 (2014).
7. B. E. Correia, J. T. Bates, R. J. Loomis, G. Baneyx, C. Carrico, J. G. Jardine, P. Rupert, C. Correnti, O. Kalyuzhniy, V. Vittal, M. J. Connell, E. Stevens, A. Schroeter, M. Chen, S. MacPherson, A. M. Serra, Y. Adachi, M. A. Holmes, Y. Li, R. E. Klevit, B. S. Graham, R. T. Wyatt, D. Baker, R. K. Strong, J. E. Crowe, P. R. Johnson, W. R. Schief, Proof of principle for epitope-focused vaccine design. *Nature*. **507**, 201–206 (2014).
8. D.-A. Silva, S. Yu, U. Y. Ulge, J. B. Spangler, K. M. Jude, C. Labão-Almeida, L. R. Ali, A. Quijano-Rubio, M. Ruterbusch, I. Leung, T. Biary, S. J. Crowley, E. Marcos, C. D. Walkey, B. D. Weitzner, F. Pardo-Avila, J. Castellanos, L. Carter, L. Stewart, S. R. Riddell, M. Pepper, G. J. L. Bernardes, M. Dougan, K. C. Garcia, D. Baker, De novo design of potent and selective mimics of IL-2 and IL-15. *Nature*. **565**, 186–191 (2019).
9. F. Sesterhenn, C. Yang, J. Bonet, J. T. Cramer, X. Wen, Y. Wang, C.-I. Chiang, L. A. Abriata, I. Kucharska, G. Castoro, S. S. Vollers, M. Galloux, E. Dheilly, S. Rosset, P. Corthésy, S. Georgeon, M. Villard, C.-A. Richard, D. Descamps, T. Delgado, E. Oricchio, M.-A. Rameix-Welti, V. Más, S. Ervin, J.-F. Eléouët, S. Riffault, J. T. Bates, J.-P. Julien, Y. Li, T. Jardetzky, T. Krey, B. E. Correia, De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science*. **368** (2020), doi:10.1126/science.aay5051.
10. C. Yang, F. Sesterhenn, J. Bonet, E. A. van Aalen, L. Scheller, L. A. Abriata, J. T. Cramer, X. Wen, S. Rosset, S. Georgeon, T. Jardetzky, T. Krey, M. Fussenegger, M. Merkx, B. E. Correia, Bottom-up de novo design of functional proteins with complex structural features. *Nat. Chem. Biol.*, 1–9 (2021).
11. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* (2020), doi:10.1073/pnas.1914677117.
12. I. Anishchenko, T. M. Chidyausiku, S. Ovchinnikov, S. J. Pellock, D. Baker, *bioRxiv*, in press, doi:10.1101/2020.07.22.211482.
13. C. Norn, B. I. M. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, B. Koepnick, I. Anishchenko, F. Players, D. Baker, S. Ovchinnikov, Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci.* **118** (2021), doi:10.1073/pnas.2017228118.
14. D. Tischer, S. Lisanza, J. Wang, R. Dong, I. Anishchenko, L. F. Milles, S. Ovchinnikov, D. Baker, *bioRxiv*, in press, doi:10.1101/2020.11.29.402743.
15. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* (2021), doi:10.1126/science.abj8754.
16. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli,

D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021).
17. R. Chowdhury, N. Bouatta, S. Biswas, C. Rochereau, G. M. Church, P. K. Sorger, M. AlQuraishi, Single-sequence protein structure prediction using language models from deep learning, 22.
18. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins Struct. Funct. Bioinforma.* **37**, 171–176 (1999).
19. J. J. Mousa, N. Kose, P. Matta, P. Gilchuk, J. E. Crowe, A novel pre-fusion conformation-specific neutralizing epitope on the respiratory syncytial virus fusion protein. *Nat. Microbiol.* **2**, 1–8 (2017).
20. T. W. Linsky, R. Vergara, N. Codina, J. W. Nelson, M. J. Walker, W. Su, C. O. Barnes, T.-Y. Hsiang, K. Esser-Nobis, K. Yu, Z. B. Reneer, Y. J. Hou, T. Priya, M. Mitsumoto, A. Pong, U. Y. Lau, M. L. Mason, J. Chen, A. Chen, T. Berrocal, H. Peng, N. S. Clairmont, J. Castellanos, Y.-R. Lin, A. Josephson-Day, R. S. Baric, D. H. Fuller, C. D. Walkey, T. M. Ross, R. Swanson, P. J. Bjorkman, M. Gale, L. M. Blancas-Mejia, H.-L. Yen, D.-A. Silva, De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science* (2020), doi:10.1126/science.abe0075.
21. F. Frolow, A. J. Kalb (Gilboa), J. Yariv, Structure of a unique twofold symmetric haem-binding site. *Nat. Struct. Biol.* **1**, 453–460 (1994).
22. A. Lombardi, F. Pirro, O. Maglio, M. Chino, W. F. DeGrado, De Novo Design of Four-Helix Bundle Metalloproteins: One Scaffold, Diverse Reactivities. *Acc. Chem. Res.* **52**, 1148–1159 (2019).
23. J. R. Calhoun, F. Nastri, O. Maglio, V. Pavone, A. Lombardi, W. F. DeGrado, Artificial diiron proteins: From structure to function. *Pept. Sci.* **80**, 264–278 (2005).
24. M. Yáñez, J. Gil-Longo, M. Campos-Toimil, in *Calcium Signaling*, Md. S. Islam, Ed. (Springer Netherlands, Dordrecht, 2012; https://doi.org/10.1007/978-94-007-2888-2_19), *Advances in Experimental Medicine and Biology*, pp. 461–482.
25. C. U. Kim, H. Song, B. S. Avvaru, S. M. Gruner, S. Park, R. McKenna, Tracking solvent and protein movement during CO2 release in carbonic anhydrase II crystals. *Proc. Natl. Acad. Sci.* **113**, 5257–5262 (2016).
26. M. R. Badger, G. D. Price, The Role of Carbonic Anhydrase in Photosynthesis. *Annu. Rev. Plant Physiol.* **45**, 369–92 (1994).
27. P. Mirjafari, K. Asghari, N. Mahinpey, Investigating the Application of Enzyme Carbonic Anhydrase for CO2 Sequestration Purposes. *Ind. Eng. Chem. Res.* **46**, 921–926 (2007).
28. H.-S. Cho, N.-C. Ha, G. Choi, H.-J. Kim, D. Lee, K. S. Oh, K. S. Kim, W. Lee, K. Y. Choi, B.-H. Oh, 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).
29. P. W. Dempsey, M. E. D. Allison, S. Akkaraju, C. C. Goodnow, D. T. Fearon, C3d of Complement as a Molecular Adjuvant: Bridging Innate and Acquired Immunity. *Science*. **271**, 348–350 (1996).
30. T. M. Ross, Y. Xu, R. A. Bright, H. L. Robinson, C3d enhancement of antibodies to hemagglutinin accelerates protection against influenza virus challenge. *Nat. Immunol.* **1**, 127–131 (2000).
31. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs* (2019) (available at http://arxiv.org/abs/1810.04805).
32. R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, M. N. Do, Semantic Image Inpainting with Deep Generative Models. *ArXiv160707539 Cs* (2017) (available at http://arxiv.org/abs/1607.07539).
33. N. Ollikainen, T. Kortemme, Computational Protein Design Quantifies Structural

