

De novo protein design with deep learning

David Juergens

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
David Baker, Chair
Frank DiMaio
Neil King

Program Authorized to Offer Degree:
Molecular Engineering and Sciences

©Copyright 2024

David Juergens

University of Washington

Abstract

De novo protein design with deep learning

David Juergens

Chair of the Supervisory Committee:

David Baker

Department of Biochemistry

The ability to predict the exact set of atoms which possess a desired chemical function lies at the heart of modern day problems in energy, health, and sustainability. A particularly useful medium for constructing these sets of atoms is proteins due to their diverse chemistry, ease of synthesis, and broad compatibility with materials and molecules from across the periodic table. The grand challenge of designing proteins to possess chemical functions is the problem of mapping a desired chemical function to the amino acid sequence which would encode it. Here I describe my efforts (in close collaboration with many colleagues) to uncover this mapping using deep learning. Specifically, through the training and deployment of various deep neural networks that generate protein structures in an intuitive and useful manner, we uncover design principles and algorithms that enable a new generation of custom-designed molecules.

Introduction

The ability to rationally create molecules which perform arbitrary chemical functions would broadly benefit our society. It is the hypothesis of our field, and my shared belief, that if we can find models that accurately predict and simulate the properties of proteins, then we can likely access a large fraction of interesting chemical function space through *de novo* protein design [1]. In the pursuit of improved rational engineering of molecular systems, I propose investigating the following conjectures:

- (1) Systems possessing chemical functions that are useful for solving many challenges that society faces today in areas such as sustainability, energy, and medicine can be rationally engineered via *de novo* protein design.
- (2) Many of the chemical functions we want to design can be specified via structure-based molecular constraints in proteins, and we can thus design many chemical functions through building proteins which satisfy these constraints or conditions.
- (3) Deep neural networks can be trained to learn aforementioned conditional distributions of proteins using data that already exists, and this will be broadly useful for designing *de novo* molecules with more complex function and with an improved success rate.

Generally, the guiding light for the research thrusts I will undertake to support (or not) these conjectures takes the following form: Using machine learning, train and employ models to approximate and sample from the conditional probability distribution

$$p(X, S | M)$$

where X denotes the atomic structure of a protein, S denotes a protein sequence, and M denotes a protein “motif” - a set of atom identities and geometries we know or theorize possesses chemical function, and/or we would like to design stabilizing interactions with. For generality, please note that if we have no conditioning information (i.e., unconditional protein generation) we can just say $M = M_0$ is a null motif, and $p(X, S | M_0) = p(X, S)$. In what

follows, I hope to detail the significance of and progress towards accomplishing the goal of modeling this distribution to the reader.

Background and significance

In this section, I provide a very brief background knowledge base which our work stands on the shoulders of.

The free energy of protein folding.

Proteins are polymers composed of 20 canonical amino acids [2] linked together via peptide bonds in a chain. As hypothesized by Christian Anfinsen [3], the 3D structure that the entire polymer adopts natively is encoded by the sequence of amino acids, and is its lowest free energy state. The relationship between protein sequence and native structure is a complex and delicate function of the energetic and entropic contributions that each amino acid in the chain imparts on a protein-solvent system.

The energetic/enthalpic contributions to this function from each amino acid are derived from fundamental physical interactions between atoms, with both classical and quantum origins. Briefly, these include (see refs [4], [5])

- London dispersion interactions, in which random fluctuations in the electron occupancy around a (neutral) atom create short-lived charge separations (dipoles) which on average tend to attract atoms toward each other.
- Pauli-style repulsion, in which the forbidden occupancy of two electrons in exactly the same quantum state around an atom sharply repels two atoms once they reach a certain (very close) distance from each other. This results in atoms effectively having a hard sphere radius, within which other atoms will likely not enter without forming chemical bonds.
- Coulombic attraction and repulsion of various forms, including
 - charge-charge interactions
 - dipole-charge interactions

- dipole-dipole, or higher order multipole interactions (Keesom interaction)
- dipole-induced-dipole interactions (Debye interaction)
- Favorable electron orbital overlap, proximity and/or alignment

Together, the London, Debye and Keesom interactions comprise what is commonly referred to as Van der Waals interactions. A famously useful function which reasonably accurately describes the combination of non-bonded intermolecular forces from Van der Waals interactions and Pauli repulsion is the Lennard-Jones (6-12) potential [6].

The entropic contributions to protein structure are equally (if not more) consequential, and must also be introduced. The main entropic driving factors for protein folding consist of interactions between the solvent (water, ions) and the polypeptide chain. Concretely, it is largely the change in apolar surface area exposed to solvent which causes the exceptionally favorable increase in solvent entropy, and thus drives a protein to bury its apolar surface in a hydrophobic core by folding up [7]. The simple relationship between exposed apolar surface area and free energy of a protein-solvent system means the polar/apolar character of each of the amino acids in a sequence can have quite a large effect on the free energy change upon transitioning from unfolded to folded, and thus the native structure (in the eyes of Anfinsen's hypothesis). Additionally, as a protein folds, its conformational entropy (the number of different conformational microstates the protein is able to access) decreases unfavorably, and must be made up for by aforementioned enthalpic and solvent entropy gains.

Structural manifestations of primary sequence contributions to free energy.

The structural manifestations of these energetic and entropic contributions to the free energy landscape of protein conformations are important and numerous. They begin to take effect the moment that an amino is released from a ribosome synthesizing a protein.

Enthalpically, Ramachandran and Sasisekharan seminally observed that there exists a joint statistical distribution for the ϕ and ψ backbone dihedral angles of amino acids in protein structures observed in nature [8]. There are two main, energetic driving factors for this: First, there are only certain ϕ and ψ combinations which make it possible to avoid overlapping of Van der Waals radii between the atoms within the backbone and sidechain of each amino acid. Second, there are certain ϕ/ψ combinations which allow the repeated formation of energetically favorable hydrogen bonds between amino acid backbone -OH groups and N atoms. Repeated regions of amino acids which have these ϕ/ψ combinations generate α -helices and β -pleated sheets. It should be also noted that because of the coupling between amino acid type and ϕ/ψ dihedral preferences, there exists individual preferences among the different amino acids to adopt α -helical and β -sheet conformations, or other less structured conformations like loops [7].

Beyond secondary structure formation, another important (enthalpically driven) characteristic of natural proteins is to have tightly but not overly-packed protein cores which don't have large voids within them. This trend is mostly driven by the need to recover the large number of favorable Van der Waals and hydrogen bonding interactions that were removed by creating a cavity in the solvent where a protein could exist in the first place.

Entropically, the need for biological systems to have proteins which don't aggregate on a large scale, along with the need to leverage powerful solvent entropy changes upon folding, means that we generally see proteins that have polar surfaces and apolar cores. It is useful to recall here that in a similar vein to the differing secondary structure propensities of the 20 canonical amino acids, the different degrees of polarization within the various amino acids then manifests as unique distributions of burial that each amino acid tends to exhibit on average in natural protein structures [9]. Overwhelmingly, apolar amino acids are more concentrated in the cores of proteins, and polar amino acids are common on their surfaces. Note that for monomeric species this is true at the tertiary level, for oligomeric species this is true at the quaternary level.

Protein structure prediction

Knowledge that sequence determines structure, and structure determines function then creates a demand for knowledge of protein structures. It is expensive to experimentally determine high resolution protein structures and so the ability to predict them directly from amino acid sequence has long been sought after [10]. Some of the first major successes in computational protein structure prediction came from *ab initio* methods, in which Markov-chain Monte-Carlo optimization of an energy function to find low free energy conformations of proteins was performed [11]. Subsequently, the structure prediction field turned to using coevolutionary information from sequencing databases to infer contact maps, further improving the accuracy of energy based structure prediction algorithms by using the maps as terms in the energy function being optimized [12], [13], [14].

After it was discovered that optimization of energy functions guided by contact maps could produce higher accuracy protein structure predictions, the field was introduced to a new and improved prediction paradigm using machine learning [15], [16], [17]. In this paradigm, contact maps were extracted from multiple sequence alignments using Potts models, and subsequently refined by deep residual convolutional neural networks (akin to image processing networks [18]) to produce accurate contact predictions. Using these contact predictions as terms in energy functions further improved structure prediction performance. Pushing further, DeepMind broke open the field of high accuracy protein structure prediction with the introduction of AlphaFold2 (AF2) [19]. With AF2, protein structure prediction at near experimental accuracy was possible the majority of the time. Following shortly after, other groups developed deep neural networks with similar capabilities [20], [21].

De novo protein design

While in the pursuit of *ab initio* protein structure prediction, the method of predicting amino acid sidechain conformations via MCMC optimization of an energy function by searching through

combinations of “rotamers” (rotameric conformers, discrete options for sidechain conformations) from an available library was developed [22]. It was noticed that instead of restricting the set of rotamers being searched through at a given sequence position to those associated with a single amino acid type (that which is in the native sequence), one could also select from rotamers of other amino acid identities to optimize the energy of the protein. In other words, one could design an amino acid sequence whose lowest energy structure was some structure of interest. Today, this problem is referred to as the “fixed backbone sequence design problem”, and its robust solution is of great interest because it would unlock the ability for scientists to genetically encode any macromolecular structures that they desire - so long as they can describe the backbone they desire in the first place.

Since the original Rosetta MCMC rotameric search for sequence design, new methods have been developed which take advantage of advances in the field of deep learning for improved protein sequence design. For example, we developed a technique which employed deep network hallucination to perform improved fixed backbone sequence design by optimizing protein sequences for their likelihood of achieving the correct conformation, as measured by a structure prediction neural network [9]. This method contrasts with the original *de novo* protein sequence design paradigm by looking for large gaps in the free energy landscape between the desired structure and close, off-target states, as opposed to optimizing for low energy on an absolute energy scale without regard for off-target states. Furthermore, it proposed circumnavigating the need to enumerate a partition function for probabilistic calculations of protein structure propensities, and instead used the structure prediction network to implicitly approximate the partition function and directly yield probabilities for different states. More recently, methods for performing autoregressive protein sequence design have further pushed the boundary of accurate and robust fixed backbone design [23], [24]. Already, we are seeing these methods transform the field of structure-based *de novo* protein design.

***De novo* backbone generation**

Fixed-backbone protein sequence design is only useful if one possesses a protein backbone they want to realize in the first place. As such, people have also been looking for ways of generating protein backbones they want to design.

One of the first examples of this was a study in which the authors designed a novel protein topology by hand [25]. It is not possible to scale protein structure design by hand to larger and more complex folds, and thus computational techniques were needed. One early example of a computational technique to do this was the Rosetta Blueprint Builder [26]. This method modestly expanded capabilities for protein designers to generate protein backbones. To get around the difficulty of generating protein backbones, researchers have also used structures already discovered by nature when needed. For example, by using the Rosetta matcher algorithm [27] to find backbones which can host enzyme active sites.

To advance the state of the art for protein backbone generation, researchers started looking for deep learning solutions to the problem. One of the first methods discovered for doing this to be experimentally verified was “deep network hallucination” [28]. In this method, the trRosetta [16] neural network was employed for protein sequence design by performing an MCMC optimization scheme in which the “confidence” (KL divergence against some reference distribution of molten globule protein sequences) of the network’s predictions are optimized by changing the input protein sequence. There were also other reports describing methods to use deep learning for this sort of task described earlier [29], [30]. More recently, [31] used the hallucination technique developed by [28] to use the AF2 structure prediction network for generating symmetric oligomers by incorporating symmetry into the loss function being optimized.

At this point, the field began searching for other techniques beyond hallucination for protein structure generation, and this is where the work of me and my colleagues begins.

Findings

In this section, I will outline the context and results for a multitude of protein design method development campaigns that I have undertaken with my colleagues to better understand the role that deep learning can play in *de novo* protein and peptide design, and to support the conjectures of my proposed thesis. In section Findings 1, I will discuss a work entitled '*Scaffolding protein functional sites using deep learning*' which was recently published in the journal Science [32]. My main contribution in this work was developing the protein inpainting neural network ("RFjoint") as well as designing and/or characterizing various proteins. The second work, in section Findings 2 is entitled '*De novo design of protein structure and function with RFdiffusion*' and is published in Nature [33]. My main contributions in this work were co-inventing and co-developing the various training and inference schemes for RoseTTAFold diffusion, characterizing the network's ability to design proteins computationally, as well as designing experimentally characterized proteins. Please note that these works were large collaborations with many colleagues, and I of course did not accomplish all of this on my own.

Since the publication of these works, my colleagues and I have done (and are actively doing) more research to expand the molecular engineering capabilities of these technologies beyond modeling "protein-only" systems. These new works include the design of macrocyclic peptides for use as drugs and biosensors, enzymes which catalyze the hydrolysis of small substrates, and symmetric macromolecular systems capable of coordinating ligands and self-assembling into larger quaternary structures. I will discuss these works which are not yet published (but in preparation) in the sections Findings 3 and Findings 4.

Finally, we have authored other published manuscripts in which we used our methodological advances to design functional proteins for creating conformation switching systems [34], binders to bioactive helical peptides [35], and binders to small molecules [36]. I will not discuss these here, but it is noteworthy that these works rely (at least in part) on the aforementioned developments.

Findings 1: Scaffolding protein functional sites using deep learning

Introduction

Deep learning is revolutionizing many fields because of its broad applicability to many tasks, and the unreasonable effectiveness of overparameterized neural networks for learning from data [37], [38]. There are two main modeling paradigms in which deep learning is employed: The first is called “supervised learning”, in which models are trained to map independent variables (such as pixel intensities from a 64x64 grid, or a string representation of a molecule [39], [40], [41]) to labels for responding variables (such as categories of animals within an image, or physical properties of some molecule [42], [43], [44]). The second major learning paradigm in deep learning is “unsupervised learning”, in which true data are corrupted and then a model is tasked with reconstructing the original data to learn underlying patterns. Unsupervised learning is an incredibly useful modeling paradigm (if framed correctly) because it does not require labeled data.

An important class of unsupervised models are large language models (LLMs) [45], [46], [47], [48], [49], [50], [51], most of which now derive their neural architectures from the transformer [52]. There are two main ways in which self-supervised learning is executed to train modern LLMs - the autoregressive (AR) modeling task, and the masked language modeling (MLM) task. The autoregressive modeling task seeks to learn the probability distribution of sequences of tokens (i.e., words, amino acids, letters, etc.) by factorizing the joint probability distribution of having some set of tokens in a series into a series of conditional distributions that are then

strung together using the chain rule of probability. That is, an autoregressive model seeks to learn

$$p(S) = p(s_n | s_{n-1}) \dots p(s_2 | s_1) p(s_1) = \prod_{i=2}^n p(s_i | s_{i-1}) p(s_1)$$

Where S is a series of tokens (e.g., an amino acid sequence, a sentence), and s_i is a single token within the series, and n is the number of tokens in the series. A categorical cross entropy (CCE; [53]) loss function is used for training the model, and typically large mini-batch sizes are used in stochastic gradient descent of the loss function. Once a model has been sufficiently trained on this task, it can then be used in a generative manner to sample from the joint distribution $p(S)$ it has learned.

The other major self-supervised task for language is termed the “masked language modeling” (MLM) task, in which some fraction of tokens are randomly masked and the network is tasked with recovering the true identity of the missing tokens. Again using a cross entropy loss function, learning this task causes the network to learn dependencies between the tokens in order to correctly predict the missing information [21]. Seminal work which demonstrated the power of learning this task was the BERT study [47], and work which has since taken this modeling paradigm and applied it to protein design and structure prediction is the ESM-family of models [21].

Motif scaffolding

To design functional proteins, they need to have groups of atoms within them that possess the ability to exhibit some chemical function. Generally, we refer to these amino acid identities and their geometries which impart chemical function as protein “motifs”. Examples of such constructs could be the active site within an enzyme, or the interface with which some protein binds to another (Fig 1A). The first step to designing *de novo* proteins which possess functional sites is to find a motif which is either known or predicted to have chemical function. The second

step is to then find a protein backbone and sequence which, when folded, accurately recapitulates the precise geometry of the motif. Previous methods have been used to scaffold functional sites, but they were not general enough to be extended to arbitrary motif scaffolding problems [54], [55], [56], [57], [58].

Training the RoseTTAFold neural network on a protein MLM task

In order to overcome the limitations of previous motif scaffolding methods, we set out to develop new methods which could take in an arbitrary set of (potentially discontinuous) amino acids and relative geometries, and then jointly find protein structures and sequences which fold to recapitulate that site (i.e., “scaffold the motif”). The first method, called protein hallucination (Fig 1B) extends beyond [28] by combining the KL divergence loss against a background distribution (“increasing prediction confidence”) with losses that aim to have the network recapitulate the structure of the motif being scaffolded. The benefit of protein hallucination is that it does not require retraining a structure prediction network to use it for design.

The second method, called protein inpainting (Fig 1C), is a method which took the RoseTTAFold [20] structure prediction network and fine-tuned it to learn the masked language modeling task over protein sequences and structures jointly. If the MLM task on protein structures and sequences is successfully learned, the model could be applied to a wide variety of protein structure prediction and design challenges because they can be similarly formulated as missing information recovery problems (Fig. 1D). Although protein inpainting has been explored before [29], [30], in this study we approach it using the power of a pretrained structure-prediction network.

