nature

https://doi.org/10.1038/s41586-023-06415-8

Accelerated Article Preview

# De novo design of protein structure and function with RFdiffusion

Received: 14 December 2022

Accepted: 7 July 2023

Accelerated Article Preview

Cite this article as: Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* https://doi.org/10.1038/s41586-023-06415-8 (2023)

**Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek & David Baker**

# De novo design of protein structure and function with RFdiffusion

Joseph L. Watson[‡1,2], David Juergens[‡1,2,3], Nathaniel R. Bennett[‡1,2,3], Brian L. Trippe[‡2,4,5], Jason Yim[‡2,6], Helen E. Eisenach[‡1,2], Woody Ahern[‡1,2,7], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,9], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,8], Isaac Sappington[1,2,8], Susana Vázquez Torres[1,2,8], Anna Lauko[1,2,8], Valentin De Bortoli[9], Emile Mathieu[10], Sergey Ovchinnikov[14,15], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[12], David Baker*[1,2,11]

‡Equal contribution
*To whom correspondence should be addressed

1. Department of Biochemistry, University of Washington, Seattle, WA 98105, USA
2. Institute for Protein Design, University of Washington, Seattle, WA 98105, USA
3. Graduate Program in Molecular Engineering, University of Washington, Seattle, WA 98105, USA
4. Columbia University, Department of Statistics, New York, NY 10027, USA
5. Irving Institute for Cancer Design, Columbia University, New York, NY 10027, USA
6. Massachusetts Institute of Technology, Cambridge, MA 02139, USA
7. Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98105, USA
8. Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA
9. Centre National de la recherche scientifique, École Normale Supérieure rue d'Ulm, Paris 75005, France
10. Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom
11. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA
12. School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea
13. Faculty of Applied Sciences, Harvard University, Cambridge, MA 01451, USA
14. John Harvard Distinguished Science Fellowship, Harvard University, Cambridge, MA 01451, USA

**There has been considerable recent progress in designing new proteins using deep learning methods[1–9]. Despite this progress, a general deep learning framework for protein design that enables solution of a wide range of design challenges, including *de novo* binder design and design of higher order symmetric architectures, has yet to be described. Diffusion models[10,11] have had considerable success in image and language generative modeling but limited success when applied to protein modeling, likely due to the complexity of protein backbone geometry and sequence-structure relationships. Here we show that by fine tuning the RoseTTAFold structure prediction network on protein structure denoising tasks, we obtain a generative model of protein backbones that**

**achieves outstanding performance on unconditional and topology-constrained protein monomer design, protein binder design, symmetric oligomer design, enzyme active site scaffolding, and symmetric motif scaffolding for therapeutic and metal-binding protein design. We demonstrate the power and generality of the method, called RoseTTAFold Diffusion (RFdiffusion), by experimentally characterizing the structures and functions of hundreds of designed symmetric assemblies, metal binding proteins and protein binders. The accuracy of RFdiffusion is confirmed by the cryo-EM structure of a designed binder in complex with Influenza hemagglutinin which is nearly identical to the design model. In a manner analogous to networks which produce images from user-specified inputs, RFdiffusion enables the design of diverse functional proteins from simple molecular specifications.**

# Main

*De novo* protein design seeks to generate proteins with specified structural and/or functional properties, for example making a binding interaction with a given target[12], folding into a particular topology[13], or stabilizing a desired functional "motif" (geometries and amino acid identities that produce a desired activity)[4]. Denoising diffusion probabilistic models (DDPMs), a powerful class of machine learning models recently demonstrated to generate novel photorealistic images in response to text prompts[14,15], have several properties well-suited to protein design. First, DDPMs generate highly diverse outputs, as they are trained to denoise data (for instance images or text) that have been corrupted with Gaussian noise. By learning to stochastically reverse this corruption, diverse outputs closely resembling the training data are generated. Second, DDPMs can be guided at each step of the iterative generation process towards specific design objectives through provision of conditioning information. Third, for almost all protein design applications it is necessary to explicitly model 3D structure; rotationally-equivariant DDPMs are able to do this in a global representation frame independent manner. Recent work has adapted DDPMs for protein monomer design by conditioning on small protein "motifs"[5,9] or on secondary structure and block-adjacency ("fold") information[8]. While promising, these attempts have shown limited success in generating sequences that fold to the intended structures *in silico*[5,16], likely due to the limited ability of the denoising networks to generate realistic protein backbones, and have not been tested experimentally.

We reasoned that improved diffusion models for protein design could be developed by taking advantage of the deep understanding of protein structure implicit in powerful structure prediction methods like AlphaFold2[17] (AF2) and RoseTTAFold[18] (RF). RF has properties well suited for use in a protein design DDPM (Fig. 1A): it generates protein structures with high precision, operates on rigid-frame representation of residues with rotational equivariance, and has an architecture enabling conditioning on design specifications at the individual residue, inter-residue distance and orientations, and 3D coordinate levels. In previous work, we fine-tuned RF to complete protein backbones around input functional motifs in a *single* step ($RF_{joint}$ Inpainting[4]). Experimental characterization showed that the method can scaffold a wide range of protein functional motifs with atomic accuracy[19], but the approach fails on minimalist site descriptions

that do not sufficiently constrain the overall fold, and because it is deterministic, can produce only a limited diversity of designs for a given problem. We reasoned that by instead fine-tuning RoseTTAFold as the denoising network in a generative diffusion model, we could overcome both problems: because the starting point is random noise, each denoising trajectory yields a different solution, and because structure is built up progressively through many denoising iterations, little to no starting structural information should be required. In this study we used an updated version of RoseTTAFold[18] as the basis for the denoising network architecture (see Methods section 1), but other equivariant structure prediction networks (AF2[17], OmegaFold[20], ESMFold[21]) could in principle be substituted into an analogous DDPM.

We construct a RoseTTAFold-based diffusion model, RFdiffusion, using the RF frame representation which comprises a Cα coordinate and N-Cα-C rigid orientation for each residue. We generate training inputs by noising structures sampled from the Protein Data Bank (PDB) for up to 200 steps[22]. For translations, we perturb Cα coordinates with 3D Gaussian noise. For residue orientations, we use Brownian motion on the manifold of rotation matrices (building on refs [23,24]). To enable RFdiffusion to learn to reverse each step of the noising process, we train the model by minimizing a mean squared error (MSE) loss between frame predictions and the *true* protein structure (without alignment), averaged across all residues (Methods 2.5). This loss drives denoising trajectories to match the data distribution at each timestep and hence to converge on structures of designable protein backbones (Extended Data Fig. 2A). MSE contrasts to the loss used in RF structure prediction training ("frame aligned point error", FAPE) in that unlike FAPE, MSE loss is not invariant to the global reference frame and therefore promotes continuity of the global coordinate frame between timesteps (Methods 2.5).

To generate a new protein backbone, we first initialize random residue frames and RFdiffusion makes a denoised prediction. Each residue frame is updated by taking a step in the direction of this prediction with some noise added to generate the input to the next step. The nature of the noise added and the size of this reverse step is chosen such that the denoising process matches the distribution of the noising process (Methods 2.2-2.3, Extended Data Fig. 2A). RFdiffusion initially seeks to match the full breadth of possible protein structures compatible with the purely random frames with which it is initialized, and hence the denoised structures do not initially appear protein-like (Fig. 1C left). However, through many such steps, the breadth of possible protein structures from which the input could have arisen narrows, and RFdiffusion predictions come to closely resemble protein structures (Fig. 1C right). We use the ProteinMPNN network[1] to subsequently design sequences encoding these structures, typically sampling 8 sequences per design, in line with previous work[5,16] (but see Supplementary Information Fig. 2A). We also considered simultaneously designing structure and sequence within RFdiffusion, but given the excellent performance of combining ProteinMPNN with the diffusion of structure alone, we did not extensively explore this possibility.

Fig. 1A highlights the similarities between RoseTTAFold structure prediction and an RFdiffusion denoising step: in both cases, the networks transform coordinates into a predicted structure, conditioned on inputs to the model. In RoseTTAFold, sequence is the primary input, with additional structural information provided as templates and initial coordinates to the model. In

RFdiffusion, the primary input is the noised coordinates from the previous step. For design tasks, we optionally provide a range of auxiliary conditioning information, including partial sequence, fold information, or fixed functional motif coordinates (Fig. 1B, Methods 3 and 5.16).

We explored two different strategies for training RFdiffusion: 1) in a manner akin to "canonical" diffusion models, with predictions at each timestep independent of predictions at previous timesteps (as in previous work[5,8,9,16]), and 2) with self-conditioning[25], where the model can condition on previous predictions between timesteps (Fig. 1A bottom row, Methods 2.4). The latter strategy was inspired by the success of "recycling" in AF2, which is also central to the more recent RF model used here (Methods 1). Self-conditioning within RFdiffusion dramatically improved performance on *in silico* benchmarks encompassing both conditional and unconditional protein design tasks (Fig. 2E, Extended Data Fig. 1E). Increased coherence of predictions within self-conditioned trajectories may, at least in part, explain these performance increases (Extended Data Fig. 1H). Fine-tuning RFdiffusion from pre-trained RF weights was far more successful than training for an equivalent length of time from untrained weights (Extended Data Fig. 1F-G, see also Supplementary Information Fig. 1) and the MSE loss was also crucial for unconditional generation (Extended Data Fig. 1D). For all *in silico* benchmarks in this paper, we use the AF2 structure prediction network[17] for validation and define an *in silico* "success" as an RFdiffusion output for which the AF2 structure predicted from a single sequence is (1) of high confidence (mean predicted aligned error, pAE, < 5), (2) globally within 2Å backbone-RMSD of the designed structure, and (3) within 1Å backbone-RMSD on any scaffolded functional-site (Methods 5.3). This measure of *in silico* success has been found to correlate with experimental success[4,7,26] and is significantly more stringent than TM-score based metrics used elsewhere (refs [5,16,27–29], Supplementary Information Fig. 2C-D).

***Unconditional protein monomer generation***

As illustrated in Fig. 2A-C, Supplementary Information Fig. 3C-D, starting from random noise, RFdiffusion can readily generate elaborate protein structures with little overall structural similarity to structures seen during training, indicating considerable generalization beyond the PDB (see Supplementary Information Table 1 for comparison of all designs in the paper to the PDB). The designs are diverse (Supplementary Information Fig. 3A), spanning a wide range of alpha-, beta- and mixed alpha-beta- topologies, with AF2 and ESMFold (Fig. 2C, Extended Data Fig. 1B-C, Supplementary Information Fig. 2B) predictions very close to the design structure models for *de novo* designs with as many as 600 residues. RFdiffusion generates plausible structures for even very large proteins, but these are difficult to validate *in silico* as they are likely generally beyond the single sequence prediction capabilities of AF2 and ESMFold. The quality and diversity of designs that are sampled is inherent to the model, and does not depend on *any* auxiliary conditioning input (for example secondary structure information[8]). We experimentally characterized 6 of the 300 amino acid designs and 3 of the 200 amino acid designs, and found that they have circular dichroism (CD) spectra consistent with the mixed alpha-beta topologies of the designs and are extremely thermostable (Extended Data Fig. 3). Physics-based protein design methodologies have struggled in unconstrained generation of diverse protein monomers due to the difficulty of sampling on the very large and rugged conformational landscape[30], and overcoming this limitation has been a primary test of deep

learning based protein design approaches[5,6,8,16,27,31]. RFdiffusion strongly outperforms Hallucination with RoseTTAFold, an experimentally validated method using Monte Carlo search or gradient descent to identify sequences predicted to fold into stable structures (Fig. 2D). RFdiffusion generation is also more compute efficient than unconstrained Hallucination with RoseTTAFold, and efficiency can be dramatically improved by taking larger steps at inference time, and by truncating trajectories early, which is possible because RF predicts the *final* structure at each timestep (Extended Data Fig. 2B-C). For example, a 100 residue protein can be generated in as little as 11s on an NVIDIA RTX A4000 GPU, in contrast to RoseTTAFold Hallucination that takes around 8.5 minutes.

It is often desirable to be able to specify a protein fold during design (such as TIM barrels or cavity-containing NTF2s for small molecule binder and enzyme design[32,33]), and thus we further fine-tuned RFdiffusion to condition on secondary structure and/or fold information, enabling rapid and accurate generation of diverse designs with the desired topologies (Fig. 2G, Extended Data Fig. 4). *In silico* success rates were 42.5% and 54.1% for TIM barrels and NTF2 folds respectively (Extended Data Fig. 4D), and experimental characterization of 11 TIM barrel designs indicated that at least 8 designs were soluble, thermostable, and had circular dichroism (CD) spectra consistent with the design model (Fig. 2G, Extended Data Fig. 4E-F).

***Design of higher order oligomers***

There is considerable interest in designing symmetric oligomers, which can serve as vaccine platforms[34], delivery vehicles[35], and catalysts[36]. Cyclic oligomers have been designed using structure prediction networks with an adaptation of Hallucination that searches for sequences predicted to fold to the desired cyclic symmetry, but this approach fails for higher order dihedral, tetrahedral, octahedral, and icosahedral symmetries, likely in part because of the much lower representation of such structures in the PDB[7].

We set out to generalize RFdiffusion to create symmetric oligomeric structures with any specified point group symmetry. Given a specification of a point group symmetry for an oligomer with N chains, and the monomer chain length, we generate random starting residue frames for a single monomer subunit as in the unconditional generation case, and then generate N-1 copies of this starting point arranged with the specified point group symmetry. Because RFdiffusion is equivariant (inherited from RF) with respect to rotation and relabelings of chains, symmetry is largely maintained in the denoising predictions; we explicitly re-symmetrize at each step but this changes the structures only slightly (compare gray and colored chains in Extended Data Fig. 5A, Methods Proposition 2). For octahedral and icosahedral architectures, we explicitly model only the smallest subset of monomers required to generate the full assembly (e.g. for icosahedra, the subunits at the five-fold, three-fold, and two-fold symmetry axes) to reduce the computational cost and memory footprint.

Despite not being trained on symmetric inputs, RFdiffusion is able to generate symmetric oligomers with high *in silico* success rates (Extended Data Fig. 5B), particularly when guided by an auxiliary inter- and intra-chain contact potential (Extended Data Fig. 5C). As illustrated in Fig. 3 and Extended Data Fig. 5E, RFdiffusion designs are nearly indistinguishable from AF2

predictions of the structures adopted by the designed sequences, and many have little resemblance to previously solved protein structures (Extended Data Fig. 5D, Supplementary Information Table 1). A number of the oligomeric topologies are not seen in the PDB, including two-layer beta barrels (Fig. 3A, C10 symmetry) and complex mixed alpha/beta topologies (Fig. 3A, C8 symmetry; closest TM align in PDB: 6BRP, 0.47; 6BRO, 0.43 respectively).

We selected 608 designs for experimental characterization and found using size exclusion chromatography (SEC) that at least 87 had oligomerization states closely consistent with the design models (within the 95% confidence interval, 126 designs within the 99% CI, as determined by SEC calibration curves; Supplementary Information Fig. 4-5). We took advantage of the increased size of these oligomers (as compared to the smaller unconditional and fold-conditioned monomers described above) and collected negative stain electron microscopy (nsEM) data on a subset of these designs across different symmetry groups. For most, distinct particles were evident with shapes resembling the design models in both the raw micrographs and subsequent 2D classifications (Fig. 3, and Extended Data Fig. 5F). nsEM characterization of a C3 design (HE0822) with 350 residue subunits (1050 residues in total) suggests that the actual structure is very close to the design, both over the 350 residue subunits and the overall C3 architecture. 2D class averages are clearly consistent with both top- and side-views of the design model, and a 3D reconstruction of the density has key features consistent with the design, including the distinctive pinwheel shape (Fig. 3B, top row). Electron microscopy 2D class averages of C5 and C6 designs with greater than 750 residues (HE0794, HE0789, HE0841) were also consistent with the respective design models (Extended Data Fig. 5F).

RFdiffusion also generated cyclic oligomers with alpha/beta barrel structures that resemble expanded TIM barrels and provide an interesting comparison between innovation during natural evolution and innovation through deep learning. The TIM barrel fold, with 8 strands and 8 helices, is one of the most abundant folds in nature[37]. nsEM confirmed the structure of two RFdiffusion designed cyclic oligomers which considerably extend beyond this fold (Fig. 3B, bottom rows). HE0626 is a C6 alpha/beta barrel composed of 18 strands and 18 helices, and HE0675 is a C8 octamer composed of an inner ring of 16 strands and an outer ring of 16 helices arranged locally in a very similar repeating pattern to the TIM barrel (1:1 helix:strand). For both HE0626 and HE0675 we obtained nsEM 3D reconstructions that are in agreement with the computational design models. The HE0600 design is also an alpha-beta barrel (Extended Data Fig. 5F), but has two strands for every helix (24 strands and 12 helices in total) and is hence locally quite different from a TIM barrel. Whereas natural evolution has extensively explored structural variations of the classic 8-strand/8-helix TIM barrel fold, RFdiffusion can more readily explore global changes in barrel curvature, enabling discovery of TIM barrel-like structures with many more helices and strands.