Constraints on Amino Acid Covariation. *PLOS Comput. Biol.* **9**, e1003313 (2013).
34. K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, Highly accurate protein structure prediction for the human proteome. *Nature* (2021), doi:10.1038/s41586-021-03828-1.
35. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
36. J. Ingraham, V. K. Garg, R. Barzilay, T. Jaakkola, Generative models for graph-based protein design, 10 (2019).
37. A. Strokach, D. Becerra, C. Corbi-Verge, A. Perez-Riba, P. M. Kim, Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Syst.* **11**, 402-411.e4 (2020).
38. S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, Low- N protein engineering with data-efficient deep learning. *Nat. Methods*. **18**, 389–396 (2021).
39. D. Repecka, V. Jauniskis, L. Karpus, E. Rembeza, J. Zrimec, S. Poviloniene, I. Rokaitis, A. Laurynenas, W. Abuajwa, O. Savolainen, R. Meskys, M. K. M. Engqvist, A. Zelezniak, Expanding functional protein sequence space using generative adversarial networks. *bioRxiv*, 789719 (2019).
40. J.-E. Shin, A. J. Riesselman, A. W. Kollasch, C. McMahon, E. Simon, C. Sander, A. Manglik, A. C. Kruse, D. S. Marks, Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 1–11 (2021).
41. Z. Wu, K. E. Johnston, F. H. Arnold, K. K. Yang, Protein sequence design with deep generative models. *Curr. Opin. Chem. Biol.* **65**, 18–27 (2021).
42. N. Anand-Achim, R. R. Eguchi, A. Derry, R. B. Altman, P.-S. Huang, “Protein sequence design with a learned potential” (preprint, Bioinformatics, 2020), , doi:10.1101/2020.01.06.895466.
43. A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, N. Naik, *bioRxiv*, in press, doi:10.1101/2021.07.18.452833.
44. S. Ovchinnikov, P.-S. Huang, Structure-based protein design with deep learning. *Curr. Opin. Chem. Biol.* **65**, 136–144 (2021).
45. N. Anand, R. Eguchi, P.-S. Huang, Fully differentiable full-atom protein backbone generation (2019) (available at https://openreview.net/forum?id=SJxnVL8YOV).
46. R. R. Eguchi, N. Anand, C. A. Choe, P.-S. Huang, *bioRxiv*, in press, doi:10.1101/2020.08.07.242347.
47. E. Jang, S. Gu, B. Poole, Categorical Reparameterization with Gumbel-Softmax. *ArXiv161101144 Cs Stat* (2017) (available at http://arxiv.org/abs/1611.01144).
48. N. Bogard, J. Linder, A. B. Rosenberg, G. Seelig, A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell*. **178**, 91-106.e23 (2019).
49. J. Linder, G. Seelig, Fast differentiable DNA and protein sequence optimization for molecular design. *ArXiv200511275 Cs Stat* (2020) (available at http://arxiv.org/abs/2005.11275).
50. D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2017) (available at http://arxiv.org/abs/1412.6980).
51. M. Jendrusch, J. O. Korbel, S. K. Sadiq, *bioRxiv*, in press, doi:10.1101/2021.10.11.463937.
52. L. Moffat, J. G. Greener, D. T. Jones, *bioRxiv*, in press, doi:10.1101/2021.08.24.457549.
53. S. K. Jha, A. Ramanathan, R. Ewetz, A. Velasquez, S. Jha, Protein Folding Neural Networks Are Not Robust. *ArXiv210904460 Cs Q-Bio* (2021) (available at http://arxiv.org/abs/2109.04460).

