Deep Learning Tools for Protein Binder Design

Nathaniel Richard Bennett

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

Doctor of Philosophy

University of Washington

2023

Reading Committee:

David Baker, Chair

Frank DiMaio

Philip Bradley

Program Authorized to Offer Degree:

Molecular Engineering and Sciences

University of Washington

Abstract

Deep Learning Tools for Protein Binder Design

Nathaniel Richard Bennett

Chair of the Supervisory Committee:

David Baker

Department of Biochemistry

The ability to design protein-binding proteins is broadly useful. In this dissertation I will show our work to develop a deep learning-based pipeline for protein binder design. I will show how we configured AlphaFold2 to classify *in silico* designs which are likely to bind from those which are not likely to bind. I will then demonstrate how we can use the ProteinMPNN model, in combination with classical Rosetta protocols, to perform efficient sequence design on binder backbones. Finally, I will show how we trained a denoising diffusion model to generate protein backbones and how this can be used to massively accelerate the binder design pipeline. This deep learning-based pipeline is faster, easier to use, and has much higher experimental success rates than the previous Rosetta-based pipeline.

Acknowledgements

I would like to thank Brian Coventry and Joe Watson for being amazing mentors, collaborators, and friends.

All of the work presented in this dissertation has been a team effort and this work would not have happened without the hard work of my collaborators.

Thanks to my coauthors on the DL binder design paper: Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, Frank Dimaio, Steven De Munck, and Savvas N. Savvides.

Thanks to my coauthors on the RF Diffusion paper: Joseph L. Watson, David Juergens, Brian L. Trippe, Jason Yim, Helen E, Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakola, Frank DiMaio, and Minkyung Baek.

Thanks to David Baker for his guidance and wisdom as an advisor.

Thanks to Luki Goldschmidt for making the digs virtually indestructible.

Thanks to my parents for encouraging me at every step of my graduate work and sharing their personal learnings from graduate school.

Thanks to Phil, Erin, and Caci for keeping me sane and commiserating when things do not go according to plan.

Thanks to my amazing community of friends in Seattle who have made my life so rich.

Table of contents:

1. Introduction:.....5

2. Improving *de novo* Protein Design with Deep Learning:.....11

3. Broadly Applicable Protein Design with RF*diffusion*:.....44

4. Conclusions:.....82

5. References:.....84

Chapter 1

Introduction

The ability to design custom protein-protein interfaces (PPIs) is broadly useful. Previous work from our lab has established a PPI design pipeline that allowed the *de novo* design of PPIs[1]. The previous PPI pipeline, however, had several drawbacks; the experimental success rates were low (<0.1%), the pipeline was complicated and computationally expensive to run, and the pipeline was limited to design within a limited set of miniprotein scaffolds (the Scaffold Set). Our research focuses on developing and incorporating novel deep learning (DL) methods into the PPI pipeline to solve these outstanding issues. We report that framing binder design as a structure prediction task and using AlphaFold2[2] (AF2) to assess the likelihood of both our binder assuming the designed structure and the binder forming the correct interface increases experimental success rate by 10 fold. We find that incorporating a novel structure-to-sequence model (ProteinMPNN[3]) also makes the PPI pipeline more computationally efficient. Finally, we present RF*diffusion,* a denoising diffusion probabilistic model which operates in protein structure space. RF*diffusion* is able to sample both dock and scaffold degrees of freedom at runtime. This massively simplifies the binder design pipeline and trivially allows the design of binders with scaffolds outside of the Scaffold Set.

Protein Interface Design

Proteins are a class of macromolecules consisting of a chain of amino acid monomers. The specific order of amino acids in the chain determines the three dimensional structure that the

chain will fold into. The three dimensional shape of the protein chain is responsible for the vast majority of the biochemical properties of each protein. The prodigious number of possible unique protein chains and the correspondingly large number of possible structures allows proteins to perform many diverse functions. Proteins mediate the majority of biological processes including DNA synthesis, production of chemical energy, and cell-to-cell signaling, among many many other processes. A protein function of particular pharmaceutical interest in recent years is the role of a certain class of proteins, called antibodies, to bind to a diverse set of proteins. It was quickly realized that by designing these antibodies to form custom PPIs, several biological processes could be modulated such as cell-to-cell signaling (important in cancer therapeutics) and virus-host recognition (important in antiviral therapeutics).