RFdiffusion also readily generated structures with dihedral, tetrahedral and icosohedral symmetries (Fig. 3C-D, Fig. Extended Data Fig. 5E,F). SEC characterization indicated that 38 D2, 7 D3, and 3 D4 designs had the expected molecular weights (these have 4, 6, and 8 chains, respectively) (Supplementary Information Fig. 5). While the D2 dihedrals are too small for nsEM, 2D class averages–and for some, 3D reconstructions of D3 and D4 designs were congruent

with the overall topologies of the design models (Fig. 3C, Extended Data Fig. 5F). Similarly, 3D reconstruction (Fig. 3C) and cryogenic electron microscopy (cryo-EM) 2D class averages (Extended Data Fig. 5G, Supplementary Information Fig. 6) of the D4 HE0537 closely match the design model, recapitulating the approximate 45° offset between tetramic subunits. 2D nsEM class averages for a 12 chain tetrahedron (HE0964) were consistent with the design model (Extended Data Fig. 5F). 48 icosahedra were selected for experimental validation, and one, HE0902, a 15nm (diameter) highly-porous assembly (Fig. 3D, left) was observed in nsEM micrographs to form homogeneous particles. 2D class averages and a 3D reconstruction very closely match the design model (Fig. 3D), with triangular hubs arrayed around the empty C5 axes. Designs such as HE0902 (and future similar large assemblies) should be useful as new nanomaterials and vaccine scaffolds, with robust assembly and (in the case of HE0902) the outward facing N- and C-termini offering multiple possibilities for antigen display.

***Functional motif scaffolding***

We next investigated the use of RFdiffusion for scaffolding protein structural motifs that carry out binding and catalytic functions, where the role of the scaffold is to hold the motif in precisely the 3D geometry needed for optimal function. In RFdiffusion, we input motifs as 3D coordinates (including sequence and sidechains) both during conditional training and inference, and build scaffolds that hold the motif atomic coordinates in place. A number of deep learning methods have been developed recently to address this problem, including $RF_{joint}$ Inpainting[4], constrained Hallucination[4], and other DDPMs[5,8,29]. To rigorously evaluate the performance of these methods in comparison to RFdiffusion across a broad set of design challenges, we established an *in silico* benchmark test (Supplementary Methods Table 9) comprising 25 motif-scaffolding design problems addressed in six recent publications encompassing several design methodologies[4,5,29,38–40]. The challenges span a broad range of motifs, including simple "inpainting" problems, viral epitopes, receptor traps, small molecule binding sites, binding interfaces and enzyme active sites.

RFdiffusion solves 23 of the 25 benchmark problems, compared to 15 for Hallucination and 19 for $RF_{joint}$ Inpainting (Fig. 4A-B). For 19/23 of the problems solved by RFdiffusion, the fraction of successful designs is higher than either Hallucination or $RF_{joint}$ Inpainting. The excellent performance of RFdiffusion required no hyperparameter tuning or external potentials; this contrasts with Hallucination, for which problem-specific optimization can be required. In 17/23 of the problems, RFdiffusion generated successful solutions with higher *in silico* success rates when noise was not added during the reverse diffusion trajectories (see Extended Data Fig. 1I for further discussion of the effect of noise on design quality, and Supplementary Information Fig. 8 for analysis of design diversity). The ability of RFdiffusion to scaffold functional motifs is not related to their presence in the RFdiffusion training set (Supplementary Information Fig. 7).

One of the benchmark problems is the scaffolding of the p53 helix that binds MDM2. Inhibiting this interaction through high-affinity competitive inhibition by scaffolding the p53 helix and making additional interactions with MDM2 is a promising therapeutic avenue[41]. *In silico* success has been described elsewhere[4], but experimental success has not been reported. We used an RFdiffusion model fine-tuned on protein complexes (Methods 5.11) to generate 96 designs

scaffolding this helix. We scaffolded the p53 helix in the presence of MDM2, so additional interactions could be designed by RFdiffusion, and experimentally identified 0.5nM and 0.7nM binders (Fig. 4C-D), three orders of magnitude higher affinity than the reported 600nM affinity of the p53 peptide alone[42]. 55 of the 96 designs showed some detectable binding at 10μM (Fig. 4E, Supplementary Information Fig. 10H).

***Scaffolding enzyme active sites***

A grand challenge in protein design is to scaffold minimal descriptions of enzyme active sites comprising a few single amino acids. While some *in silico* success has been reported previously[4], a general solution that can readily produce high-quality, orthogonally-validated outputs remains elusive. Following fine-tuning on a task mimicking this problem (Methods 4.2), RFdiffusion was able to scaffold enzyme active sites comprising multiple sidechain and backbone functional groups with high accuracy and *in silico* success rates across a range of enzyme classes (Fig. 4F, Extended Data Fig. 6A-D; *in silico* successes were not present without fine-tuning). While RFdiffusion is currently unable to *explicitly* model bound small molecules (see conclusion), the substrate can be *implicitly* modeled using an external potential to guide the generation of "pockets" around the active site. As a demonstration, we scaffold a retroaldolase active site triad while implicitly modeling its substrate (Extended Data Fig. 6E-H).

***Symmetric functional-motif scaffolding***

A number of important design challenges involve the scaffolding of multiple copies of a functional motif in symmetric arrangements. For example, many viral glycoproteins are trimeric, and symmetry matched arrangements of inhibitory domains can be extremely potent[43–46]. Conversely, symmetric presentation of viral epitopes in an arrangement that mimics the virus could induce new classes of neutralizing antibodies[47,48]. To explore this general direction, we sought to design trimeric multivalent binders to the SARS-CoV-2 spike protein. In previous work, flexible linkage of a binder to the ACE2 binding site (on the spike protein receptor binding domain) to a trimerization domain yielded a high-affinity inhibitor that had potent and broadly neutralizing antiviral activity in animal models[43]. Ideally, however, symmetric fusions to binders would be rigid, so as to reduce the entropic cost of binding while maintaining the avidity benefits from multivalency. We used RFdiffusion to design C3 symmetric trimers which rigidly hold three binding domains (the functional motif in this case) such that they exactly match the ACE2 binding sites on the SARS-CoV-2 spike protein trimer. Design models were confidently predicted by AF2 to both assemble as C3-symmetric oligomers, and to scaffold the AHB2 SARS-CoV-2 binder interface with high accuracy (Fig. 5A).

The ability to scaffold functional sites with any desired symmetry opens up new approaches to designing metal-coordinating protein assemblies[49,50]. Divalent transition metal ions exhibit distinct preferences for specific coordination geometries (e.g., square planar, tetrahedral, and octahedral) with ion-specific optimal sidechain–metal bond lengths. RFdiffusion provides a general route to building up symmetric protein assemblies around such sites, with the symmetry of the assembly matching the symmetry of the coordination geometry. As a first test, we sought to design square planar $Ni^{2+}$ binding sites. We designed C4 protein assemblies with four central histidine imidazoles arranged in an ideal $Ni^{2+}$-binding site with square planar coordination

geometry (Fig. 5B). Diverse designs starting from distinct C4-symmetric histidine square planar sites had good *in silico* success with the histidine residues in near ideal geometries for coordinating metal in the AF2 predicted structures (Supplementary Information Fig. 9).

We expressed and purified 44 designs in *E. coli.*, and found that 37 had SEC chromatograms consistent with the intended oligomeric state (Extended Data Fig. 7B). 36 were tested for $Ni^{2+}$ coordination by isothermal titration calorimetry, and 18 were found to bind $Ni^{2+}$ with dissociation constants ranging from low nanomolar to low micromolar (Fig. 5C,D and Extended Data Fig. 7A). The inflection points in the wild-type isotherms indicate binding with the designed stoichiometry, a 1:4 ratio of ion:monomer. While most of the designed proteins displayed exothermic metal coordination, in a few cases binding was endothermic (Fig. 5D, left, Extended Data Fig. 7A: NiB2.9, NiB2.10, NiB2,15, NiB2.23), suggesting that $Ni^{2+}$ coordination is entropically driven in these assemblies. To confirm that $Ni^{2+}$ binding was indeed mediated by the scaffolded histidine 52, we mutated this residue to alanine, which abolished or dramatically reduced binding in 17/17 cases with successful expression (Extended Data Fig. 7A,C and Fig. 5C,D; one mutant did not express). We structurally characterized by nsEM a subset of the designs – NiB1.12, NiB1.15, NiB1.17, and NiB1.20 – that displayed histidine-dependent binding. All four designs exhibited clear 4-fold symmetry both in the raw micrographs and in 2D class averages (Fig. 5C-D), with design NiB1.17 also clearly displaying 2-fold axis “side-views” with a measured diameter approximating the design model. A 3D reconstruction of NiB1.17 was in close agreement to the design model (Fig. 5C).

***Design of protein-binding proteins***

The design of high-affinity binders to target proteins is a grand challenge in protein design, with numerous therapeutic applications[51]. A general method to *de novo* binder design from target structure information alone using the physically-based Rosetta method was recently described[12], and subsequently, utilizing ProteinMPNN for sequence design and AF2 for design filtering was found to improve design success rates[26]. However, experimental success rates were low, still requiring many thousands of designs to be screened for each design campaign[12], and the approach relied on pre-specifying a particular set of protein scaffolds as the basis for the designs, inherently limiting the diversity and shape complementarity of possible solutions[12]. To our knowledge, no deep-learning method has yet demonstrated experimental general success in designing completely *de novo* binders.

We reasoned that RFdiffusion might be able to address this challenge by directly generating binding proteins in the context of the target. For many therapeutic applications, for example blocking a protein-protein interaction, it is desirable to bind to a particular site on a target protein. To enable this, we fine-tuned RFdiffusion on protein complex structures, providing as input a subset of the residues on the target chain (called “interface hotspots”) to which the diffused chain binds (Fig. 6A, Extended Data Fig. 8A,B). For design cases where a particular binder fold might be especially compatible, we enabled coarse-grained control over binder scaffold topology by fine-tuning an additional model to condition binder diffusion on secondary structure and block-adjacency information, in addition to conditioning on interface hotspots (Fig. Extended Data Fig. 8C-D, Methods 4.3).

To compare RFdiffusion to previous binder design methods, we performed binder design campaigns against 5 targets: Influenza A H1 Hemagglutinin (HA)[52], Interleukin-7 Receptor-α (IL-7Rα)[12], Programmed Death-Ligand 1 (PD-L1)[12], Insulin Receptor, and Tropomyosin Receptor Kinase A (TrkA)[12]. We designed putative binders to each target, both with and without conditioning on compatible fold information, with high *in silico* success rates (Extended Data Fig. 8E,F). Designs were filtered by AF2 confidence in the interface and monomer structure[26], and 95 were selected for each target for experimental characterization.

The designed binders were expressed in *E. coli* and purified, and binding was assessed through single point biolayer interferometry (BLI) screening at 10μM binder concentration (Extended Data Fig. 8G). The overall experimental success rate, defined as binding at or above 50% of the maximal response for the positive control, was 19% (this is a conservative estimate as some designs which showed binding had insufficient material to permit screening at 10μM (Extended Data Fig. 8G)); an increase of approximately 2 orders-of-magnitude over our previous Rosetta-based method on the same targets (Fig. 6B). Binders were identified for all 5 targets, with fewer than 100 designs tested per target compared to thousands in previous studies. Full BLI titrations for a subset of the designs showed nanomolar affinities with no further experimental optimization, including HA and IL-7Rα binders with affinities of approximately 30nM (Fig. 6C). Binding interfaces were often highly distinct from interfaces to these targets in the PDB (Supplementary Information Figs. 11, 12). To assess binder specificity, 6 of the highest affinity IL-7Rα binders were assessed via competition BLI, and all 6 competed for binding with a structurally validated positive control binding to the same site (Supplementary Information Fig. 10A; further work is required to fully characterize proteome-wide specificity).

We solved the structure of the highest affinity Influenza binder, *HA_20*, in complex with Iowa43 HA using cryo electron microscopy. Raw electron micrographs revealed a well-folded HA glycoprotein with clearly discernible side, top, and tilted view orientations suspended in a thin layer of vitreous ice (Extended Data Fig. 9A). 2D class averages further show clear secondary structure elements corresponding to both Iowa43 HA (Extended Data Fig. 9B), as well as the *HA_20* binder bound to the stem (Fig 6E). 3D heterogenous refinement without symmetry revealed full occupancy of all three HA stem epitopes by the *HA_20* binder. A final non-uniform 3D refinement reconstruction with C3 symmetry yielded a 2.9 Å map of the HA/*HA_20* protein-protein complex (Fig 6F) and corresponding 3D structure which nearly perfectly matches the computational design model (0.63Å, Fig 6F,G; the sidechain interactions at the interface are very different from the closest structure in the PDB; Extended Data Fig. 9H). Over the binder alone, the experimental structure deviates from the RFdiffusion design by only 0.6Å (Fig 6H). These results demonstrate the ability of RFdiffusion to generate new proteins with atomic level accuracy, and to precisely target functionally relevant sites on therapeutically important proteins.

## Discussion

RFdiffusion is a comprehensive improvement over current protein design methods. RFdiffusion readily generates diverse unconditional designs up to 600 residue structures that are accurately predicted by AF2, far exceeding the complexity and accuracy achieved by previous methods

(although during review of this manuscript, a hallucination-based approach also achieved high unconditional performance[53]). Half of our tested unconditional designs express solubly and exhibit CD spectra consistent with the design models and high thermostability. Despite their substantially increased complexity, the ideality and stability of RFdiffusion designs is akin to that of previous *de novo* design methods. RFdiffusion enables generation of higher order architectures with any desired symmetry - surpassing Hallucination methods, which have so far been limited to cyclic symmetries. Electron microscopy confirmed that structures of these oligomers are very similar to the design models, and in many cases show little global similarity to known protein oligomers.

There has been recent progress in scaffolding protein functional motifs using deep learning methods (RF Hallucination, $RF_{joint}$ Inpainting, and diffusion), but Hallucination is slow for large systems, inpainting fails when insufficient starting information is provided, and previous diffusion methods had quite low accuracy. RFdiffusion outperforms these previous methods in the complexity of the motifs that can be scaffolded, the precision with which sidechains are positioned (for catalysis and other functions), and the accuracy of motif recapitulation by AF2. The design of MDM2 binding proteins with three orders of magnitude higher affinities than the scaffolded P53 motif demonstrates the robustness of RFdiffusion motif-scaffolding. Combining accurate motif-scaffolding with the design of symmetric assemblies, enabled consistent and atomically precise positioning of sidechains to coordinate $Ni^{2+}$ ions across diverse tetramers.

For binder design from target structural information alone, previous work required screening testing tens of thousands of sequences[12]. RFdiffusion, when combined with improved filtering[26] raises experimental success rates by two orders of magnitude; high affinity binders can be identified from dozens of designs, in many cases eliminating the requirement for slow and expensive high-throughput screening (at least for the somewhat non-polar sites targeted here; further studies will be required to assess success rates on more polar target sites and sites without native binding partners). A high resolution cryo-EM structure of one of these designs in complex with influenza hemagglutinin further shows that RFdiffusion can design functional proteins with atomic accuracy. Vázquez Torres *et al.* demonstrate the ability of RFdiffusion to design picomolar affinity binders to flexible helical peptides[54], further highlighting its utility for *de novo* binder design. Vázquez Torres *et al.* also show how RFdiffusion can be extended for protein model refinement by partial noising and denoising, which enables tunable sampling around a given input structure. For peptide binder design, this enabled increases in affinity of nearly three orders of magnitude without high-throughput screening.

The breadth and complexity of problems solvable with RFdiffusion and the robustness and accuracy of the solutions far exceeds what has been achieved previously. In a manner reminiscent of the generation of images from text prompts, RFdiffusion makes possible, with minimal specialist knowledge, the generation of functional proteins from minimal molecular specifications (for example, high affinity binders to a user-specified target protein, and diverse protein assemblies from user-specified symmetries).

The power and scope of RFdiffusion can be extended in several directions. RF has recently been extended to nucleic acids and protein-nucleic acid complexes[55], which should enable RFdiffusion to design nucleic acid binding proteins, and perhaps folded RNA structures. Extension of RF to incorporate ligands should similarly enable extension of RFdiffusion to explicitly model ligand atoms, and allow the design of protein-ligand interactions. The ability to customize RFdiffusion to specific design challenges by addition of external potentials and by fine-tuning (as illustrated here for catalytic site scaffolding, binder-targeting and fold-specification), along with continued improvements to the underlying methodology, should enable *de novo* protein design to achieve still higher levels of complexity, to approach and – in some cases – surpass what natural evolution has achieved.

# References


1. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
2. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
3. Singer, J. M. *et al.* Large-scale design and refinement of stable proteins using sequence-only models. *PLOS ONE* **17**, e0265020 (2022).
4. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
5. Trippe, B. L. *et al.* Diffusion Probabilistic Modeling of Protein Backbones in 3D for the motif-scaffolding problem. in *The International Conference on Learning Representations* (2023).
6. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
7. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
8. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. Preprint at https://doi.org/10.48550/ARXIV.2205.15019 (2022).
9. Luo, S. *et al.* Antigen-Specific Antibody Design and Optimization with Diffusion-Based

Generative Models. in *Advances in Neural Information Processing Systems* (2022).

10. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. & Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. in *International Conference on Machine Learning* **37**, 2256–2265 (PMLR, 2015).
11. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. in *Advances in Neural Information Processing Systems* vol. 33 6840–6851 (2020).
12. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).
13. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
14. Ramesh, A. *et al.* Zero-Shot Text-to-Image Generation. in *International Conference on Machine Learning* **139**, 8821–8831 (PMLR, 2021).
15. Saharia, C. *et al.* Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. in *Advances in Neural Information Processing Systems* (2022).
16. Wu, K. E. *et al.* Protein structure generation via folding diffusion. Preprint at https://doi.org/10.48550/arXiv.2209.15611 (2022).
17. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
18. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
19. Watson, J. L., Bera, A., Juergens, D., Wang, J. & Baker, D. X-ray crystallographic validation of design from this paper. (2022). ??
20. Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. 2022.07.21.500999 Preprint at https://doi.org/10.1101/2022.07.21.500999 (2022).
21. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *Science* **379**, 1123–1130 (2023).

22. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
23. De Bortoli, V. *et al.* Riemannian Score-Based Generative Modelling. in *Advances in Neural Information Processing Systems* (2022).
24. Leach, A., Schmon, S. M., Degiacomi, M. T. & Willcocks, C. G. Denoising Diffusion Probabilistic Models On SO(3) For Rotational Alignment. in *ICLR 2022 Workshop on Geometrical and Topological Representation Learning* (2022).
25. Chen, T., Zhang, R. & Hinton, G. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. in *International Conference on Learning Representations* (2023).
26. Bennett, N. *et al.* Improving de novo Protein Binder Design with Deep Learning. *Nat. Commun.* **14**, 2022.06.15.495993 (2023).
27. Anand, N. & Huang, P. Generative modeling for protein structures. in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).
28. Ingraham, J. *et al.* Illuminating protein space with a programmable generative model. 2022.12.01.518682 Preprint at https://doi.org/10.1101/2022.12.01.518682 (2022).
29. Lee, J. S. & Kim, P. M. ProteinSGM: Score-based generative modeling for de novo protein design. *bioRxiv* (2022) doi:10.1101/2022.07.13.499967.
30. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).
31. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. AlphaDesign: A de novo protein design framework based on AlphaFold. 2021.10.11.463937 Preprint at https://doi.org/10.1101/2021.10.11.463937 (2021).
32. Basanta, B. *et al.* An enumerative algorithm for de novo design of proteins with diverse pocket structures. *Proc. Natl. Acad. Sci.* **117**, 22135–22145 (2020).
33. Pan, X. *et al.* Expanding the space of protein geometries by computational design of de novo fold families. *Science* **369**, 1132–1136 (2020).

34. Marcandalli, J. *et al.* Induction of Potent Neutralizing Antibody Responses by a Designed Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. *Cell* **176**, 1420-1431.e17 (2019).
35. Butterfield, G. L. *et al.* Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).
36. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).
37. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the (βα)8-Barrel Enzyme Fold. *Chem. Rev.* **105**, 4038–4055 (2005).
38. Sesterhenn, F. *et al.* De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science* **368**, (2020).
39. Yang, C. *et al.* Bottom-up de novo design of functional proteins with complex structural features. *Nat. Chem. Biol.* **17**, 492–500 (2021).
40. Glasgow, A. *et al.* Engineered ACE2 receptor traps potently neutralize SARS-CoV-2. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 28046–28055 (2020).
41. Chène, P. Inhibiting the p53-MDM2 interaction: an important target for cancer therapy. *Nat. Rev. Cancer* **3**, 102–109 (2003).
42. Kussie, P. H. *et al.* Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **274**, 948–953 (1996).
43. Hunt, A. C. *et al.* Multivalent designed proteins neutralize SARS-CoV-2 variants of concern and confer protection against infection in mice. *Sci. Transl. Med.* **14**, eabn1252 (2022).
44. Silverman, J. *et al.* Multivalent avimer proteins evolved by exon shuffling of a family of human receptor domains. *Nat. Biotechnol.* **23**, 1556–1561 (2005).
45. Detalle, L. *et al.* Generation and Characterization of ALX-0171, a Potent Novel Therapeutic Nanobody for the Treatment of Respiratory Syncytial Virus Infection. *Antimicrob. Agents Chemother.* **60**, 6–13 (2016).

46. Strauch, E.-M. *et al.* Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nat. Biotechnol.* **35**, 667–671 (2017).
47. Boyoglu-Barnum, S. *et al.* Quadrivalent influenza nanoparticle vaccines induce broad protection. *Nature* **592**, 623–628 (2021).
48. Walls, A. C. *et al.* Elicitation of Potent Neutralizing Antibody Responses by Designed Protein Nanoparticle Vaccines for SARS-CoV-2. *Cell* **183**, 1367-1382.e17 (2020).
49. Salgado, E. N., Lewis, R. A., Mossin, S., Rheingold, A. L. & Tezcan, F. A. Control of protein oligomerization symmetry by metal coordination: C2 and C3 symmetrical assemblies through Cu(II) and Ni(II) coordination. *Inorg. Chem.* **48**, 2726–2728 (2009).
50. Salgado, E. N. *et al.* Metal templated design of protein interfaces. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1827–1832 (2010).
51. Quijano-Rubio, A., Ulge, U. Y., Walkey, C. D. & Silva, D.-A. The advent of de novo proteins for cancer immunotherapy. *Curr. Opin. Chem. Biol.* **56**, 119–128 (2020).
52. Chevalier, A. *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
53. Frank, C. *et al.* Efficient and scalable de novo protein design using a relaxed sequence space. 2023.02.24.529906 Preprint at https://doi.org/10.1101/2023.02.24.529906 (2023).
54. Torres, S. V. *et al.* De novo design of high-affinity protein binders to bioactive helical peptides. 2022.12.10.519862 Preprint at https://doi.org/10.1101/2022.12.10.519862 (2022).
55. Baek, M., McHugh, R., Anishchenko, I., Baker, D. & DiMaio, F. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. 2022.09.09.507333 Preprint at https://doi.org/10.1101/2022.09.09.507333 (2022).
56. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
57. Ribeiro, A. J. M. *et al.* Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **46**, D618–D623 (2018).

58. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).

# Figure Legends

**Figure 1: Protein design using RFdiffusion**
**A)** Top panel: Diffusion models for proteins are trained to recover corrupted (noised) protein structures and to generate new structures by reversing the corruption process through iterative denoising of initially random noise $X_T$ into a realistic structure $X_0$. Middle panel: RoseTTAFold (RF, left) can be fine-tuned as the denoising network in a DDPM. RFdiffusion (right) is fine-tuned from a *pre-trained* RF network with minimal architectural changes. In RF, the primary input to the model is sequence. In RFdiffusion, the primary input is diffused residue frames. In both cases, the model predicts final 3D coordinates (denoted $\hat{X}_0$ in RFdiffusion). Bottom panel: In RFdiffusion, the model receives its previous prediction as a template input ("self-conditioning", see Methods 2.4). At each timestep "t" of a trajectory (typically 200 steps), RFdiffusion takes $\hat{X}_0^{t+1}$ from the previous step and $X_t$ and then predicts an updated $X_0$ structure ($\hat{X}_0^{t}$). The next coordinate input to the model ($X_{t-1}$) is generated by a noisy interpolation toward $\hat{X}_0^{t}$. **B)** RFdiffusion is broadly applicable for protein design. RFdiffusion generates protein structures either without additional input (top row), or by conditioning on (top to bottom): symmetry specifications; binding targets; protein functional motifs; symmetric functional motifs. In each case random noise, along with conditioning information, is input to RFdiffusion, which iteratively refines that noise until a final protein structure is designed. **C)** An example of an unconditional design trajectory for a 300-residue chain, depicting the input to the model ($X_t$) and the corresponding $\hat{X}_0$ prediction. At early timesteps (high t), $\hat{X}_0$ bears little resemblance to a protein but is refined into a protein structure.

**Figure 2: Outstanding performance of RFdiffusion for monomer generation.**
**A)** RFdiffusion can generate new monomeric proteins of different lengths (left: 300, right: 600) with no conditioning information. Gray=design model; colors= AlphaFold2 (AF2) prediction. RMSD AF2 vs design (Å), left to right: 0.90, 0.98, 1.15, 1.67. **B)** Unconditional designs from RFdiffusion are novel and not present in the training set as quantified by highest TM score to the protein data bank (PDB). Designs are increasingly novel with increasing length. **C)** Unconditional samples are closely re-predicted by AF2. Beyond 400 amino acids, the recapitulation by AF2 deteriorates. **D)** RFdiffusion significantly outperforms Hallucination (with RoseTTAFold) at unconditional monomer generation (two-proportion z-test of *in silico* success: *n*=400 designs per condition, *z=9.5, p=1.6e-21*). While Hallucination successfully generates designs up to 100 amino acids in length, *in silico* success rates rapidly deteriorate beyond this length. **E)** Ablating pre-training (by starting from untrained RF), RFdiffusion fine-tuning (i.e., using original RF structure prediction weights as the denoiser), self-conditioning, or MSE losses (by training with FAPE) each dramatically decrease the performance of RFdiffusion. RMSD between design and AF2 is shown, for the unconditional generation of 300 amino acid proteins

(see Methods 5.8). **F)** Two example 300 amino acid proteins that expressed as soluble monomers. Designs (gray) overlaid with AF2 predictions (colors) are shown on the left, alongside CD spectra (top) and melt curves (bottom) on the right. The designs are highly thermostable. **G)** RFdiffusion can condition on fold information. An example TIM barrel is shown (bottom left), conditioned on the secondary structure and block-adjacency of a previously designed TIM barrel, PDB: 6WVS (top left). Designs have very similar CD spectra to 6WVS (top right) and are highly thermostable (bottom right). See also Extended Data Fig. 3 for additional traces. Boxplots represent median ± IQR; tails: min/max excluding outliers (±1.5x IQR).

**Figure 3: Design and experimental characterization of symmetric oligomers. A)** RFdiffusion-generated assemblies overlaid with the AF2 structure predictions based of the designed sequences; in all 5 cases they are nearly indistinguishable (for the octahedron (bottom), the prediction was for the C3 substructure). Symmetries are indicated to the left of the design models. **B-C)** Designed assemblies characterized by negative stain electron microscopy. Model symmetries: **B)** Cyclic: C3 (HE0822, 350AA/chain); C6 (HE0626, 100AA/chain); C8 (HE0675, 60AA/chain) **C)** Dihedral: D3 (HE0490, 80AA/chain); and D4 (HE0537, 100AA/chain). From left to right: 1) symmetric design model, 2) AF2 prediction of design following sequence design with ProteinMPNN, 3) 2D class averages showing both top and side views (scale bar = 60Å for all class averages), 4) 3D reconstructions from class averages with the design model fit into the density map. The overall shapes are consistent with the design models, and confirm the intended oligomeric state. As in **A**), AF2 predictions of each design are nearly indistinguishable from the design model (backbone RMSDs (Å) for HE0822, HE0626, HE0490, HE0675, and HE0537, are 1.33, 1.03, 0.60, 0.74, and 0.75, respectively). **D)** nsEM characterization of an icosahedral particle (HE0902, 100AA/chain). The design model, including the AF2 prediction of the C3 subunit are shown on the left. nsEM data are shown on the right: on top, a representative micrograph is shown alongside 2D class averages along each symmetry axis (C3, C2, and C5, from left to right) with the corresponding 3D reconstruction map views shown directly below overlaid on the design model.

**Figure 4**: **Scaffolding of diverse functional sites with** RFdiffusion**. A)** RFdiffusion outperforms other methods across 25 benchmark motif scaffolding problems collected from six recent publications (Supplementary Methods Table 9). *In silico* success is defined as AF2 RMSD to design model < 2Å, AF2 RMSD to the native functional motif < 1 Å, and AF2 pAE < 5. 100 designs were generated per problem, with no prior optimization on the benchmark set (some optimization was necessary for Hallucination). Supplementary Methods Table 10 presents full results. *In silico* success rates on the problems are correlated between the methods, and RFdiffusion can still struggle on challenging problems where all methods have low success. B) Four examples of designs where RFdiffusion significantly outperforms existing methods. Teal: native motif; colors: AF2 prediction of a design. Metrics (RMSD AF2 vs design / vs native motif (Å), AF2 pAE): 5TRV Long: 1.17/0.57, 4.73; 6E6R Long: 0.89/0.27, 4.56; 7MRX Long: 0.84/0.82 4.32; 5TPN: 0.59/0.49 3.77. C) RFdiffusion can scaffold the p53 helix that binds MDM2 (left) and makes additional contacts with the target (right, average 31% increased surface area. Design: *p53_design_89*). Designs were generated with an RFdiffusion model fine-tuned on complexes. D) BLI measurements demonstrate high affinity binding to MDM2

(*p53_design_89*: 0.7nM, *p53_design_53*: 0.5nM*)*; the native affinity is 600nM[42]. E) 55/95 designs showed binding to MDM2 (> 50% of maximum response). 32 of these were monomeric (Supplementary Information Fig. 10H.) F) After fine-tuning (Methods 4.2), RFdiffusion can scaffold enzyme active sites. An oxidoreductase example (EC1) is shown (PDB 1A4I); catalytic site (teal); RFdiffusion output (gray: model, colors: AF2 prediction); zoom of active site. AF2 vs design backbone RMSD 0.88Å, AF2 vs design motif backbone RMSD 0.53Å, AF2 vs design motif full-atom RMSD 1.05Å, AF2 pAE 4.47. G) *In silico* success rates on active sites derived from EC1-5 (AF2 Motif RMSD vs native: backbone < 1Å, backbone and sidechain atoms < 1.5Å, RMSD AF2 vs design < 2Å, AF2 pAE < 5).

**Figure 5: Symmetric motif scaffolding with RFdiffusion. A)** Design of symmetric oligomers scaffolding the binding interface of ACE2 mimic AHB2 (left, teal) against the SARS-CoV-2 spike trimer (left, gray). Three AHB2 copies are input to RFdiffusion along with C3 noise (middle); output are C3-symmetric oligomers holding the three AHB2 copies in place to engage all spike subunits. AF2 predictions (right) recapitulate the AHB2 structure with 0.6Å RMSD over the assymetric unit and 2.9Å RMSD over the C3 assembly. **B)** Design of C4-symmetric oligomers to scaffold a $Ni^{2+}$ binding motif (left). Starting from square-planar histidine rotamers within helical fragments (Methods 5.9), RFdiffusion generates a C4 oligomer scaffolding the binding domain (middle). AF2 predictions (color) agree closely with the design model (gray), with backbone RMSD < 1.0 Å (right). **C)** nsEM 2D class averages (scale bar = 60 Å) and 3D reconstruction density are consistent with the symmetry and structure of the NiB1.17 design model shown superimposed on the density in ribbon representation (top). Isothermal titration calorimetry binding isotherm of design NiB1.17 (blue) indicates a dissociation constant < 20 nM at a metal:monomer stoichiometry of 1:4. The H52A mutant isotherm (pink) ablates binding, indicating scaffolded histidine residues are critical for metal binding. **D)** Additional experimentally characterized $Ni^{2+}$ binders NiB2.15 (left), NiB1.12 (middle), and NiB1.20 (right). Metal coordinating sidechains in the design models (top, teal) are closely recapitulated in the AF2 predictions (colors). 2D nsEM class averages (middle, scale bar = 60Å) are consistent with design models. Binding isotherms for wild-type and H52A mutant (bottom) indicate $Ni^{2+}$ binding mediated directly by the scaffolded histidines at the designed stoichiometry. Note that for ITC plots, points represent single measurements.

**Figure 6: *De novo* design of protein-binding proteins.**
**A)** RFdiffusion generates protein binders given a target and specification of interface hotspot residues. **B)** *De novo* binders were designed to five protein targets; Influenza A H1 Hemagglutinin (HA), Interleukin-7 Receptor-α (IL-7Rα), Insulin Receptor (InsR), Programmed Death-Ligand 1 (PD-L1), and Tropomyosin Receptor Kinase A (TrkA) and hits with BLI response ≥ 50% of the positive control were identified for all targets. For IL-7Rα, InsR, PD-L1, and TrkA, RFdiffusion has success rates ~2 orders-of-magnitude higher than the original design campaigns. We attribute one order-of-magnitude to RFdiffusion, and the second to filtering with AF2 (estimated success rates for previous campaigns if AF2 confidence had been used: HA: No designs passed AF2 filter, IL-7Rα: 2.2%, InsR: 5.5%, PD-L1: 3.7%, TrkA: 1.5%). **C)** For IL-7Rα, InsR, PD-L1 and TrkA, the highest affinity binder is shown above a BLI titration series. Reported $K_D$s are based on global kinetic fitting with fixed global $R_{max}$. **D)** The highest affinity HA binder,

*HA_20*, binds with a $K_D$ of 28nM. **C-D)** Yellow/orange: target/hotspot residues; gray: design model; purple: AF2 prediction (RMSD AF2 vs design **C)** left to right: *IL7Ra_55* (2.1Å), *InsulinR_30* (2.6Å), *PDL1_77* (1.5Å), *TrkA_88* (1.4Å) **D)** *HA_20* (1.7Å). **E)** Cryo-EM 2D class averages of *HA_20* bound to Influenza Hemagglutinin, strain A/USA:Iowa/1943 H1N1 (scale bar = 10nm). **F)** 2.9Å cryo-EM 3D reconstruction of the complex viewed along two orthogonal axes. *HA_20* (purple) is bound to H1 along the stem of all three subunits. **G)** The cryo-EM structure of the *HA_20* binder in complex closely matches the design model (RMSD to RFdiffusion design: 0.63Å, yellow: Influenza Hemagglutinin). **H)** Structure of the *HA_20* binder alone superimposed on the design model viewed along two orthogonal axes. For cryo-EM panels, yellow: Influenza H1 map/structure; gray: *HA_20* binder design model; purple: *HA_20* binder map/structure.

## Acknowledgements

We thank N. Anand and D. Tischer for helpful discussions, and I. Kalvet and Y. Kipnis for providing helpful Rosetta scripts. We thank A. Dosey for the provision of purified Influenza Hemagglutinin protein. We thank R. Wu, J. Mou, K. Choi, L. Wu, and D. Blei for valuable feedback during writing. We thank I. Haydon for help with graphics. We also thank L. Goldschmidt and K. VanWormer, respectively, for maintaining the computational and wet lab resources at the Institute for Protein Design.