54. A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, A. Madry, Adversarial Examples Are Not Bugs, They Are Features. *ArXiv190502175 Cs Stat* (2019) (available at http://arxiv.org/abs/1905.02175).
55. R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, *bioRxiv*, in press, doi:10.1101/2021.02.12.430858.
56. A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, J. Carreira, Perceiver: General Perception with Iterative Attention. *ArXiv210303206 Cs Eess* (2021) (available at http://arxiv.org/abs/2103.03206).
57. W. Kabsch, A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*. **32**, 922–923 (1976).
58. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning. *Nature*, 1–5 (2020).
59. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J. Gray, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
60. S. E. Boyken, Z. Chen, B. Groves, R. A. Langan, G. Oberdorfer, A. Ford, J. M. Gilmore, C. Xu, F. DiMaio, J. H. Pereira, B. Sankaran, G. Seelig, P. H. Zwart, D. Baker, De novo design of protein homo-oligomers with modular hydrogen-bond network--mediated specificity. *Science*. **352**, 680–687 (2016).
61. N. Hiranuma, H. Park, M. Baek, I. Anishchenko, J. Dauparas, D. Baker, Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* **12**, 1340 (2021).
62. D.-A. Silva, B. E. Correia, E. Procko, in *Computational Design of Ligand Binding Proteins*, B. L. Stoddard, Ed. (Springer, New York, NY, 2016; https://doi.org/10.1007/978-1-4939-3569-7_17), *Methods in Molecular Biology*, pp. 285–304.
63. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
64. T. Brunette, F. Parmeggiani, P.-S. Huang, G. Bhabha, D. C. Ekiert, S. E. Tsutakawa, G. L. Hura, J. A. Tainer, D. Baker, Exploring the repeat protein universe through computational protein design. *Nature*. **528**, 580–584 (2015).
65. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinforma.* **65**, 712–725 (2006).
66. H. Park, P. Bradley, P. Greisen, Y. Liu, V. K. Mulligan, D. E. Kim, D. Baker, F. DiMaio, Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
67. R. Pascolutti, X. Sun, J. Kao, R. L. Maute, A. M. Ring, G. R. Bowman, A. C. Kruse, Structure and Dynamics of PD-L1 and an Ultra-High-Affinity PD-1 Receptor Mutant. *Structure*. **24**, 1719–1728 (2016).
68. J. S. McLellan, M. Chen, A. Kim, Y. Yang, B. S. Graham, P. D. Kwong, Structural basis of respiratory syncytial virus neutralization by motavizumab. *Nat. Struct. Mol. Biol.* **17**, 248–250 (2010).
69. J. Shang, G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara, Q. Geng, A. Auerbach, F. Li, Structural basis of receptor recognition by SARS-CoV-2. *Nature*. **581**, 221–224 (2020).
70. J. L. Fallon, F. A. Quiocho, A Closed Compact Structure of Native Ca2+-Calmodulin. *Structure*. **11**, 1303–1307 (2003).
71. G. Szakonyi, J. M. Guthridge, D. Li, K. Young, V. M. Holers, X. S. Chen, Structure of

complement receptor 2 in complex with its C3d ligand. *Science*. **292**, 1725–1728 (2001).
72. J.-C. D. Schwartz, X. Zhang, A. A. Fedorov, S. G. Nathenson, S. C. Almo, Structural basis for co-stimulation by the human CTLA-4/B7-2 complex. *Nature*. **410**, 604–608 (2001).

Acknowledgements

We would like to thank Luki Goldschmidt for maintaining the computational resource in the IPD; Christoffer Norn for general discussions about trRosetta; Brian Coventry and Nathaniel Bennett for advice on interface design; Bruno Correia, Casper Goverde, and Karla Castro for advice on RSV-F epitopes and motif grafting methods; Ta-yi Yu, Gyu Rie Lee, Linna An, and Xinru Wang for advice on flow cytometry; Runze Dong and Varshan Muhunthan for exploratory analyses; Brian Trippe for feedback on the manuscript; Sam Pellock for expertise on enzyme design.

Funding

We thank Microsoft for support and for providing Azure computing resources. J.W. is supported by a postdoctoral fellowship from the Washington Research Foundation. D.T. is supported by The Open Philanthropy Project Improving Protein Design Fund. S.L. is supported by Amgen. L.F.M. is supported by a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C) and an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019). D.J. is supported by Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. M.E. is supported by the “la Caixa” Foundation. I.A. is supported by the National Institute of Allergy and Infectious Diseases (NIAID, Federal Contract HHSN272201700059C). S.O. supported by NIH grant DP5OD026389. D.B. is supported by the Howard Hughes Medical Institute.

Author contributions

Designed the research: JW, SL, DJ, DT, SO, DB

Developed the hallucination method: JW, SL, DT, IA, SO, JD

Developed the inpainting method: DJ, JLW, JW, DT, SL

Generated designs using hallucination: JW, SL, DT, SO

Generated designs using inpainting: DJ, JLW, JW, AS, SL

Analyzed data: JW, SL, DJ, DT, JLW, ME

Trained neural networks: DJ, MB, JLW

Performed experiments: JW, SL, LFM, JC, WY

Wrote the manuscript: JW, SL, DJ, DT, JLW, DB

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

All code will be made publicly available upon publication.