The first protocols to design binders with custom PPIs involved immunizing an animal with a target protein and isolating antibody binders from the serum of the animal. While powerful and still the method-of-choice for discovering novel antibodies in the pharmaceutical industry, this approach is slow, expensive, and provides no control over the orientation of the complex or binding site. This approach is also limited in scaffold type to only antibodies which are often unstable or immunogenic. Methods to isolate binding proteins via directed evolution have also proven powerful and decrease the time and expense required to generate binding proteins, these methods share the shortcomings of immunization in that they offer no control over the orientation of the complex or binding site. *De novo* binder design holds the promise of precise control over binding site as well as binder fold and orientation. The first attempts at *de novo* binder design relied on using a native interface and building a *de novo* scaffold to support this interface, this protocol call is called motif grafting. Motif grafting was initially done using Rosetta-based tools[4,5] and more recently is done using DL based tools[6]. Motif grafting can

reliably generate *de novo* binders to target proteins where a structure of an appropriate complex is already available; this criterion, however, is seldom met for target proteins of interest. Recently, a protocol for fully *de novo* protein binder design was developed by our lab[1]. This pipeline offered, for the first time, a path to completely *de novo* protein binders. This protocol, however, is complicated and time-consuming to run, yielded a low experimental success rate (<0.01%) and is limited to designing binders of the folds contained in a set of minibinder scaffolds.

The Protein Interface Design Pipeline

As much of this work builds on and extends the design pipeline previously developed by our lab[1], a thorough description of the design pipeline is warranted.

The Scaffold Set: The PPI pipeline begins with a dataset of high confidence small proteins called the scaffold set. These scaffolds are selected to fulfill computational criteria that correlate with experimental design success or pass a high throughput experimental stability assay[7]. The process to generate new scaffolds is computationally intensive and, to thoroughly search fold space, requires the identification of the characteristic parameters of a fold[8–10]. Examples of the characteristic parameters of the three-helical bundle fold include: the number of residues in the first helix, the number of residues in the first loop, and the angle between the first and second helices. Characteristic parameters have enabled the search of simple fold spaces such as three helical bundles, four helical bundles, and ferredoxins. For larger and more complex scaffolds, the identification of characteristic parameters is more challenging and, as such, a brute force search through fold space is not feasible.

Docking: A subset of the scaffold set is chosen to be docked against the target protein. Each scaffold in the subset is mutated to contain only Valine residues (Valine is a residue with the average volume of all residues). These poly-Valine scaffolds are then coarsely docked to the target protein using a software package called PatchDock[11]. PatchDock rapidly matches shape-complementary patches on the surface of the scaffold to patches on the surface of the target protein. The coarse docks output from PatchDock are then used as initial docks for the RifDock[9] program. RifDock performs a fine-grained search around the initial dock proposed by PatchDock. RifDock finds docks that allow the design of favorable contacts between the scaffold and the target. At the conclusion of the RifDock protocol, the number of scaffold-target docks is on the order of 10 million.

Filtering of Docks: Performing a full design trajectory on the entire set of 10 million RifDocks outputs is computationally inefficient, as many of these outputs are poor quality. Thus, a rapid design calculation is performed on the entire set of docks to efficiently determine which docks are promising candidates for further design and which are not. This rapid design calculation, called The Predictor[1], uses a coarse energy function and a truncated sampling trajectory to quickly generate a plausible binding interface structure. The quality of this interface is then assessed using a higher accuracy energy function. The top scoring scaffold-target docks according to this interface quality score are advanced to the next step of the pipeline; on the order of one million docks are normally chosen.

