Published in final edited form as:

Science. 2022 July 22; 377(6604): 387–394. doi:10.1126/science.abn2100.

Scaffolding protein functional sites using deep learning

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

[*]To whom correspondence should be addressed. dabaker@uw.edu, so@fas.harvard.edu.
[†]These authors contributed equally to this work.

Author contributions
Designed the research: JW, SL, DJ, DT, JLW, SO, DB
Developed the motif-constrained hallucination method: JW, DT, SL, IA, SO
Contributed code and ideas for hallucination: MB, JD
Generated designs using hallucination: JW, SL, DT, SO
Developed the inpainting method: DJ, JLW
Contributed code and ideas for inpainting: MB, JW, SL, DT
Generated designs using inpainting: DJ, JLW, AS
Analyzed data: JW, SL, DJ, DT, JLW, ME
Trained neural networks: DJ, JLW, MB
Performed RSV-F experiments: KMC, RR, LFM, JW
Performed Di-iron experiments: JLW, DJ
Performed EF-hand experiments: AS, JLW
Performed PD-L1 experiments: WY, DRH, JW, SL, DJ
Contributed reagents and technical expertise: TS, JHC, LFM, NB, BIMW, BC, AM, FD
Wrote the manuscript: JW, DJ, JLW, SL, DT, SO, DB

Competing interests
Authors declare that they have no competing interests.

Supplementary materials
Materials and Methods
Supplementary Text
Figures S1 – S21
Tables S1 – S3
Algorithm S1
Data S1 – S2
References 59–87

[g]Molecular Engineering Graduate Program, University of Washington, Seattle, WA 98105, USA

[h]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland

Abstract

The binding and catalytic functions of proteins are generally mediated by a small number of functional residues held in place by the overall protein structure. We describe deep learning approaches for scaffolding such functional sites without needing to pre-specify 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 the 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 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--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). 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. 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 (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 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) three dimensional structure during Monte Carlo sampling in sequence space (12). We refer to this process as “hallucination” as it produces solutions that the network considers ideal proteins but do not correspond to any known natural protein; 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). Given this ability to design both sequence and structure, we reasoned that trRosetta could be adapted to tackle the functional site scaffolding problem.

Partially constrained hallucination using a multi-objective 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; Methods). While 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 PD-L1 generated by constrained hallucination with binding motifs derived from PD-1 (Table S1) (15), followed by Rosetta design, were found to have binding affinities in the mid-nanomolar range (Fig. S1B–E). While 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 residue-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 (MSA) are masked and the network is trained to recover this missing sequence information in addition to predicting structure. This ability to recover both sequence and structural information provides a second solution to the functional site scaffolding problem: given a functional site description, a forward pass through the network can be used to complete, or “inpaint”, both protein sequence and structure in a missing/masked region of protein (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 (17) or 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), here we approach it using the power of a pre-trained structure-prediction network.

We began from a RosettaFold 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 (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; Algorithm S1). The resulting model was able to inpaint missing regions with high fidelity (Fig. 1E, S4) and performed well at sequence design (32% native sequence recovery during training, Fig. S4C) and structure prediction (Fig. S4C). We call this network RF_{joint} and use it to generate all inpainted designs below except 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, 24). We used AF to compare the ability of hallucination and inpainting to rebuild missing protein regions (Fig. 1F–G, S5). Inpainting yielded solutions with more accurately predicted fixed regions (“AF-RMSD”; Fig. 1G, S5B) and structures overall more confidently predicted from their amino acid sequences (“AF pLDDT”, Fig. 1F, S5A), and required only 1–10 seconds per design on an NVIDIA RTX2080 GPU (hallucination requires 5–20 minutes per design). However, hallucination gave better results when the missing region was large (Fig. S5) and generated greater structural diversity (Fig. S8, 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 (Fig. 2–5, Table S1). For almost all problems, we obtained designs that are closely recapitulated by AF with overall and motif (functional site) RMSD typically <2 Å and <1 Å respectively, with high model confidence (pLDDT > 80; Table S2); such recapitulation suggests the designed sequences encode the designed structures (although it should be noted that AF has limited ability to predict protein stability (25) or mutational effects (26, 27)). More critically, we assessed the activities of the designs experimentally (with the exception of those labeled “in silico” in Fig. 2–5).

Designing immunogen candidates and receptor traps

The goal of immunogen design is to scaffold a native epitope recognized by a neutralizing antibody as accurately as possible, in order to elicit antibodies binding the native protein upon immunization. Additional interactions with the antibody are undesirable because the goal 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 3-helix bundle (7), as well as RSV-F site V, a 19-residue helix-loop-strand motif that has not yet been scaffolded successfully (28). We were able to hallucinate designs recapitulating both epitopes to sub-angstrom backbone RMSD in a variety of folds (Fig. 2A, Fig. S9; structures and sequences for all designs below are in Data S1–2 and differ considerably from native proteins (Table S2); RF and AF models are in Fig. S9, S11, 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 E. coli and found that three bound the neutralizing antibody hRSV90 (28) with Kd's of 0.9–1.3 uM (Fig. 2C, S11; Methods; Supplementary Text). The Kd for the RSVF trimer is lower (23nM), but the interface is larger encompassing both sites II and V (28). 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, S11A), and circular dichroism spectra were consistent with the designed scaffold structures for designs (Fig. 2D) and their point mutants (Fig. S11C). Four of the inpainted designs bound hRSV90 by yeast display, but were poorly expressed in E. coli (Fig. S11C–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 which 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 (RBD) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein (29). 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, S9C).

Designing metal-coordinating proteins

Di-iron sites are important in biological systems for iron storage (30) and can mediate catalysis (31, 32). 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. 3A–E, S13) and hallucination (Fig. S12; the latter were not tested due to buried polar residues; Supplementary Text). The designs had diverse helix connectivities and low structural similarity to the parent (Fig. S13B, S12; TM-score 0.55–0.71 to 1BCF_A). We chose 96 inpainted designs to test experimentally, and found that 76 had soluble expression, at least 8 (Supplementary Text) had a spectroscopic shift indicative of Co^{2+}-binding (a proxy for iron binding) (33, 34), and 3 (dife_inp_1–3, Fig. 3B, S13E) had CD spectra consistent with the designed fold (Fig. 3D, S13F) and were stabilized by metal binding (Fig. 3E, S13G). Mutation of the metal binding residues abolished binding (Fig. 3B, 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 (35), a 12-residue loop flanked by helices. Both constrained hallucination and inpainting readily generated scaffolds recapitulating either 1 or 2 EF-hand motifs to within 1.0 Å AF-RMSD of the native motif (Fig. 3F, Fig S14A,B, table S2). We chose 20 hallucinations and 55 inpaints to display on yeast and screen for calcium binding using tryptophan-enhanced terbium fluorescence (36). 6 hallucinations and 4 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*, was purified from E. coli as a monomer (Fig. S14C), had the expected CD spectrum (Fig. 3G) and a clear terbium binding signal (Fig. 3H) which 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 (32–34). The active site consists of 3 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, 3-segment site, hallucination was able to generate designs with sub-angstrom motif AF-RMSDs with correct His placement for Zn^{2+} coordination (Fig. 4A, S9D); these are less than 100 residues, significantly smaller than the 261 residue native protein.