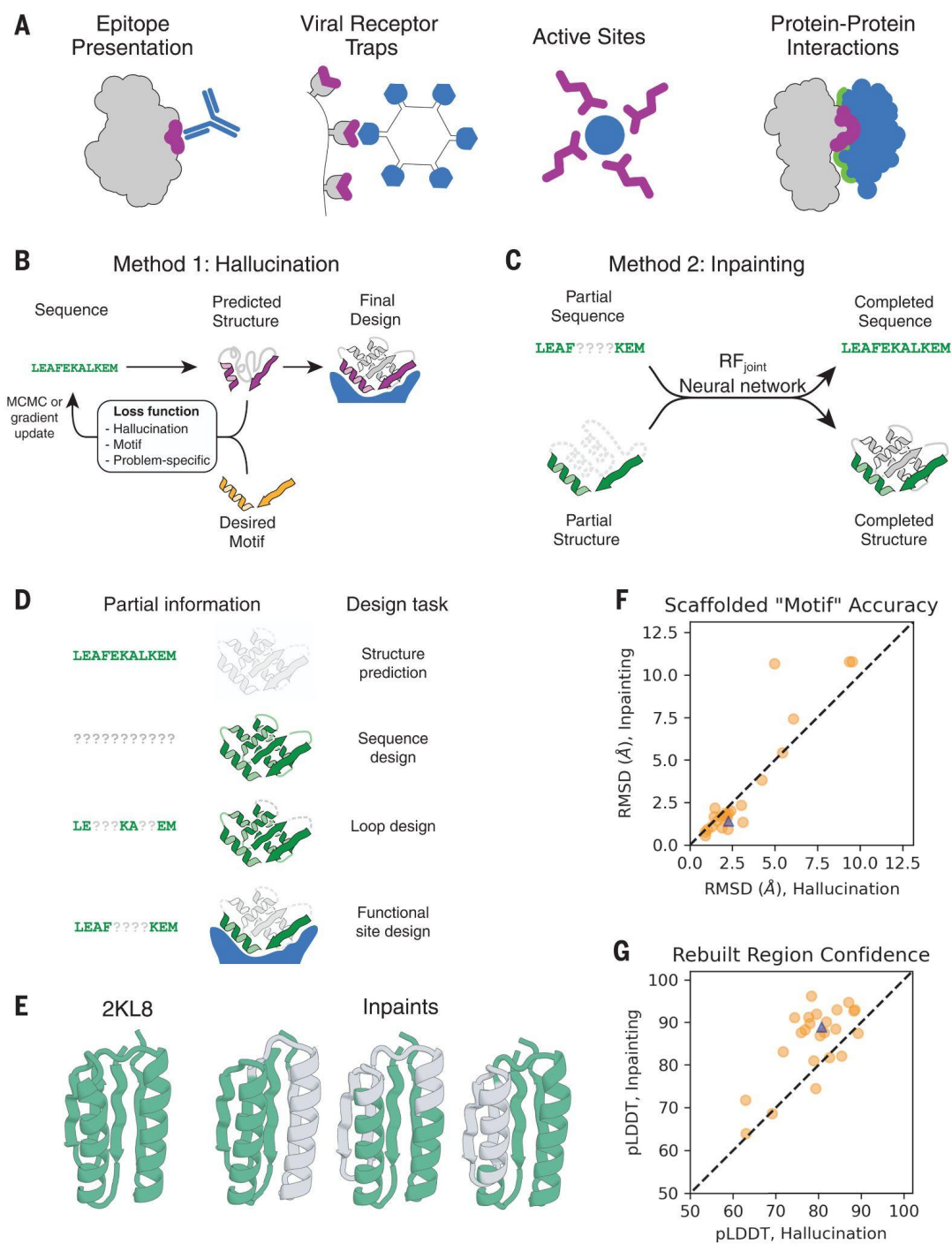


Figure. 1. Methods for protein function design.

Figure and legend from [32]: (A) Applications of functional-site scaffolding. (B and C) Design methods. (B) Constrained hallucination. At each iteration, a sequence is passed to the trRosetta or RoseTTAFold neural network, which predicts 3D coordinates and inter-residue distances and

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

We began from a RoseTTAFold (RF) model trained for structure prediction [20] and carried out further training on fixed-backbone sequence design in addition to the standard fixed-sequence structure prediction task to avoid model degradation. This model, denoted RFimplicit, was able to recover small, contiguous regions missing both

sequence and structure. Encouraged by this result, we trained a model explicitly on inpainting segments with missing sequence and structure given the surrounding protein context, in addition to sequence design and structure prediction tasks (Fig. 2 A-B). The resulting model was able to inpaint missing regions with high fidelity (Fig. 1E) and performed well at sequence design (32% native sequence recovery during training) and structure prediction (Fig. 2C). We call this network RFjoint.

A more detailed description of the training procedure for RFjoint is as follows: Starting with pre-trained RosettaFold, we fine-tuned this model for inpainting, for an additional 27 epochs on three tasks, training only on the PDB training set. For tasks 1 and 2 (fixed backbone sequence design, and inpainting respectively, chosen 33% of the time each) were masked in essentially the same manner. Contiguous regions of 10-35 amino acids comprising at least one full secondary structure element (helix, loop or strand) were masked out (Task 1: only sequence masked; Task 2: sequence and structure masked). The sequence and structure of a further 3-6 'flanking' residues were masked out either side of this contiguous region (Fig. 2A, red). The distograms (but not angle maps or amino acid identity) were provided for the residue immediately N- and C-terminal to the central contiguous masked region (Fig. 2A, asterisks). Noise was also applied to these two positions, by randomly translating them following a normal distribution ($\mu = 0 \text{ \AA}$, $\sigma = 1 \text{ \AA}$), such that at inference time, coordinates would be provided to the network as a "guide" rather than as absolute positions. Losses were not applied to the flanking regions either side of these two coordinates. The masking of flanking sequence and structure modestly improved the performance of the network in the benchmarking test,

compared to just masking a 10-35 residue window (Fig. 2D). The final task (structure prediction from MSA information) was the original task the pre-trained RosettaFold was trained on, which differs slightly from the original RosettaFold network training [20]. Specifically, in this task, 15% of the MSA (excluding the input sequence) was randomly masked or corrupted (following the strategy used by AlphaFold [19], of this 15% of residues, 70% of residues were replaced with a ‘mask’ token, 10% were mutated to a random amino acid, 10% were mutated to another amino acid in the MSA column, and 10% were not replaced). Homologous template structural inputs were unchanged from the original network. The applied loss function was the same for all three tasks, and was formulated as

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

Where $\mathcal{L}_{dist}$ is a cross entropy loss over the distogram and anglegram predictions as described in [20], $\mathcal{L}_{aa}$ is a cross entropy loss over any masked positions in the input MSA, $\mathcal{L}_{tors}$ is a cross entropy loss on binned backbone dihedral angle predictions, $\mathcal{L}_{FAPE}$ is a backbone level frame aligned point error, as described in [19], with a relu cutoff of 20. $\mathcal{L}_{ddt}$ is the IDDT loss as calculated in [20]. Note that structure related losses are applied over the entire predicted protein, and the sequence cross entropy loss is only applied at masked (Tasks 1 and 2) and/or corrupted (Task 3) regions. For the fixed-backbone sequence design task (Fig. 2A, Task 1) and for the inpainting task (Fig. 2A, Task 2), no loss was applied on the ‘flanking’ region of protein N- and C-terminal to the central masked region. The learning rate was set to 0.0003 throughout the training of these three tasks, with a batch size of 512. We refer to this fine-tuned

RosettaFold inpainting model as RFjoint, and selected training curves from this model are shown in Fig. 2B,C.

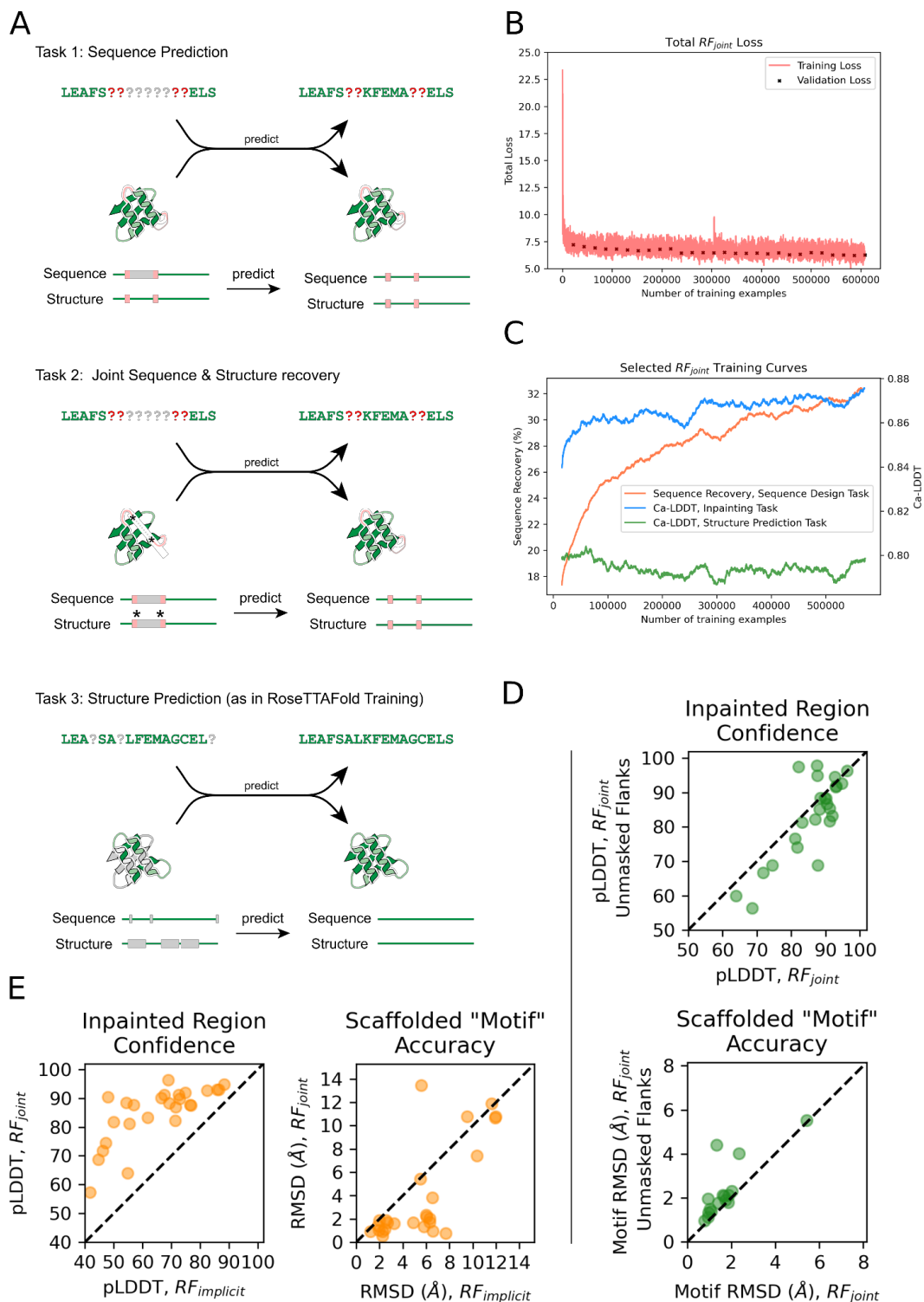


Figure 2. Training of joint sequence-structure recovery RosettaFold (RFjoint).

Figure and legend from [32]: (A) Depiction of the three tasks used to train RFjoint, which were trained with equal likelihood (see Algorithm 1). Task 1 comprised a fixed-backbone sequence design task of a continuous segment of a given protein, without the immediate up- and downstream protein visible (see Methods). Task 2 comprised an inpainting task, where the model was tasked with predicting the sequence and structure of a continuous section of protein, also without up- and downstream protein visible. Asterisks indicate “guiding points” provided as inputs during inpainting to Task 3 is the structure prediction task originally used to train RosettaFold. (B) Training curve for RFjoint, showing total training (red) and validation (black crosses) losses decreasing. (C) A selection of different losses associated with each of the three tasks. RFjoint does not severely deteriorate in its ability to predict protein structures (task 3, green line), but its ability to inpaint structure (task 2) improves dramatically (blue line). The model also learns to predict the sequence of a fixed backbone (task 1, orange line). (D) Masking out the structure and sequence of the flanking regions (depicted in (A), Tasks 1 and 2) improves inpainting performance. RFjoint was compared to an identically-trained model, except that flanking regions were not masked during training, on a benchmarking task for motif scaffolding. Both AlphaFold pLDDT in the inpainted region (top), and the “Motif” RMSD of the AlphaFold predictions (bottom) were marginally better for RFjoint. (E) RFjoint outperforms RFimplicit, both in terms of the AlphaFold pLDDT in the inpainted region (left), and in the “Motif” RMSD of the AlphaFold prediction (right). Graphs in D and E correspond to a masked window of 30 residues.

Designing metal coordinating proteins with RFjoint

Di-iron sites are important in biological systems for iron storage [59] and can mediate catalysis [60], [61]. We were able to recapitulate the di-iron site from E. coli bacterioferritin, composed of four parallel helical segments, to sub-angstrom AF-RMSD using both inpainting (Fig. 3, A to E) and hallucination (the hallucinations were not

tested owing to buried polar residues. The designs had diverse helix connectivities and low structural similarity to the parent [template modeling (TM)–score 0.55 to 0.71 to PDB ID 1BCF_A]. We chose 96 inpainted designs to test experimentally and found that 76 had soluble expression, at least eight had a spectroscopic shift indicative of Co²⁺ binding (a proxy for iron binding) [62], [63], and three (dife_inp_1, dife_inp_2, and dife_inp_3; Fig. 3B) had CD spectra consistent with the designed fold (Fig. 3D) and were stabilized by metal binding (Fig. 3E). Mutation of the metal binding residues abolished binding (Fig. 3B), and titration analysis of dife_inp_1 suggested that both metal binding sites were successfully scaffolded (Fig. 3C).

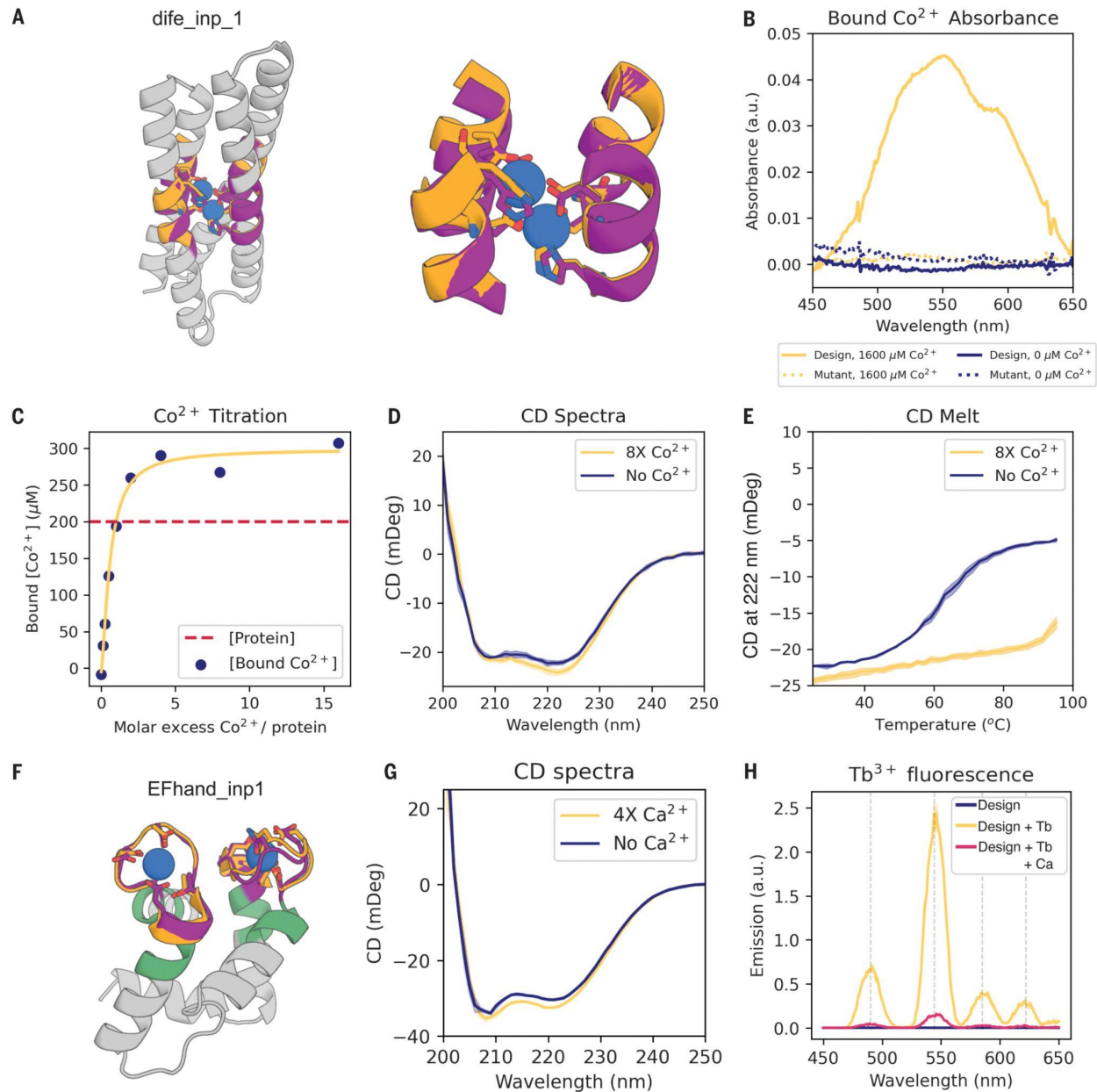


Figure. 3. Design of metal binding proteins with RFjoint.

Figure and legend from [32]: (A) Scaffolding of di-iron binding site from *E. coli* cytochrome *b1* (PDB ID 1BCF chain A residues 18 to 25, 27 to 54, 94 to 97, and 123 to 130) using inpainting. Colors: native protein scaffold, light yellow; native functional motif, orange; hallucinated scaffold, gray; hallucinated motif, purple; and bound metal, blue. (B) Absorbance spectra of dife_inp_1 (or mutant) in the presence (or absence) of an eight-fold molar excess of Co^{2+} . Peaks at 520, 555, and 600 nm,

consistent with Co^{2+} binding to the scaffolded motif (32). In the mutant, the six coordinating residues [side chains shown in (A)] are mutated to alanine (E16A, E55A, H58A, E89A, H92A, E115A). Protein concentration: 200 μM . (C) dife_inp_1 Co^{2+} titration (protein concentration: 200 μM). Quantification of the absorbance at 550 nm, using a predicted extinction coefficient of 155 for Co^{2+} binding the motif (32), is consistent with both binding sites being recapitulated. (D) CD spectra of dife_inp_1 in the presence and absence of Co^{2+} are both consistent with the predicted helical structure. (E) Temperature dependence of dife_inp_1 CD signal in the presence and absence of Co^{2+} . Coordination of Co^{2+} in the core stabilizes the protein. Protein concentration: 6.7 μM ; Co^{2+} concentration: 53.3 μM . (F) Inpainted design EFhand_inp_1 scaffolding the double EF-hand motif with input motif residues in purple, input nonmotif residues in green, and overlaid with the native motif from PDB ID 1PRW (orange). (G) CD spectra of EFhand_inp_1 incubated with and without CaCl_2 suggest stabilization of the protein upon binding calcium. (H) Tryptophan-enhanced terbium fluorescence spectra of EFhand_inp_1 suggests that the design binds terbium [64]. Terbium binding signal is competed by 1 mM CaCl_2 (red). Design metrics (AF pLDDT, motif AF-RMSD): dife_inp_1 (92, 0.65 Å) and EFhand_inp1 (84, 0.7 Å)

We next scaffolded the calcium-binding EF-hand motif [65], a 12-residue loop flanked by helices. Both constrained hallucination and inpainting readily generated scaffolds recapitulating either one or two EF-hand motifs to within 1.0 Å AF-RMSD of the native motif (Fig. 3F). We chose 20 hallucinations and 55 inpaints to display on yeast and screen for calcium binding using tryptophan-enhanced terbium fluorescence [66]. Six hallucinations and four inpaintings had fluorescence consistent with ion binding [one of these proteins (EFhand_inp_2) was designed using RFimplicit]. The top hit from yeast, the inpainted EFhand_inp_1, purified from *E. coli* as a monomer, had the expected CD spectrum (Fig. 3G) and a clear terbium binding signal (Fig. 3H) that was eliminated by CaCl_2 competition (Fig. 3H).