Supplementary materials

- Materials and Methods
- Supplementary Text
- Figures S1 - S21
- Tables S1 - S3
- Algorithm S1
- Data S1

Figures

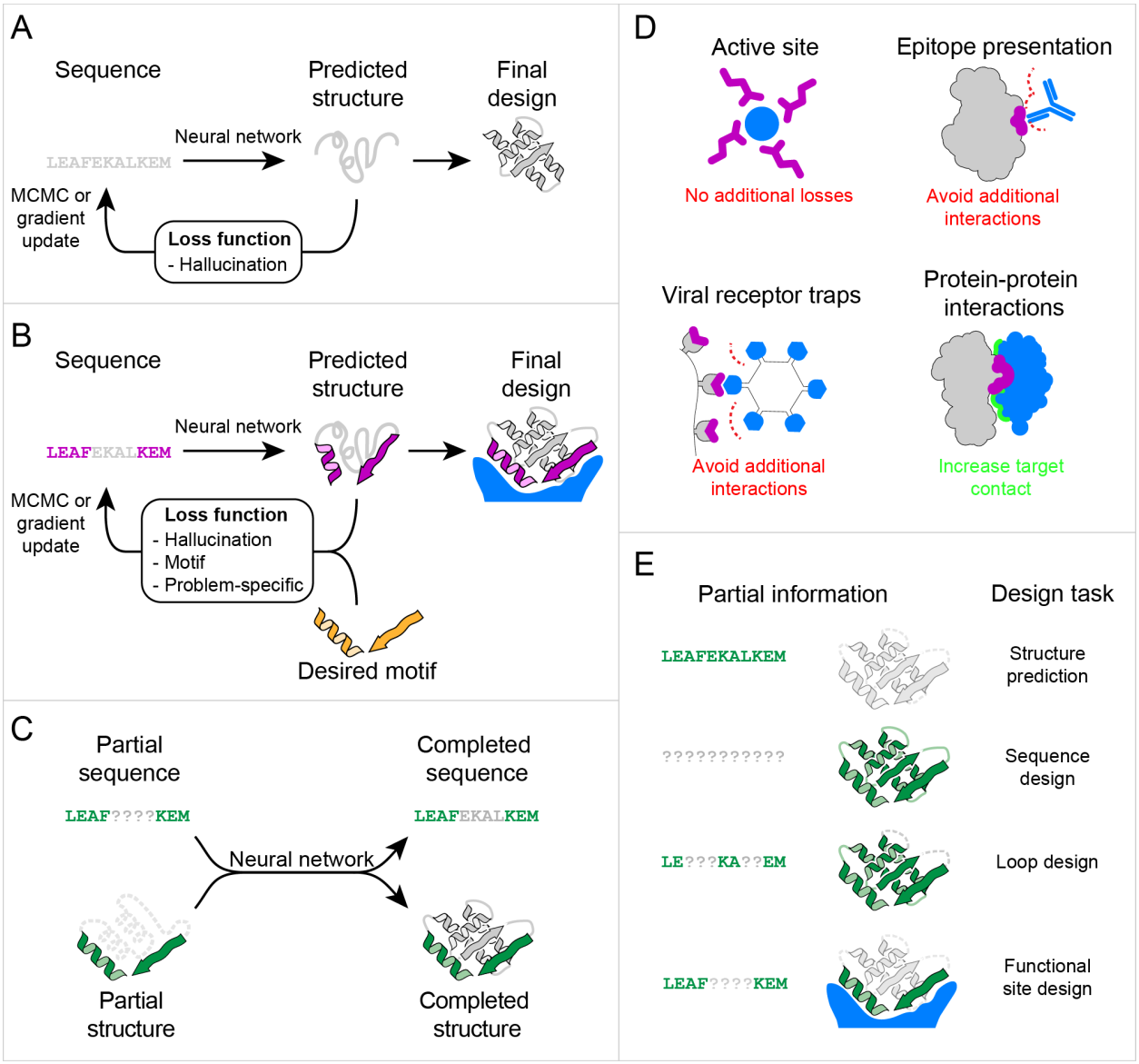

Figure 1. Methods for protein function design

(A) Free hallucination. At each iteration, a sequence is passed to the trRosetta or RoseTTAFold neural network, which predicts 3D coordinates and residue-residue distances and orientations (Fig. S3) which are scored by a loss function that rewards certainty of the predicted structure. The sequence is updated either by back propagating the gradient of the loss to the inputs or by MCMC, and passed back into the network for the next iteration. (B) Constrained hallucination. Same approach as in (A) but the loss function rewards motif recapitulation and other task-specific functions in addition to structural certainty. (C) Missing information recovery. Partial sequence and/or structural information is input into the network, and complete sequence and structure are output. (D) Design problems that can be addressed by constrained hallucination, and the corresponding loss functions (Fig. S3; Methods). (E) Protein design challenges

formulated as missing information recovery problems. Colors in all panels: native functional motif (orange); hallucinated scaffold (gray); constrained motif (purple); binding partner (blue); non-masked region (green); masked region (light gray, dotted lines)

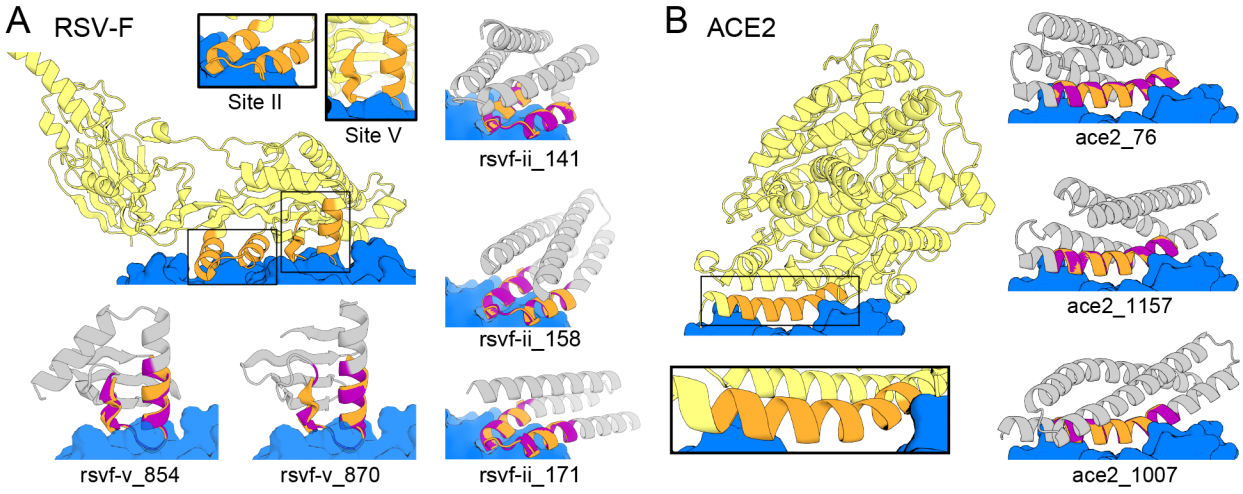

Figure 2. Hallucination of epitope scaffolds and receptor traps.