We next scaffolded the catalytic sidechains of Δ^5-3-ketosteroid isomerase (Table S1) involved in steroid hormone biosynthesis (37). We attempted to use gradient descent by backpropagation through AF (Materials and Methods; a sidechain-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 sidechain-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 sidechain geometry (Fig. 4C–D, 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 2 discontiguous segments of the interfacial beta-sheet from a high-affinity mutant of PD-1 (Fig 5A; 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 inter-chain predicted alignment error (inter-PAE) of <10 Å (Materials and Methods). Unlike 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. 5B–C), worse than HAC PD-1 (K_D = 110 pM) (38) but better than WT PD-1 (K_D = 3.9 μM) (38). *pdl1_inp_1* expressed as a monomer (Fig. S15E), was thermostable, and had a CD spectrum consistent with that of a mixed alpha-beta fold (Fig. S15F). Unlike native PD-1, which has a immunoglobulin family beta-sandwich fold, *pdl1_inp_1* has 2 helices buttressing the interfacial beta sheet, as well as an additional 5th inpainted strand extending the interface (Fig. S15 A,B). The closest PDB hit had a TM-score of 0.61 and the closest BLAST NR hit had a sequence identity of 25.4%.

We next used inpainting to design ligands engaging multiple receptor binding sites. The nerve growth factor 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. 5D–E). A design predicted to be well-structured (AF pLDDT > 80) and interact with TrkA (inter-PAE < 10 Å) was expressed, purified and bound TrkA as assessed by biolayer interferometry (BLI) (Fig. 5F). A double mutant that knocked out both designed binding sites abolished TrkA binding, while single mutants knocking out either one of the binding sites maintained partial binding (Fig. 5F; 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, 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 SAP score < 35 (Fig. S17D–E); 3 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 inter-chain 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 55–80-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. S17F,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 which 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; 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 on par with structurally validated *de novo* designed proteins (Fig. S7A–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 (Methods), while 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 due to the difficulty of efficiently representing and learning structure (50). Generative adversarial networks (GANs) and variational autoencoders (VAEs) 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Luki Goldschmidt and Kandise VanWormer, respectively, for maintaining the computational and wet lab resources in the IPD; Christoffer Norn for general discussions about trRosetta; Brian Coventry for advice on interface design; Casper Goverde 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; Naozumi Hiranuma for exploratory RoseTTAFold training sessions; Brian Trippe for feedback on the manuscript; Sam Pellock for expertise on enzyme design; Andrew Fitzgibbon for conceptual discussions on training RosettaFold; Chris 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 (DB, AS); a Microsoft gift (MB, JD); Eric and Wendy Schmidt by recommendation of the Schmidt Futures (DJ); the DARPA Synergistic Discovery and Design project HR001117S0003 contract FA8750-17-C-0219 (DB, WY); the DARPA Harnessing Enzymatic Activity for Lifesaving Remedies project HR001120S0052 contract HR0011-21-2-0012 (NB); the Washington Research Foundation (JW); the Open Philanthropy Project Improving Protein Design Fund (DB, DT); Amgen (SL); the Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C) and EMBO Non-Stipendiary Fellowship (ALTF 1047-2019) (LFM); the EMBO Fellowship (ALTF 191-2021) (TS); European Molecular Biology Organization Grant (ALTF 139-2018) (BIMW); the “la Caixa” Foundation (ME); the National Institute of Allergy and Infectious Diseases (NIAID) Federal Contract HHSN272201700059C (IA), NIH grant DP5OD026389 (SO); the National Science Foundation MCB 2032259 (SO); the Howard Hughes Medical Institute (DB, RR, KMC), the National Institute on Aging grant 5U19AG065156 (DB, JLW, DRH, ME); the National Cancer Institute grant R01CA240339 (DB, JHC); Swiss National Science Foundation (KMC, BC); Swiss National Center of Competence for Molecular Systems Engineering (KMC, BC); Swiss National Center of Competence in Chemical Biology (KMC, BC); European Research Council grant 716058 (KMC, BC).

Data and materials availability

Code and neural network weights are available at https://github.com/RosettaCommons/RFDesign and archived at Zenodo (doi: 10.5281/zenodo.6673001). Plasmids of designed proteins are available upon request.

References

1. Khersonsky O, Wollacott AM, Jiang L, Dechancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D, Kemp elimination catalysts by computational enzyme design. 453 (2008), doi:10.1038/nature06879.
2. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D, De Novo Computational Design of Retro-Aldol Enzymes. Science. 319, 1387–1391 (2008). [PubMed: 18323453]
3. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St. Clair JL, Gallaher J, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D, Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. Science. 329 (2010), doi:10.1126/science.1190239.
4. Cao L, Coventry B, Goreshnik I, Huang B, Park JS, Jude KM, Marković I, Kadam RU, Verschueren KHG, Verstraete K, Walsh STR, Bennett N, Phal A, Yang A, Kozodoy L, DeWitt M, Picton L, Miller L, Strauch E-M, DeBouver ND, Pires A, Bera AK, Halabiya S, Hammerson B, Yang W, Bernard S, Stewart L, Wilson IA, Ruohola-Baker H, Schlessinger J, Lee S, Savvides SN, Garcia KC, Baker D, Design of protein binding proteins from target structure alone. Nature (2022), doi:10.1038/s41586-022-04654-9.
5. Chevalier AA, Silva D, Rocklin GJ, Derrick R, Vergara R, Murapa P, Bernard SM, Zhang L, Yao G, Bahl CD, Miyashita S, Goreshnik I, James T, Bryan M, Fernández-velasco DA, Stewart L, Dong M, Huang X, Massively parallel de novo protein design for targeted therapeutics. Nat. Publ. Group (2017), doi:10.1038/nature23912.
6. Procko E, Berguig GY, Shen BW, Song Y, Frayo S, Convertine AJ, Margineantu D, Booth G, Correia BE, Cheng Y, Schief WR, Hockenbery DM, Press OW, Stoddard BL, Stayton PS, Baker D,

A Computationally Designed Inhibitor of an Epstein-Barr Viral Bcl-2 Protein Induces Apoptosis in Infected Cells. Cell. 157, 1644–1656 (2014). [PubMed: 24949974]

7. Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, Rupert P, Correnti C, Kalyuzhniy O, Vittal V, Connell MJ, Stevens E, Schroeter A, Chen M, MacPherson S, Serra AM, Adachi Y, Holmes MA, Li Y, Klevit RE, Graham BS, Wyatt RT, Baker D, Strong RK, Crowe JE, Johnson PR, Schief WR, Proof of principle for epitope-focused vaccine design. Nature. 507, 201–206 (2014). [PubMed: 24499818]
8. Silva D-A, Yu S, Ulge UY, Spangler JB, Jude KM, Labão-Almeida C, Ali LR, Quijano-Rubio A, Ruterbusch M, Leung I, Biary T, Crowley SJ, Marcos E, Walkey CD, Weitzner BD, Pardo-Avila F, Castellanos J, Carter L, Stewart L, Riddell SR, Pepper M, Bernardes GJL, Dougan M, Garcia KC, Baker D, De novo design of potent and selective mimics of IL-2 and IL-15. Nature. 565, 186–191 (2019). [PubMed: 30626941]
9. Sesterhenn F, Yang C, Bonet J, Cramer JT, Wen X, Wang Y, Chiang C-I, Abriata LA, Kucharska I, Castoro G, Vollers SS, Galloux M, Dheilly E, Rosset S, Corthésy P, Georgeon S, Villard M, Richard C-A, Descamps D, Delgado T, Oricchio E, Rameix-Welti M-A, Más V, Ervin S, Eléouët J-F, Riffault S, Bates JT, Julien J-P, Li Y, Jardetzky T, Krey T, Correia BE, De novo protein design enables the precise induction of RSV-neutralizing antibodies. Science. 368 (2020), doi:10.1126/science.aay5051.
10. Yang C, Sesterhenn F, Bonet J, van Aalen EA, Scheller L, Abriata LA, Cramer JT, Wen X, Rosset S, Georgeon S, Jardetzky T, Krey T, Fussenegger M, Merkx M, Correia BE, Bottom-up de novo design of functional proteins with complex structural features. Nat. Chem. Biol, 1–9 (2021). [PubMed: 33328655]
11. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D, Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. (2020), doi:10.1073/pnas.1914677117.
12. Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK, DiMaio F, Carter L, Chow CM, Montelione GT, Baker D, De novo protein design by deep network hallucination. Nature. 600, 547–552 (2021). [PubMed: 34853475]
13. Norn C, Wicky BIM, Juergens D, Liu S, Kim D, Tischer D, Koepnick B, Anishchenko I, Players F, Baker D, Ovchinnikov S, Protein sequence design by conformational landscape optimization. Proc. Natl. Acad. Sci. 118 (2021), doi:10.1073/pnas.2017228118.
14. Tischer D, Lisanza S, Wang J, Dong R, Anishchenko I, Milles LF, Ovchinnikov S, Baker D, bioRxiv, in press, doi:10.1101/2020.11.29.402743.
15. Pascolutti R, Sun X, Kao J, Maute RL, Ring AM, Bowman GR, Kruse AC, Structure and Dynamics of PD-L1 and an Ultra-High-Affinity PD-1 Receptor Mutant. Structure. 24, 1719–1728 (2016). [PubMed: 27618663]
16. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D, Accurate prediction of protein structures and interactions using a three-track neural network. Science (2021), doi:10.1126/science.abj8754.
17. Devlin J, Chang M-W, Lee K, Toutanova K, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs (2019) (available at http://arxiv.org/abs/1810.04805).
18. Yeh RA, Chen C, Lim TY, Schwing AG, Hasegawa-Johnson M, Do MN, Semantic Image Inpainting with Deep Generative Models. ArXiv160707539 Cs (2017) (available at http://arxiv.org/abs/1607.07539).
19. Li Z, Nguyen SP, Xu D, Shang Y, in 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI) (2017), pp. 1085–1091.
20. Anand N, Huang P, in Advances in Neural Information Processing Systems 31, Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, Eds. (Curran Associates, Inc., 2018; http://papers.nips.cc/paper/7978-generative-modeling-for-protein-structures.pdf), pp. 7494–7505.
21. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes

B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D, Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583–589 (2021). [PubMed: 34265844]

22. Chowdhury R, Bouatta N, Biswas S, Rochereau C, Church GM, Sorger PK, AlQuraishi M, Single-sequence protein structure prediction using language models from deep learning, 22.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000). [PubMed: 10592235]
24. Simons KT, Bonneau R, Ruczinski I, Baker D, Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Struct. Funct. Bioinforma. 37, 171–176 (1999).
25. Kim T-E, Tsuboyama K, Houliston S, Martell CM, Phoumyvong CM, Haddox HK, Arrowsmith CH, Rocklin GJ, Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation (2021), p. 2021.12.17.472837, , doi:10.1101/2021.12.17.472837.
26. Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, Kondrashov FA, Ivankov DN, Using AlphaFold to predict the impact of single mutations on protein stability and function (2021), p. 2021.09.19.460937, , doi:10.1101/2021.09.19.460937.
27. Buel GR, Walters KJ, Can AlphaFold2 predict the impact of missense mutations on structure? Nat. Struct. Mol. Biol. 29, 1–2 (2022). [PubMed: 35046575]
28. Mousa JJ, Kose N, Matta P, Gilchuk P, Crowe JE, A novel pre-fusion conformation-specific neutralizing epitope on the respiratory syncytial virus fusion protein. Nat. Microbiol. 2, 1–8 (2017).
29. Linsky TW, Vergara R, Codina N, Nelson JW, Walker MJ, Su W, Barnes CO, Hsiang T-Y, Esser-Nobis K, Yu K, Reneer ZB, Hou YJ, Priya T, Mitsumoto M, Pong A, Lau UY, Mason ML, Chen J, Chen A, Berrocal T, Peng H, Clairmont NS, Castellanos J, Lin Y-R, Josephson-Day A, Baric RS, Fuller DH, Walkey CD, Ross TM, Swanson R, Bjorkman PJ, Gale M, Blancas-Mejia LM, Yen H-L, Silva D-A, De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. Science (2020), doi:10.1126/science.abe0075.
30. Frolow F, Kalb Gilboa AJ, Yariv J, Structure of a unique twofold symmetric haem-binding site. Nat. Struct. Biol. 1, 453–460 (1994). [PubMed: 7664064]
31. Lombardi A, Pirro F, Maglio O, Chino M, DeGrado WF, De Novo Design of Four-Helix Bundle Metalloproteins: One Scaffold, Diverse Reactivities. Acc. Chem. Res. 52, 1148–1159 (2019). [PubMed: 30973707]
32. Calhoun JR, Nastri F, Maglio O, Pavone V, Lombardi A, DeGrado WF, Artificial diiron proteins: From structure to function. Pept. Sci. 80, 264–278 (2005).
33. Keech AM, Brun NEL, Wilson MT, Andrews SC, Moore GR, Thomson AJ, Spectroscopic Studies of Cobalt(II) Binding to Escherichia coli Bacterioferritin*. J. Biol. Chem. 272, 422–429 (1997). [PubMed: 8995278]
34. Marsh ENG, DeGrado WF, Noncovalent self-assembly of a heterotetrameric diiron protein. Proc. Natl. Acad. Sci. 99, 5150–5154 (2002). [PubMed: 11959963]
35. Yáñez M, Gil-Longo J, Campos-Toimil M, in Calcium Signaling, Islam Md. S., Ed. (Springer Netherlands, Dordrecht, 2012; 10.1007/978-94-007-2888-2_19), Advances in Experimental Medicine and Biology, pp. 461–482.
36. Caldwell SJ, Haydon IC, Piperidou N, Huang P-S, Bick MJ, Sjöström HS, Hilvert D, Baker D, Zeymer C, Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion. Proc. Natl. Acad. Sci. 117, 30362–30369 (2020). [PubMed: 33203677]
37. Cho H-S, Ha N-C, Choi G, Kim H-J, Lee D, Oh KS, Kim KS, Lee W, Choi KY, Oh B-H, 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). [PubMed: 10551849]
38. Maute RL, Gordon SR, Mayer AT, McCracken MN, Natarajan A, Ring NG, Kimura R, Tsai JM, Manglik A, Kruse AC, Gambhir SS, Weissman IL, Ring AM, Engineering high-affinity PD-1