Designing a PDL1-binding protein with RFjoint

To design binders to the cancer checkpoint protein PD-L1, we scaffolded two discontinuous segments of the interfacial β sheet from a high-affinity mutant of PD-1 (Fig. 4A) [67]. Inpainting yielded designs with not only good AF predictions of the binder monomer (AF pLDDT > 80, motif AF-RMSD < 1.4 Å) but also of the complex between the binder and PD-L1, with an interchain predicted alignment error (AF inter-PAE) of <10 Å. Of 31 designs selected for experimental testing, one design, pdl1_inp_1, bound PD-L1 with a K_d of 326 nM (Fig. 4, B and C), worse than high-affinity consensus (HAC) PD-1 (K_d = 110 pM) [68] but better than wild-type PD-1 (K_d = 3.9 μ M) [68].

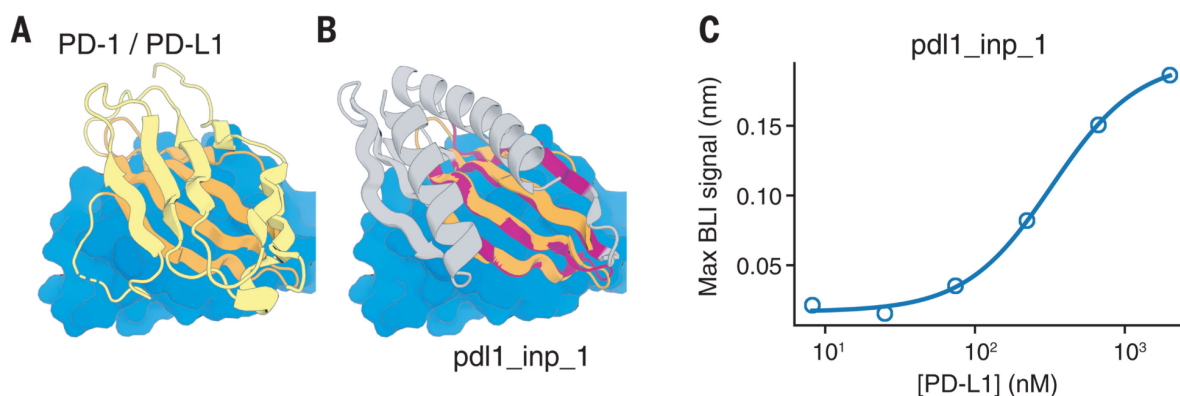


Figure 4. Design of a PDL1-binder with RFjoint

Figure and caption from [32]: Designs containing target-binding interfaces built around native-complex-derived binding motifs. Targets are in blue, native scaffolds in yellow or pink, native motifs in orange, designed scaffolds in gray, and designed motifs in purple. (A) Crystal structure of HAC PD-1 in complex with PD-L1. (B) Inpainted PD-L1 binder superimposed on PD-1 interface motif. (C) BLI binding signal versus PD-L1 concentration. K_d = 326 nM.

Unlike native PD-1, which has an immunoglobulin family β -sandwich fold, pdl1_inp_1 has two helices buttressing the interfacial β sheet, as well as an additional fifth inpainted strand extending the interface (fig. 5, A and B). The pdl1_inp_1 design expressed as a monomer (fig. 5D), was thermostable, and had a CD spectrum consistent with that of a mixed α - β fold (fig. 5E). The closest Protein Data Bank (PDB; [69]) hit had a TM-score [70] of 0.61, and the closest Basic Local Alignment Search Tool (BLAST; [71]) NR hit had a sequence identity of 25.4%.

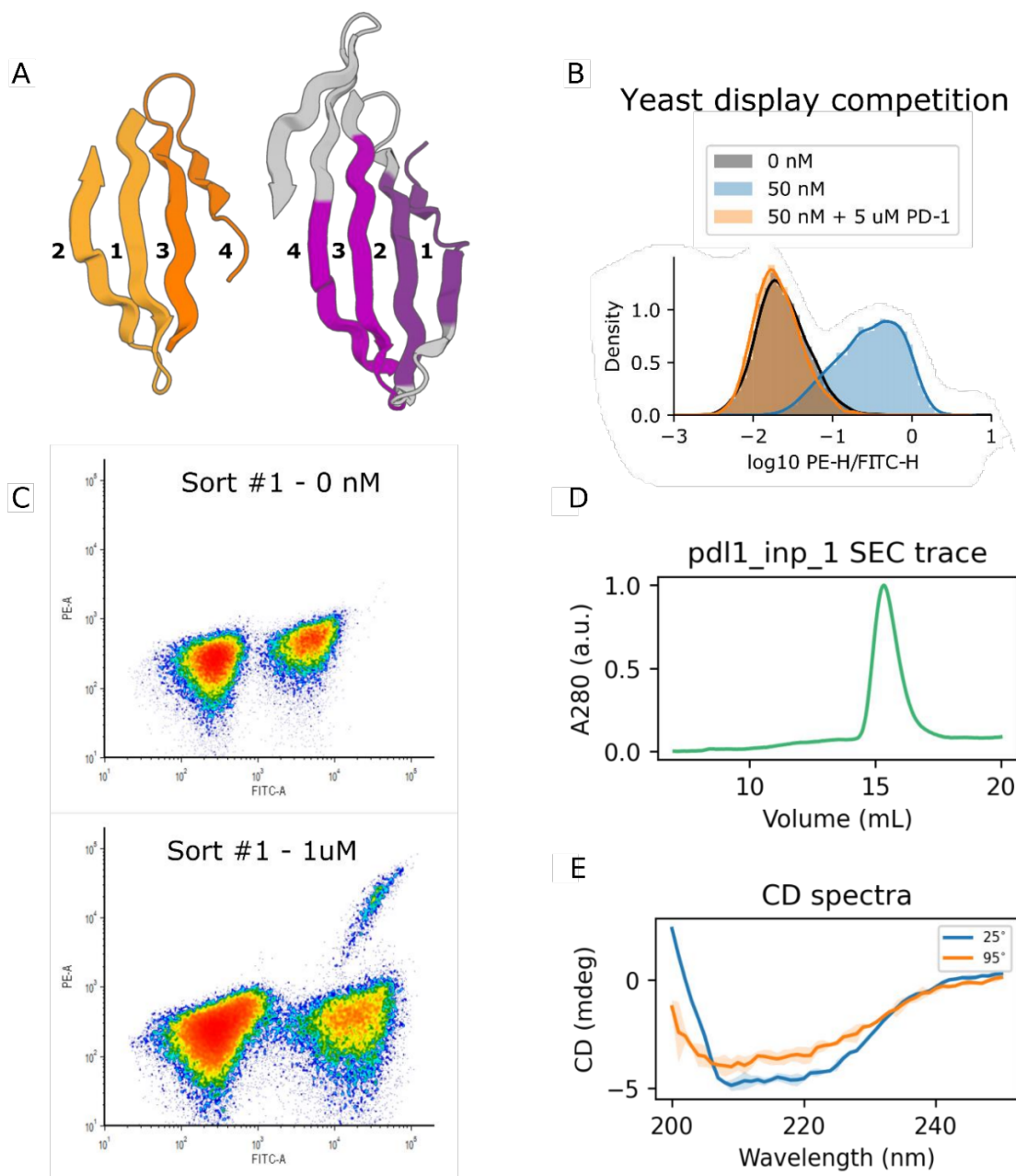


Figure 5. Further experimental characterization of inpainted PD-L1 binder

Figure and legend from [32]: (A) The looping order of the interfacial beta strands in the design (purple / dark purple) has changed dramatically from the HAC PD-1, demonstrating the ease of relooping secondary structure elements while maintaining the desired motif with inpainting. Notably, the order in which the two discontinuous strand-loop-strand submotifs appear in primary structure

has switched, as well as the order in which strands 3 and 4 from HAC PD-1, which become strands 1 and 2 in the design, respectively. (B) Binding signal (PE-H) normalized to yeast surface expression (FITC-H) of clonal yeast population displaying pdl1_inp_1 labeled with 0 or 50 nM PD-L1, or 50 nM PD-L1 + 5 μ M unlabeled PD-1. Loss of binding upon PD-1 competition suggests that pdl1_inp_1 binds PD-L1 at the native PD-1 binding site. (C) Fluorescence activated cell sorting data from yeast display binding experiments. Titles denote the concentration of a disulfide linked homodimeric PD-L1 target present in the binding reaction. Sort #1 denotes the first pooled sort of 31 designs.

Conclusions from training a protein inpainting model

The inpainting approach for scaffolding functional sites presented here requires no inputs other than the structure and sequence of the desired functional site and, unlike previous methods, does not require specifying the secondary structure or topology of the scaffold and can simultaneously generate both sequence and structure. The data reported here (along with the rest in ref [32]) support the three conjectures I raised at the beginning of this document, in that it demonstrates

1. *De novo* protein design is a molecular engineering approach that can be used to make functional molecular systems relevant for solving challenges that society faces today.
2. By using structural constraints and sampling protein structures and sequences conditioned on some motif (i.e., $p(X,S|M)$, which RFjoint seeks to predict), we can design proteins that possess useful chemical functions.
3. Training deep neural networks (such as RFjoint) to learn conditional distributions like $p(X,S|M)$ can serve as an avenue for generating functional proteins without extensive experimental search and optimization.

There are indeed limitations to the inpainting approach, the biggest of which is that it is not a generative model of the data we train it on, but rather a maximum likelihood estimator [53]. The repercussions of this are that RFjoint tends to struggle when not enough context of the overall protein fold being designed is provided (e.g., extremely small motif scaffolding tasks, unconditional protein generation). Mathematically, this corresponds to not truly modeling the full distribution of possible protein structures available for solving a given design challenge. In the next section, I will detail the most recent methods development campaign my colleagues and I have embarked on, which is to make a truly generative model over protein structures.

Findings 2: Broadly applicable protein design with RoseTTAFold

Diffusion

Introduction

To improve upon protein inpainting, we would like to design a model which does not simply produce the mean solution when presented with some conditioning information. This is incredibly important because the mean of many valid solutions to a motif scaffolding problem may not (is likely not) a solution itself. As such, my colleagues and I turned to the field of deep generative modeling to find modeling paradigms which we thought could solve the problem of sampling more fully the distribution of all possible protein structures. Here, we were inspired by the fidelity of diffusion generative models.

Denoising diffusion probabilistic models (DDPMs), a powerful class of machine learning models recently demonstrated to generate novel photorealistic images in response to text prompts [72], [73] have several properties well-suited to protein design. First, DDPMs generate highly diverse outputs – DDPMs 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; SE(3)-equivariant DDPMs are able to do this in a representation-frame independent manner. Recent work has adapted DDPMs for protein monomer design by conditioning on small protein “motifs” [74], [75] or on secondary structure and block-adjacency (“fold”) information [76]. While promising, these attempts have shown limited success in generating sequences that fold to the intended structures in silico [74], [77], likely due to the limited ability of the denoising networks to generate realistic protein backbones, and have not been tested experimentally.

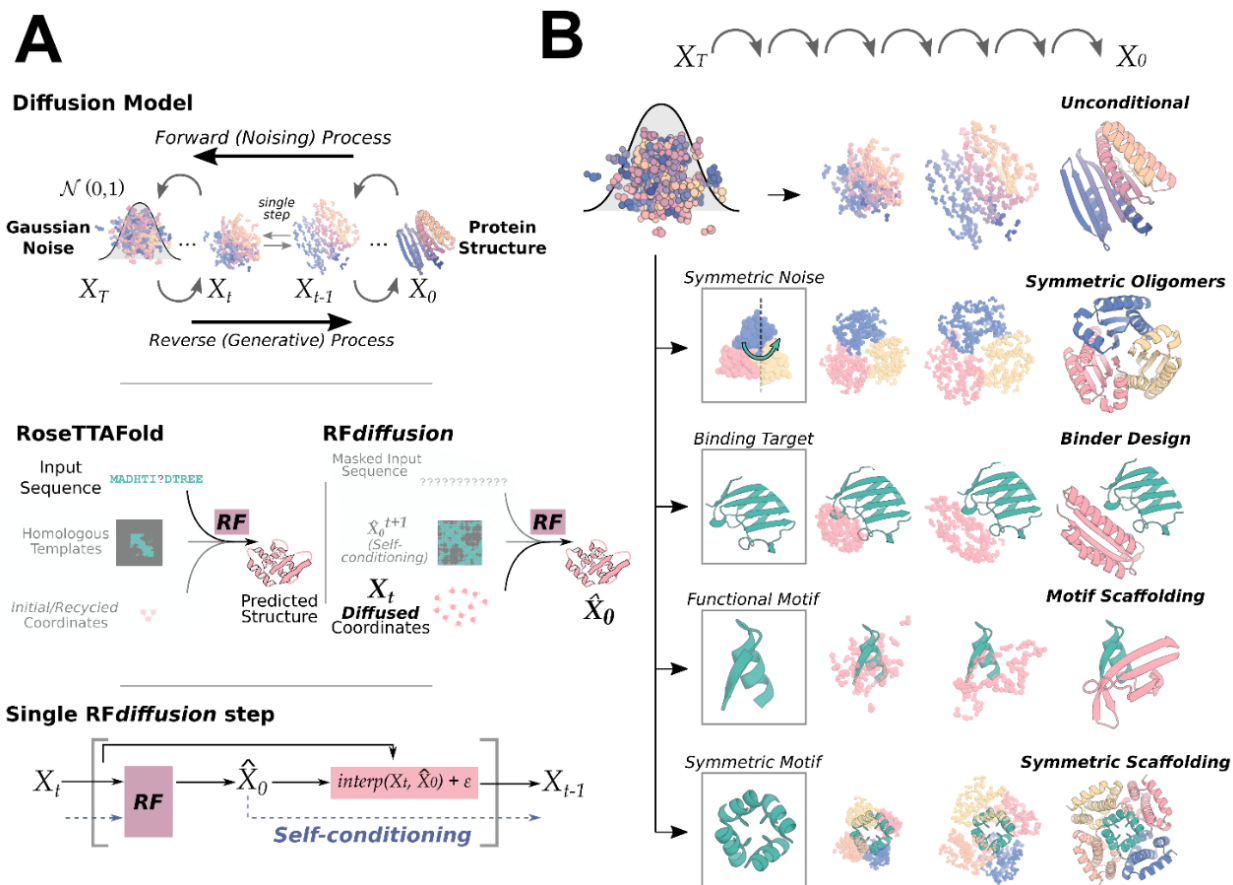


Figure 6: RFdiffusion is a denoising diffusion probabilistic model with RoseTTAFold fine-tuned as the denoising network.

Figure and legend from [33]: A) Top panel: Diffusion models for proteins are trained to recover structures of proteins corrupted with noise, and 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 trained from a pre-trained RF network with minimal architectural changes. While 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 directly (denoted X_0 in RFdiffusion). In RFdiffusion, the model receives its previous prediction as a template input (“self-conditioning”). Bottom panel: At each timestep “ t ” of a design trajectory (typically 200 steps), RFdiffusion takes X_t

and $\hat{X}_{0^{t+1}}$ from the previous step and then predicts an updated X_0 structure ($\hat{X}_{0^t}$). The coordinate input to the model at the next time step (X_{t-1}) is generated by a noisy interpolation toward X_{0^t} . B) RFdiffusion is of broad applicability to protein design. RFdiffusion generates protein structures either without additional input (top row), or by conditioning on: symmetric inputs to design symmetric oligomers (second row); a binding target (third row); protein functional motifs (fourth row); symmetric functional motifs to design symmetric oligomers scaffolds (bottom row). 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.

We construct a RoseTTAFold-based diffusion model, RFdiffusion, using the RF frame representation which comprises a C α coordinate and N-C α -C rigid orientation for each residue. We generate training inputs by simulating the noising process for a random number of steps (up to 200) on structures sampled from the Protein Data Bank (PDB) [69] (Figure 6A top, noising process). For translations, we perturb C α coordinates with 3D Gaussian noise. For residue orientations, we use Brownian motion on the manifold of rotation matrices (building on refs [[78], [79]]). To enable RFdiffusion to learn to reverse each step of the noising process (Figure 6A top, generative 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. This loss drives denoising trajectories to match the data distribution at each timestep and hence to converge on structures of designable protein backbones (Fig. 7). 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.

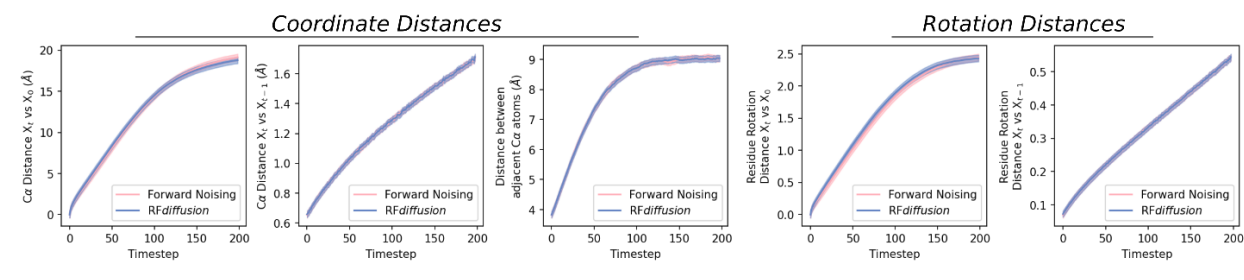


Figure 7: RFdiffusion learns the distribution of the denoising process.

Figure and legend from [33]: Analysis of simulated forward (noising) and reverse (denoising) trajectories shows that the distribution of C_α 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 C_α coordinate at X_t and its position in X_0 ; ii) average distance between a C_α coordinate at X_t and X_{t-1} ; iii) average distance between adjacent C_α 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} .

To generate a new protein backbone, we first initialize random residue frames and RFdiffusion makes a denoised prediction (Fig 6A bottom, 6B). 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 (Fig 7). 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. 8A 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. 8A right). We use the ProteinMPNN network1 to subsequently design sequences encoding these structures, typically sampling 8 sequences per design, in line with previous work [74], [77]. 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.