(A) Design of proteins scaffolding immunogenic epitopes on RSV protein F (site II: PDB 3IXT chain P residues 254-277; site V: 5TPN chain A residues 163-181). Comparisons of the RF hallucinated models to unbiased AF2 structure predictions from the design sequence are in Fig. S8; 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 which 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 RMSD AF versus native). 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); rsvf-v_854 (81.5, 0.75 Å); rsvf-v_870 (80.4, 0.76 Å). (B) Design of COVID-19 receptor trap based on ACE2 interface helix (6VW1 chain A residues 24-42). Design metrics: ace2_76 (89.1, 0.55 Å); ace2_1157 (80.4, 0.47 Å); ace2_1007 (83.3, 0.57 Å). Colors: native protein scaffold (light yellow); native functional motif (orange); hallucinated scaffold (gray); hallucinated motif (purple); binding partner (blue). See Table S2 for additional metrics on each design.

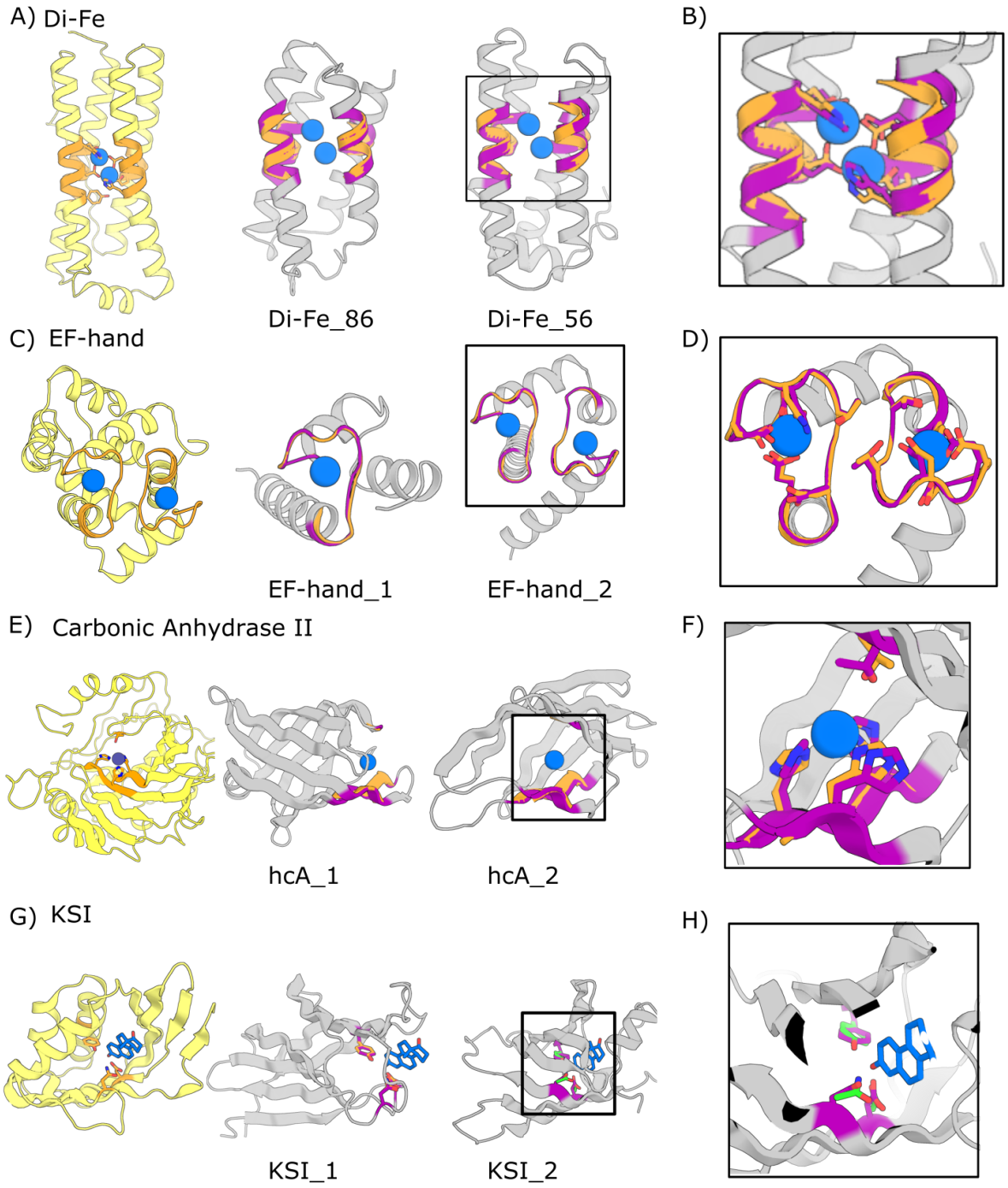

Figure 3. Hallucination of metal binding and enzyme active sites.