This work was supported by gifts from Microsoft (D.J., M.B., D.B.), Amgen (J.L.W.), the Audacious Project at the Institute for Protein Design (B.L.T., I.S., J.Y., H.E., D.B.), the Washington State General Operating Fund supporting the Institute for Protein Design (P.V., I.S.), grant INV-010680 from the Bill and Melinda Gates Foundation Grant (W.B.A., D.J., J.W., D.B.), grant DE-SC0018940 MOD03 from the U.S. Department of Energy Office of Science (A.J.B., D.B.), grant 5U19AG065156-02 from the National Institute for Aging (S.V.T., D.B.), an EMBO long-term fellowship ALTF 139-2018 (B.I.M.W.), the Open Philanthropy Project Improving Protein Design Fund (R.J.R., D.B.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research (N.R.B.), a Washington Research Foundation Fellowship (S.J.P.), a Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C, L.F.M.), an EMBO Non-Stipendiary Fellowship (ALTF 1047-2019, L.F.M.), the Defense Threat Reduction Agency grants HDTRA1-19-1-0003 (N.H., D.B.) and HDTRA12210012 (F.D.), the Institute for Protein Design Breakthrough Fund (A.C., D.B.), an EMBO Postdoctoral Fellowship (ALTF 292-2022, J.L.W.) and the Howard Hughes Medical Institute (A.C., W.S., R.J.R., D.B.), an NSF-GRFP (J.Y), an NSF Expeditions grant (1918839, J.Y, R.B., T.S.J.), the Machine Learning for Pharmaceutical Discovery and Synthesis consortium (J.Y, R.B., T.S.J.), the Abdul Latif Jameel Clinic for Machine Learning in Health (J.Y, R.B., T.S.J), the DTRA Discovery of Medical Countermeasures Against New and Emerging threats program (J.Y, R.B., T.S.J), and EPSRC Prosperity Partnership EP/T005386/1 (E.M.), the DARPA Accelerated Molecular Discovery program and the Sanofi Computational Antibody Design grant (J.Y, R.B., T.S.J.). We thank Microsoft and AWS for generous gifts of cloud computing resources.

## Author Contributions

Conceived the study: J.L.W., D.J., N.R.B, B.L.T., J.Y., D.B.; Trained RFdiffusion: J.L.W., D.J., N.R.B, W.A., B.L.T., J.Y.; Extended diffusion to residue orientations: B.L.T., J.Y. with assistance from V.D.B., E.M.; Generated experimentally characterized designs: H.E.E., D.J., J.L.W., N.R.B., N.H., W.S., P.V., I.S.; Generated computational designs: W.A., B.L.T., J.Y., D.J., J.L.W., N.R.B.; Experimentally characterized designs: H.E.E., A.J.B., R.J.R., L.F.M., B.I.M.W., S.J.P., N.H., A.C., S.V.T., J.L.W., B.L.T.; Contributed additional code: J.W., A.L., W.S.; Implemented RFdiffusion on Google Colab: S.O.; Trained RF: M.B., F.D.; Offered supervision throughout the project: D.B., T.S.J. and R.B.; Wrote the manuscript: J.L.W., D.J., B.L.T., N.R.B., J.Y., H.E., D.B. All authors read and contributed to the manuscript. J.L.W. and D.J. agree that the order of their respective names may be changed for personal pursuits to best suit their own interests.

## Data Availability

Design structures, AlphaFold2 models and experimental measurements are available at https://figshare.com/s/439fdd59488215753bc3. Cryo-EM maps and corresponding atomic models for the Influenza HA binder in Figure 6D-H have been deposited in the PDB and the Electron Microscopy Data Bank under accession codes 8SK7 and EMDB-40557, respectively. Electron microscopy data collected for the HE0537 oligomer is available at EMDB-40602.

## Code Availability

Code for running RFdiffusion has been released on GitHub, free for academic, personal, and commercial use at https://github.com/RosettaCommons/RFdiffusion. It is also available as a Google Colab notebook, accessible via GitHub.

## Competing Interests

The authors declare no competing interests.

Correspondence should be addressed to D.B. (dabaker@uw.edu).

## Extended Data Legends

**Extended Data Table 1: Cryo-EM data collection, refinement and validation statistics**

**Extended Data Figure 1: Training ablations reveal determinants of RFdiffusion success. A-C)** RFdiffusion can generate high quality large unconditional monomers. Designs are routinely accurately recapitulated by AF2 (see also Fig. 2C), with high confidence (**A**) for proteins up to approximately 400 amino acids in length. **B)** Further orthogonal validation of designs by

ESMFold. **C)** Recapitulation of the design structure is often better with ESMFold compared with AF2. For each backbone, the best of 8 ProteinMPNN sequences is plotted, with points therefore paired by backbone rather than sequence. **D)** Comparing RFdiffusion trained with MSE loss on Cα atoms and N-Cα-C backbone frames (Methods 2.5), rather than with FAPE loss[8,17]. The MSE loss is not invariant to the global coordinate frame, unlike FAPE loss, and is required for good performance at unconditional generation (left, two-proportion z-test of *in silico* success rate, *n*=400 designs per condition*, z=4.1, p=4.1e-5*). For motif scaffolding problems, where the "motif" provides a means to align the global coordinate frame between timesteps, FAPE loss performs approximately as well as MSE loss, suggesting the L2 nature of MSE loss (as opposed to the L1 loss in FAPE) is not empirically critical for performance. **E)** Allowing the model to condition on its $X_0$ prediction at the previous timestep (see Methods 2.4) improves designs. Designs with self-conditioning (pink) have improved recapitulation by AF2 (left) and better AF2 confidence in the prediction (right). Two-proportion z-test of *in silico* success rate, *n*=800 designs per condition *z=11.4, p=6.1e-30.* **F)** RFdiffusion leverages the representations learned during RF pre-training. RFdiffusion fine-tuned from pre-trained RF (pink) comprehensively outperforms a model trained for an equivalent amount of time, from untrained weights (gray). For context, sequences generated by ProteinMPNN on these output backbones are little better than sampling ProteinMPNN sequences from random Gaussian-sampled coordinates (white). Two-proportion z-test of *in silico* success rate, pre-training vs without pre-training (or vs random noise; both have zero success rate), *n=800 designs per condition, z=23.0, p=3.1e-117*. Note that the data in pink in **D-F** is the same data, reproduced in each plot for clarity. **G)** The median (by AF2 RMSD vs design) 300 amino acid unconditional sample highlighting the importance of self-conditioning and pre-training. Without pre-training, RFdiffusion outputs bear little resemblance to proteins (gray, left). Without self-conditioning, outputs show characteristic protein secondary structures, but lack core-packing and ideality (gray, middle). With pre-training and self-conditioning, proteins are diverse and well-packed (pink, right). **H)** Greater coherence during unconditional denoising may partly explain the effect of self-conditioning. Successive $X_0$ predictions are more similar when the model can self-condition (lower RMSD between $X_0$ predictions, pink curve). Data are aggregated from unconditional design trajectories of 100, 200 and 300 residues. **I)** During the reverse (generation) process, the noise added at each step can be scaled (reduced). Reducing the noise scale improves the *in silico* design success rates (left, middle; two-proportion z-test of *in silico* success rate*, n*=800 designs per condition, 0 vs 0.5: *z=1.7, p=0.09*, 0 vs 1: *z=6.5, p=6.8e-11*; 0.5 vs 1: *z=4.8, p=1.4e-6*). This comes at the expense of diversity, with the number of unique clusters at a TM score cutoff of 0.6 reduced when noise is reduced (right). Note throughout this figure the 6EXZ_long benchmarking problem is abbreviated to 6EXZ for brevity. Boxplots represent median±IQR; tails: min/max excluding outliers (±1.5xIQR).

**Extended Data Figure 2: RFdiffusion learns the distribution of the denoising process, and inference efficiency can be improved**. **A)** Analysis of simulated forward (noising) and reverse (denoising) trajectories shows that the distribution of Cα coordinates and residue orientations closely match, demonstrating that RFdiffusion has learned the distribution of the denoising process as desired. Left to right: i) average distance between a Cα coordinate at $X_t$ and its

position in $X_0$; ii) average distance between a Cα coordinate at $X_t$ and $X_{t-1}$; iii) average distance between adjacent Cα coordinates at $X_t$; iv) average rotation distance between a residue orientation at $X_t$ and $X_0$; v) average rotation distance between a residue orientation at $X_t$ and $X_{t-1}$. **B-C)** While RFdiffusion is trained to generate samples over 200 timesteps, in many cases, trajectories can be shortened to improve computational efficiency. **B)** Larger steps can be taken between timesteps at inference. Decreasing the number of timesteps speeds up inference, and often does not decrease *in silico* success rates (left) (for example, on an NVIDIA A4000 GPU, 100 amino acid designs can be generated with 15 steps, in ~11s, with an *in silico* success rate of over 60%). With as few as 50 timesteps when normalized for compute budget (center) it is often much more efficient to run more trajectories with fewer timesteps. This can be done without loss of diversity in samples (right). For harder problems (e.g. unconditional 300 amino acids), one must strike an intermediate number of total timesteps (e.g., T=50) for optimal compute efficiency. Note that for all other analyses in the paper, 200 inference steps were used, in line with how RFdiffusion is trained. **C)** An alternative to taking larger steps is to stop trajectories early (possible because RFdiffusion predicts $X_0$ at every timestep). In many cases, trajectories can be stopped at timestep 50-75 with little effect on the final *in silico* success rate of designs (left), and when normalized by compute budget (center), success rates per unit time are typically higher generating more designs with early-stopping. Again, this can be done without a significant loss in diversity (right).

**Extended Data Figure 3: Unconditionally-generated designs are folded and thermostable.** **A)** Four 200 amino acid and fourteen 300 amino acid proteins were tested for expression and stability. 9/18 designs expressed, with a major peak at the expected elution volume. Blue: 300 amino acid proteins; Purple: 200 amino acid proteins. **B)** Colored AF2 predictions overlaid on gray design models (left), circular dichroism spectra at 25ºC (blue) and 95ºC (pink) (middle) and circular dichroism melt curves (right) for all 9 designs passing expression thresholds. In all cases, proteins remain well folded even at 95ºC. Note that data on *300aa_3* and *300aa_8* are duplicated from Fig. 2F, reproduced here for clarity.

**Extended Data Figure 4: RFdiffusion can condition on fold information to generate specific, thermostable folds. A)** 6WVS is a previously-described *de novo* designed TIM barrel (left). A fine-tuned RFdiffusion model can condition on 1D and 2D inputs representing this protein fold, specifically secondary structure (middle, bottom) and block-adjacency information (middle, top) (see Methods 4.3.2). RFdiffusion then generates proteins that closely recapitulate this course-grained fold information (right). **B)** Outputs are diverse with respect to each other. With this coarse-grained fold specification, *in silico* successful designs are much more diverse (as quantified by pairwise TM scores) compared to diversity generated through simply sampling many sequences for the original PDB backbone (6WVS). **C)** NTF2 folds are useful scaffolds for *de novo* enzyme design[56], and can also be readily generated with fold-conditioning in RFdiffusion. Designs are diverse and closely recapitulated by AF2. **D)** *In silico* success rates are high with fold-conditioned diffusion. TIM barrels are generated with an AF2 *in silico* success rate of 42.5% (left bar, pink) with *in silico* success incorporating both AF2 metrics and a TM score vs 6WVS > 0.5. NTF2 folds are generated with an AF2 *in silico* success rate of 54.1% (right bar,

pink), with *in silico* success incorporating both AF2 metrics and a TM score vs PDB: 1GY6 > 0.5. *In silico* success was further validated with ESMFold (blue bars), where a plDDT > 80 was used as the confidence metric for success. Gray: RFdiffusion design, colors: AF2 prediction. **E)** 11 TIM barrel designs were purified alongside the 6WVS positive control. Ten of these express and elute predominantly as monomers (note that the designs are approximately 4kDa larger than 6WVS). **F)** Eight designs expressed sufficiently for analysis by circular dichroism. All designs are folded, with circular dichroism spectra consistent with the designed structure (middle), and similar to 6VWS. Designs were also all highly thermostable, with CD melt analyses demonstrating designs were folded even at 95°C (right). Designs are shown in gray, with the AF2 predictions overlaid in colors (left). Note that data on *6WVS* and *TIM_barrel_6* are duplicated from Fig. 2G, reproduced here for clarity.

**Extended Data Figure 5: Symmetric oligomer design with RFdiffusion. A)** Due to the (near-perfect - see Methods 3.1) equivariance properties of RFdiffusion, $X_0$ predictions from symmetric inputs are also symmetric, even at very early timepoints (and becoming increasingly symmetric through time; RMSD vs symmetrized: *t=200* 1.20Å; *t=150* 0.40Å; *t=50* 0.06Å; *t=0* 0.02Å). Gray: symmetrized (top left) subunit; colors: RFdiffusion $X_0$ prediction. **B)** *In silico* success rates for symmetric oligomer designs of various cyclic and dihedral symmetries. *In silico* success is defined here as the proportion of designs for which AF2 yields a prediction from a single sequence that has mean plDDT > 80 and backbone RMSD over the oligomer between the design model and AF2 < 2Å. Note that 16 sequences per RFdiffusion design were sampled. **C)** Box plots of the distribution of backbone RMSDs between AF2 and the RFdiffusion design model with and without the use of external potentials during the trajectory. The external potentials used are the “inter-chain” contact potential (pushing chains together), as well as the “intra-chain” contact potential (making chains more globular). Using these potentials dramatically improves *in silico* success (Two-proportion z-test of *in silico* success rate: *n*=100 designs per condition, *z=4.3, p=1.9e-5)*. **D)** Designs are diverse with respect to the training dataset (the PDB). While the monomers (typically 60-100aa) show reasonable alignment to the PDB (median 0.72), the whole oligomeric assemblies showed little resemblance to the PDB (median 0.50). **E)** Additional examples of design models (left) against AF2 predictions (right) for C3, C5, C12, and D4 symmetric designs (the symmetries not displayed in Fig. 3) with backbone RMSDs against their AF2 predictions of 0.82, 0.63, 0.79, and 0.78 with total amino acids 750, 900, 960, 640. **F)** Additional nsEM data for symmetric designs. The model is shown on the left and the 2D class averages on the right for each design. **G)** Two orthogonal side views of HE0537 by cryo-EM. Representative 2D class averages from the cryo-EM data are shown to the right of 2D projection images of the computational design model (lowpass filtered to 8 Å), which appear nearly identical to the experimental data. Scale bars shown (white) are 60Å. Boxplot represents median±IQR; tails: min/max excluding outliers (±1.5xIQR).

**Extended Data Figure 6: External potentials for generating pockets around substrate molecules. A-D)** Example *in silico* successful designs for enzyme classes 2-5 (ref [57], see also Fig. 4). Native enzyme (PDB: 1CWY, 1DE3, 1P1X, 1SNZ); catalytic site (teal); RFdiffusion output (gray: model, colors: AF2 prediction). Metrics (AF2 vs design backbone RMSD, AF2 vs

design motif backbone RMSD, AF2 vs design motif full-atom RMSD, AF2 pAE): EC2: 0.93Å, 0.50Å, 1.29Å, 3.51; EC3: 0.92Å, 0.60Å, 1.07Å, 4.59; EC4: 0.93Å, 0.80Å, 1.03Å, 4.41; EC5: 0.78Å, 0.44Å, 1.14Å, 3.32. **E-H)** Implicit modeling of a substrate for while scaffolding a retroaldolase active site triad [TYR1051-LYS1083-TYR1180] from PDB: 5AN7. **E)** The potential used to implicitly model the substrate, which has both a repulsive and attractive field (see Methods 4.4). **F)** Left: Kernel densities demonstrate that without using the external potential (pink), designs often fall into two failure modes: (1) no pocket, and (2) clashes with the substrate. Right: clashes (substrate < 3Å of the backbone) & pockets (no clash and > 16 Cα within 3-8Å of substrate) with and without the potential. Two-proportion z-test: *n*=71/51 +/- potential; clashes *z=-2.053, p=0.020*, pocket *z=-2.274, p=0.011*. Each datapoint represents a design already passing the stringent *in silico* success metrics (AF2 motif RMSD < 1Å, AF2 backbone RMSD < 2Å, AF2 pAE < 5). Note that the potential and clash definition pertain only to backbone Cα atoms, and do not currently include sidechain atoms. **G)** Designs close to the labeled local maxima of the kernel density estimate. Without the potential, the catalytic triad is predominantly (*1*) exposed on the surface with no residues available to provide substrate stabilization or (*2*) buried in the protein core, preventing substrate access. With the potential, the catalytic triad is predominantly (*3*), partially buried in a concave pocket with shape complementary to the substrate. Backbone atoms within 3Å of the substrate are shown in red. **H)** A variety of diverse designs with pockets made using the potential, with no clashes between the substrate and the AF2-predicted backbone. The functional form and parameters used for the pocket potential are detailed in Methods 4.4. In each case the substrate is superimposed on the AF2 prediction of the catalytic triad.