Design of Interfaces: The filtered scaffold-target docks are now put through an entire Rosetta design trajectory. All residues of the scaffold that are close to the interface are allowed to change residue identity to yield the interface with an optimal energy score, determined by Rosetta. This design protocol uses Rosetta FastDesign[12] to perform joint sequence and structure

design and then uses Rosetta FastRelax[12] to locally search for the lowest energy structure of the interface. This design step is the most computationally intensive step in the pipeline, requiring ~450 CPU-seconds to design each dock. At the conclusion of this step, the docks have been refined and assigned a sequence and may be understood as full binder designs. These binder designs are then passed to the final step in the pipeline.

Filtering Binder Designs: While powerful, the Rosetta design protocol is often unsuccessful in discovering a high-quality interface. To identify and select only those binder designs which contain a high-quality interface, several *in silico* metrics are calculated for each binder design. These *in silico* metrics attempt to quantify the likelihood that the binder design will fold as expected in solution and will form the designed interface; the metrics are chosen based on correlation with experimental success rates of past binder design campaigns.

Deep Learning Methods for Protein Modeling and Design

In parallel to these advances in protein interface design, deep learning (DL) methods have become increasingly powerful in the fields of protein modeling and protein design. DL methods attempt to learn a function that maps an input (eg. an English sentence) to an output (eg. a French translation). These functions are learned by tuning the parameters of the function to minimize the value of some loss function (eg. translation error) over the members of a dataset (eg. a set of English sentences and their French translations). The dominant domains of DL have, until recently, been in the fields of speech[13] and vision[14]. The development of geometric deep learning techniques[15,16] that leverage the symmetries of the physical world has catalyzed the introduction of DL methods into the field of protein modeling and design[17]. The prime examples of DL methods being effective in protein modeling are AlphaFold2 (AF2) and RoseTTAFold[18] (RF),

DL models that solved the long-standing grand challenge of predicting the structure a protein sequence will assume. At the time our work began, however, the usefulness of these structure prediction models for binder design was unclear.

Chapter 2

Improving *de novo* Protein Binder Design with Deep Learning

In this chapter we explore the augmentation of energy based protein binder design using deep learning methods trained on large numbers of protein structures to complement the assessment of both the probability that a designed sequence adopts the designed monomer structure, and the probability that this structure binds the target as designed. We find in retrospective analysis of experimental datasets of 1 million designs against 13 different targets that deep learning structure prediction methods are considerably more effective in discriminating binding from non binding designs than energy calculations or accuracy prediction methods that consider only the folded state, with an average enrichment of successful designs of nearly 10 fold. Guided by these retrospective results, we use the deep learning augmented pipeline prospectively to compute sequences predicted to bind to four additional targets of considerable therapeutic interest; we again observe that incorporating deep learning structure prediction increases design success rate nearly 10 fold. Our results indicate how physically based and deep learning methods can be integrated to solve outstanding design challenges, and the combined method provides a powerful route to high affinity binding proteins to arbitrary protein targets.

This chapter presents work with Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, Frank Dimaio, Steven De Munck, Savvas N. Savvides, and David Baker. It was published on BioRxiv[19] in 2022. N.R.B., B.C., and D.B. designed the research; N.R.B. and B.C. contributed

equally; N.R.B. and B.C. developed the method; N.R.B. and B.C. designed the binders; I.G., B.H., A.A., Y.P.P., and D.V. performed the yeast screening; J.D. developed ProteinMPNN; M.B. and F.D. developed RoseTTAFold2; S.D.M. and S.N.S. solved the structures of ALK and LTK; all authors analyzed data; L.S. and D.B. supervised research; N.R.B., B.C., and D.B. wrote the manuscript with input from the other authors. All authors revised the manuscript.