variants for optimized immunotherapy and immuno-PET imaging. Proc. Natl. Acad. Sci. 112, E6506–E6514 (2015). [PubMed: 26604307]

39. Wiesmann C, Ultsch MH, Bass SH, de Vos AM, Crystal structure of nerve growth factor in complex with the ligand-binding domain of the TrkA receptor. Nature. 401, 184–188 (1999). [PubMed: 10490030]
40. Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde SR, Stancheva VG, Li X-H, Liu K, Zheng Z, Barrero DJ, Roy U, Kuper J, Fernández IS, Szakal B, Branzei D, Rizo J, Kisker C, Greene EC, Biggins S, Keeney S, Miller EA, Fromme JC, Hendrickson TL, Cong Q, Baker D, Computed structures of core eukaryotic protein complexes. Science. 0, eabm4805.
41. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Potapenko A, Ballard AJ, Romera-Paredes B, Nikolov S, Jain R, Clancy E, Reiman D, Petersen S, Senior AW, Kavukcuoglu K, Birney E, Kohli P, Jumper J, Hassabis D, Highly accurate protein structure prediction for the human proteome. Nature (2021), doi:10.1038/s41586-021-03828-1.
42. Ingraham J, Garg VK, Barzilay R, Jaakkola T, Generative models for graph-based protein design, 10 (2019).
43. Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim PM, Fast and Flexible Protein Design Using Deep Graph Neural Networks. Cell Syst. 11, 402–411.e4 (2020). [PubMed: 32971019]
44. Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM, Low- N protein engineering with data-efficient deep learning. Nat. Methods. 18, 389–396 (2021). [PubMed: 33828272]
45. Repecka D, Jauniskis V, Karpus L, Rembeza E, Zrimec J, Poviloniene S, Rokaitis I, Laurynenas A, Abuajwa W, Savolainen O, Meskys R, Engqvist MKM, Zelezniak A, Expanding functional protein sequence space using generative adversarial networks. bioRxiv, 789719 (2019).
46. Shin J-E, Riesselman AJ, Kollasch AW, McMahon C, Simon E, Sander C, Manglik A, Kruse AC, Marks DS, Protein design and variant prediction using autoregressive generative models. Nat. Commun. 12, 1–11 (2021). [PubMed: 33397941]
47. Wu Z, Johnston KE, Arnold FH, Yang KK, Protein sequence design with deep generative models. Curr. Opin. Chem. Biol. 65, 18–27 (2021). [PubMed: 34051682]
48. Anand-Achim N, Eguchi RR, Derry A, Altman RB, Huang P-S, “Protein sequence design with a learned potential” (preprint, Bioinformatics, 2020), , doi:10.1101/2020.01.06.895466.
49. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos JL, Xiong C, Sun ZZ, Socher R, Fraser JS, Naik N, bioRxiv, in press, doi:10.1101/2021.07.18.452833.
50. Ovchinnikov S, Huang P-S, Structure-based protein design with deep learning. Curr. Opin. Chem. Biol. 65, 136–144 (2021). [PubMed: 34547592]
51. Anand N, Eguchi R, Huang P-S, Fully differentiable full-atom protein backbone generation (2019) (available at https://openreview.net/forum?id=SJxnVL8YOV).
52. Eguchi RR, Anand N, Choe CA, Huang P-S, bioRxiv, in press, doi:10.1101/2020.08.07.242347.
53. Lin Z, Sercu T, LeCun Y, Rives A, Deep generative models create new and diverse protein structures, 17.
54. Jendrusch M, Korbel JO, Sadiq SK, bioRxiv, in press, doi:10.1101/2021.10.11.463937.
55. Moffat L, Greener JG, Jones DT, bioRxiv, in press, doi:10.1101/2021.08.24.457549.
56. Moffat L, Kandathil SM, Jones DT, Design in the DARK: Learning Deep Generative Models for De Novo Protein Design (2022), p. 2022.01.27.478087, , doi:10.1101/2022.01.27.478087.
57. Li L, Liu Y, Tao J, Zhang M, Pan H, Xu X, Tang R, Surface Modification of Hydroxyapatite Nanocrystallite by a Small Amount of Terbium Provides a Biocompatible Fluorescent Probe. J. Phys. Chem. C 112, 12219–12224 (2008).
58. Anishchenko I, Chidyausiku TM, Ovchinnikov S, Pellock SJ, Baker D, bioRxiv, in press, doi:10.1101/2020.07.22.211482.
59. Jang E, Gu S, Poole B, Categorical Reparameterization with Gumbel-Softmax. ArXiv161101144 Cs Stat (2017) (available at http://arxiv.org/abs/1611.01144).