**Extended Data Figure 7: Additional $Ni^{2+}$ binding C4 oligomers. A)** AF2 predictions of a subset of the experimentally verified $Ni^{2+}$ binding oligomers, with corresponding isothermal titration calorimetry (ITC) binding isotherms for the wild-type (blue) and H52A mutant (pink) below. Note that these, with Figure 5, encompass all of the experimentally validated outputs deriving from unique RFdiffusion backbones. Wild-type dissociation constants are displayed in each plot. We observe a mixture of endothermic (NiB2.10, NiB2.23, NiB2.15) and exothermic isotherms. For all cases displayed we observe no binding to the ion for H52A mutants, indicating the scaffolded histidine at position 52 is critical for ion binding. $K_D$ values in the isotherms indicate binding of the ion with the designed stoichiometry (1:4 $Ni^{2+}$:protein). Note that each backbone depicted is from a unique RFdiffusion sampling trajectory, and that models and data for designs NiB2.15, NiB1.12, NiB1.20 and NiB1.17 from Figure 5 are duplicated here for ease of viewing. **B)** Size exclusion chromatograms for elutions from the 44 purifications suggest the vast majority of designs are soluble and have the correct oligomeric state. **C)** Size exclusion chromatograms for 20 H52A mutants show that the mutants remain soluble and retain the intended oligomeric state. Note that for ITC plots, points represent single measurements.

**Extended Data Figure 8: Targeted unconditional and fold-conditioned protein binder design.**

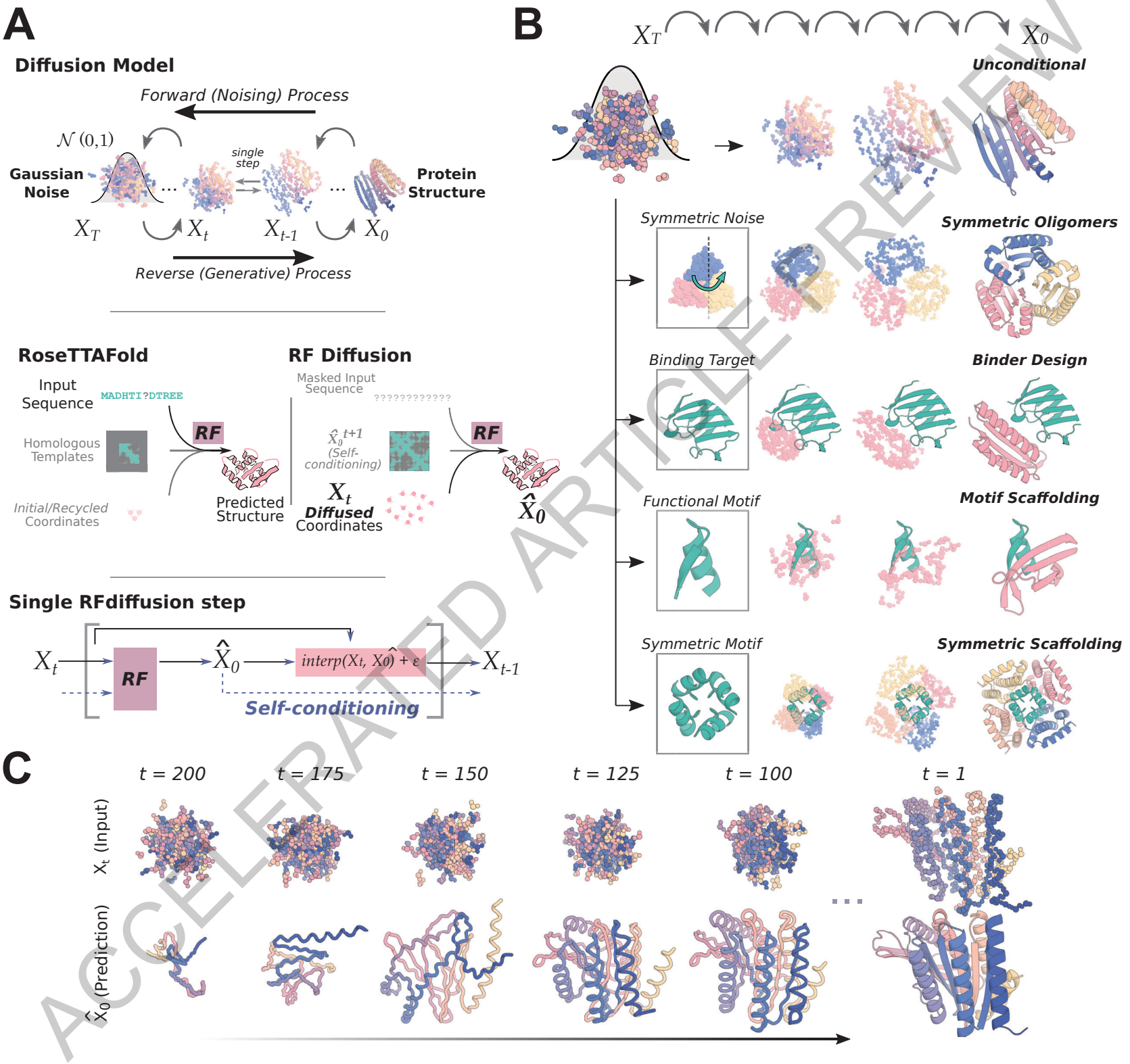
**A-B)** The ability to specify where on a target a designed binder should bind is crucial. Specific "hotspot" residues can be input to a fine-tuned RFdiffusion model, and with these inputs, binders almost universally target the correct site. **A)** IL-7Rα (PDB: 3DI3) has two patches that are

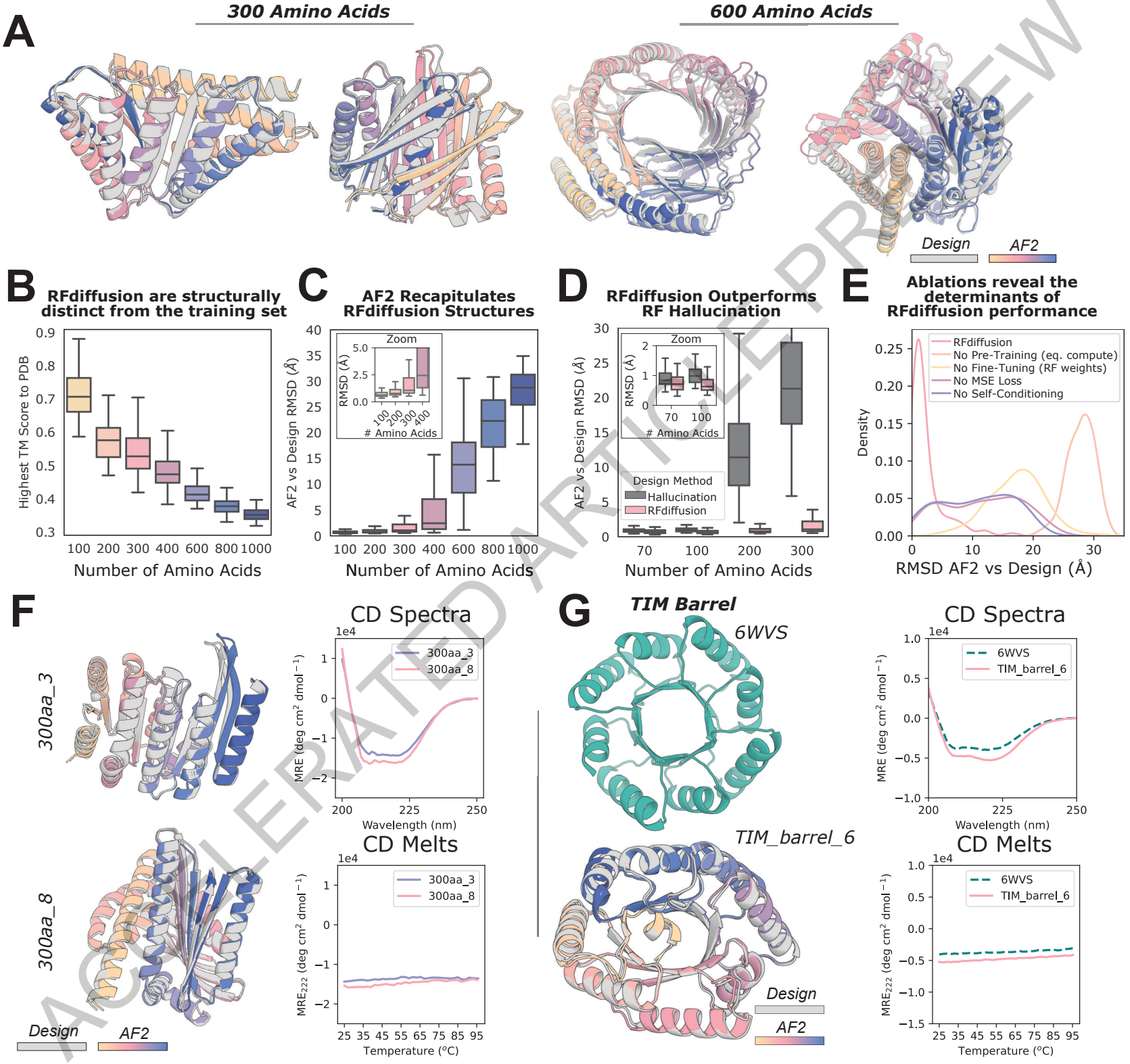
optimal for binding, denoted Site 1 and Site 2 here. For each site, 100 designs were generated (without fold-specification). **B)** Without guidance, designs typically target Site 1 (left bar, gray), with contact defined as Cα-Cα distance between binder and hotspot reside < 10Å. Specifying Site 1 hotspot residues increases further the efficiency with which Site 1 is targeted (left bar, pink). In contrast, specifying the Site 2 hotspot residues can completely redirect RFdiffusion, allowing it to efficiently target this site (right bar, pink). **C-D)** As well as conditioning on hotspot residue information, a fine-tuned RFdiffusion model can also condition on input fold information (secondary structure and block-adjacency information - see Methods 4.5). This effectively allows the specification of a (for instance, particularly compatible) fold that the binder should adopt. **C)** Two examples showing binders can be specified to adopt either a ferredoxin fold (left) or a particular helical bundle fold (right). **D)** Quantification of the efficiency of fold-conditioning. Secondary structure inputs were accurately respected (top, pink). Note that in this design target and target site, RFdiffusion without fold-specification made generally helical designs (right, gray bar). Block-adjacency inputs were also respected for both input folds (bottom, pink). **E)** Reducing the noise added at each step of inference improves the quality of binders designed with RFdiffusion, both with and without fold-conditioning. As an example, the distribution of AF2 interaction pAEs (known to indicate binding when pAE < 10[26]) is shown for binders designed to PD-L1. In both cases, the proportion of designs with interaction pAE < 10 is high (blue curve), and improved when the noise is scaled by a factor 0.5 (pink curve) or 0 (yellow curve). **F)** Full *in silico* success rates for the protein binders designed to five targets. In each case, the best fold-conditioned results are shown (i.e. from the most target-compatible input fold), and the success rates at each noise scale are separated. In line with current best practice[26], we tested using Rosetta FastRelax[58] before designing the sequence with ProteinMPNN, but found that this did not systematically improve designs. *In silico* success is defined in line with current best practice[26]: AF2 plDDT of the monomer > 80, AF2 interaction pAE < 10, AF2 RMSD monomer vs design < 1Å. **G)** Experimentally-validated *de novo* protein binders were identified for all five of the targets. Designs that bound at 10 μM during single point BLI screening with a response equal to or greater than 50% of the positive control were considered binders. Concentration is denoted by hue for designs that were screened at concentrations less than 10 μM and thus may be false negatives.

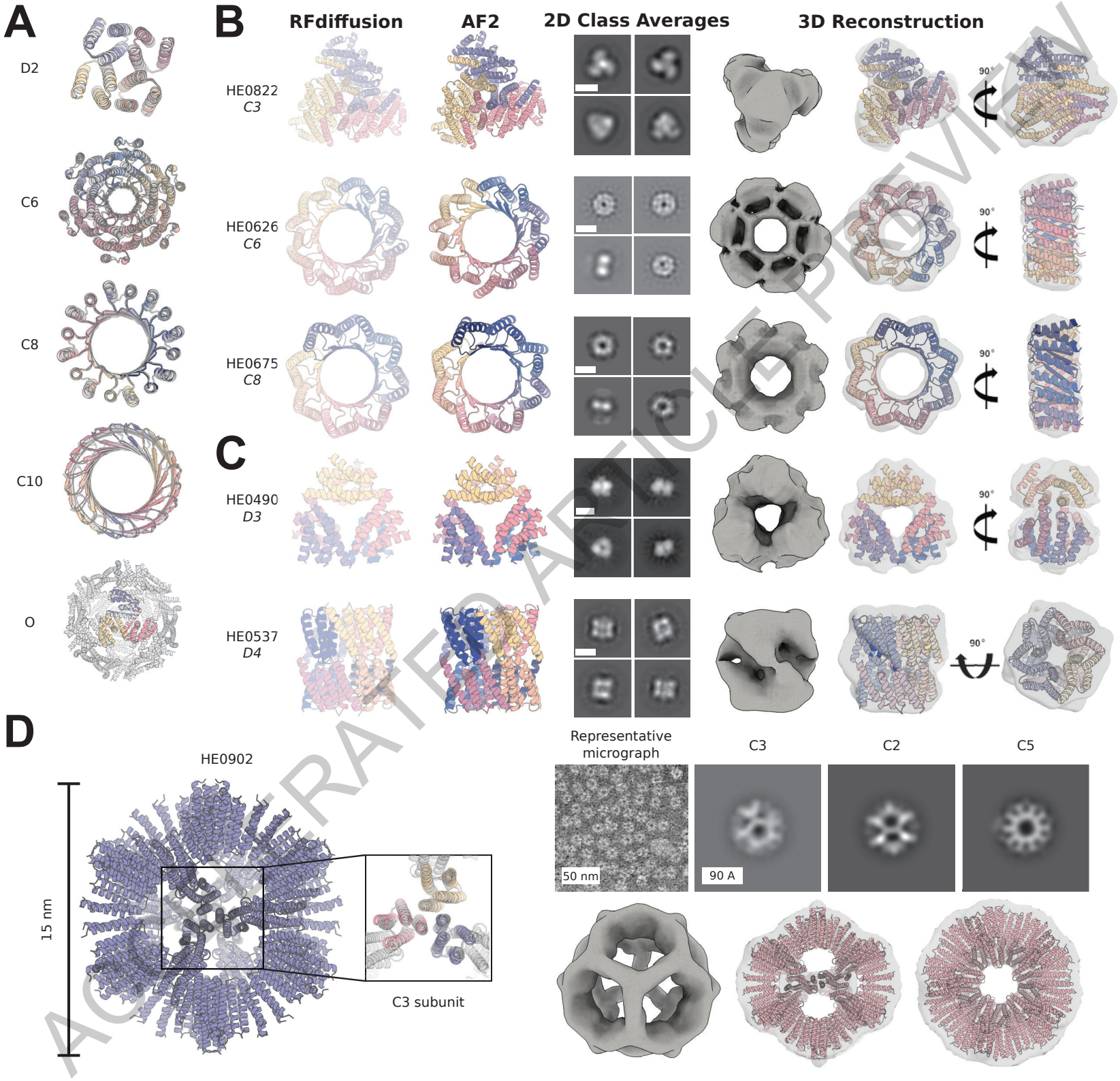
**Extended Data Figure 9: Cryo-electron microscopy structure determination of designed Influenza HA binder.**

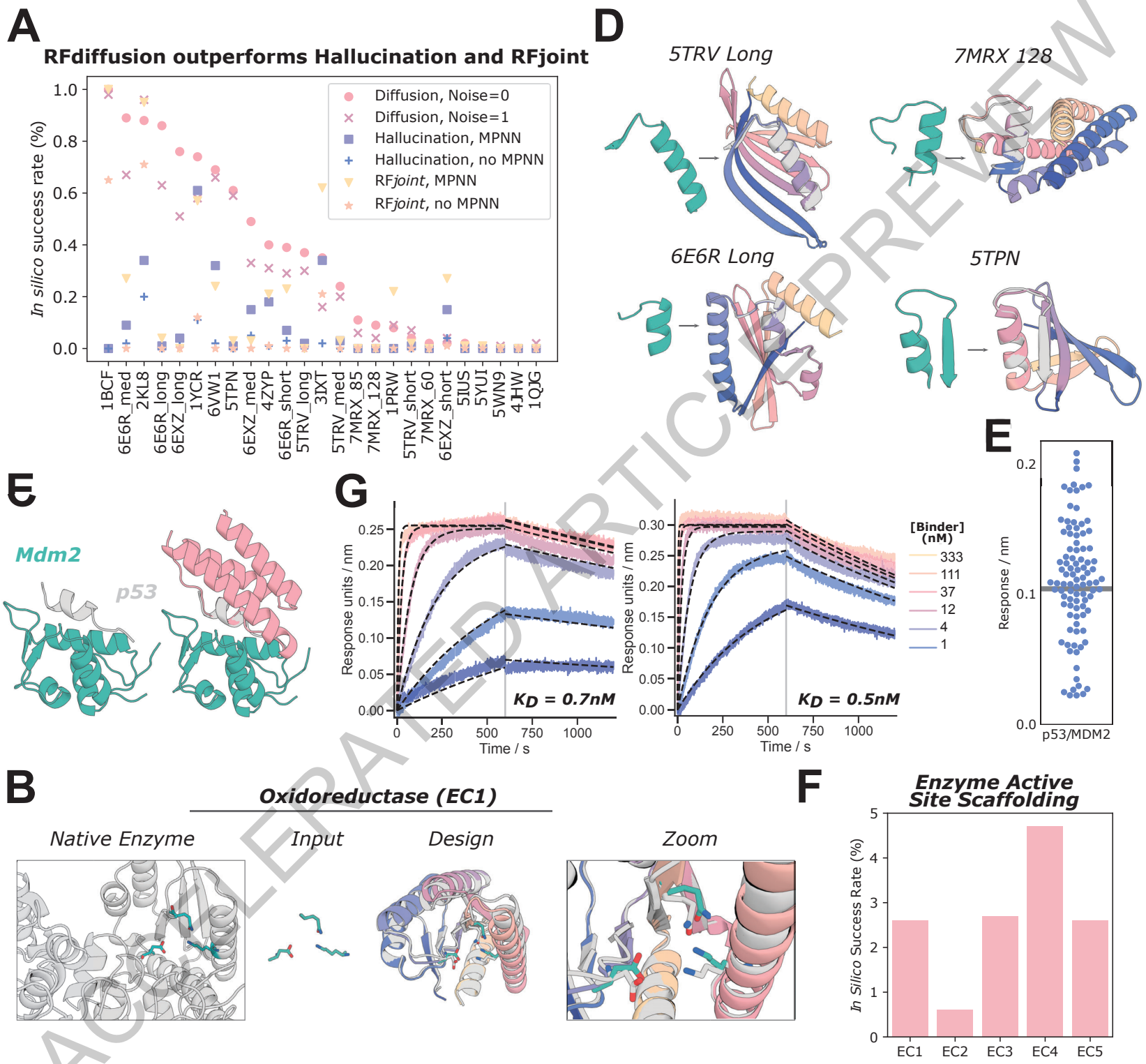
**A)** Representative raw micrograph showing ideal particle distribution and contrast. **B)** 2D Class averages of Influenza H1+*HA_20* binder with clearly defined secondary structure elements and a full-sampling of particle view angles (scale bar = 10 nm). **C)** Cryo-EM local resolution map calculated using an FSC value of 0.143 viewed along two different angles. Local resolution estimates range from ~2.3Å at the core of H1 to ~3.4Å along the periphery of the N-terminal helix of the *HA_20* binder. **D)** Cryo-EM structure of the full H1+*HA_20* binder complex (purple: *HA_20*; yellow: H1; teal: glycans). **E)** Global resolution estimation plot. **F)** Orientational distribution plot demonstrating complete angular sampling. **G)** 3D *ab initio* (left) and 3D heterogenous refinement (right - unsharpened) outputs, performed in the absence of applied symmetry, and showing clear density of the *HA_20* binder bound to all three stem epitopes of