(A-F) Hallucinations using backbone description of site using RF. (G-H) Hallucination using sidechain description of site using trRosetta followed by AF2. (A) Di-iron binding site from E. coli cytochrome b1 (1BCF chain A residues 18-25, 27-54, 94-97, 123-130). (C) EF-hand Calcium binding site. (E) Carbonic anhydrase II active site (5YUI chain A residues 62-65, 93-97, 118-120). (G) Δ^5-3-ketosteroid Isomerase active site (1QJG chain A residues 14, 38, 99).

Colors: native protein scaffold (light yellow); native functional motif (orange); hallucinated scaffold (gray); hallucinated motif (purple); bound metal (blue). Active site residues shown for boxed designs in panel B, D, F, and H for di-iron, EF-hand, carbonic anhydrase II, and Δ^5-3-Ketosteroid Isomerase respectively. Design metrics (AF pLDDT, motif RMSD AF versus native): Di-Fe_86 (84, 0.90 Å), Di-Fe_56 (84, 0.86 Å) EF-hand_1 (84, 0.37 Å), EF-hand_2 (80, 0.37 Å), hcA_1 (73, 1.04 Å), hcA_2 (71, 0.62 Å), KSI_1 (84, 0.30 Å Cb), KSI_2 (72, 0.53 Å Cb)

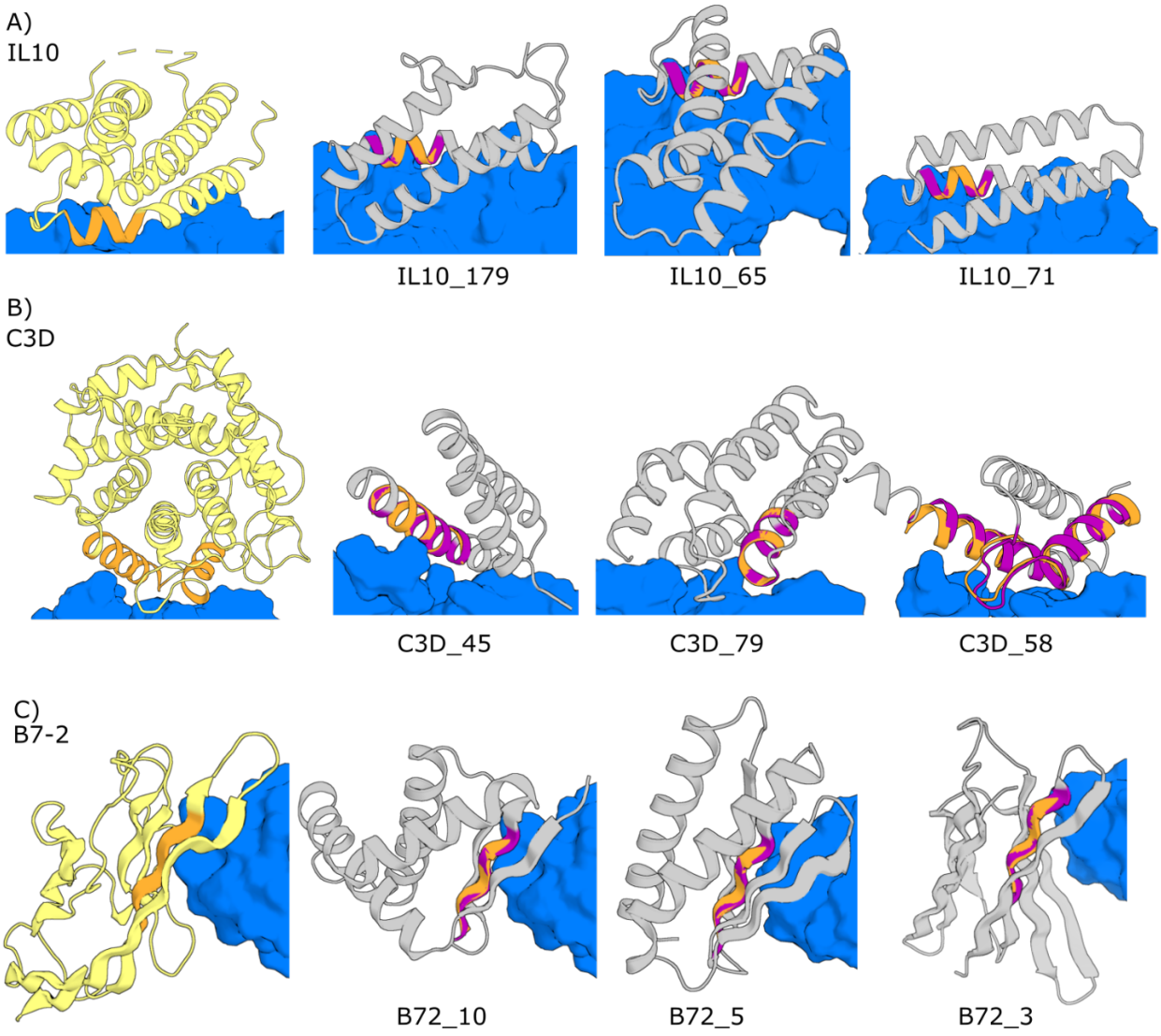

Figure 4. Hallucination of protein-protein interactions.