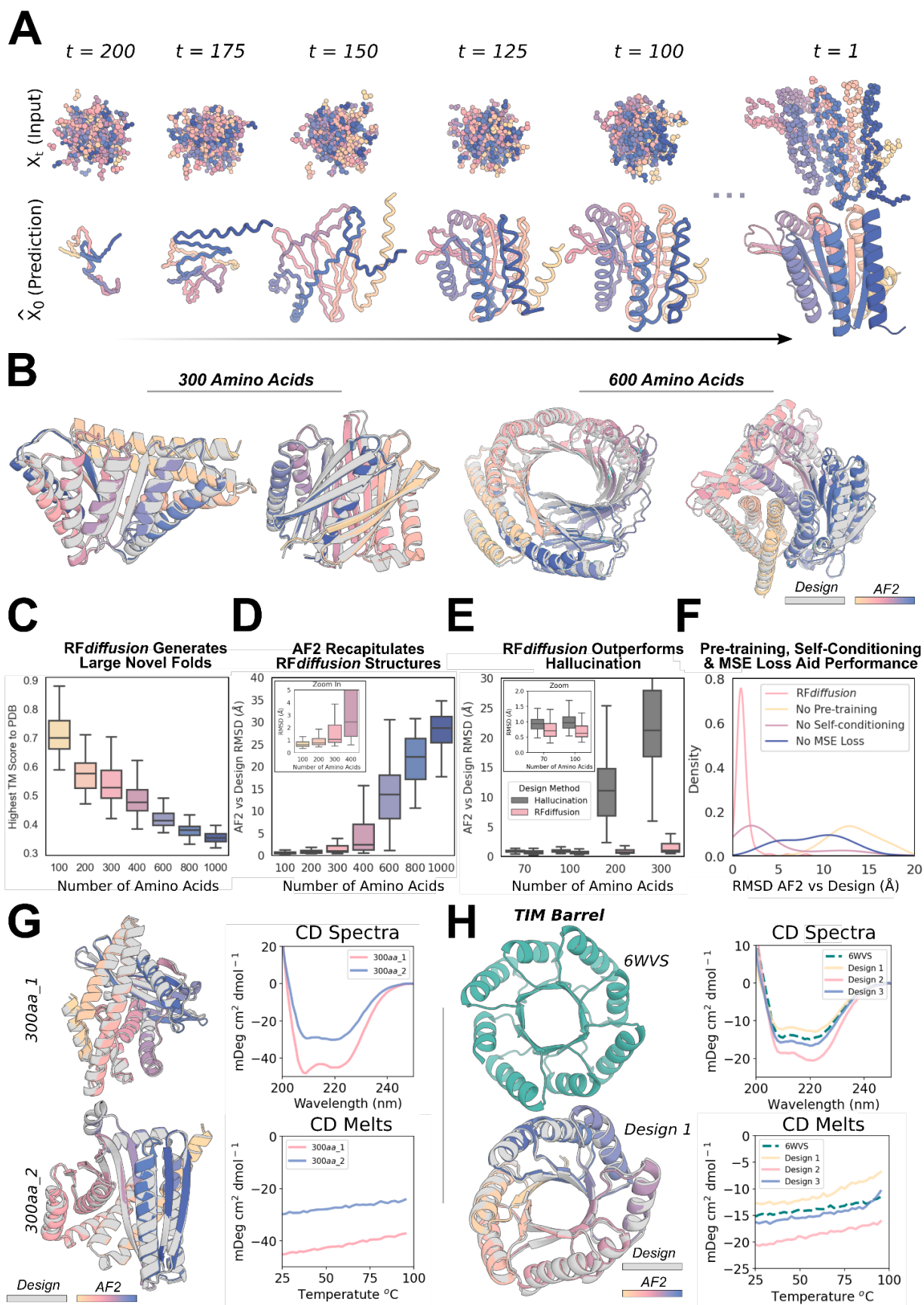


Figure 8: Outstanding performance of RFdiffusion for monomer generation.

Figure and legend from [33]: A) An example trajectory of an unconditional 300 amino acid design, depicting the input to the model (X_t) and the corresponding X_0 prediction. At early timesteps (high t), X_0 predictions bear little resemblance to a protein, but are gradually refined into a protein structure. B) 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. C) 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. D) Unconditional samples are closely re-predicted by AF2. Beyond 400 amino acids, the recapitulation by AF2 deteriorates. E) RFdiffusion significantly outperforms Hallucination (with RoseTTAFold) at unconditional monomer generation (two-way ANOVA & Tukey's test, $p < 0.001$). While Hallucination successfully generates designs up to 100 amino acids in length, success rates rapidly deteriorate beyond this length. F) Ablating pre-training (by starting from untrained RF), 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 motif-scaffolding problem "5TPN" G) 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. H) 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).

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 [74], [75], [76], [77], and 2) with self-conditioning [80], where the model can condition on previous predictions between

timesteps (Fig. 6A bottom row). The latter strategy was inspired by the success of “recycling” in AF2, which is also central to the more recent RF model used here. Self-conditioning within RFdiffusion dramatically improved performance on in silico benchmarks encompassing both conditional and unconditional protein design tasks (Fig 8F). Increased coherence of predictions within self-conditioned trajectories may, at least in part, explain these performance increases. Fine-tuning RFdiffusion from pre-trained RF weights was far more successful than training for an equivalent length of time from untrained weights (Fig. 8F) and the MSE loss was also crucial (Fig. 8F). For all in silico benchmarks in this work, we use the AF2 structure prediction network [19] 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. This definition of success is significantly more stringent than those described elsewhere (refs [[74], [76], [77], [81]]) but is a good predictor of experimental success [31], [32], [82].

Unconditional protein monomer generation

As illustrated in Fig. 8B-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. The designs are diverse, spanning a wide range of alpha-, beta- and mixed alpha-beta- topologies, with AF2 predictions (Fig. 8D) 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 we speculate they are often 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). 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 (Fig. 8G showing two).

Fold conditioned monomer generation

It is often desirable to be able to specify a particular protein fold during design (such as TIM barrels or cavity-containing NTF2s for small molecule binder and enzyme design), 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. 8H). *In silico* success rates were 42.5% and 54.1% for TIM barrels and NTF2 folds respectively, 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. 8H showing two).

Design of higher order oligomers

There is considerable interest in designing symmetric oligomers, which can serve as vaccine platforms [83], delivery vehicles [84], and catalysts [85]. 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.

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. 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, particularly when guided by an auxiliary potential which biases trajectories to make more inter- and intra-chain contact. As illustrated in Fig. 9, RFdiffusion designs are nearly indistinguishable from AF2 predictions of the structures adopted by the designed sequences (predictions of the full assemblies for the cyclic and dihedral designs, and trimeric substructures of the octahedral and icosahedral designs), and many have little resemblance to previously

solved protein structures. A number of the oligomeric topologies are not seen in the PDB, including two-layer beta strand barrels (Fig. 9A, C10 symmetry) and complex mixed alpha/beta topologies (Fig. 9A, 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). 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. 9).

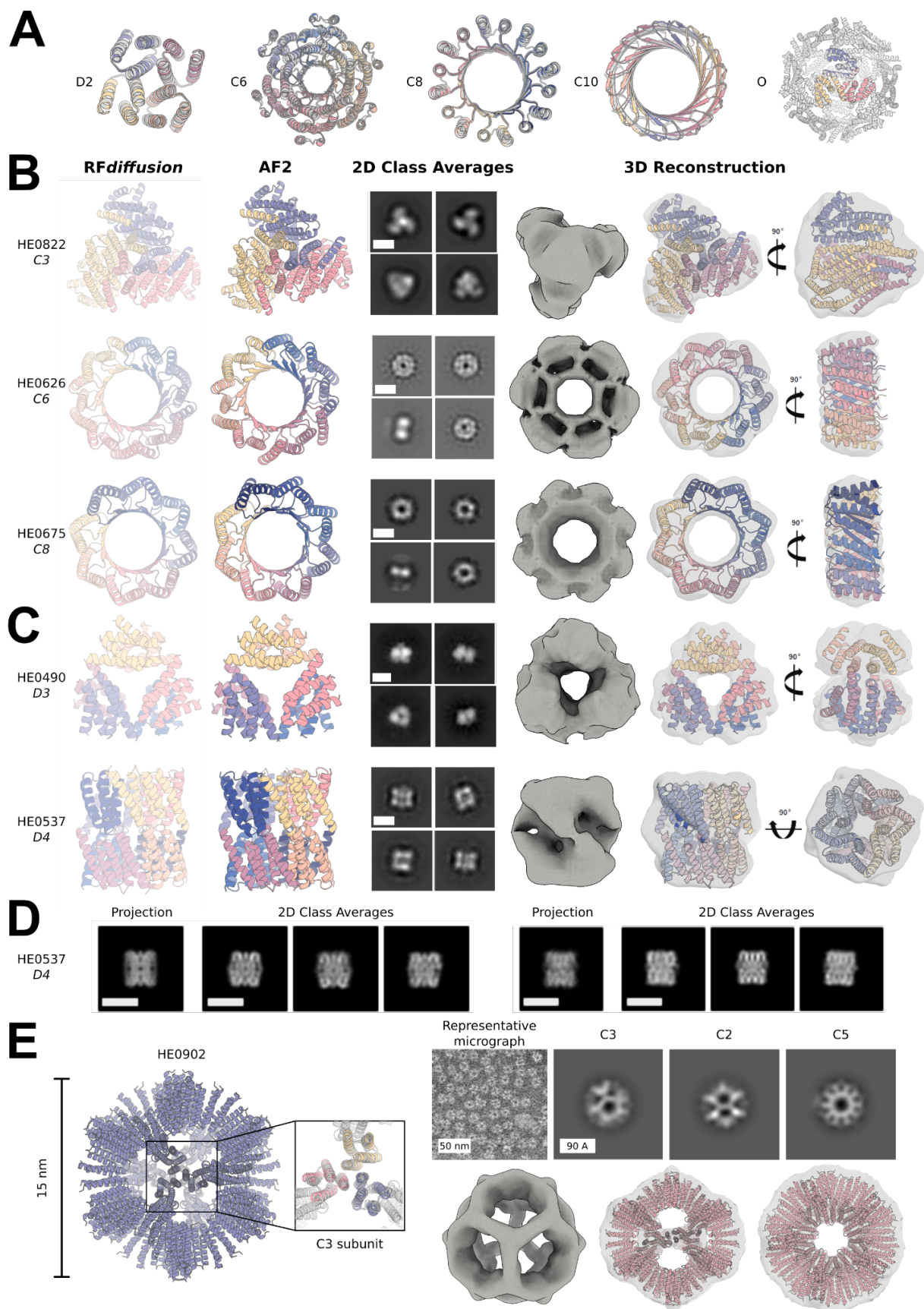


Figure 9: Design and experimental characterization of high-order symmetric oligomers.

Figure and caption from [33]: A) RFdiffusion-generated assemblies overlaid with the AF2 structure predictions based on the designed sequences; in all 5 cases they are nearly indistinguishable. Symmetries are indicated to the left of the design models. The octahedral symmetries were validated by their C3 subunits only, as shown in panel A. B-C) Designed assemblies characterized by negative stain electron microscopy. Model symmetries: B) Cyclic: C3 (HE0822, 350 AA/chain); C6 (HE0626, 100 AA/chain); C8 (HE0675, 60 AA/ chain) C) Dihedral: D3 (HE0490, 80 AA/ chain); and D4 (HE0537, 100 AA/ chain). From left to right: 1) symmetric design model, 2) AF2 prediction of design following sequence design with ProteinMPNN, 3) 2D class averages showing a combination of (at minimum) top and side views (scale bar = 60 Å for all class averages), 4) 3D reconstructions from class averages with the design model fit into the density map. The overall shapes are closely consistent with the design models, and confirm the intended oligomeric state. As in A), the AF2 predictions of each design are nearly indistinguishable from the original diffusion 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) Two orthogonal side views of HE0537 by cryo-EM. Representative 2D class averages from the cryo-EM data are shown to the right of the predicted 2D projection images of the computational design model (lowpass filtered to 8 Å), which appear nearly identical to the experimental data. Scale bar shown is 60 Å for all images. E) Characterized icosahedral particle (HE0902, 100 AA/ chain) by negative stain electron microscopy. 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 representing each axis of symmetry (C3, C2, and C5, from left to right) with their corresponding 3D reconstruction map views shown directly below and demonstrating high agreement to the design model.

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 RFjoint Inpainting [32], constrained Hallucination [86], and other DDPMs [74], [81]. 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 comprising 25 motif-scaffolding design problems addressed in six recent publications encompassing several design methodologies [32], [54], [74], [86], [87]. 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 RFjoint Inpainting (Fig. 10A-B). For 19/23 of the problems solved by RFdiffusion, the fraction of successful designs is higher than either Hallucination or RFjoint 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. Furthermore, the ability of RFdiffusion to scaffold functional motifs is not related to their presence in the

RFdiffusion training set. 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 40. In silico success has been described elsewhere⁴, but experimental success has not been reported. We used an RFdiffusion model fine-tuned on protein complexes 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. 10C-D), three orders of magnitude higher affinity than the reported 600nM affinity of the p53 peptide alone [88]. The experimental success rate for this problem was particularly striking with 55/95 designs showing some detectable binding at 10 μ M (Fig. 10E) and multiple designs with affinities in the low- to sub- nanomolar range (Fig. 10D).

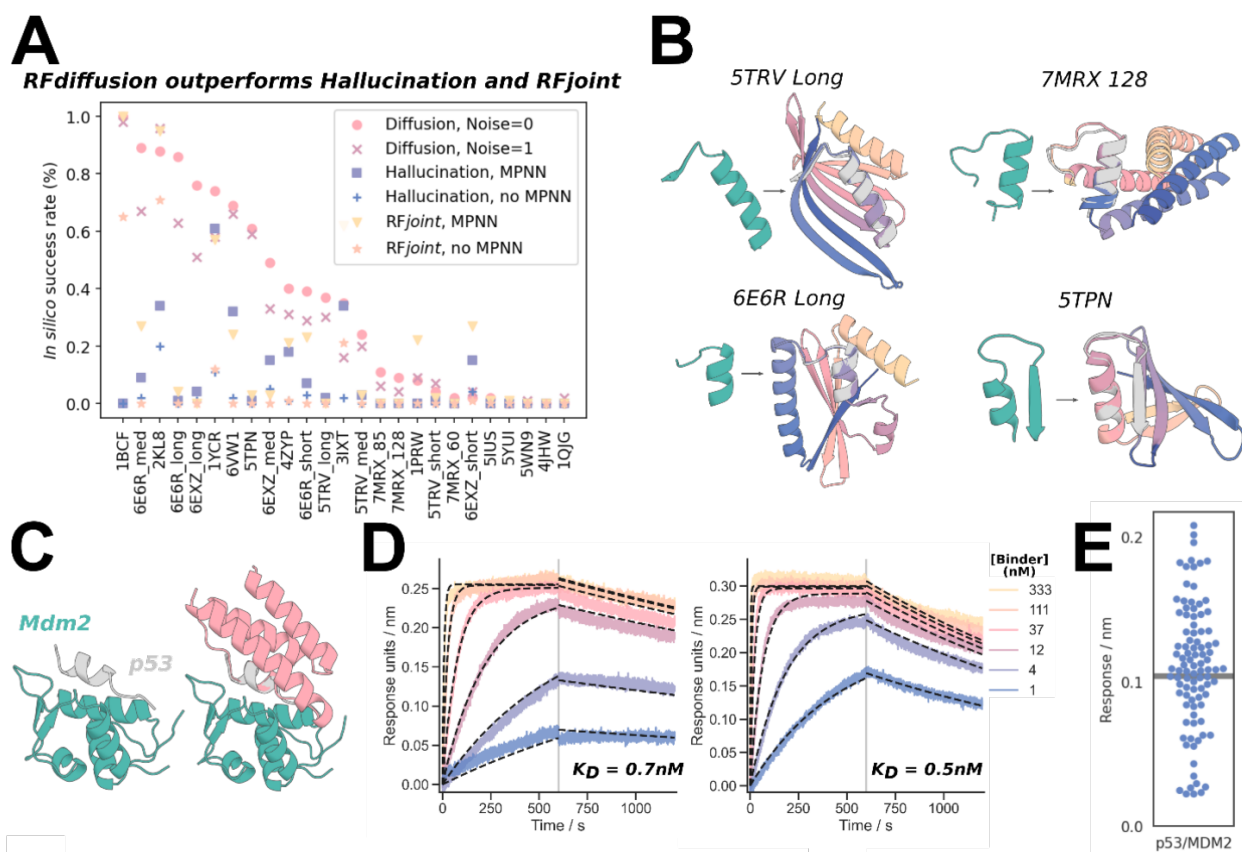


Figure 10: Scaffolding of diverse functional-sites with RFdiffusion.

Figure and caption from [33]: **A)** RFdiffusion has state of the art performance across 25 benchmark motif scaffolding problems collected from six recent publications, encompassing a broad range of motifs (Supplementary Table 9). In silico success was defined as AF2 RMSD to design model $< 2\text{\AA}$, AF2 RMSD to the native functional site (the “motif”) $< 1\text{\AA}$, and AF2 predicted alignment error (pAE) < 5 , and the examples are ordered by in silico success rate with RFdiffusion (with noise scale = 0). 100 designs were generated per problem, with no prior optimization on the benchmark set (some optimization was necessary for the Hallucination results). Supplementary Table 10 presents full results. Note that while RFdiffusion robustly outperforms both RF hallucination and RF_{joint} inpainting, in silico success rates on the problems are correlated between the three methods, and RFdiffusion can still struggle on the more challenging problems (i.e., where all three methods have low success). **B)** Four examples of designs for benchmarking problems where RFdiffusion significantly outperforms existing

methods. Teal: native motif; colors: AF2 prediction of an RFdiffusion 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 native p53 helix that binds to MDM2 (left) and makes additional contacts with the target (right, average 31% increased surface area). Note that these designs were generated with an RFdiffusion model fine-tuned on protein complexes. **D)** Biolayer interferometry (BLI) measurements demonstrate high affinity (0.7nM and 0.5nM) binding to MDM2 for the two designs shown in **C**; the native p53 helix affinity is 600nM. **E)** Experimental success rates were high, with 55/95 designs showing significant binding to MDM2 (> 50% of maximum response), and 32 of these eluted as monomers by SEC.

Symmetric motif scaffolding for metal coordinating oligomers

The ability to scaffold functional sites with any desired symmetry opens up new approaches to designing metal-coordinating protein assemblies [89], [90]. 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 Ni²⁺ binding sites. We designed C₄ protein assemblies with four central histidine imidazoles arranged in an ideal Ni²⁺-binding site with square planar coordination geometry. Diverse designs starting from various different C₄-symmetric histidine square planar sites (Fig. 11A) had good in silico success, with the histidine residues in near ideal geometries for coordinating metal in the AF2 predicted structures (Fig. 11D). We

expressed and purified 44 designs in *E. coli.*, and found that 37 had SEC chromatograms consistent with the intended oligomeric state. 36 of these designs were tested for Ni²⁺ coordination by isothermal titration calorimetry. 18 designs bound Ni²⁺ with dissociation constants ranging from low nanomolar to low micromolar (Fig. 11B,D). 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. 11D, left), suggesting that Ni²⁺ coordination is entropically driven in these assemblies. To confirm that Ni²⁺ 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 (Fig. 11C,D; one mutant did not express). We structurally characterized by nsEM a subset of the designs – NiB1.17, NiB1.12, NiB1.15, 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. 11C,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. 11C).