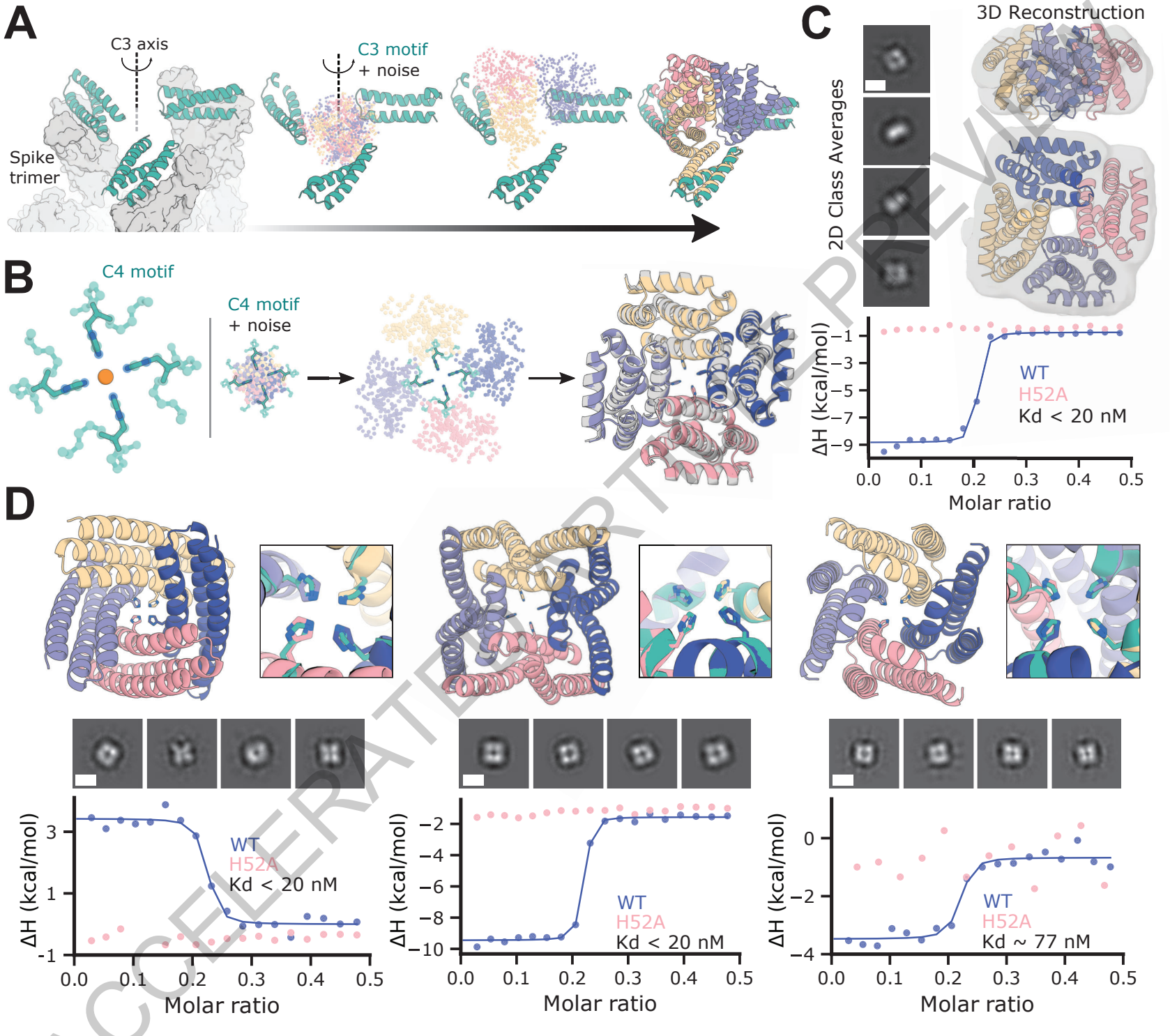
the Iowa43 HA glycoprotein trimer, in all maps. **H)** The designed binder has topological similarity to 5VLI, a protein in the PDB, but binds with very different interface contacts.