INTRODUCTION

Methods for designing proteins which bind with high affinity and specificity to protein targets of interest are of considerable importance in biomedicine for generating new candidate therapeutics[4], diagnostics[5], and imaging reagents[6,7]. Currently, the most widely used methods involve immunization of an animal with the target to elicit antibodies[8], or screening high complexity random libraries of antibody[9] or other scaffolds[10] for binding activities. Although powerful, these methods require considerable experimental effort and do not provide substantial control over the properties of the resulting binding molecules. Methods for computationally designing binders could potentially provide much faster routes to affinity reagents having desired biophysical properties that target specific surface patches, and there has been considerable progress in computational design of protein binding proteins based on extension of binding motifs observed in protein structures[11–15]. Recently, a general Rosetta-based approach to designing binding proteins using only the structure of the target was developed and used to design binding proteins to 13 different target sites[1]. Given a specified region on a target of interest, the method designs sequences predicted to fold up into protein structures that have shape and chemical complementarity to the region. While providing for the first time a general

computational route to designing binders to arbitrary protein targets, the method requires screening of large numbers of computationally designed binders to identify hits as only a small fraction typically have sufficiently high affinity for experimental detection.

In parallel with advances in physical model based protein binder design, deep learning methods have achieved unprecedented accuracy in protein structure prediction. In contrast to Rosetta and other physically-based molecular mechanics methods, which employ energy functions with one or two thousand parameters obtained from structural and thermodynamic data on proteins and small molecules[16], the deep learning structure prediction methods AlphaFold2[2] (AF2) and RoseTTAFold[17] (RF) have hundreds of millions of parameters obtained by training on very large datasets of protein sequences and structures, and make no assumptions about pairwise decomposability or functional form. In place of energy-guided stochastic conformational sampling approaches utilized by physically-based approaches – molecular dynamics in many protein dynamics studies or Monte Carlo plus minimization – the deep learning methods learn iterative transformations of representation of the sequence and possible structure that very rapidly converge on often quite accurate models (the successive transformations are analogous to the structure updates in traditional simulation, but more are concerted, more directed to the likely correct structure, and with a more accurate stopping criterion[18]). For accurate prediction of the structures of naturally occurring proteins, both AF2 and RF generally require multiple sequence alignments (which contain rich co-evolutionary information on residues likely to be in contact, etc), but for de novo designed sequences, which are generally more stable and more regular than naturally occurring proteins, accurate predictions can be obtained from single sequences[19,20]. There has also been progress in accuracy prediction for protein structure models; for example

DeepAccuracyNet (DAN), which uses a representation consisting of 3D convolutions of local atomic environments[21], achieved state-of-the-art performance in accuracy prediction in CASP14.

We reasoned that these newly-developed DL methods could increase the success rate of Rosetta-based protein binder design. As noted above, while providing a general computational route to designing binders to arbitrary protein targets, the overall success rate is quite low. The approach has two primary failure modes (Fig 1.1a): first, the designed sequence may not fold to the intended monomer structure, and second, the designed monomer structure may not actually bind the target (Fig 1.1b). The physically based Rosetta approach frames both the folding and binding problems in energetic terms; for the approach to succeed, the designed sequence must have as its lowest energy state in isolation the designed monomer structure, and the complex between this designed monomer structure and the target must have sufficiently low energy to drive formation of the design-target protein complex. The primary challenges in accurate design of both the monomer structure and the protein-protein interface are inaccuracies of the energy function which for computational tractability is generally represented as a sum of pairwise decomposable terms (in Rosetta, Lennard Jones, hydrogen bonding, electrostatic, solvation, and bonded geometry), and the very large size of the space which must be sampled: if the energy function is inaccurate, or conformational sampling is incomplete, the designed sequence may not fold to the intended monomer structure and/or the monomer may not bind to the target as intended.