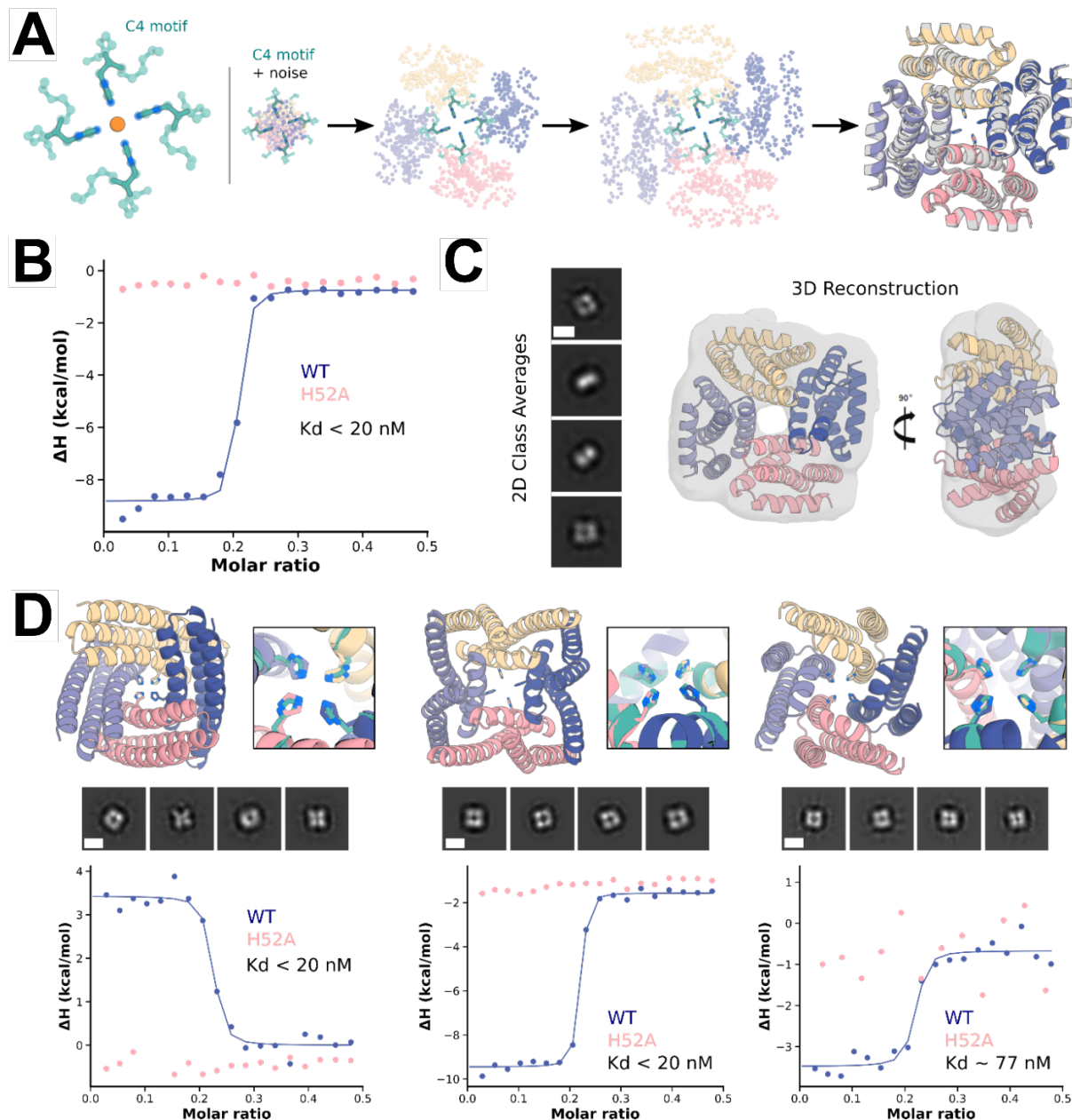


Figure 11: Symmetric motif scaffolding with RFdiffusion.

Figure and caption from [33]: A) Design of C4-symmetric oligomers to scaffold a theoretical Ni²⁺ binding motif (left). Starting from a square-planar set of histidine rotamers within three-residue helical fragments and C4-symmetric noise, an RFdiffusion trajectory iteratively builds a symmetric oligomer scaffolding the theoretical Ni²⁺ binding domain (middle). AF2 predictions (color) overlaid with the RFdiffusion design model (gray) agree closely, with backbone RMSD for the particular example < 1.0

Å (right). B) Isothermal titration calorimetry (ITC) binding isotherm of design NiB1.17 and corresponding H52A mutant. The inflection point of the wild-type isotherm (blue) displays an estimated dissociation constant of less than 20 nM at the designed metal:monomer stoichiometry of 1:4. Importantly, the H52A mutant isotherm (pink) displays complete ablation of binding, indicating the scaffolded histidine at position 52 of each protomer is critical for metal binding. C) 2D class averages (left) and corresponding 3D reconstruction with the model of design NiB1.17 docked into the 3D reconstructed density (right). The four-fold symmetry and general shape of the designed oligomer can be readily identified in the 2D class averages, with both top-down views and side views captured (scale bar = 60 Å). D) Additional experimentally characterized Ni²⁺ binding oligomers NiB2.15 (left), NiB1.12 (middle), and NiB1.20 (right) from RFdiffusion show structural diversity in successful designs. Design models and binding-site zoom (top, AF2 in colors and ideal motif in teal) show close recapitulation of the motif sidechains by AF2. 2D nsEM class averages (middle, scale bar = 60 Å), and binding isotherms for wild-type and H52A mutant (bottom) indicate tight Ni²⁺ binding mediated directly by the scaffolded histidines at the designed 1:4 stoichiometry.

Design of de novo protein-binding proteins

The design of high-affinity binders to target proteins is a grand challenge in protein design, with numerous therapeutic applications [91]. A general method to de novo design binders from target structure information alone using the physically-based Rosetta method was recently described [92]. Subsequently, utilizing ProteinMPNN for sequence design and AF2 for design filtering was found to improve design success rates[82]. However, experimental success rates were low, still requiring many thousands of designs to be screened for each design campaign[82], 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[82]. 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. 12A). 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.

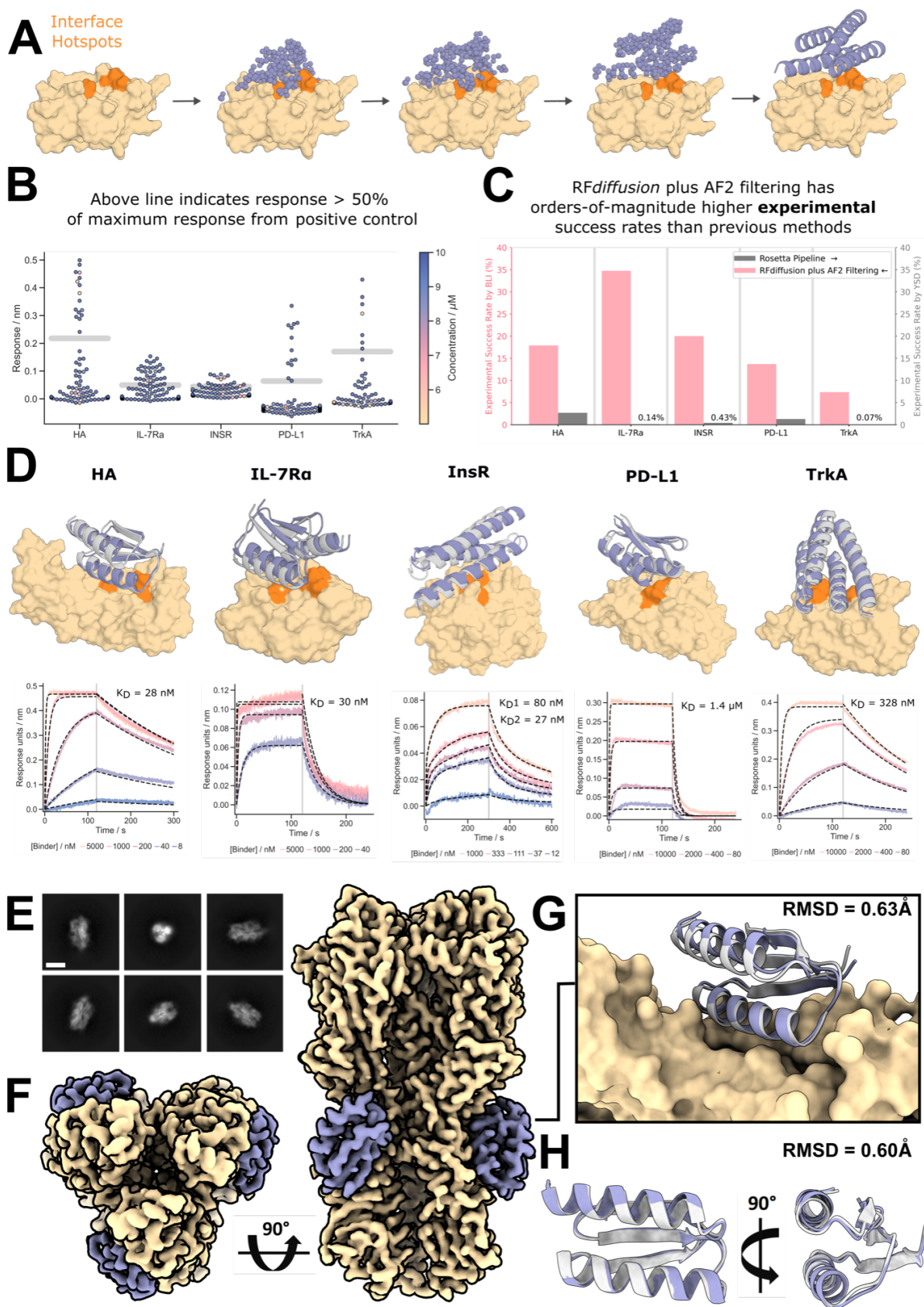


Figure 12: Design of de novo protein-binding proteins.

Figure and caption from [33]: A-B) De novo binders were designed to five protein targets; Influenza A H1 Hemagglutinin (HA), Interleukin-7 Receptor- α (IL-7R α), Insulin Receptor (InsR), Programmed Death-Ligand 1 (PD-L1), and Tropomyosin Receptor Kinase A (TrkA.) RFdiffusion generates protein binders by conditioning on interface hotspot residues. Additionally, the general topology of the binders generated by RFdiffusion can be controlled using fold-conditioning. B) De novo protein binders were identified for all five of the targets. Designs that bound at 10 μ M 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 μ M and thus may be false negatives. C) RFdiffusion designed binders have very high experimental success rates compared to the previous design campaigns against the same targets. For IL-7R α , Insulin Receptor, PD-L1, and TrkA, RFdiffusion has success rates ~2 orders-of-magnitude higher than the original design campaigns. We attribute one order-of-magnitude in success rate to RFdiffusion, and the second order of magnitude to filtering by AF2 confidence (estimated success rates for previous campaigns if AF2 confidence had been used for filtering: HA: No designs passed AF2 confidence filter, IL-7R α : 2.2%, InsR: 5.5%, PD-L1: 3.7%, TrkA: 1.5%). D) For each target, the highest affinity binder is shown alongside a BLI titration series. Reported KDs are based on global kinetic fitting with fixed global R_{max}. Yellow/orange: target/hotspot residues; gray: design model; purple: AF2 prediction (RMSD AF2 vs design, left to right: 0.7Å, 0.9Å, 1.2Å, 0.7Å). E) Cryo-EM 2D class averages of the RFdiffusion binder, HA_20, bound to Influenza Hemagglutinin, strain A/USA:Iowa/1943 H1N1 (scale bar = 10 nm). F) 2.9Å cryo-EM 3D reconstruction of the corresponding complex viewed along two orthogonal axes. The RFdiffusion binder can be seen bound to H1 along the stem of all three subunits. G) The cryo-EM structure of the HA_20 binder in complex with the H1 stem almost identically matches the computational design model (RMSD to RFdiffusion design: 0.63Å). 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.

To compare RFdiffusion to previous binder design methods, we performed binder design campaigns against 5 targets: Influenza A H1 Hemagglutinin (HA) [93], Interleukin-7 Receptor- α (IL-7R α) [92], Programmed Death-Ligand 1 (PD-L1) [92], Insulin Receptor, and Tropomyosin Receptor Kinase A (TrkA)[92]. We designed putative binders to each target, both with and without conditioning on compatible fold information, with high in silico success rates. Designs were filtered by AF2 confidence in the interface and monomer structure [82], 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 10 μ M binder concentration (Fig. 12B). In each case, a positive control was included that binds to the site targeted by the designs on the target protein [92]. 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 10 μ M (Fig. 12B)); an increase of approximately 2 orders-of-magnitude over our previous Rosetta-based method on the same targets (Fig. 12C). 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 moderate to high affinities with no further experimental optimization, including HA and IL-7R α binders with affinities of approximately 30nM (Fig. 12D). To assess binder specificity, 6 of the highest affinity

IL-7R α binders were assessed via competition BLI, and all 6 competed for binding with the structurally validated positive control.

Conclusions from training a diffusion generative model for protein design.

It is quite clear that *RFdiffusion* is a step function improvement over previous protein design methods in many different aspects. The high *in silico* and experimental success rate of designing with *RFdiffusion* means protein designers have to spend less time trying to make the systems they want to design, and less time screening designs to find functional ones. Surely this will enable the design of next generation complexity in functional molecular systems.

Indeed, *RFdiffusion* supports the three conjectures I raised for my thesis. We have shown protein design being used to create molecules with exceptional chemical functions that can help solve modern day problems, and we encoded these functions with structural molecular constraints. By training deep neural networks, we were able to learn conditional probability distributions which enabled users to generate functional proteins by describing structural constraints.

A component of this story that is not completely understood is why *RFdiffusion* generates proteins that are so much more “designable” (in the eyes of ProteinMPNN and AlphaFold) than those in the PDB, even though it was trained on structures in the PDB. We leave that to future work to understand.

Findings 3: *De novo* design of protein-binding macrocycles

Introduction

RFdiffusion [33] is very useful for designing proteins with various functions, however, for drug-design or biosensing applications, small cross-linked peptides or peptide-like molecules (macrocycles) are an attractive alternative to proteins due to their favorable stability *in vivo*, cell membrane permeability, and pharmacological properties [94], [95]. These advantages over traditional biologics (proteins) stem from the chemical customizability of macrocycles and their size. The stability of macrocycles is in part caused by the inefficiency of naturally occurring proteases for cleaving bonds in peptides with unnatural chirality (D-amino acids) alongside their constrained conformational freedom -- both of these prevent efficient binding of a catalytically competent protease active site to the macrocycles. The membrane permeability of macrocycles is enabled in part by their small size, and their ability to be chemically customized for penetration into lipid bilayers (e.g., via N-methylation) [95]. This permeability grants drug and biosensor designers the ability to target intracellular molecules, which is self-evidently extremely important and not possible with traditional biologics.

Due to the clear advantages that macrocycles have over proteins in certain scenarios, we set out to adapt our diffusion modeling pipeline for the design of protein-binding macrocycles. Broadly, there are two main ways in which we have done this. First, we incorporated an “inference-time” alteration to the forward pass of RFdiffusion enabling the generation of cyclized/cyclic peptides, which is inspired by ref [96] and adapted for

protein binder design. Subsequently, to overcome the inability of cyclic peptides to be genetically encoded which prevents their library-scale design and screening, we trained a new version of RFdiffusion that produces disulfide-linked peptide loops.

RFpeptides: Design of protein-binding cyclic peptides with an RFdiffusion inference-time alteration

In ref [96], Rettie et al introduce an alteration to the input relative position encoding features in AF2 that enables the network to predict the structure of cyclic peptides. Specifically, the relative position encoding was altered to encode the fact that the termini of any peptide were no longer far apart in the sequence, but actually neighboring.

We added the cyclic positional encoding scheme to RFdiffusion (Fig 13A) and observed a robust generation of diverse macrocyclic peptides (Fig 13B). Encouraged by the transferability of the cyclic positional encoding, we set out to employ RFdiffusion for the *de novo* design of protein-binding macrocycles for several compelling reasons: (1) We expected the high experimental success rate of RFdiffusion [33], [35] for protein binder design to carry over to macrocycle binder design; (2) RFdiffusion should be considerably faster than hallucinating macrocycle binders AFCycDesign [96] due to the need for multiple backward passes of gradient descent involved in the latter; (3) the method can still take advantage of current built-in functionalities of RFdiffusion, such as “hotspot” targeting and “motif” scaffolding; and (4) the method should be directly transferable to other current and future RoseTTAFold-based design networks, such as RFdiffusion All Atom [97].

We modified the RFdiffusion protein binder design pipeline to use cyclic relative position encodings for the generated chain, and standard positional encodings for the target and inter

binder-target indices (Fig 13C). RFdiffusion readily generated macrocycles with diverse secondary structure content against target proteins, and the inclusion of standard RFdiffusion hotspot features clearly shifted the distribution of generated binders toward desired residues (Fig 13D). We then completed our design pipeline by using ProteinMPNN to design amino acid sequences compatible with the backbones generated by RFdiffusion. Designs at this point could then be downselected for experimental testing with AFCycDesign and Rosetta DDG calculations (Fig 13E).