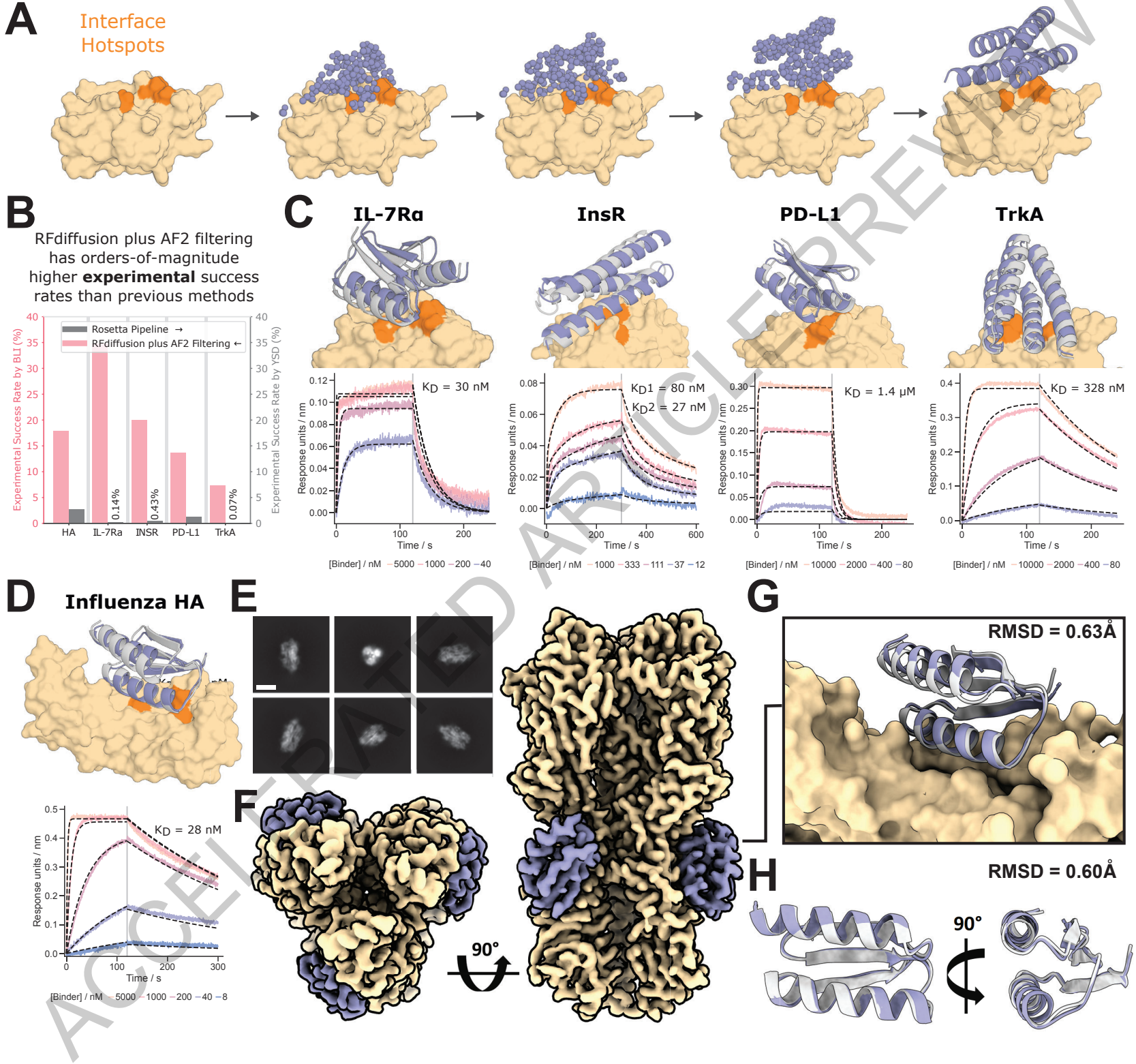


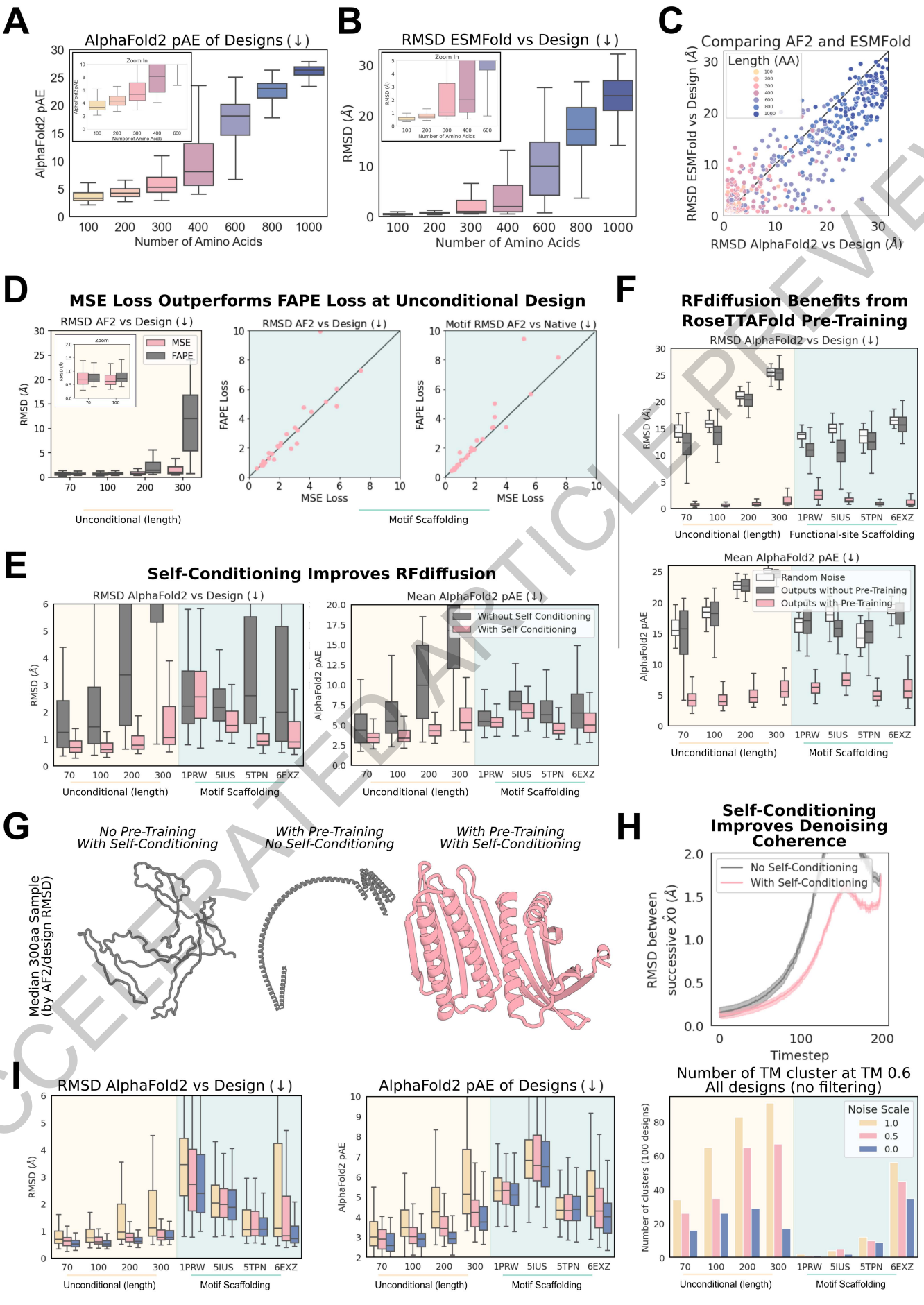






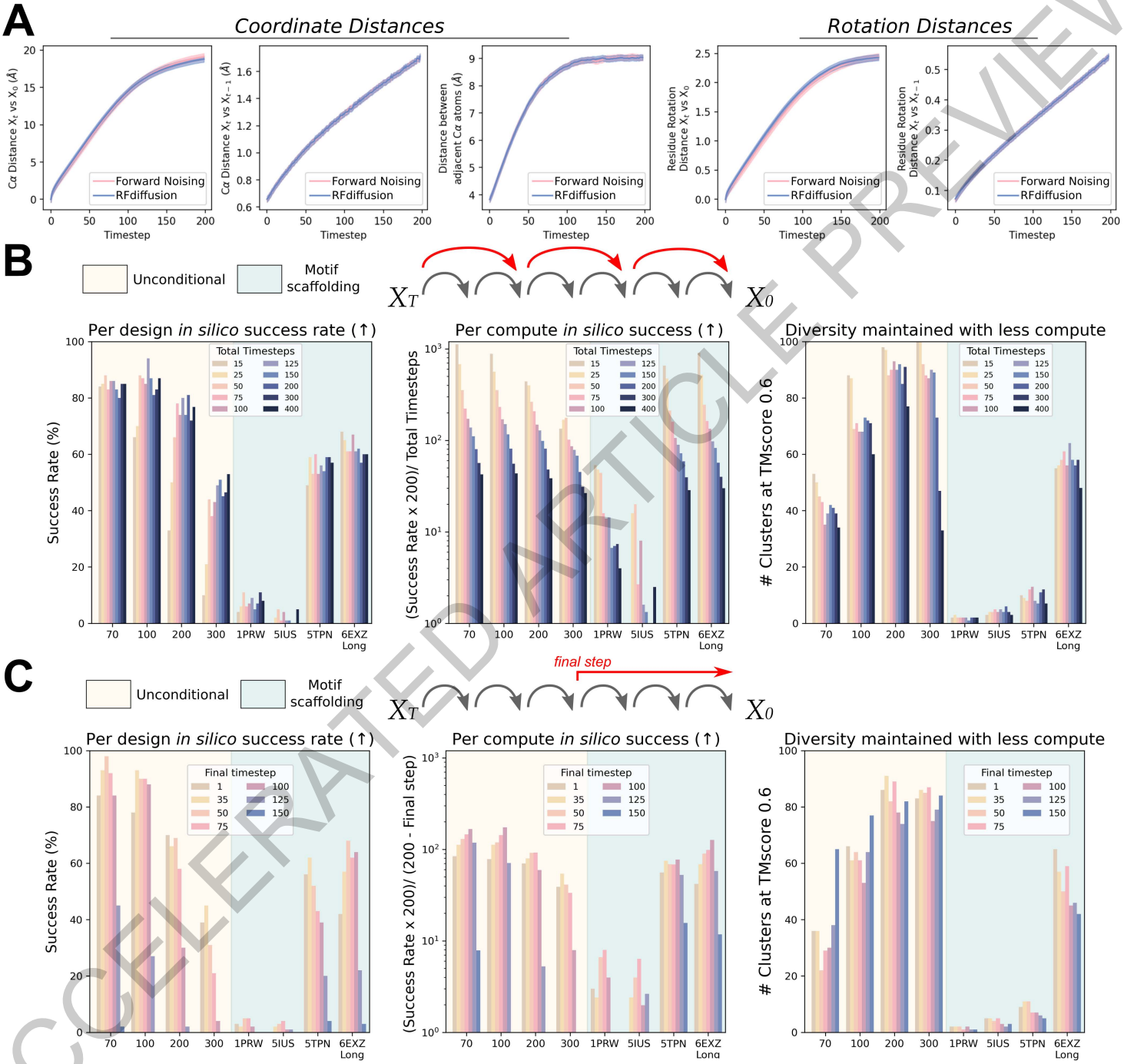




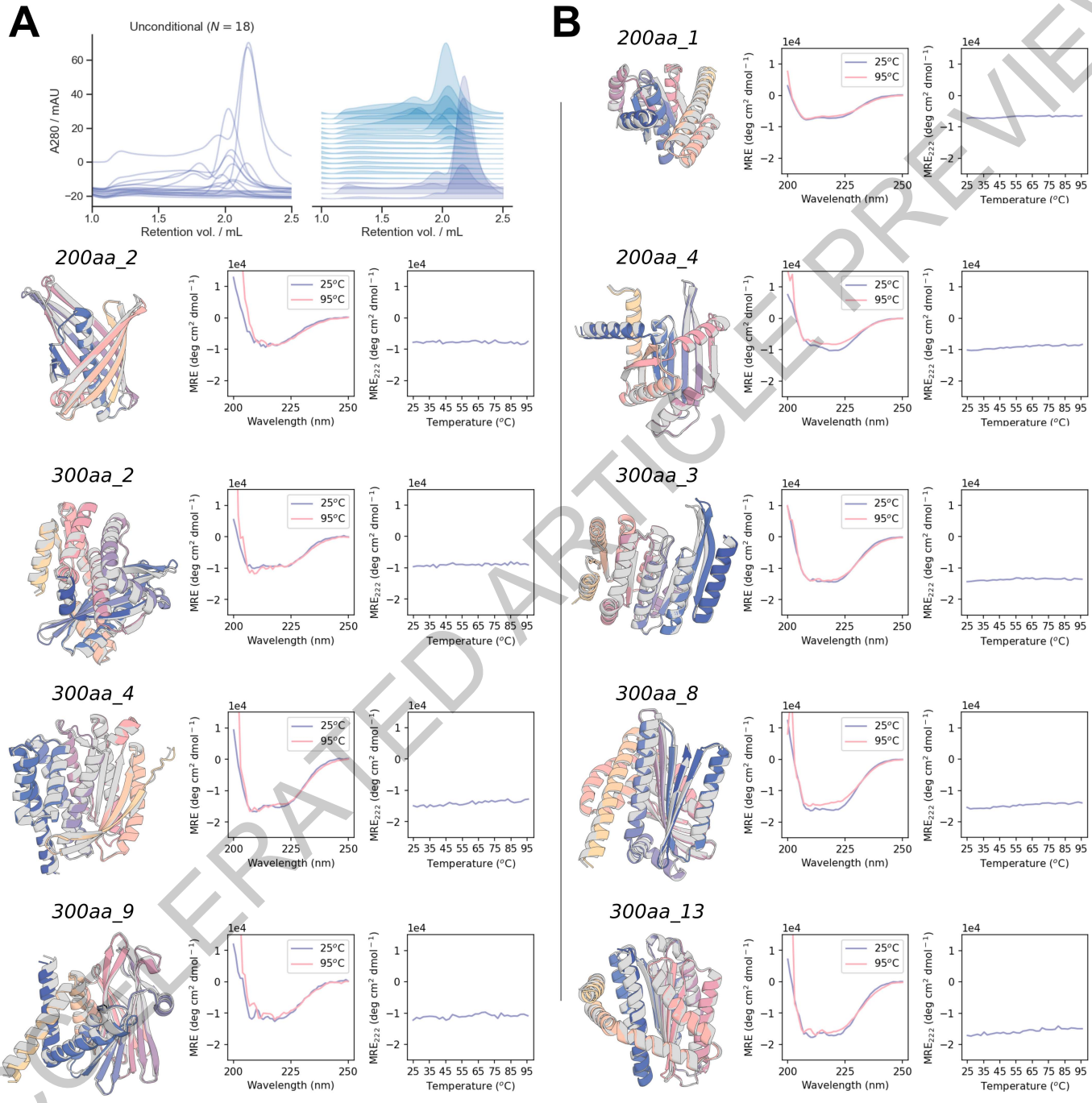




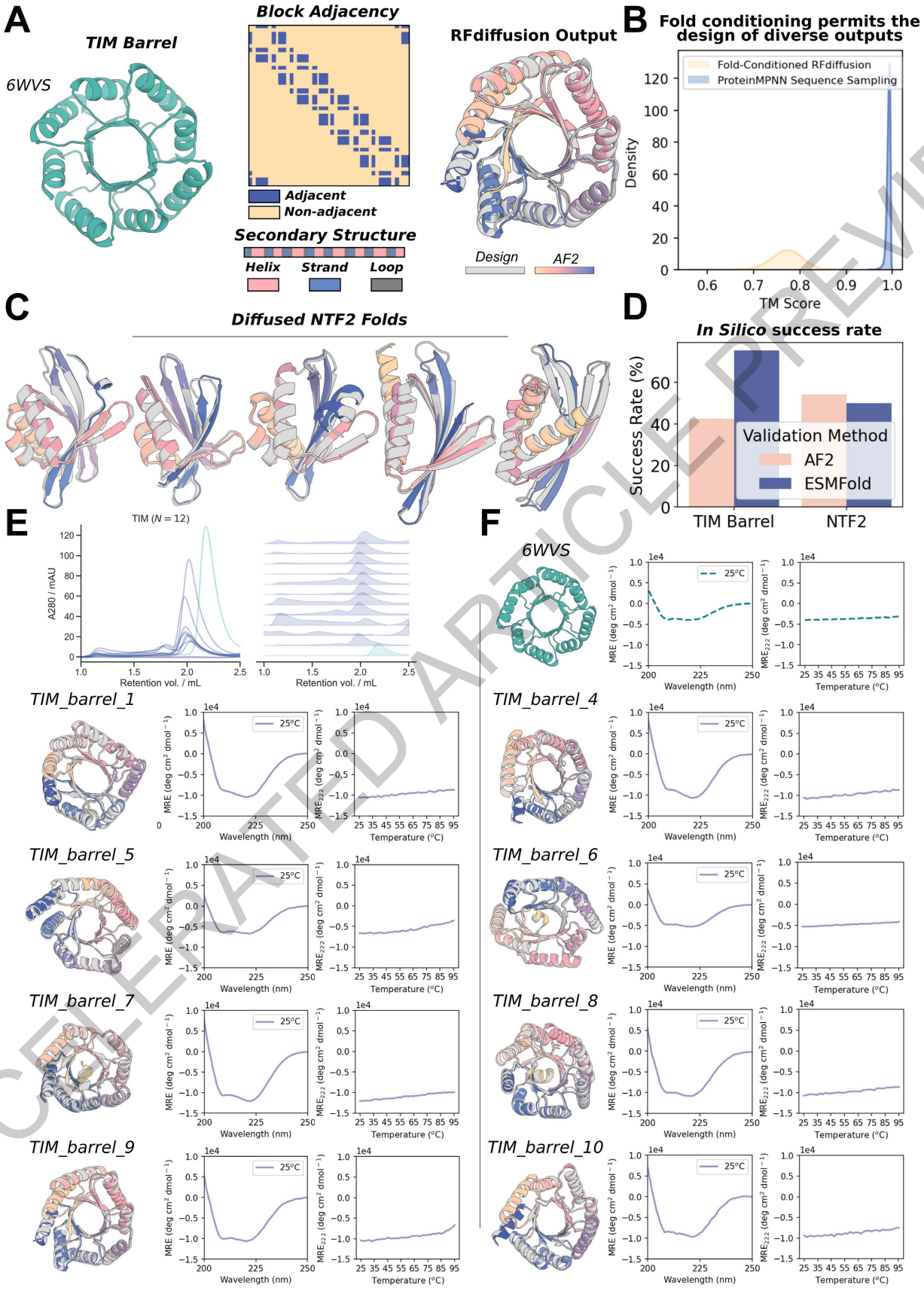

**Extended Data Fig. 1**




**Extended Data Fig. 2**

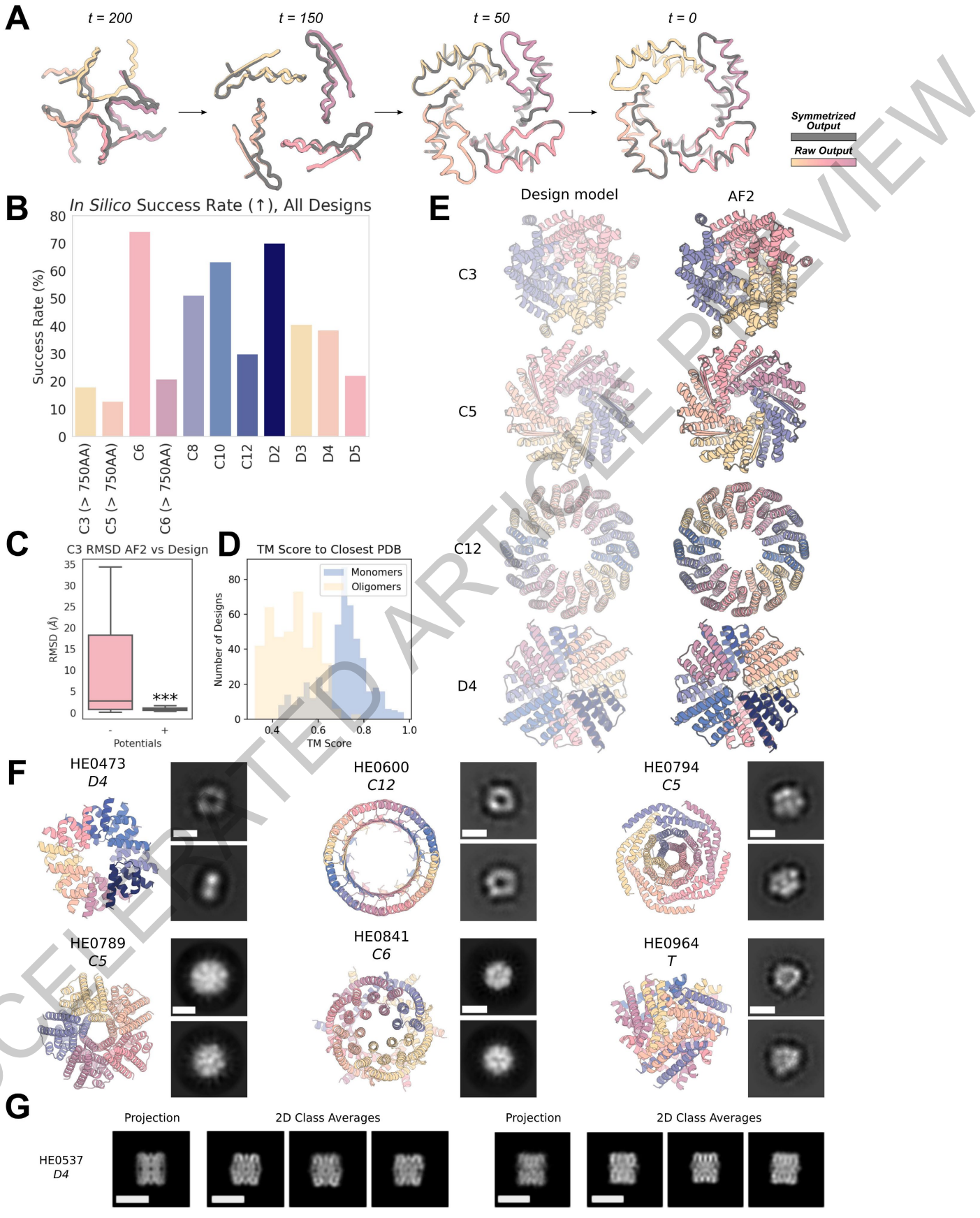



**Extended Data Fig. 3**



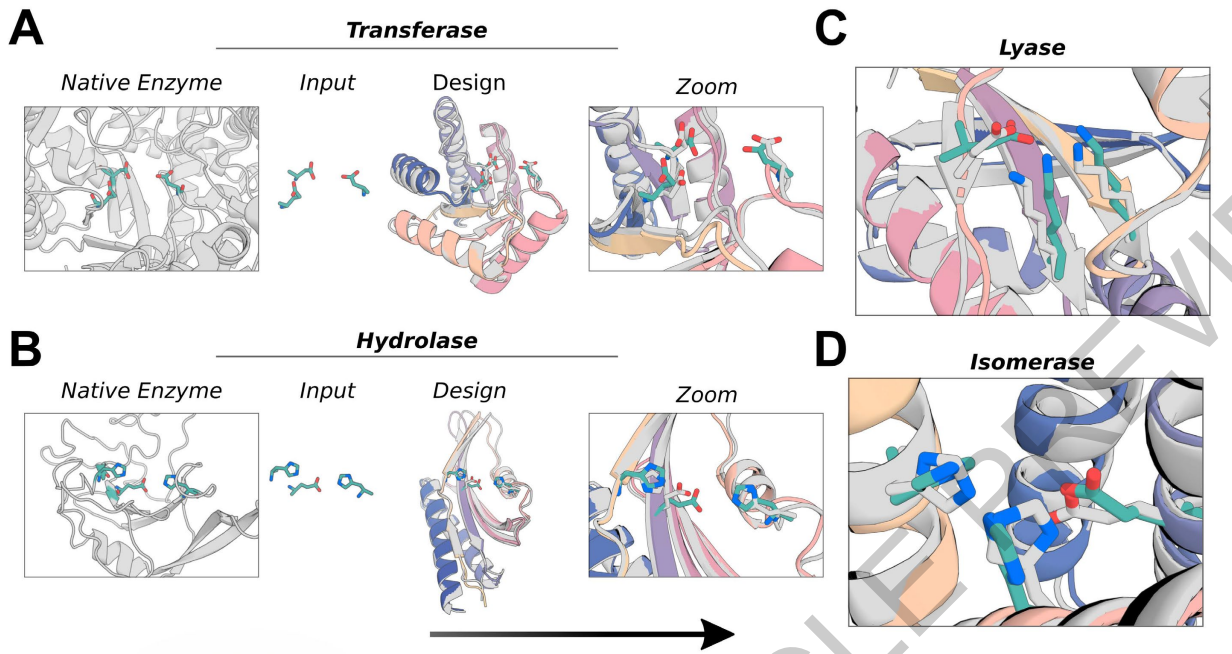


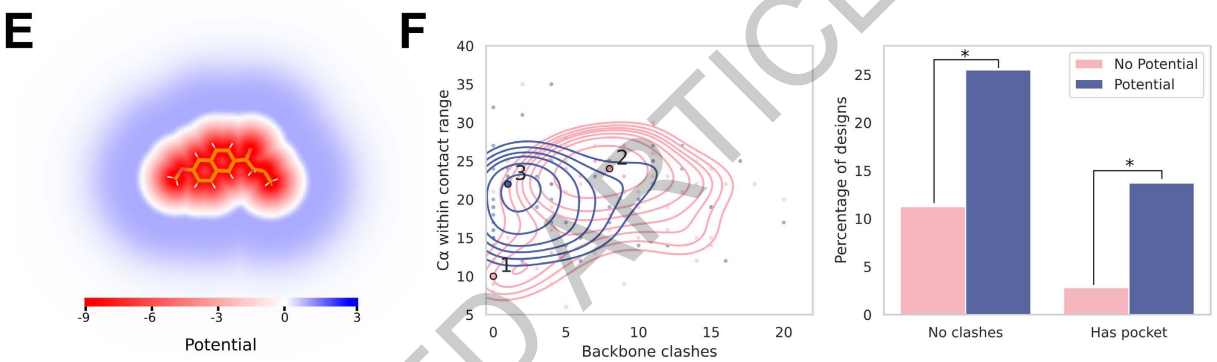

**Extended Data Fig. 4**

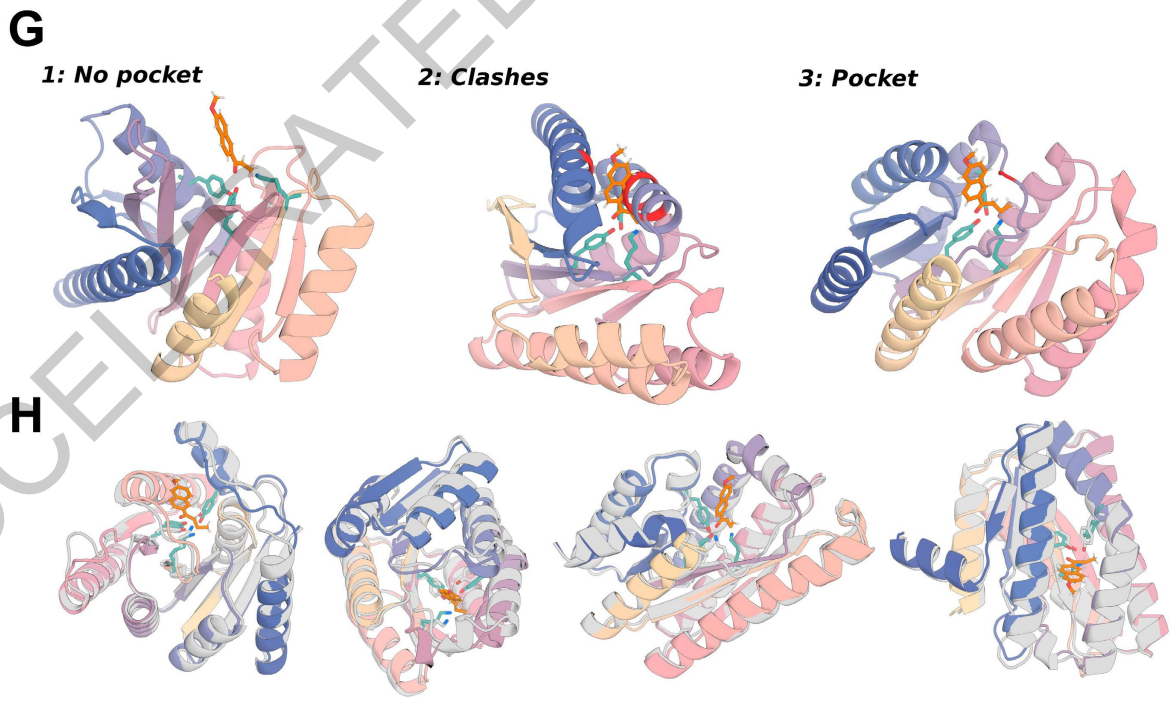





**Extended Data Fig. 5**

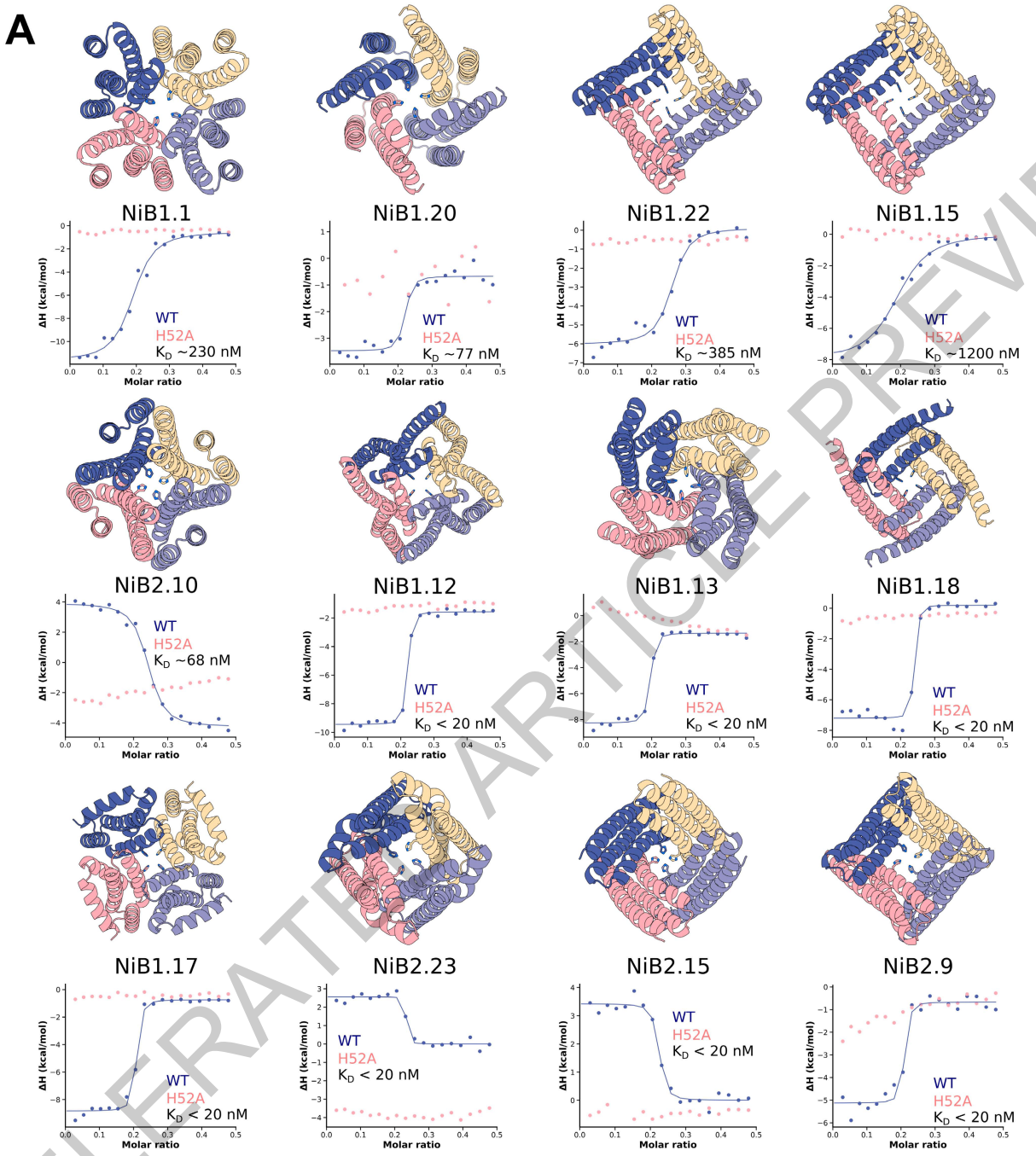


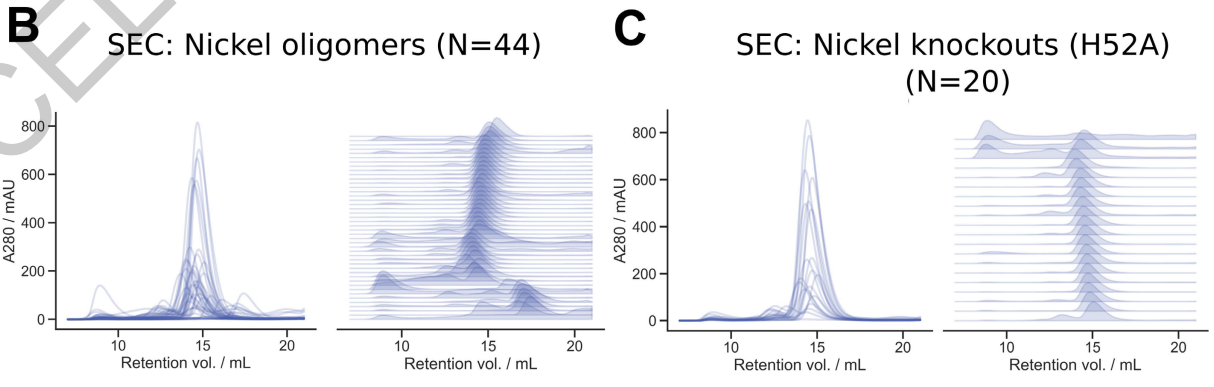





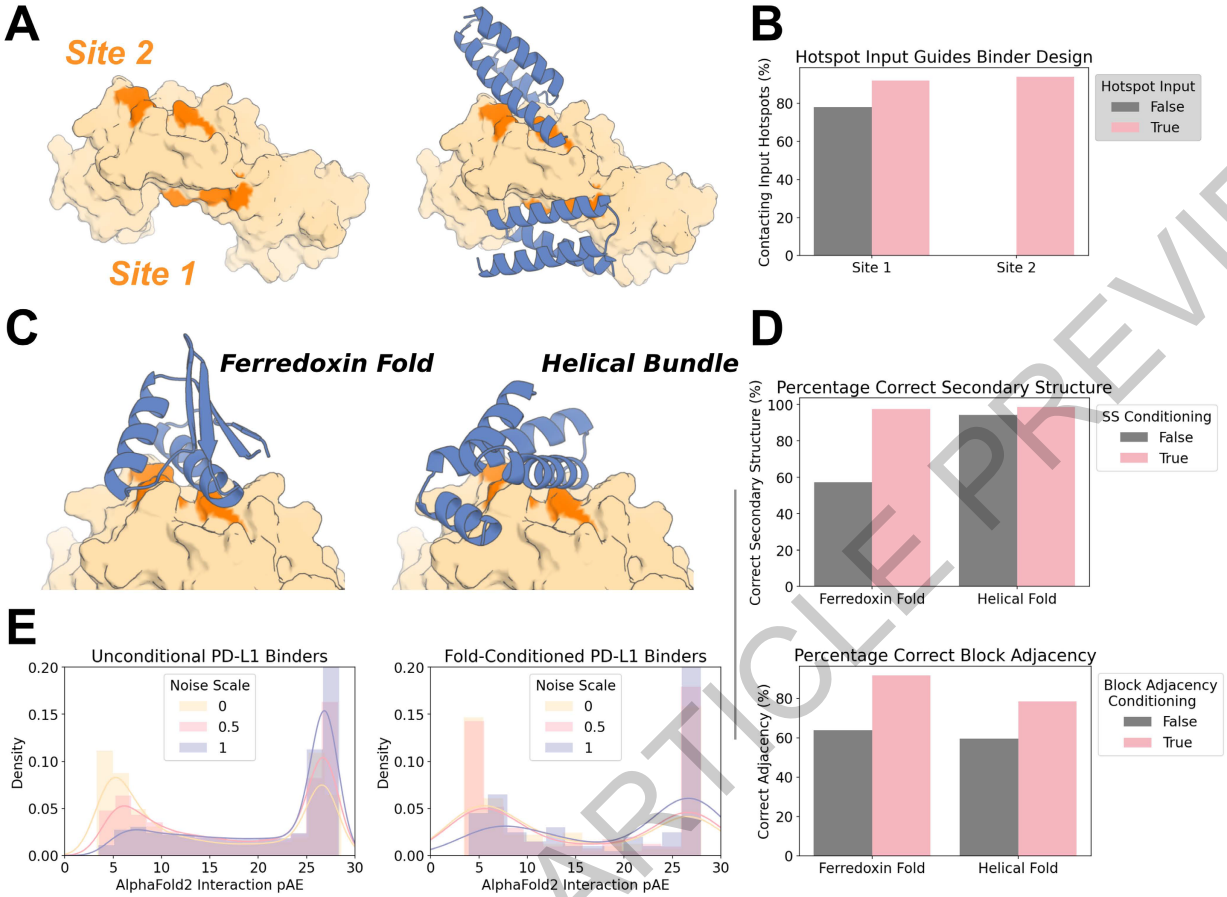

**Extended Data Fig. 6**





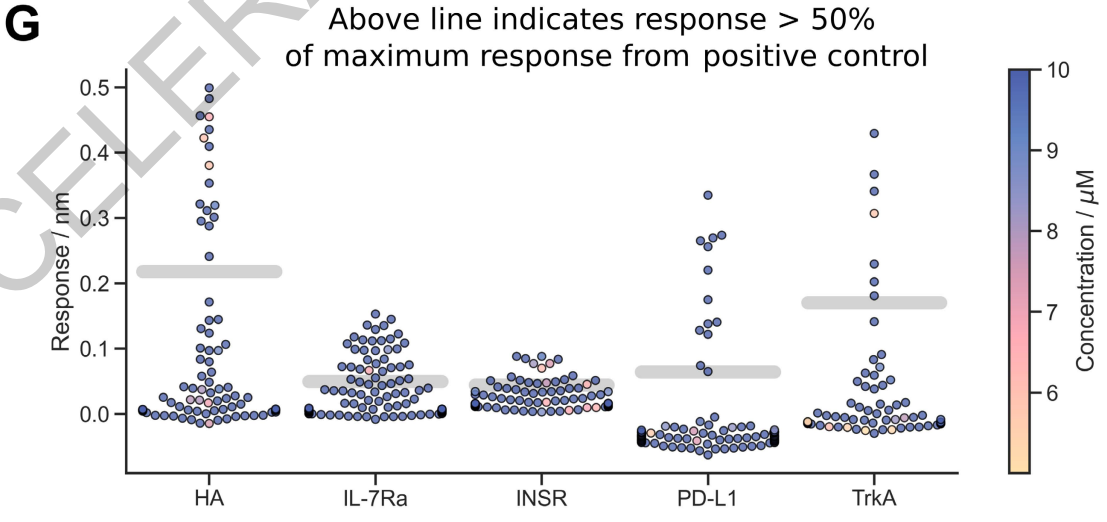

**Extended Data Fig. 7**



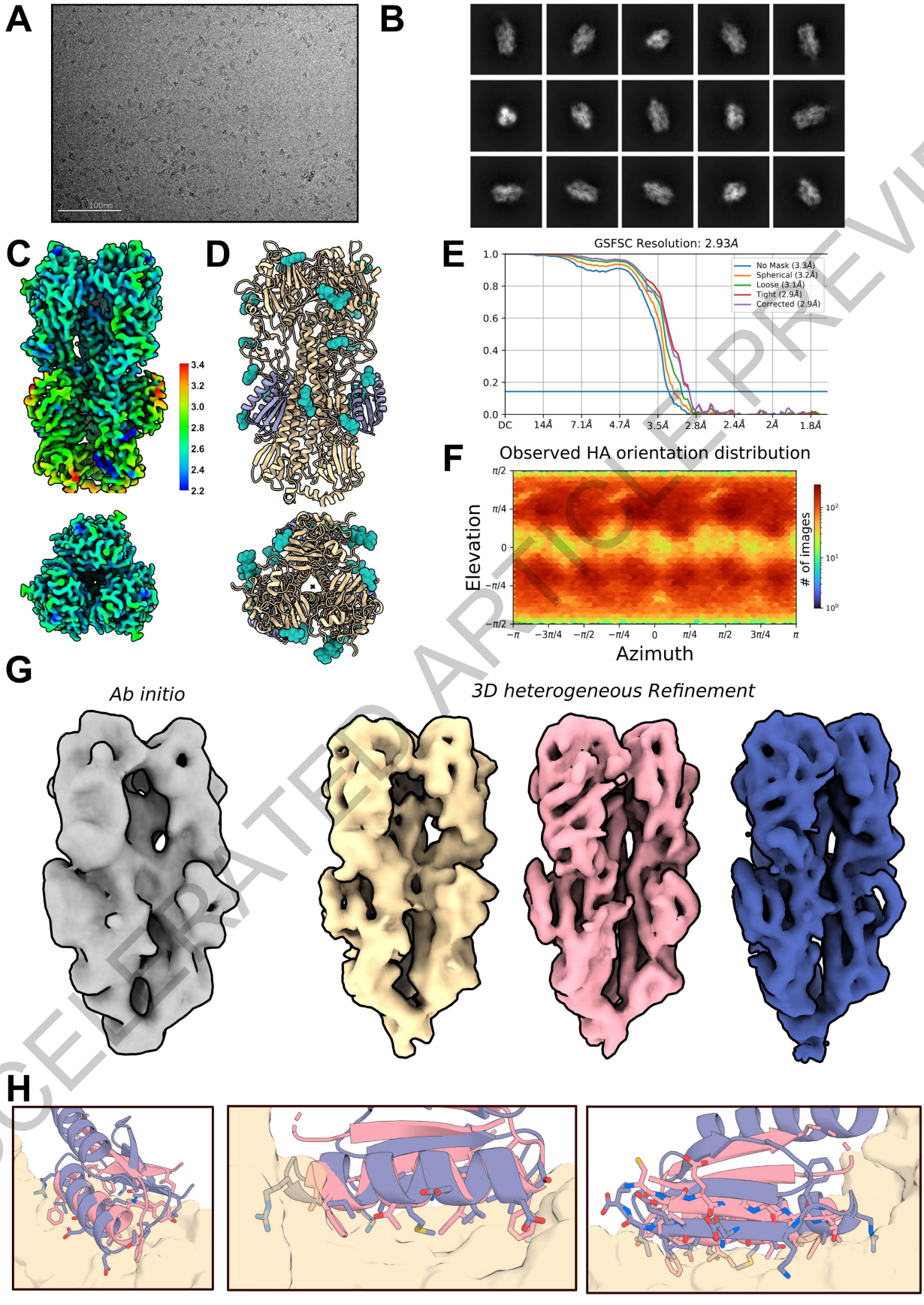

**F**

***In Silico* Success Rate (%) [raw / +FastRelax]**

| | *Noise Scale* | PD-L1 | IL7 Receptor $\alpha$ | Insulin Receptor | TrkA Receptor | Hemagglutinin |
|---|---|---|---|---|---|---|
| Unconditional | 0.0 | 39.3/47.6 | 9.1/18.8 | **17.9**/17.9 | 14.7/12.1 | 4.2/3.9 |
| | 0.5 | 25.4/29.9 | 6.6/9.9 | 7.1/6.7 | 6.5/6.0 | 0.6/1.0 |
| | 1.0 | 10.7/14.4 | 2.1/4.4 | 2.8/3.2 | 1.5/1.6 | 0.1/0.1 |