60. Bogard N, Linder J, Rosenberg AB, Seelig G, A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. Cell. 178, 91–106.e23 (2019). [PubMed: 31178116]
61. Linder J, Seelig G, Fast differentiable DNA and protein sequence optimization for molecular design. ArXiv200511275 Cs Stat (2020) (available at http://arxiv.org/abs/2005.11275).
62. Kingma DP, Ba J, Adam: A Method for Stochastic Optimization. ArXiv14126980 Cs (2017) (available at http://arxiv.org/abs/1412.6980).
63. Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, Sercu T, Rives A, bioRxiv, in press, doi:10.1101/2021.02.12.430858.
64. Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, Lerer A, Rives A, Learning inverse folding from millions of predicted structures (2022), p. 2022.04.10.487779, , doi:10.1101/2022.04.10.487779.
65. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL, Design of therapeutic proteins with enhanced stability. Proc. Natl. Acad. Sci. 106, 11937–11942 (2009). [PubMed: 19571001]
66. Jha SK, Ramanathan A, Ewetz R, Velasquez A, Jha S, Protein Folding Neural Networks Are Not Robust. ArXiv210904460 Cs Q-Bio (2021) (available at http://arxiv.org/abs/2109.04460).
67. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A, Adversarial Examples Are Not Bugs, They Are Features. ArXiv190502175 Cs Stat (2019) (available at http://arxiv.org/abs/1905.02175).
68. Demontis A, Melis M, Pintor M, Jagielski M, Biggio B, Oprea A, Nita-Rotaru C, Roli F, Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks. ArXiv180902861 Cs Stat (2019) (available at http://arxiv.org/abs/1809.02861).
69. Dang B, Mravic M, Hu H, Schmidt N, Mensa B, DeGrado WF, SNAC-tag for sequence-specific chemical protein cleavage. Nat. Methods. 16, 319–322 (2019). [PubMed: 30923372]
70. Studier FW, Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif. 41, 207–234 (2005). [PubMed: 15915565]
71. Zhang Y, Skolnick J, TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309 (2005). [PubMed: 15849316]
72. Jaegle A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J, Perceiver: General Perception with Iterative Attention. ArXiv210303206 Cs Eess (2021) (available at http://arxiv.org/abs/2103.03206).
73. Kabsch W, A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A. 32, 922–923 (1976).
74. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D, Improved protein structure prediction using potentials from deep learning. Nature, 1–5 (2020).
75. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MS, Bonneau R, Bradley P, Dunbrack RL, Das R, Baker D, Kuhlman B, Kortemme T, Gray JJ, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. 13, 3031–3048 (2017). [PubMed: 28430426]
76. Boyken SE, Chen Z, Groves B, Langan RA, Oberdorfer G, Ford A, Gilmore JM, Xu C, DiMaio F, Pereira JH, Sankaran B, Seelig G, Zwart PH, Baker D, De novo design of protein homo-oligomers with modular hydrogen-bond network--mediated specificity. Science. 352, 680–687 (2016). [PubMed: 27151862]
77. Silva D-A, Correia BE, Procko E, in Computational Design of Ligand Binding Proteins, Stoddard BL, Ed. (Springer, New York, NY, 2016; 10.1007/978-1-4939-3569-7_17), Methods in Molecular Biology, pp. 285–304.
78. Steinegger M, Söding J, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028 (2017). [PubMed: 29035372]
79. Park H, Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F, Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J. Chem. Theory Comput. 12, 6201–6212 (2016). [PubMed: 27766851]

80. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C, Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins Struct. Funct. Bioinforma. 65, 712–725 (2006).

81. Pascolutti R, Sun X, Kao J, Maute RL, Ring AM, Bowman GR, Kruse AC, Structure and Dynamics of PD-L1 and an Ultra-High-Affinity PD-1 Receptor Mutant. Structure. 24, 1719–1728 (2016). [PubMed: 27618663]

82. McLellan JS, Chen M, Kim A, Yang Y, Graham BS, Kwong PD, Structural basis of respiratory syncytial virus neutralization by motavizumab. Nat. Struct. Mol. Biol. 17, 248–250 (2010). [PubMed: 20098425]

83. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, Geng Q, Auerbach A, Li F, Structural basis of receptor recognition by SARS-CoV-2. Nature. 581, 221–224 (2020). [PubMed: 32225175]

84. Fallon JL, Quiocho FA, A Closed Compact Structure of Native Ca2+-Calmodulin. Structure. 11, 1303–1307 (2003). [PubMed: 14527397]

85. Kim CU, Song H, Avvaru BS, Gruner SM, Park S, McKenna R, Tracking solvent and protein movement during CO2 release in carbonic anhydrase II crystals. Proc. Natl. Acad. Sci. 113, 5257–5262 (2016). [PubMed: 27114542]

86. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, Pavletich NP, Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. Sci. New Ser 274, 948–953 (1996).

87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990). [PubMed: 2231712]