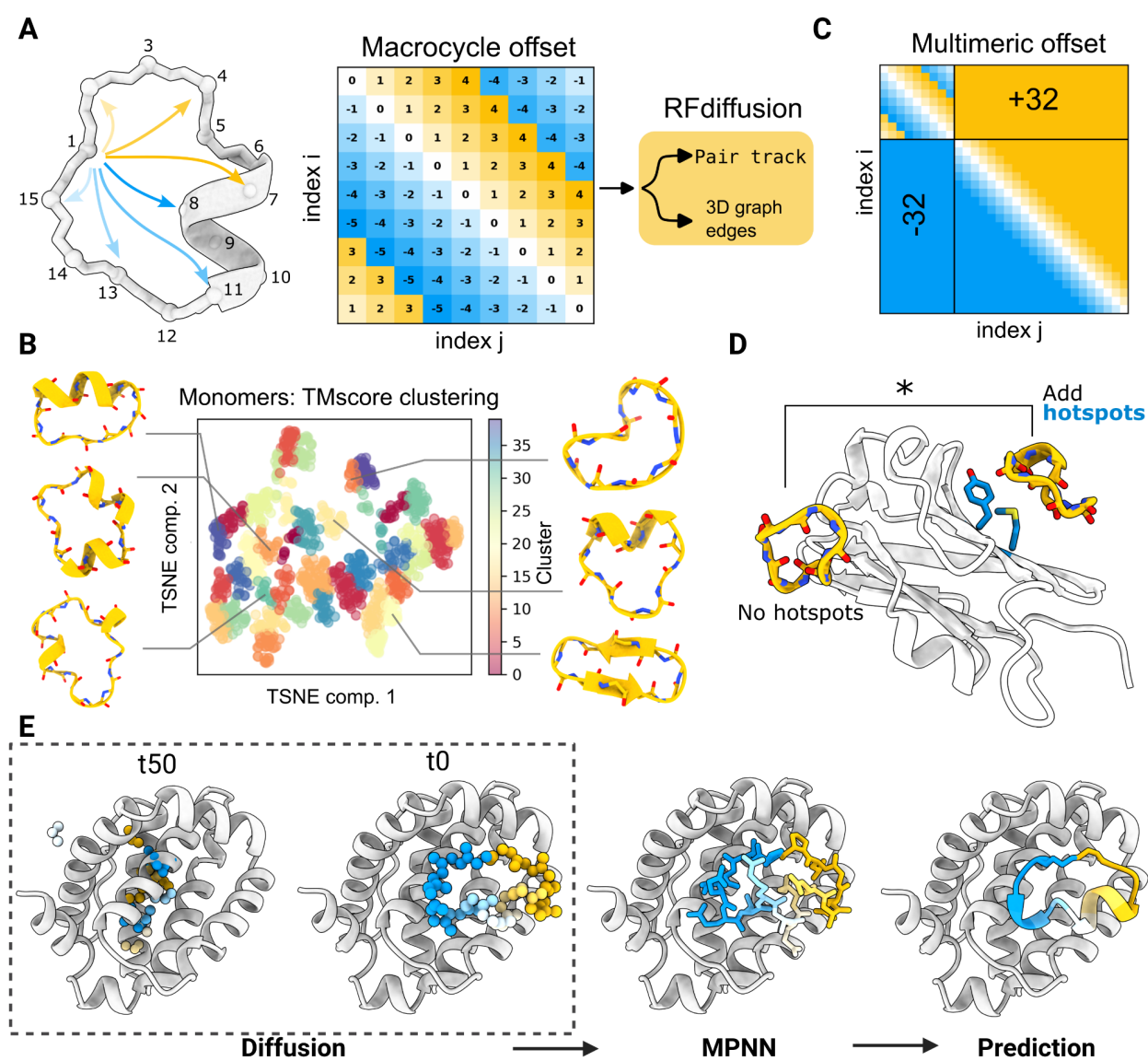


Figure 13: Introduction of cyclic relative position encodings into RFdiffusion enables the computational design of macrocyclic peptides

(A) Introduction of a cyclically symmetric relative position encoding to RFdiffusion enables the generation of macrocyclic peptide backbones in which would-be termini are covalently linked via a peptide bond. The relative position encodings are cyclized by switching from positive relative position encodings (i.e., “to the right”) to negative encodings (i.e., “to the left”) when index j is more than halfway around the peptide relative to index i . (B) Multi-chain diffusion trajectories are also made possible by combining the cyclic relative position encodings for the binder chain with standard positional encodings with the target (C) For multi-chain diffusion trajectories (e.g., macrocycle binder design), the relative position encodings for the macrocycle chain are cyclized, and that for the target chain and inter-chain are kept as standard. (D) When performing multi-chain (i.e., binder design) macrocycle diffusion, the RFdiffusion “hotspots” [33] can still be used for epitope specific targeting. These features bias the distribution of outputs more toward the desired epitope ($p < 0.05$ using a Kolmogorov-Smirnov test for changing the distribution of the smallest CA-CA distance between residue A99 in PDL1 (pdbID 5O45) and the designed peptide).

Once it was clear that RFdiffusion was generating plausible macrocycle conformations against targets in desired locations, we set out to experimentally test the protocol on a variety of targets. Figure 14A displays the design model for one such macrocycle design campaign against protein GABARAP. Notably, the RFpeptides pipeline designed a macrocycle with good shape complementarity to the target, and nontrivial beta-strand secondary structure. The peptide, *GABA_D8*, binds its target with high affinity (Fig 14B). In a second design campaign against protein Rbta, we were able to solve an X-ray crystal structure of one experimentally validated binder against the target (Fig 14C). The crystal structure exhibits remarkable similarity to the design model, deviating by 1.3 Angstroms when aligned on the target (a measure of how conserved the designed interface was) and 0.34 Angstroms when aligned on the macrocycle alone (Fig 14D).

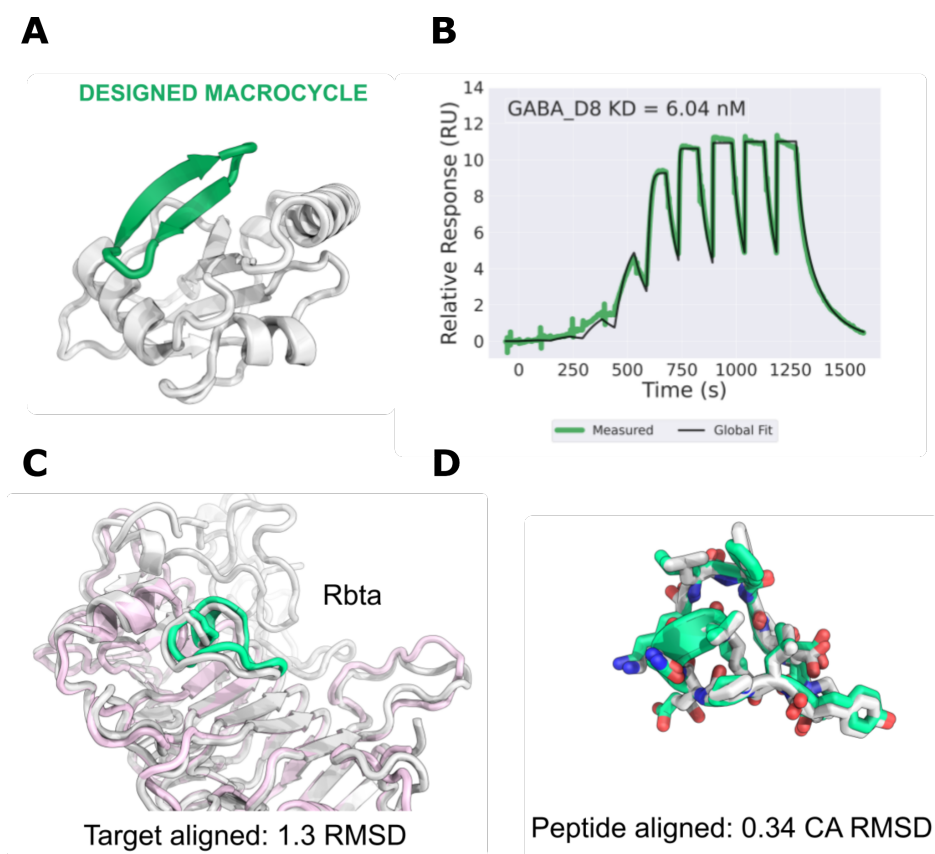


Figure 14: Design of de novo protein-binding macrocycles.

Figure adapted from Rettie et al., 2024 (in prep): (A) Design model of peptide GABA_D8 against target GABARAP, designed and characterized by Stephen Rettie using RFpeptides. (B) Surface plasmon resonance response trace and global kinetic fit for GABA_D8 binding to GABARAP displays a dissociation constant of 6nM. (C) Crystal structure of Rbta bound with de novo designed macrocycle from RFpeptides design campaign (performed by Rettie; crystal solved by Asim K. Bera). The binding pose of the peptide in the crystal structure closely matches the design model (1.3 Angstrom RMSD). (D) The conformation of the peptide itself is in exceptionally close agreement to the design model, with almost no deviation in CA atomic positions (0.34 Angstrom RMSD).

Training new versions of RFdiffusion All-Atom for designing disulfide-linked linear peptides

Although extremely promising, the RFpeptides pipeline is inherently limited in its experimental throughput because cyclic peptides are not genetically encodable, and thus must be chemically synthesized. This means that platforms for high-throughput screening of designed peptides, such as yeast surface display [98], cannot be used for screening cyclic peptides. It is possible that more potent macrocyclic drugs and sensors could be discovered with higher throughput screening, and so it is interesting to possess a version of RFdiffusion which could produce genetically encodable macrocycles. One such way of doing this is through disulfide-linked linear peptides. Disulfide bonds are able to provide similar rigidity and stability gains to proteins and peptides that cyclization with a peptide bond provides, but because they only involve sidechain groups of canonical amino acids (cysteine) the bond is perfectly genetically encodable.

We set out to train a version of RFdiffusion which produces disulfide-linked peptides, and began training from a version of RoseTTAFold All-Atom [97] which was specifically trained to predict the structure of loops with disulfide bonds in them. From here, we trained this new version of RFdiffusion All-Atom (henceforth referred to as RFpeptides-disulf) similar to ref [33], with some notable differences:

1. In addition to training on protein monomers and complexes from PDB like [33], we add three new datasets
 - a. Loops from protein monomers that are linked via disulfide bond in monomers whose sequence clusters are upsampled at a rate 200x relative to [33]. The pdb monomers were sampled 40% of the time during training, and if the monomer contained a disulfide bonded loop, the loop itself was sampled as an example with probability 0.995. If a disulfide loop was sampled, the cysteine residues in the loop were atomized.

- b. “Atomized” monomer examples, sampled at 20% of all examples and featurized as in [97]
- c. “Atomized” complex examples, sampled at 20% and featurized as in [97]
- d. Regular (non atomized) complexes were sampled for the final 20% of examples.
- e. Losses are identical to [33], but do not have the rotation component of the loss function for atomized tokens (as they do not have an orientation).

The model was trained for 11 epochs of size 25600, with a learning rate of 0.0005 and an effective batch size of 16. The loss function was identical to [33]. Figure 15A shows training curves for the first 5,000 steps (i.e., 5000x16 examples) of training, with the total loss decreasing steadily. After 6.5 epochs of training, the model readily generated plausible peptide structures with qualitatively good disulfide bond geometries (Figure 15B) spanning coil/loop, alpha, and beta secondary structures. In addition to this model, we trained a second model which received additional penalties from the atomized frame aligned point error (FAPE) discovered by [97] applied on any atomized regions from the aforementioned datasets. The intuition behind this training session was to encourage RFpeptides-disulf to generate more ideal bond geometries in the disulfide bond regions. These models were also trained to perform motif scaffolding identically to [33], and subsequent to training the models were used in a design campaign to design binders for target LRP6 [99].

During this design campaign, we analyzed the ideality of disulfide-bond geometries as well as *in silico* success rates (defined as a design from RFpeptides-disulf which, after sequence design with proteinMPNN [23] and complex structure prediction with RF2aa, yielded a target-aligned RMSD between the design model of the peptide and the prediction of the peptide < 2.0 Angstroms RMSD, and an RF2aa predicted aligned error (PAE) < 10). Figure 15 displays the results of the analysis, showing that although the model trained with FAPE error on atomized regions constructs disulfide bonds with higher ideality (15C), its success rate is marginally

worse than the models trained without FAPE (15D; 3.1% *in silico* success rate vs 3.7% for non-FAPE).

In order to prospectively test the RFpeptides-disulf pipeline experimentally, the Bhardwaj group at UW-IPD has used the network in concert with ProteinMPNN and RF2aa to design putative disulfide-linked peptide binders against various targets. Using yeast display for screening, Gizem Gokce and others in the Bhardwaj lab have discovered and validated (via next generation sequencing) various RFpeptides-disulf designs that bind a desired target protein (while on yeast; Fig 15E). Together, these computational and experimental results show promise for the RFpeptides-disulf pipeline.

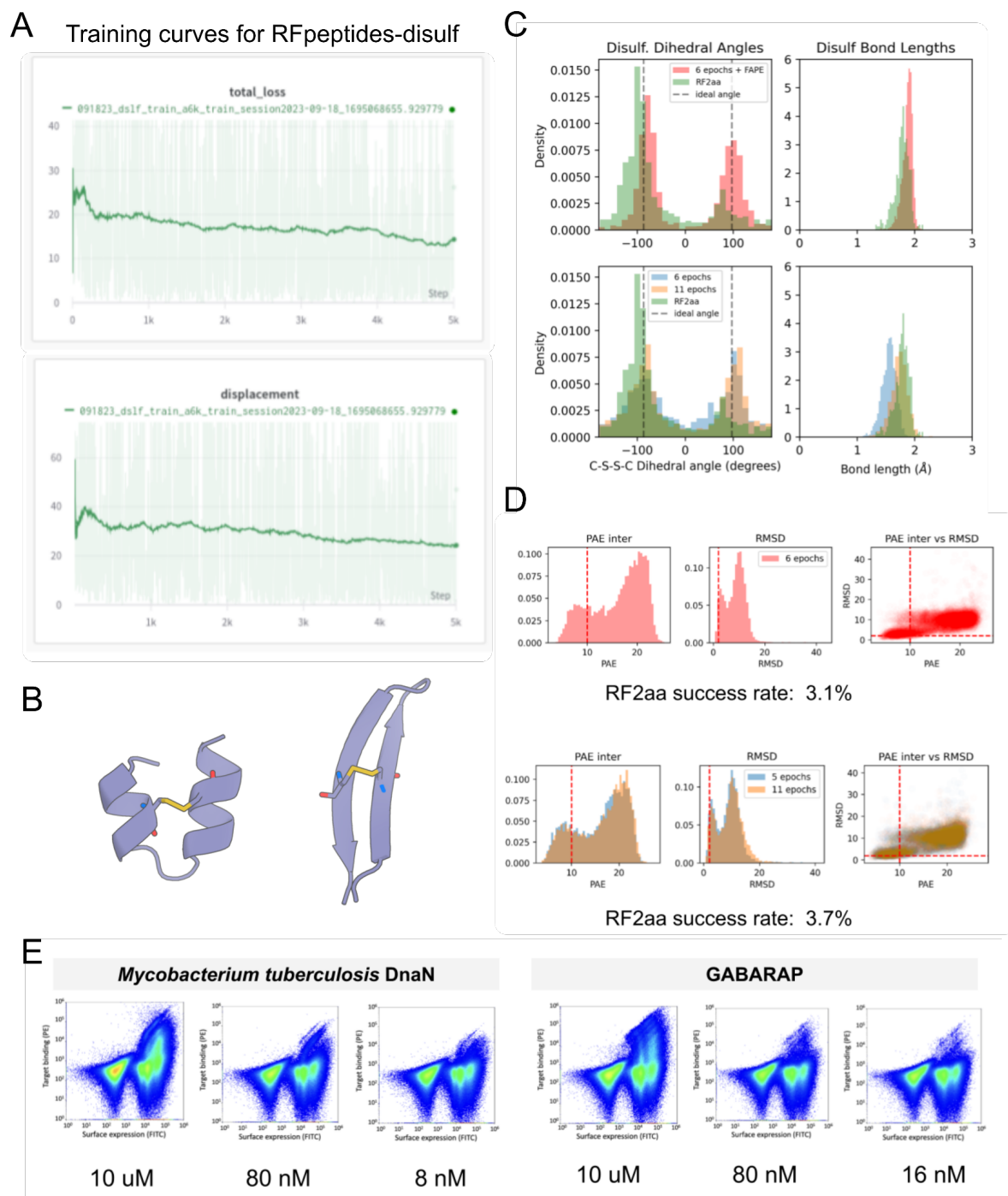


Figure 15: Training, *in silico* benchmarking, and experimental validation of RFpeptides-disulf

(A) Training curves from initial RFpeptides-disulf training session. Total loss (top) and displacement loss (bottom) are decreasing steadily. (B) Example outputs from running inference with RFpeptides-

disulf after 6 epochs of training. Peptide outputs are well-structured, containing favorable alpha-helical and beta-strand secondary structure, as well as disulfide bonds with qualitatively plausible geometry. (C) Quantification of disulfide bond geometry ideality in three RFpeptides-disulf models: The standard model (no FAPE) trained for 6 and 11 epochs of training (blue, orange respectively) and the model trained with FAPE loss for 6 epochs (red). C-S-S-C torsion angles (left) and S-S distances (right) are distributed close to the ideal angles (-87 and +97 degrees [100]) and distance (~2.03 Angstroms [101]). (D) In silico success rates for the three models when performing motif scaffolding binder design against LRP6 (using pdbID 3s8v). (E) Fluorescence activated cell sorting (FACS) plots of target binding (PE) signal vs surface expression (FITC) signal. Plots for target DnaN (left) and GABARAP (right) with titrated target concentration show binding signal even down to low concentrations of target. Designs and experiments performed by the Bhardwaj lab.

Conclusions on design of protein-binding macrocycles with RFpeptides(-disulf)

The ability to design macrocycles with atomic accuracy and with increasingly general chemistry will enable scientists to bridge the gap between small molecule drugs and large biologics, combining the best of both into a new class of therapeutics. The introduction of cyclic positional offsets in RFdiffusion to create the RFpeptides pipeline is a wonderful proof of concept both *in silico* and experimentally that atomically accurate design of cyclic peptides is possible. Additionally, it showcases that even though the theoretical size-limit of a macrocycle-protein interface is less than that of a protein-protein interface, designed macrocycles can still confer tight binding to their targets ($K_d = 6\text{nM}$ in the case of GABA_8 peptide).

When training a new model to generate disulfide-linked peptides (RFpeptides-disulf), we saw the ability to recapitulate the training in [\(Watson et al. 2023\)](#) while also adding new functionality to the model. Namely, RFpeptides-disulf can produce macrocycles which are linked via disulfide bond, usually with good bond geometry. Models trained using this strategy can be employed for protein binder design, as shown by the preliminary experimental results in Figure 15E. Given

these results, we expect to be able to find soluble versions of such peptides that bind their target in solution.

Together, the ability to design cyclic peptides and the ability to design disulfide-linked linear peptides promises a future pipeline whereby many candidate sequences can be identified as binders to a target through yeast surface display, then critical domains of those sequences can be scaffolded into cyclic peptides. The scaffolding of such domains into cyclic peptides brings with it opportunities to further stabilize these macrocycles using non-canonical amino acids.

Findings 4: Training diffusion models for all-atom motif scaffolding with arbitrary degrees of freedom

Though useful, RFDiffusion (+/- All-atom) is limited to scaffolding motifs in which all degrees of freedom (DOFs) between all atoms in a motif must be specified. In general, there are scenarios in which chemical function is conferred in molecular systems without exact constraints on all DOFs. For example, Figure 16A (left) shows that for electron transport down the length of a theoretical protein nanowire, it is possible that one would like to constrain only the distances between ions in the nanowire while leaving the geometry of the coordinating groups unconstrained. Another scenario like this could be the scaffolding of multiple enzyme active sites into an oligomer, in which the user would like to specify the *internal* geometry of each active site copy, but not specify where the active site copies are relative to the axis of symmetry or each other (Fig 16A, right).

To pursue complex function design challenges such as those described above, we need a version of RFDiffusion which accurately performs motif scaffolding but enables precise control over which degrees of freedom are required to be specified by the user. To do this, we devised a

method for performing motif scaffolding with RFdiffusion by conditioning the network on *internal* coordinates (i.e., intra-motif distances and angles) instead of specifying the motif in 3D space.

Specifically, we trained the network to perform conditional protein generation around motifs by encoding them as binned distograms and anglograms ([Yang et al. 2020](#)) which specify all 6 possible degrees of freedom between two residues or atoms. These “2D templated” motifs were generated during training by selecting between 2-8 discontinuous fragments of a protein monomer example, and creating boolean masks (Figure 16 B) that specified whether fragment *j* “could see” fragment *i* (i.e., whether or not the inter-fragment distances and angles between *i* and *j* were revealed). Intra-fragment geometries were always revealed, by definition of being a motif, and inter-fragment geometries were revealed 33% of the time.