Designs containing extended target binding interfaces built around native complex derived binding motifs. Targets are in blue and native scaffolds in yellow. (A) Target: IL10 receptor; scaffold: Interleukin 10 (1Y6K chain A residues 23-29). (B) Target: complement receptor; scaffold: Complement protein C3d (1GHQ chain A 104-126, 170-185). (C) B7-2 (1I85 chain B residues 84-88). Native functional motifs (orange); hallucinated scaffold (gray); hallucinated motif (purple). Design metrics (AF pLDDT, motif RMSD AF versus native): IL10_179 (82, 0.35 Å), IL10_65 (88, 0.37 Å), IL10_71 (75, 0.45 Å), C3D_45 (81, 0.71 Å), C3D_79 (70, 0.28 Å), C3D_58 (86, 0.47 Å), B72_10 (81, 0.29 Å), B72_5 (87, 0.23 Å), B72_3 (81, 0.25 Å)

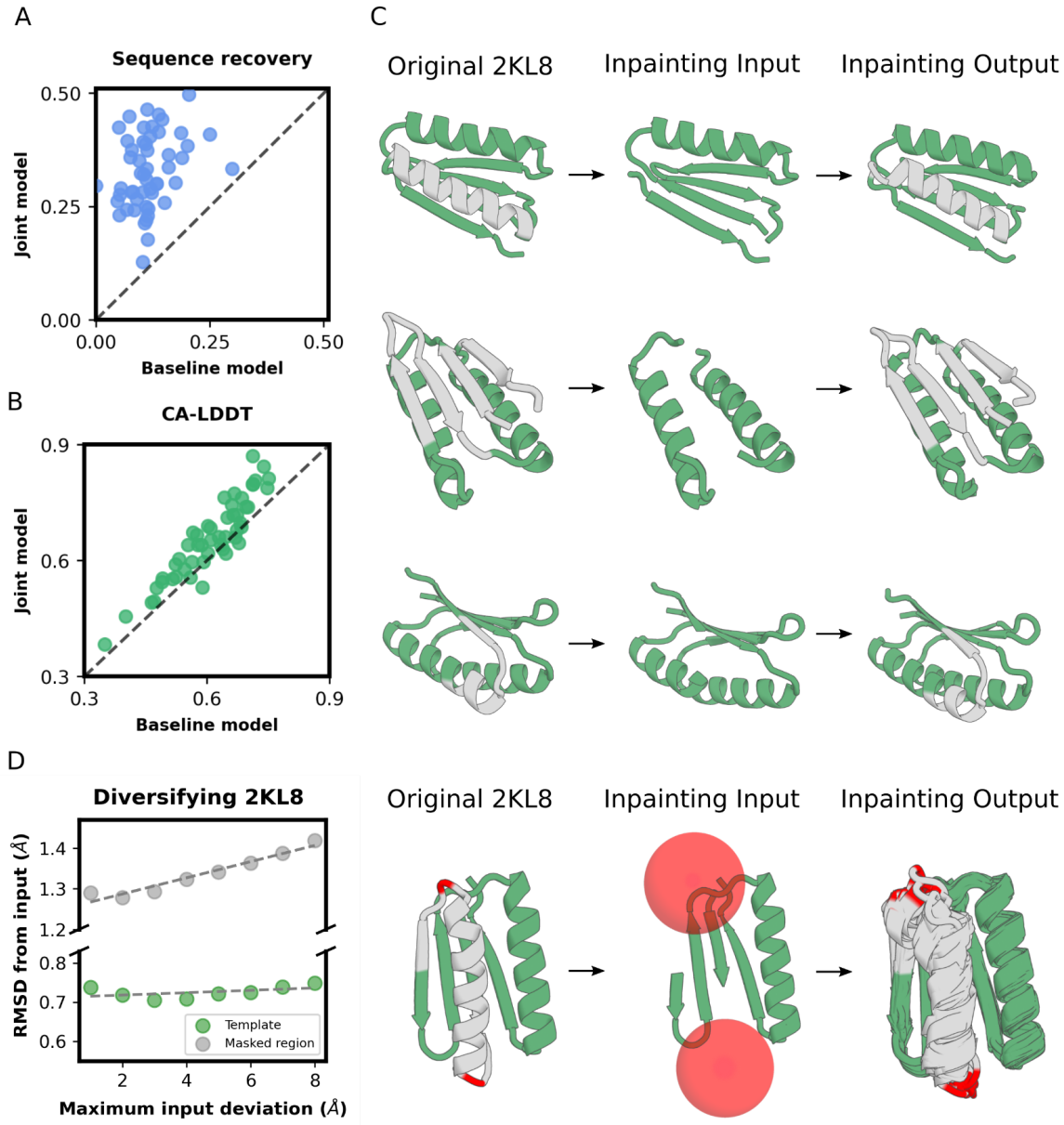

Figure 5. Joint sequence-structure recovery using RosettaFold

(A) Joint RoseTTAFold (RF_{joint}) outperforms baseline RF in fixed-backbone sequence design on a held out set of *de novo* designed proteins. (B) RF_{joint} preserves or exceeds the baseline model structure prediction quality on the *de novo* protein set. (C) Given a template sequence and structure (green) with regions of both sequence and structure masked (gray), RF_{joint} can recover the missing sequence and structure in a single forward pass. The sequence and structure in contiguous regions of test set protein 2KL8 were both masked prior to input into RF_{joint}. Top row: alpha helix. Middle row: four strand beta sheet. Bottom row: a 10-residue loop. (D) RF_{joint2} builds sequence/structure between two given residue coordinates which enables tunable diversification of rebuilt segments. The depicted gray region was masked from 2KL8, and the two coordinates shown in red were randomly translated up to 8Å in any direction (within the illustrated red spheres). RF_{joint2} is able to build back an ensemble of helical inpainted regions

(right panel, AF2 predictions, AF2 pLDDT > 0.8 for all designs shown). Increasing structural diversity could be achieved in the central inpainted region (in both the RF inpainted structure models and the AF2 structure predictions of the inpainted sequences) by increasing the distance by which the red coordinates could be translated (left graph, gray points) without substantial disruption to the remainder of the template structure (left graph, green points, n=5000 structures/point).