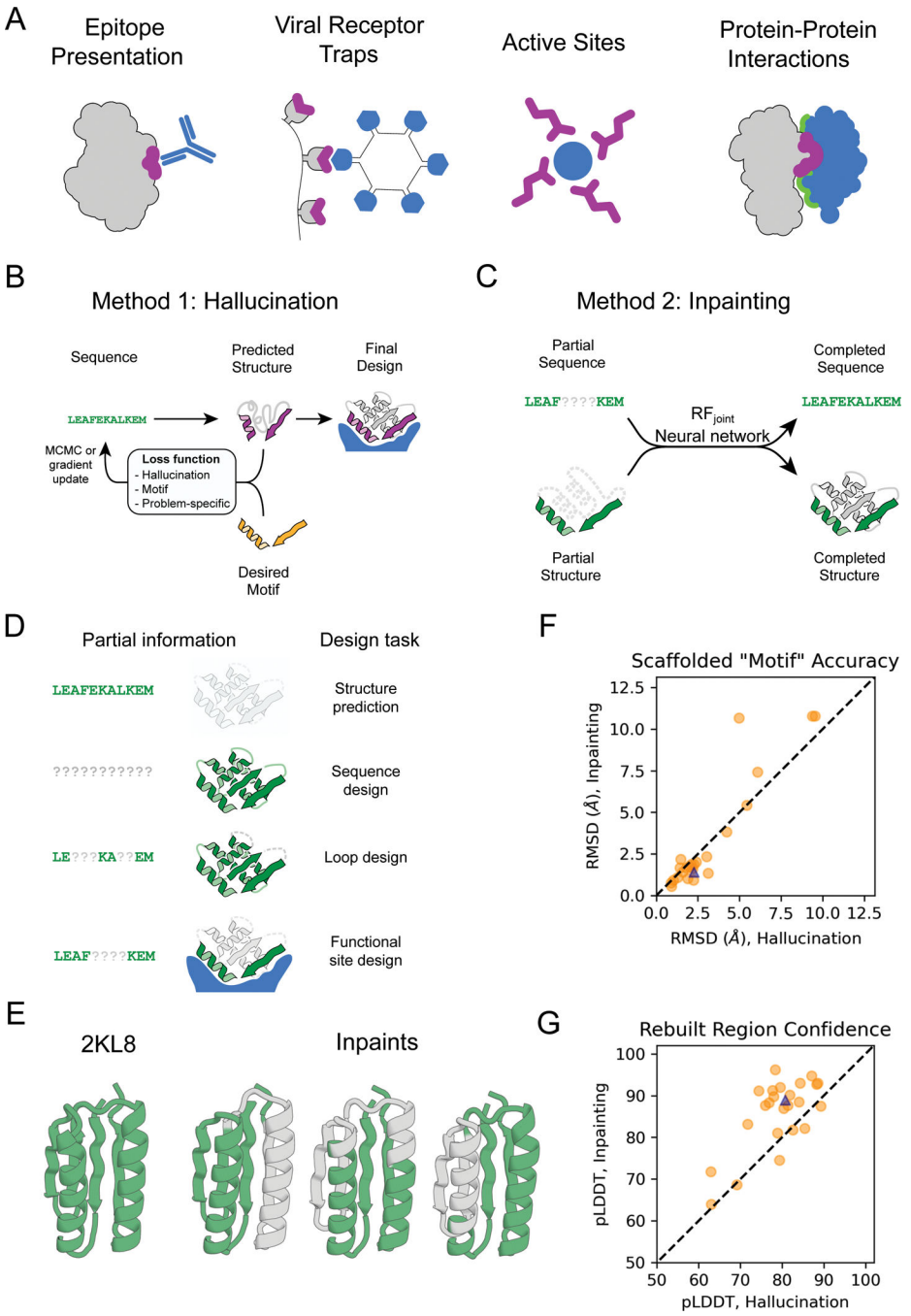

Figure 1. Methods for protein function design

(A) Applications of functional-site scaffolding. (B-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 residue-residue distances and orientations (Fig. S2) which are scored by a loss function that rewards certainty of the predicted structure along with motif recapitulation and other task-specific functions. (C) Missing information recovery ("Inpainting"). Partial sequence and/or structural information is input into a modified RosettaFold network (termed RF_{joint}), and complete sequence and structure

are output. (D) Protein design challenges formulated as missing information recovery problems. (E) Joint RosettaFold (RF_{joint}) can simultaneously recover structure and sequence of a masked region of protein. 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: 91.6/0.91, 92.0/0.69, 90.4/0.82 respectively). (F-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: Average of 20 outputs for each of the benchmarking proteins. Triangle: 2KL8. Colors in all panels: native functional motif (orange); hallucinated/inpainted scaffold (gray); constrained motif (purple); binding partner (blue); non-masked region (green); masked region (light gray, dotted lines).

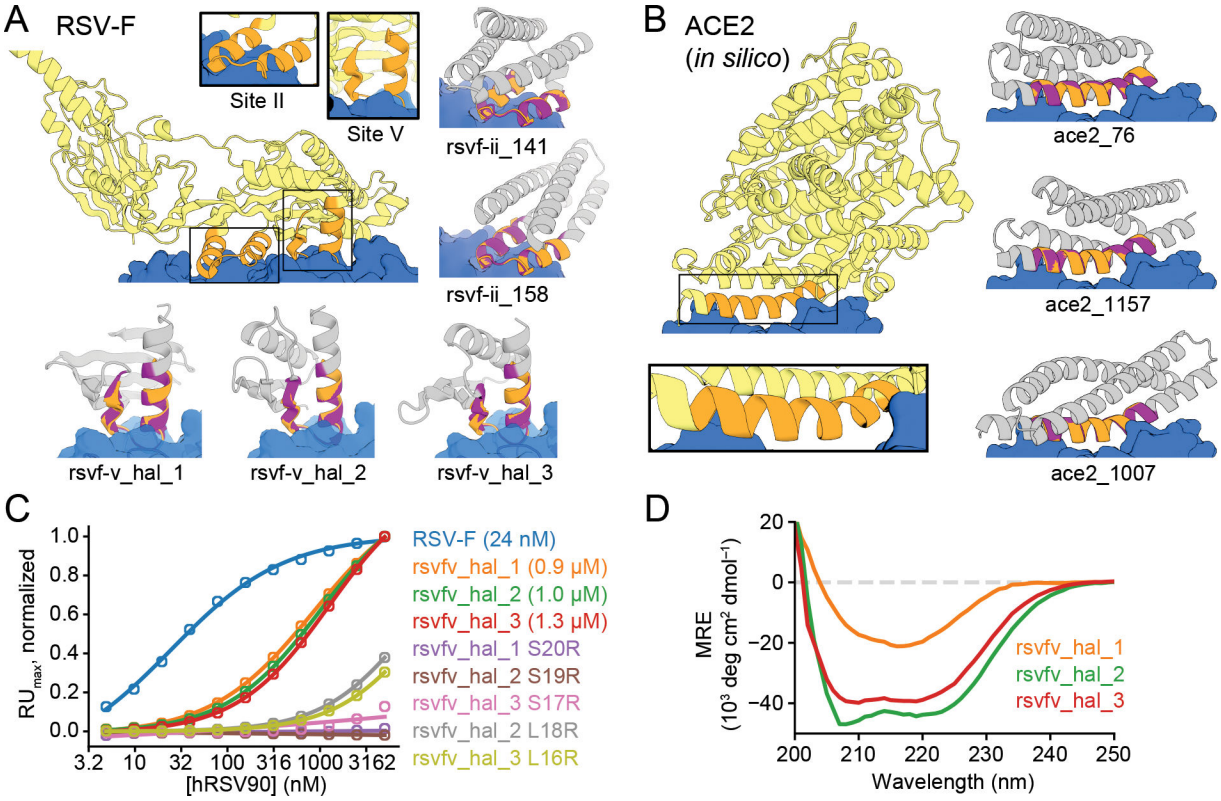

Figure 2. Design of epitope scaffolds and receptor traps.

(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 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 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); rsvfv_hal_1 (82, 0.7 Å); rsvfv_hal_2 (88, 0.64 Å); rsvfv_hal_3 (86, 0.65 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. (C) Normalized maximum SPR 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 for each design are shown in legend. (D) Mean residue ellipticity (MRE) versus wavelength, from CD spectroscopy, for the 3 RSV-F site V hallucinations with binding activity.