One challenge with encoding motifs in the 2D template of RFdiffusion instead of in 3D is that the model must learn to robustly reconstruct the 3D structure of the motif which has been encoded as distances and angles. In order to encourage our new network to reconstruct motifs well, we applied a FAPE loss ([Jumper et al. 2021](#)) on intra- and inter-motif construction. This loss is known to be very well behaved and thus should be quickly learned by the network (as long as it is expressive enough), as the true structure of the motif is fully provided in the input to the network. While training, we also empirically found that removing information about N-CA-C frame orientation (all residues) from the RFdiffusion inputs as well as not scoring the frame orientations in the loss function (only for non-motif residues) enabled the diffusion models to more robustly recapitulate motifs (Fig 16C) as training progressed. Because of the removal of frame orientation information in the models output predictions (since the orientations are not scored), we trained a separate “refinement” model (identical neural architecture to RFdiffusion2d) to take in the backbones produced by RFdiffusion2D and orient the frames properly, yielding a full backbone that is then designable with ProteinMPNN.

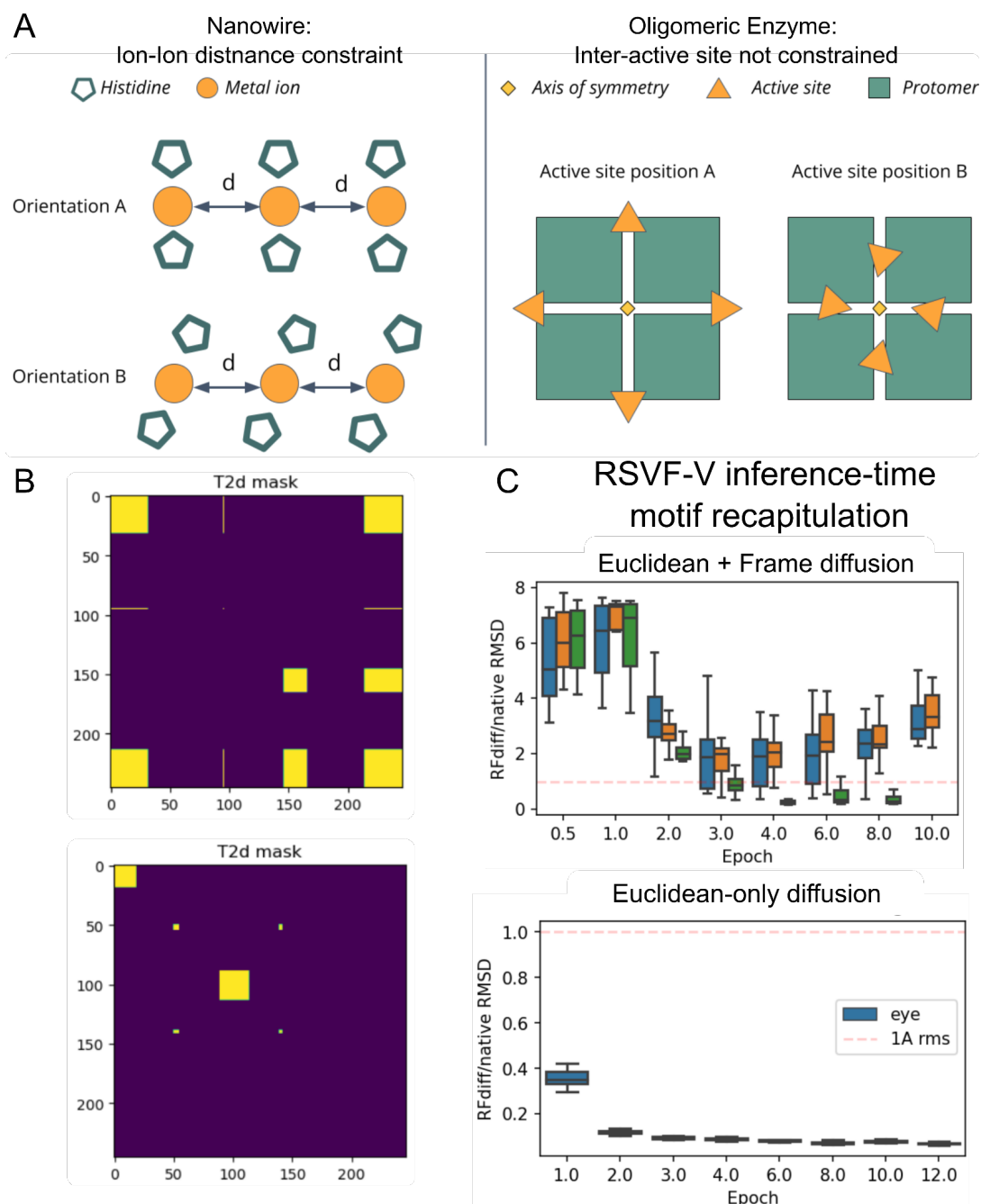


Figure 16: Introduction of inter residue distance and angle motif inputs to RFdiffusion for arbitrary degrees of freedom.

(A) Examples of motif scaffolding challenges in which it is desirable to not have to specify all possible degrees of freedom in a motif. For a nanowire that repeatedly binds metal ions in order to conduct electricity down the length of the wire, we may only want to specify the the distance

between ions and not orientation of the binding sites relative to each other (left). In the case of symmetric oligomeric enzymes with multiple copies of an active site, one wants to specify the internal geometry of each site, but not where it is relative to the symmetry axis (right). (B) Examples of boolean masks (True yellow, False purple) which define the existence of motifs (yellow on diagonal) and whether or not the pairwise distances/angles are defined between them (yellow, off diagonal). (C) Removing N-CA-C frame orientation information in the inputs to RFdiffusion, and frame loss penalties in the output (non motif regions) enables RFdiffusion to more robustly reconstruct the templated motif (RSVF site V) during inference. Top: Comparison of spherical linear interpolation (SLERP) frame noising (green), vs IGSO3 frame diffusion (orange, green). Bottom: Removing any frame noising process and only performing diffusion on CA atoms enables robust reconstruction of the motif structure.

After training this new network, it was benchmarked on the functional motif scaffolding challenge presented in [\(Watson et al. 2023\)](#). We observed that this new version of RFdiffusion (“RFdiffusion2D”) was competitive at motif scaffolding success rates *in silico* with classic RFdiffusion (Fig 17A), and exceeded at certain motif scaffolding challenges previously unsolvable by RFdiffusion (notably, an enzyme active site scaffolding challenge).

As a proof of concept and stepping stone toward our goal of making oligomeric enzymes, we sought to test RFdiffusion2D’s ability to scaffold a serine hydrolase active site into monomers (Fig 17B,C). The *in silico* results for a motif scaffolding benchmark are displayed in Figure 17C, and suggest that with enough computational sampling, one should be able to find amino acid sequences for which AF2 predicts the serine hydrolase active site will be nearly perfectly recapitulated (Fig 17D). For some problems perfect all-atom recapitulation of sidechains was rare but possible (e.g., H1/H2; Fig 17D middle), however finding sequences for which the motif was recapitulated on the backbone was easier (Fig 17D right). Interestingly, using 2 refinement

outputs (see aforementioned “refinement model”) and designing 4 amino acid sequences per refinement output instead 8 sequences for a single output (which is the standard) subtly increases computational success rate on the active site scaffolding task (Fig 17D, yellow vs green).

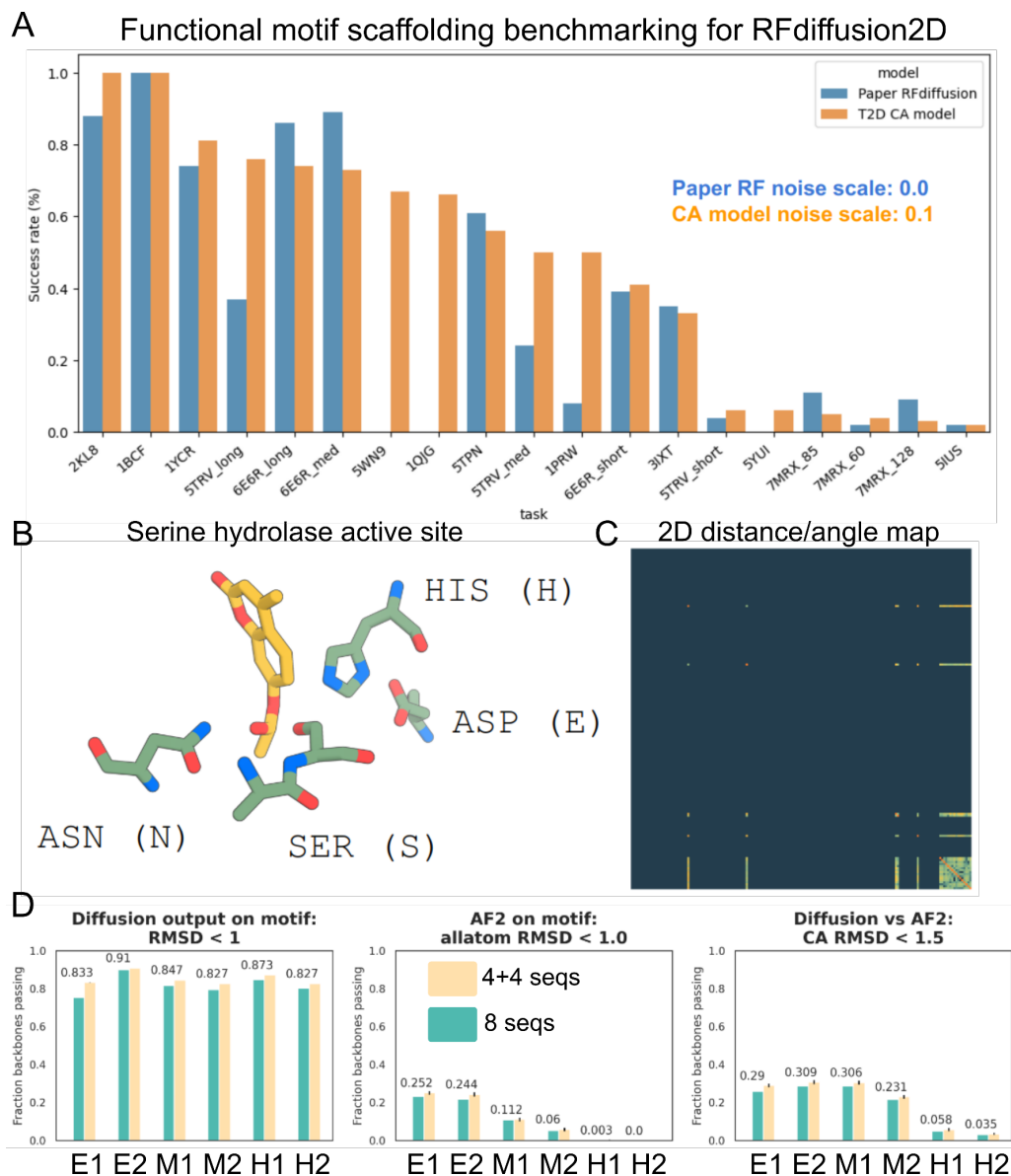


Figure 17: Benchmarking RFdiffusion2D on motif scaffolding problems

(A) RFdiffusion2D (orange) outperforms standard RFdiffusion (blue) on motif scaffolding by solving previously unsolved motif scaffolding challenges. (B) The serine hydrolase active site scaffolding

challenge. The easiest subproblem (denoted E in panel (D)) is to scaffold only the serine and histidine. The medium ('M') difficulty challenge is to scaffold S+H+E, and hard is S+H+E+N. (C) Distance/Angle map pairwise input to RFdiffusion2D for the hard serine hydrolase motif scaffolding challenge. Dark blue denotes unspecified distances/angles, light green/yellow/orange denotes supplied distances/angles. (D) Computational success rates for the E/M/H motif scaffolding problem for a synthetic (1), and native (2) set of sidechains. Green denotes the success rate for designing 8 sequences for 1 RFdiffusion2D output, and yellow denotes the success rate for designing 4 sequences for two different RFdiffusion2D outputs (8 sequences total). RFdiffusion2D readily recapitulates the motif in its own output (left), and after MPNN+AF2, there are sequences discovered for which the all-atom RMSD on the motif is <1.0 angstroms (except for H2), although success here is rare for H1 (middle). There is higher frequency of success for finding sequences for which the AF2 prediction matches well at the backbone level (right).

Given the computational success for scaffolding the serine hydrolase motif, we set out to experimentally test designs from RFdiffusion2D. Please refer to Lauko, Pellock, et al., (in prep) for design campaign details. In short, large scale inference runs with RFdiffusion2D followed by sequence design using ProteinMPNN and AF2 structure prediction, with repeated resampling of successful backbones, yielded a set of ~96 high quality designs to test. After designing the enzymes, Lauko and Pellock experimentally screened the designs and discovered active, multi-turnover serine hydrolases in the set. Figure 18A displays the predicted structures and kinetic parameters for three of the discovered enzymes. Notably, they are extremely different folds.

A

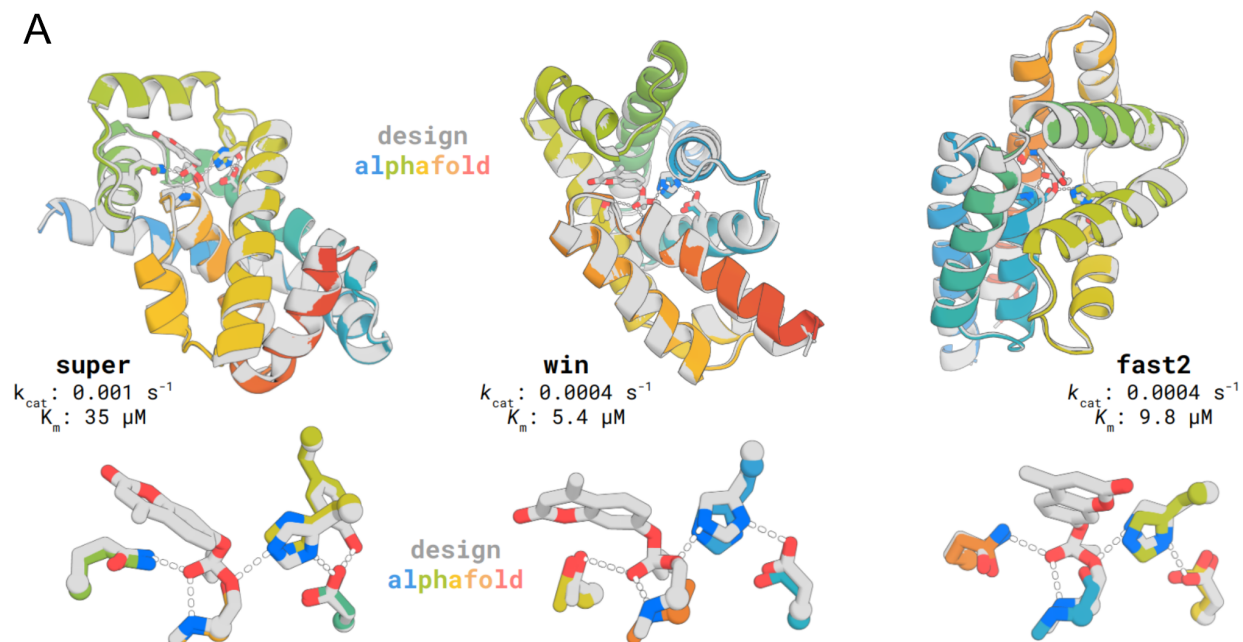


Figure 18: Experimental validation of RFdiffusion2D enzymes designed by Lauko and Spellock

(A) Predicted structures, active site zoom, and Michaelis-Menten fitted parameters for three active serine hydrolases discovered by Lauko and Pellock in a large scale computational design campaign.

Conclusions on the training of RFdiffusion2D and motif scaffolding for enzyme and small-molecule binding sites.

In pursuit of the goal of creating oligomeric enzymes, we devised a scheme and trained a model to perform the motif scaffolding task exceptionally well, both computationally and experimentally. We showed that by encoding motifs via a pairwise distance/angle template to RFdiffusion, we can improve our ability to perform the motif scaffolding task. The magnitude of improvement is sufficient to solve previously unsolved motif scaffolding challenges, such as enzyme active site scaffolding.

RFdiffusion2D serves as an important milestone for protein diffusion models on the path to creating functional molecules such as those shown in Figure 16. More broadly, the encoding of scaffolded motifs in a 2D distance map should enable users to encode arbitrary constraints on degrees of freedom between any two atoms being incorporated into a *de novo* protein with diffusion models.

Overall conclusions from the current body of work

From the aforementioned findings, we can make conclusions regarding the three proposed conjectures in the introduction.

Conclusions on conjecture 1

“Systems possessing chemical functions that are useful for solving many challenges that society faces today in areas such as sustainability, energy, and medicine can be rationally engineered via de novo protein design.”

The findings here provide overwhelming evidence that *de novo* protein design can be used to engineer molecular systems that solve (or represent progress towards solving) important challenges that society faces today. Through RFdiffusion and its successors, we have shown that systems which bind to therapeutically relevant protein targets can be designed on the computer faster and with higher success rates than ever before. We showed that systems capable of coordinating metals with intricate coordination geometries can be designed with similar facility. This opens up a wealth of opportunities for design of proteins which enable the bioremediation of contaminated environments through sequestration of toxic ions, or the sustainable harvesting of rare earth metals. Our massively improved ability to design proteins which present antigens in their native conformation opens up a wide variety of new avenues for engineering vaccines for humans and livestock. And the ability to generate enzymes that catalyze the hydrolysis of ester bonds can be used to improve our recycling, bioremediation,

and chemical production capacity and efficiency. These are simply a few of the many ways in which the first conjecture has been supported by the present work.

Conclusions on conjecture 2

“Many of the chemical functions we want to design can be specified via structure-based molecular constraints in proteins, and we can thus design many chemical functions through building proteins which satisfy these constraints or conditions.”

The findings here also provide overwhelming support for the second conjecture. Every neural network trained in the current work was trained to produce the crystal structure of a protein, and every motif scaffolded was the known conformation in a crystal structure. Through the generation of structures and sequences with a single well-defined low energy (or high probability) state, we encoded molecular binding events and catalysis into amino acid sequences. Although static structure does not define all physical properties of a molecular system, it is an exceptionally good starting point. I believe that future work should investigate the importance and design of subtle conformational molecular shifts in proteins and peptides to improve molecular function design using the techniques described here.

Conclusions on conjecture 3

“Deep neural networks can be trained to learn aforementioned conditional distributions of proteins using data that already exists, and this will be broadly useful for designing de novo molecules with more complex function and with an improved success rate.”