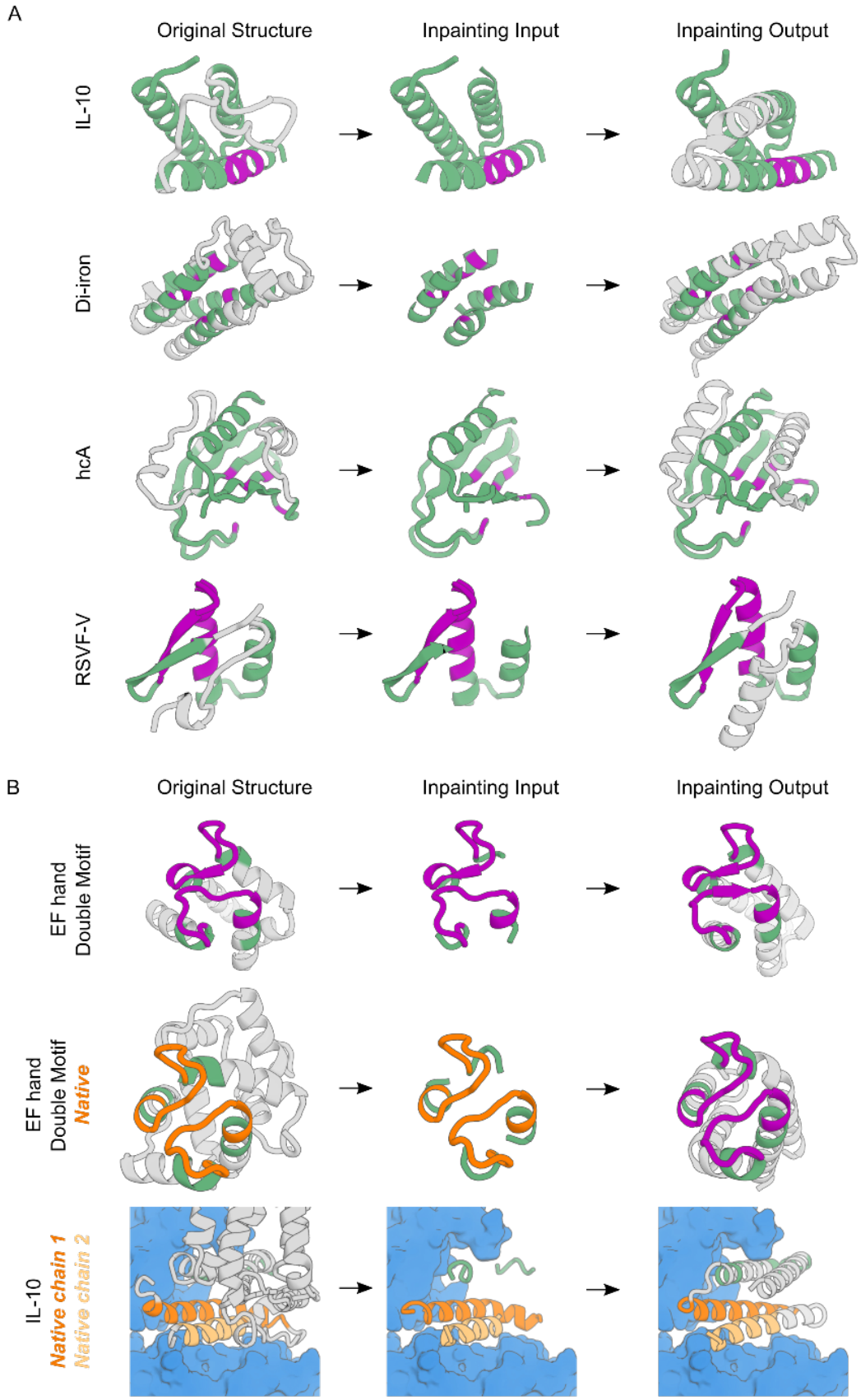
A
Original Structure
Inpainting Input
Inpainting Output
IL-10
Di-iron
hcA
RSVF-V
B
Original Structure
Inpainting Input
Inpainting Output
EF hand
Double Motif
EF hand
Double Motif
Native
IL-10
Native chain 1
Native chain 2

Figure 6. Protein function design by joint sequence-structure information recovery.

Design of proteins harboring functional motifs via information recovery using RF_{joint} and RF_{joint2}. All structures of designs shown are the AF2 prediction of that design. In all cases, template inputs (sequence and structure) that are functional and their corresponding outputs are colored in purple, template inputs that are not directly related to function are in green, along with their corresponding outputs. Functional template inputs derived from a native structure are in orange, with corresponding outputs in purple. Depicted in gray are the regions of sequence and structure masked from the original protein (input column) or that were generated via RF_{joint}/RF_{joint2} (output column). (A) RF_{joint} functional motif design examples. From top to bottom row with (AF2 motif RMSD to native, AF2 pLDDT): IL-10 (93.1, 0.57 Å), Di-Iron (91.0, 0.49 Å) carbonic anhydrase (78.8, 1.09 Å), RSVF-V (81.8, 1.39 Å). (B) $RF_{joint,\ 2}$ functional motif design examples. From top to bottom row with (AF pLDDT, motif RMSD AF vs native): EF hand double motif starting from a hallucination (85.4, 0.69 Å motif #1, 0.86 Å motif #2), EF hand double motif starting from native crystal structure (PDB: 1PRW, chain A 16-35, 52-71) (78.7, 1.13 Å motif #1, 1.10 Å motif #2), IL10 motif (light/dark orange) starting from native crystal structure (PDB: 6X93, chain A 16-41, 83-88, chain D 96-101,143-156) (75.6, 1.16 Å).

Data and code availability

All source code will be made freely available upon publication.