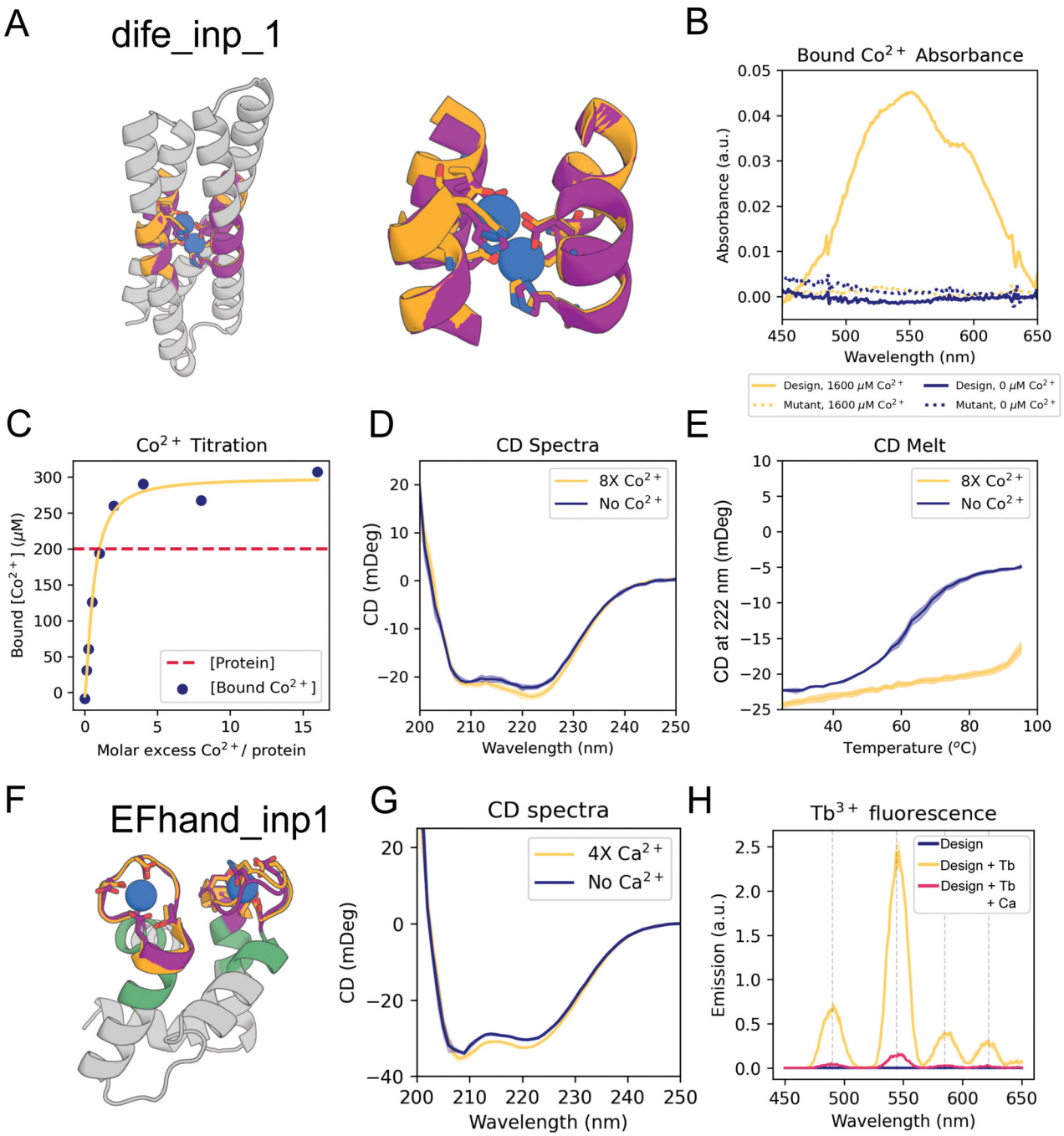

Figure 3. Design of metal binding

(A) Di-iron binding site from E. coli cytochrome b1 (1BCF chain A residues 18–25, 27–54, 94–97, 123–130). Colors: native protein scaffold (light yellow); native functional motif (orange); hallucinated scaffold (gray); hallucinated motif (purple); bound metal (blue). Active site residues shown in boxes for di-iron and EF-hand respectively. (B) Absorbance spectra showing of dife_inp_1 (or mutant) in the presence (or not) of an 8-fold molar excess of Co^{2+}. Note the peaks at 520 nm, 555 nm and 600 nm, consistent with Co^{2+} binding to the desired scaffolded motif (33). The mutant design was the same sequence but with the

6 coordinating residues (sidechains shown in (A)) mutated to alanine [E16A, E55A, H58A, E89A, H92A, E115A]). Protein concentration was 200 μM. (C) Titration analysis of Co^{2+} against the design (protein concentration = 200 μM). Quantification of the absorbance at 550 nm, using a predicted extinction coefficient of 155 for Co^{2+} binding the motif (33), is consistent with both binding sites being recapitulated in the dife_inp_1 design. (D) CD spectra of design in the presence and absence of Co^{2+}. Both spectra are consistent with the predicted helical structure. (E) CD melt curve in the presence and absence of Co^{2+}. Note that the coordination of Co^{2+} in the protein core significantly stabilizes dife_inp_1 (protein concentration in CD experiments = 6.7 μM, Co^{2+} concentration = 53.3 μM). (F) AF2 prediction of inpainted design EFhand_inp_1 scaffolding the double EF-hand motif with input motif residues in purple, input non-motif residues in green, and overlaid with the native motif from 1PRW (orange). (G) Tryptophan-enhanced terbium fluorescence spectra of EFhand_inp_1 matches known spectra (57) and suggests the design can bind terbium. (H) CD spectra of EFhand_inp_1 incubated with (4X protein concentration) and without $CaCl_2$ suggest stabilization of the protein upon binding calcium. Design metrics (AF pLDDT, motif RMSD AF versus native): dife_inp_1 (92 /0.65 Å), EFhand_inp1 (84, 0.7 Å).