Throughout this work, we have employed deep learning to learn models of a practically useful probability distribution

$$p(X, S | M)$$

Where X is a protein structure, S is a protein sequence, and M is a functional motif structure or target binding partner. RFdiffusion can be seen as modeling the distribution

$$p(X | M)$$

Which is, the probability of a protein backbone structure conditioned on some motif. Through the use of ProteinMPNN to design sequences on RFdiffusion backbones, we achieve robust modeling of our desired joint distribution:

$$p(S | X, M)p(X | M) = p(X, S | M)$$

We showed in our findings that by modeling this distribution, we achieve the design of a broad range of chemically interesting proteins, with functions relevant for solving modern day challenges. Furthermore, we achieved design and validation of these functions at speeds which were previously untenable. It is highly likely that deep learning will continue to progress the field of de novo molecular design at an unprecedented rate, due to the exceptional learning capabilities of these models.

Citations

- [1] P.-S. Huang, S. E. Boyken, and D. Baker, "The coming of age of de novo protein design," *Nature*, vol. 537, no. 7620, pp. 320–327, Sep. 2016.
- [2] F. Carey and R. Giuliano, *Organic chemistry*. McGraw-Hill Education, 2013.
- [3] N. Kresge, R. D. Simoni, and R. L. Hill, "The Thermodynamic Hypothesis of Protein Folding: the Work of Christian Anfinsen," *J. Biol. Chem.*, vol. 281, no. 14, pp. e11–e13, Apr. 2006.
- [4] D. A. McQuarrie and J. D. Simon, *Physical chemistry: a molecular approach*, vol. 1. University science books Sausalito, CA, 1997.
- [5] K. Dill and S. Bromberg, *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. Garland Science, 2010.
- [6] J. E. Lennard-Jones, "Cohesion," *Proc. Phys. Soc. London*, vol. 43, no. 5, p. 461, Sep. 1931.
- [7] P. Koehl and M. Levitt, "Structure-based conformational preferences of amino acids," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 22, pp. 12524–12529, Oct. 1999.
- [8] G. N. Ramachandran and V. Sasisekharan, "Conformation of polypeptides and proteins," *Adv. Protein Chem.*, vol. 23, pp. 283–438, 1968.
- [9] C. Norn *et al.*, "Protein sequence design by conformational landscape optimization," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 11, Mar. 2021, doi: 10.1073/pnas.2017228118.
- [10] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, no. 5540, pp. 93–96, Oct. 2001.
- [11] K. T. Simons, C. Strauss, and D. Baker, "Prospects for ab initio protein structural genomics," *J. Mol. Biol.*, vol. 306, no. 5, pp. 1191–1199, Mar. 2001.
- [12] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, "Three-dimensional structures of membrane proteins from genomic sequencing," *Cell*, vol. 149, no. 7, pp. 1607–1621, Jun. 2012.
- [13] T. Kosciółek and D. T. Jones, "De novo structure prediction of globular proteins aided by sequence variation-derived contacts," *PLoS One*, vol. 9, no. 3, p. e92197, Mar. 2014.
- [14] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, "Genomics-aided structure prediction," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 26, pp. 10340–10345, Jun. 2012.
- [15] A. W. Senior *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, Jan. 2020.
- [16] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, "Improved protein structure prediction using predicted interresidue orientations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 3, pp. 1496–1503, Jan. 2020.
- [17] R. Pearce, Y. Li, G. S. Omenn, and Y. Zhang, "Fast and accurate Ab Initio Protein structure prediction using deep learning potentials," *PLoS Comput. Biol.*, vol. 18, no. 9, p. e1010539, Sep. 2022.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv [cs.CV]*, Dec. 10, 2015. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [19] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021.
- [20] M. Baek *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, Aug. 2021.
- [21] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, Mar. 2023.
- [22] J. K. Leman *et al.*, "Macromolecular modeling and design in Rosetta: recent methods and frameworks," *Nat. Methods*, vol. 17, no. 7, pp. 665–680, Jul. 2020.
- [23] J. Dauparas *et al.*, "Robust deep learning-based protein sequence design using ProteinMPNN," *Science*, vol. 378, no. 6615, pp. 49–56, Oct. 2022.
- [24] N. Anand *et al.*, "Protein sequence design with a learned potential," *Nat. Commun.*, vol. 13,

- no. 1, p. 746, Feb. 2022.
- [25] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a novel globular protein fold with atomic-level accuracy," *Science*, vol. 302, no. 5649, pp. 1364–1368, Nov. 2003.
 - [26] L. An and G. R. Lee, "De Novo Protein Design Using the Blueprint Builder in Rosetta," *Curr. Protoc. Protein Sci.*, vol. 102, no. 1, p. e116, Dec. 2020.
 - [27] A. Zanghellini *et al.*, "New algorithms and an in silico benchmark for computational enzyme design," *Protein Sci.*, vol. 15, no. 12, pp. 2785–2794, Dec. 2006.
 - [28] I. Anishchenko *et al.*, "De novo protein design by deep network hallucination," *Nature*, vol. 600, no. 7889, pp. 547–552, Dec. 2021.
 - [29] Z. Li, S. P. Nguyen, D. Xu, and Y. Shang, "Protein Loop Modeling Using Deep Generative Adversarial Network," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov. 2017, pp. 1085–1091.
 - [30] N. Anand and P. Huang, "Generative modeling for protein structures," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, Accessed: May 02, 2023. [Online]. Available: <https://proceedings.neurips.cc/paper/7978-generative-modeling-for-protein-structures>
 - [31] B. I. M. Wicky *et al.*, "Hallucinating symmetric protein assemblies," *Science*, vol. 378, no. 6615, pp. 56–61, Oct. 2022.
 - [32] J. Wang *et al.*, "Scaffolding protein functional sites using deep learning," *Science*, vol. 377, no. 6604, pp. 387–394, Jul. 2022.
 - [33] J. L. Watson *et al.*, "De novo design of protein structure and function with RFdiffusion," *Nature*, vol. 620, no. 7976, pp. 1089–1100, Aug. 2023.
 - [34] F. Praetorius *et al.*, "Design of stimulus-responsive two-state hinge proteins," *Science*, vol. 381, no. 6659, pp. 754–760, Aug. 2023.
 - [35] S. Vázquez Torres *et al.*, "De novo design of high-affinity binders of bioactive helical peptides," *Nature*, vol. 626, no. 7998, pp. 435–442, Feb. 2024.
 - [36] L. An *et al.*, "De novo design of diverse small molecule binders and sensors using Shape Complementary Pseudocycles," *bioRxiv*, Dec. 2023, doi: 10.1101/2023.12.20.572602.
 - [37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
 - [38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
 - [39] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation," *arXiv [cs.LG]*, May 31, 2019. [Online]. Available: <http://arxiv.org/abs/1905.13741>
 - [40] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, Feb. 1988.
 - [41] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "InChI, the IUPAC International Chemical Identifier," *J. Cheminform.*, vol. 7, p. 23, May 2015.
 - [42] E. Komp and S. Valleau, "Machine Learning Quantum Reaction Rate Constants," *J. Phys. Chem. A*, vol. 124, no. 41, pp. 8607–8613, Oct. 2020.
 - [43] R. Gómez-Bombarelli *et al.*, "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules," *ACS Cent Sci*, vol. 4, no. 2, pp. 268–276, Feb. 2018.
 - [44] G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. F. Da Silva, and M. G. Quiles, "Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset," *J. Phys. Chem. A*, vol. 124, no. 47, pp. 9854–9866, Nov. 2020.
 - [45] J. Hoffmann *et al.*, "Training Compute-Optimal Large Language Models," *arXiv [cs.CL]*, Mar. 29, 2022. [Online]. Available: <http://arxiv.org/abs/2203.15556>
 - [46] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv [cs.CL]*, Feb. 27, 2023. [Online]. Available: <http://arxiv.org/abs/2302.13971>
 - [47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv [cs.CL]*, Oct. 11, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>

- [48] S. Bubeck *et al.*, “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” *arXiv [cs.CL]*, Mar. 22, 2023. [Online]. Available: <http://arxiv.org/abs/2303.12712>
- [49] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” *arXiv [cs.CL]*, May 28, 2020. [Online]. Available: <http://arxiv.org/abs/2005.14165>
- [50] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” *arXiv [cs.CL]*, Mar. 04, 2022. [Online]. Available: <http://arxiv.org/abs/2203.02155>
- [51] J. Kaplan *et al.*, “Scaling Laws for Neural Language Models,” *arXiv [cs.LG]*, Jan. 23, 2020. [Online]. Available: <http://arxiv.org/abs/2001.08361>
- [52] A. Vaswani *et al.*, “Attention is all you need,” *arXiv [cs.CL]*, Jun. 12, 2017. Accessed: May 02, 2023. [Online]. Available: <https://proceedings.neurips.cc/paper/7181-attention-is-all>
- [53] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- [54] C. Yang *et al.*, “Bottom-up de novo design of functional proteins with complex structural features,” *Nat. Chem. Biol.*, vol. 17, no. 4, pp. 492–500, Apr. 2021.
- [55] F. Sesterhenn *et al.*, “De novo protein design enables the precise induction of RSV-neutralizing antibodies,” *Science*, vol. 368, no. 6492, May 2020, doi: 10.1126/science.aay5051.
- [56] D.-A. Silva *et al.*, “De novo design of potent and selective mimics of IL-2 and IL-15,” *Nature*, vol. 565, no. 7738, pp. 186–191, 2019.
- [57] B. E. Correia *et al.*, “Proof of principle for epitope-focused vaccine design,” *Nature*, vol. 507, no. 7491, pp. 201–206, Mar. 2014.
- [58] E. Procko *et al.*, “A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells,” *Cell*, vol. 157, no. 7, pp. 1644–1656, Jun. 2014.
- [59] F. Frolov, A. J. Kalb, and J. Yariv, “Structure of a unique twofold symmetric haem-binding site,” *Nat. Struct. Biol.*, vol. 1, no. 7, pp. 453–460, Jul. 1994.
- [60] A. Lombardi, F. Pirro, O. Maglio, M. Chino, and W. F. DeGrado, “De Novo Design of Four-Helix Bundle Metalloproteins: One Scaffold, Diverse Reactivities,” *Acc. Chem. Res.*, vol. 52, no. 5, pp. 1148–1159, May 2019.
- [61] J. R. Calhoun, F. Natri, O. Maglio, V. Pavone, A. Lombardi, and W. F. DeGrado, “Artificial diiron proteins: from structure to function,” *Biopolymers*, vol. 80, no. 2–3, pp. 264–278, 2005.
- [62] E. N. G. Marsh and W. F. DeGrado, “Noncovalent self-assembly of a heterotetrameric diiron protein,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 8, pp. 5150–5154, Apr. 2002.
- [63] A. M. Keech, N. E. Le Brun, M. T. Wilson, S. C. Andrews, G. R. Moore, and A. J. Thomson, “Spectroscopic studies of cobalt (II) binding to Escherichia coli bacterioferritin,” *J. Biol. Chem.*, vol. 272, no. 1, pp. 422–429, 1997.
- [64] L. Li *et al.*, “Surface Modification of Hydroxyapatite Nanocrystallite by a Small Amount of Terbium Provides a Biocompatible Fluorescent Probe,” *J. Phys. Chem. C*, vol. 112, no. 32, pp. 12219–12224, Aug. 2008.
- [65] M. Yáñez, J. Gil-Longo, and M. Campos-Toimil, “Calcium Binding Proteins,” in *Calcium Signaling*, M. S. Islam, Ed., Dordrecht: Springer Netherlands, 2012, pp. 461–482.
- [66] S. J. Caldwell *et al.*, “Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 48, pp. 30362–30369, Dec. 2020.
- [67] R. Pascolutti *et al.*, “Structure and Dynamics of PD-L1 and an Ultra-High-Affinity PD-1 Receptor Mutant,” *Structure*, vol. 24, no. 10, pp. 1719–1728, Oct. 2016.
- [68] R. L. Maute *et al.*, “Engineering high-affinity PD-1 variants for optimized immunotherapy and immuno-PET imaging,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 47, pp. E6506–14, Nov. 2015.
- [69] H. Berman, K. Henrick, and H. Nakamura, “Announcing the worldwide Protein Data Bank,” *Nat. Struct. Biol.*, vol. 10, no. 12, p. 980, Dec. 2003.
- [70] Y. Zhang and J. Skolnick, “TM-align: a protein structure alignment algorithm based on the TM-score,” *Nucleic Acids Res.*, vol. 33, no. 7, pp. 2302–2309, Apr. 2005.
- [71] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

- [72] C. Saharia *et al.*, “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” *arXiv [cs.CV]*, May 23, 2022. [Online]. Available: <http://arxiv.org/abs/2205.11487>
- [73] A. Ramesh *et al.*, “Zero-Shot Text-to-Image Generation,” in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 8821–8831.
- [74] B. L. Trippe *et al.*, “Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem,” *arXiv [q-bio.BM]*, Jun. 08, 2022. [Online]. Available: <http://arxiv.org/abs/2206.04119>
- [75] S. Luo, Y. Su, X. Peng, S. Wang, J. Peng, and J. Ma, “Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures,” *bioRxiv*, p. 2022.07.10.499510, Oct. 16, 2022. doi: 10.1101/2022.07.10.499510.
- [76] N. Anand and T. Achim, “Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models,” *arXiv [q-bio.QM]*, May 26, 2022. [Online]. Available: <http://arxiv.org/abs/2205.15019>
- [77] K. E. Wu, K. K. Yang, R. van den Berg, J. Y. Zou, A. X. Lu, and A. P. Amini, “Protein structure generation via folding diffusion,” *arXiv [q-bio.BM]*, Sep. 30, 2022. [Online]. Available: <http://arxiv.org/abs/2209.15611>
- [78] A. Leach, S. M. Schmon, M. T. Degiacomi, and C. G. Willcocks, “Denoising diffusion probabilistic models on so (3) for rotational alignment,” 2022, [Online]. Available: <https://openreview.net/pdf?id=BY88eBbkpe5>
- [79] V. De Bortoli, E. Mathieu, M. Hutchinson, J. Thornton, Y. W. Teh, and A. Doucet, “Riemannian Score-based generative modelling,” *arXiv [cs.LG]*, Feb. 06, 2022. [Online]. Available: <http://arxiv.org/abs/2202.02763>
- [80] T. Chen, R. Zhang, and G. Hinton, “Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning,” *arXiv [cs.CV]*, Aug. 08, 2022. [Online]. Available: <http://arxiv.org/abs/2208.04202>
- [81] J. S. Lee, J. Kim, and P. M. Kim, “ProteinSGM: Score-based generative modeling for de novo protein design,” *bioRxiv*, p. 2022.07.13.499967, Feb. 04, 2023. doi: 10.1101/2022.07.13.499967.
- [82] N. Bennett *et al.*, “Improving de novo Protein Binder Design with Deep Learning,” *bioRxiv*, p. 2022.06.15.495993, Jun. 17, 2022. doi: 10.1101/2022.06.15.495993.
- [83] G. L. Butterfield *et al.*, “Evolution of a designed protein assembly encapsulating its own RNA genome,” *Nature*, vol. 552, no. 7685, pp. 415–420, Dec. 2017.
- [84] D. S. Goodsell and A. J. Olson, “Structural symmetry and protein function,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 29, pp. 105–153, 2000.
- [85] R. Sterner and B. Höcker, “Catalytic versatility, stability, and evolution of the (betaalpha)8-barrel enzyme fold,” *Chem. Rev.*, vol. 105, no. 11, pp. 4038–4055, Nov. 2005.
- [86] M. Jendrusch, J. O. Korbel, and S. Kashif Sadiq, “AlphaDesign: A de novo protein design framework based on AlphaFold,” *bioRxiv*, p. 2021.10.11.463937, Oct. 12, 2021. doi: 10.1101/2021.10.11.463937.
- [87] P. Chène, “Inhibiting the p53-MDM2 interaction: an important target for cancer therapy,” *Nat. Rev. Cancer*, vol. 3, no. 2, pp. 102–109, Feb. 2003.
- [88] P. H. Kussie *et al.*, “Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain,” *Science*, vol. 274, no. 5289, pp. 948–953, Nov. 08, 1996.
- [89] E. N. Salgado, X. I. Ambroggio, J. D. Brodin, R. A. Lewis, B. Kuhlman, and F. A. Tezcan, “Metal templated design of protein interfaces,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 5, pp. 1827–1832, Feb. 2010.
- [90] E. N. Salgado, R. A. Lewis, S. Mossin, A. L. Rheingold, and F. A. Tezcan, “Control of protein oligomerization symmetry by metal coordination: C2 and C3 symmetrical assemblies through Cu(II) and Ni(II) coordination,” *Inorg. Chem.*, vol. 48, no. 7, pp. 2726–2728, Apr. 2009.
- [91] A. Quijano-Rubio, U. Y. Ulge, C. D. Walkey, and D.-A. Silva, “The advent of de novo

- proteins for cancer immunotherapy," *Curr. Opin. Chem. Biol.*, vol. 56, pp. 119–128, Jun. 2020.
- [92] L. Cao *et al.*, "Design of protein-binding proteins from the target structure alone," *Nature*, vol. 605, no. 7910, pp. 551–560, May 2022.
 - [93] A. Chevalier *et al.*, "Massively parallel de novo protein design for targeted therapeutics," *Nature*, vol. 550, no. 7674, pp. 74–79, Oct. 2017.
 - [94] G. Bhardwaj *et al.*, "Accurate de novo design of hyperstable constrained peptides," *Nature*, vol. 538, no. 7625, pp. 329–335, Oct. 2016.
 - [95] G. Bhardwaj *et al.*, "Accurate de novo design of membrane-traversing macrocycles," *Cell*, vol. 185, no. 19, pp. 3520–3532.e26, Sep. 2022.
 - [96] S. A. Rettie *et al.*, "Cyclic peptide structure prediction and design using AlphaFold," *bioRxiv*, Feb. 2023, doi: 10.1101/2023.02.25.529956.
 - [97] R. Krishna *et al.*, "Generalized biomolecular modeling and design with RoseTTAFold All-Atom," *Science*, vol. 384, no. 6693, p. eadl2528, Apr. 2024.
 - [98] G. M. Cherf and J. R. Cochran, "Applications of Yeast Surface Display for Protein Engineering," *Methods Mol. Biol.*, vol. 1319, pp. 155–175, 2015.
 - [99] S. D. Brown *et al.*, "Isolation and characterization of LRP6, a novel member of the low density lipoprotein receptor gene family," *Biochem. Biophys. Res. Commun.*, vol. 248, no. 3, pp. 879–888, Jul. 1998.
 - [100] A. A. Dombkowski, K. Z. Sultana, and D. B. Craig, "Protein disulfide engineering," *FEBS Lett.*, vol. 588, no. 2, pp. 206–212, Jan. 2014.
 - [101] J. D. Lee and M. W. R. Bryant, "The crystal structure of diphenyl disulphide," *Acta Crystallogr. B*, vol. 25, no. 10, pp. 2094–2101, Oct. 1969.