| Fold-Conditioned | 0.0 | 44.0/46.0 | 11.7/**21.9** | 10.6/10.8 | **29.2**/24.2 | **14.6**/13.9 |
| | 0.5 | 37.0/**49.0** | 11.9/15.0 | 3.7/3.7 | 16.3/15.2 | 3.2/4.7 |
| | 1.0 | 22.0/26.0 | 5.3/10.7 | 2.5/3.0 | 6.6/7.2 | 0.2/2.3 |




**Extended Data Fig. 8**




**Extended Data Fig. 9**

| | HE0537 (EMDB-40602) | H1+HA_20 (EMDB-40557) (PDB 8SK7) |
|---|---|---|
| **Data collection and processing** | | |
| Magnification | 36,000 | 105000 |
| Voltage (kV) | 200 | 300 |
| Electron exposure (e–/Å$^2$) | 65 | 64.273 |
| Defocus range (μm) | -0.8: -2 | 0.7-1.8 |
| Pixel size (Å) | 0.883 | 0.84 |
| Symmetry imposed | D4 | C3 |
| Initial particle images (no.) | 184,703 | 2,396,954 |
| Final particle images (no.) | 36,827 | 308,846 |
| Map resolution (Å) | 6.06 | 2.93 |
| FSC threshold | | |
| Map resolution range (Å) | 5.8-8.47 | 2.2-3.4 |
| | | |
| **Refinement** | | |
| Initial model used (PDB code) | | 3LZG |
| Model resolution (Å) | | 119.6 |
| FSC threshold | | |
| Validation | | |
| MolProbity score | | 0.92 |
| Clashscore | | 1.67 |
| Poor rotamers (%) | | 6 |
| Ramachandran plot | | |
| Favored (%) | | 98.72 |
| Allowed (%) | | 1.28 |
| Disallowed (%) | | 0.00 |

**Extended Data Table 1**