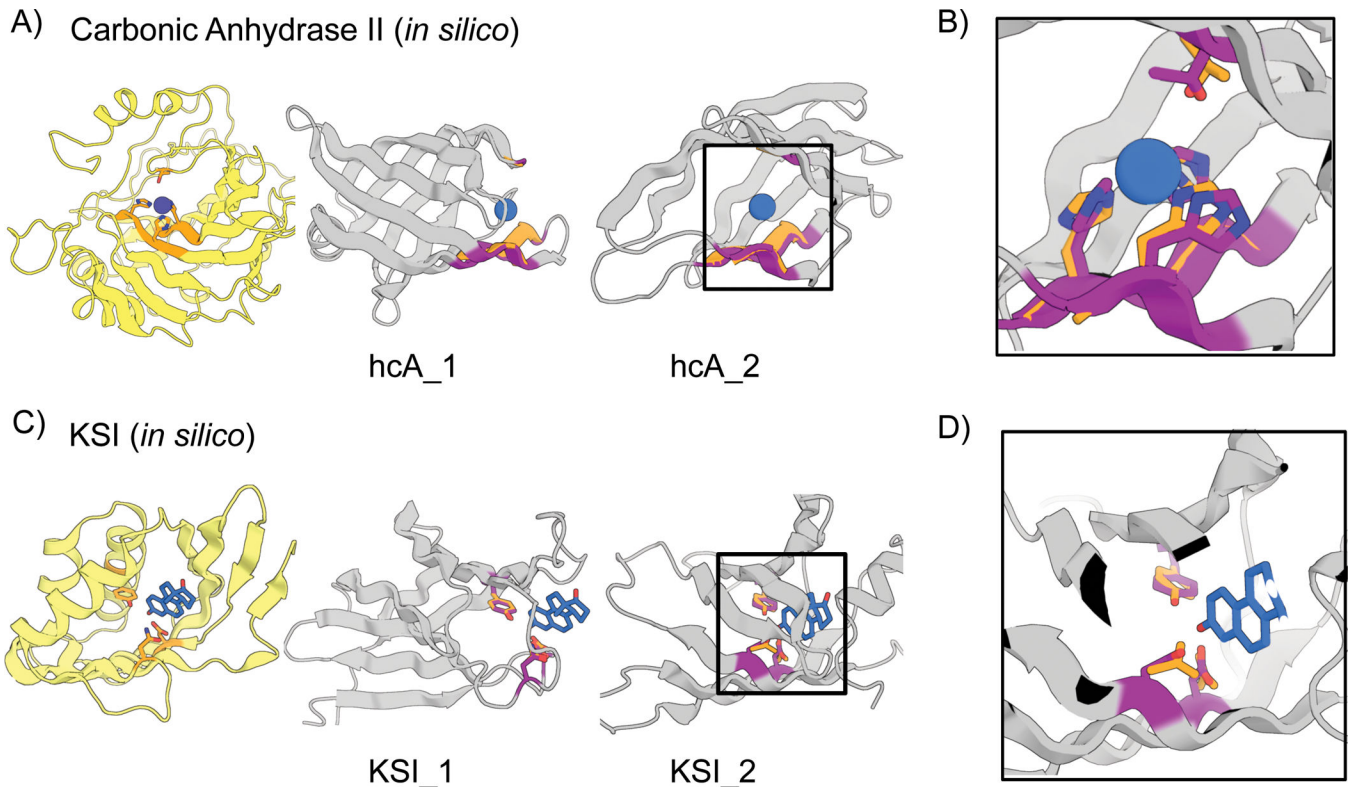

Figure 4. *In silico* design of enzyme active sites.

(A-B) Hallucinations using backbone description of site using RF. (C-D) Hallucination using sidechain description of site using AF2 augmented with trRosetta (Materials and Methods). (A) Carbonic anhydrase II active site (5YUI chain A residues 62–65, 93–97, 118–120). (B) 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 and for carbonic anhydrase II, and Δ^5-3-Ketosteroid Isomerase respectively. Design metrics (AF pLDDT, motif RMSD AF versus native): 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)

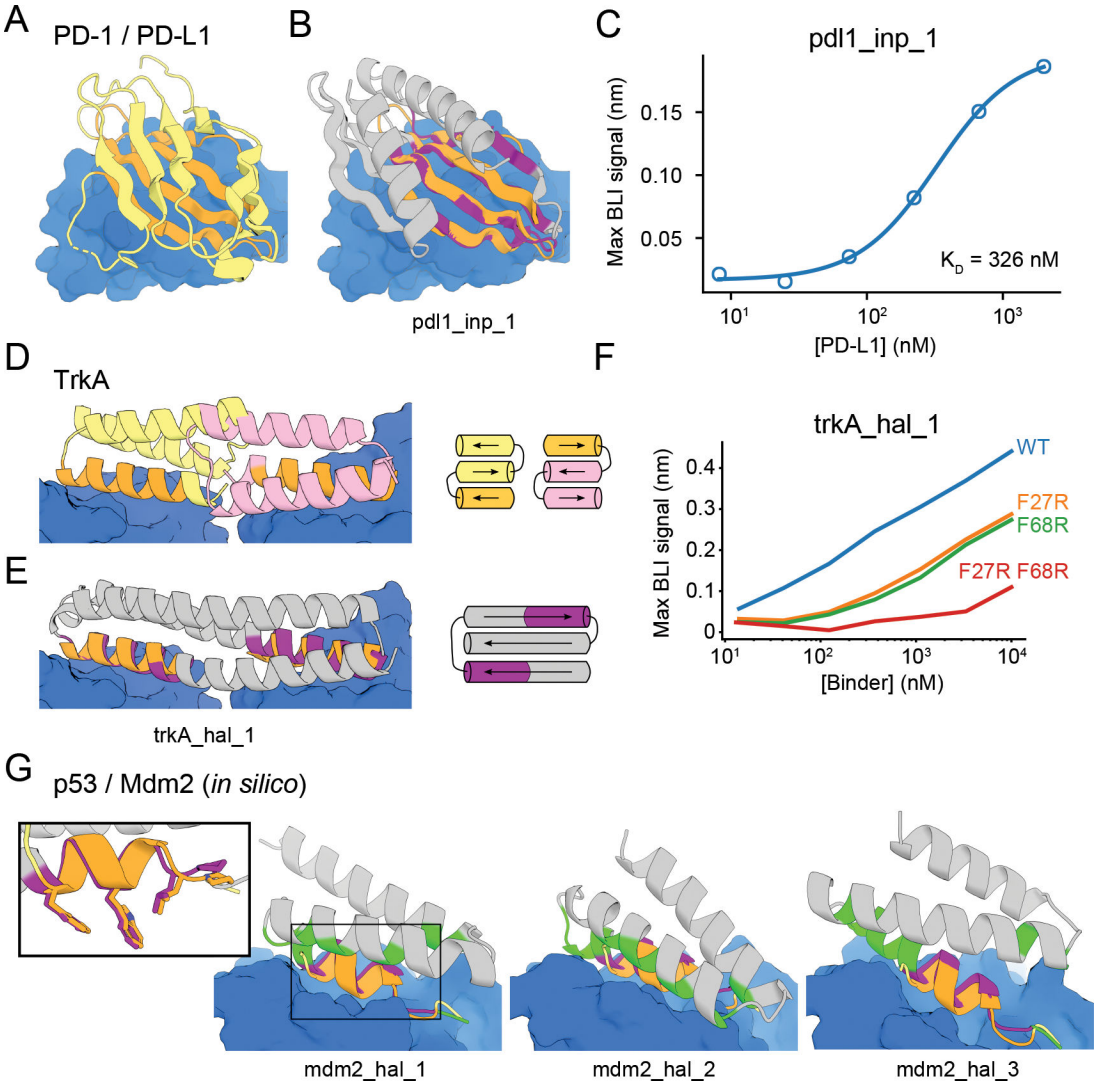

Figure 5. Design of protein-binding proteins.

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 high-affinity consensus (HAC) PD-1 in complex with PD-L1. (B) Inpainted PD-L1 binder superimposed on PD-1 interface motif. (C) Max BLI binding signal versus PD-L1 concentration. (D) Crystal structure of previously designed TrkA minibinder in complex with TrkA, superimposed on TrkA receptor dimer. (E) Hallucinated bivalent TrkA binder. Protein topologies of (D-E) are shown to the right. (F) Max BLI binding signal versus TrkA concentration, showing that both binding sites bind TrkA. (G) Hallucinated Mdm2 binder designs superimposed on native p53 helix in complex with Mdm2 (see also Fig. S17D–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 sidechains for binding.