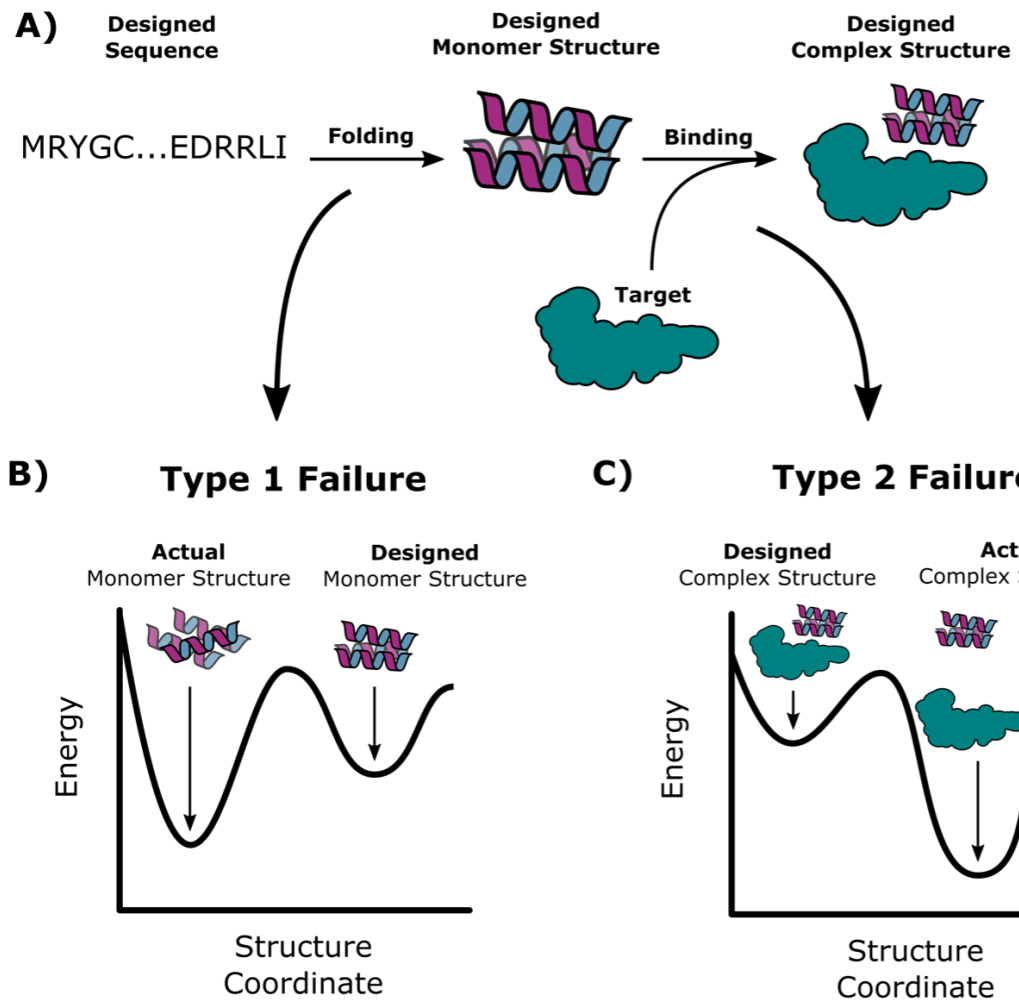
In this work, we develop a deep learning-augmented *de novo* protein binder design protocol. We show retrospectively and prospectively that this improved protocol has nearly 10-fold higher success rate than the original energy-based method.

RESULTS

Retrospective Analysis of Type I Failures

We began by investigating the ability of deep learning methods to discriminate binders from non-binders (a task we call filtering) in the set of ~1 million experimentally characterized designs for 10 different targets described in Cao et al. 15,000-100,000 designs were experimentally tested for each target, and the number of actual binders ranged from 1 to 584.

We first focused on identifying Type I failures (Fig 1.1) in which the designed sequence does not fold to the intended monomer structure. As a baseline, we used the Rosetta energy of the monomer, normalized by chain length (since energy is an extensive quantity). Not surprisingly as this metric was already used as a stringent filter in generating the input scaffold set for the Rosetta interface design calculations[22], it provided little discriminatory power (Fig 1.1d). In contrast, the deep learning based accuracy prediction method DAN was able to partially discriminate binders from non-binders (Fig 1.1d).


A)
Designed Sequence
Designed Monomer Structure
Designed Complex Structure
MRYGC...EDRRLI
Folding
Binding
Target
B)
Type 1 Failure
Actual Monomer Structure
Designed Monomer Structure
Energy
Structure Coordinate
C)
Type 2 Failure
Designed Complex Structure
Actual Complex Structure
Energy
Structure Coordinate

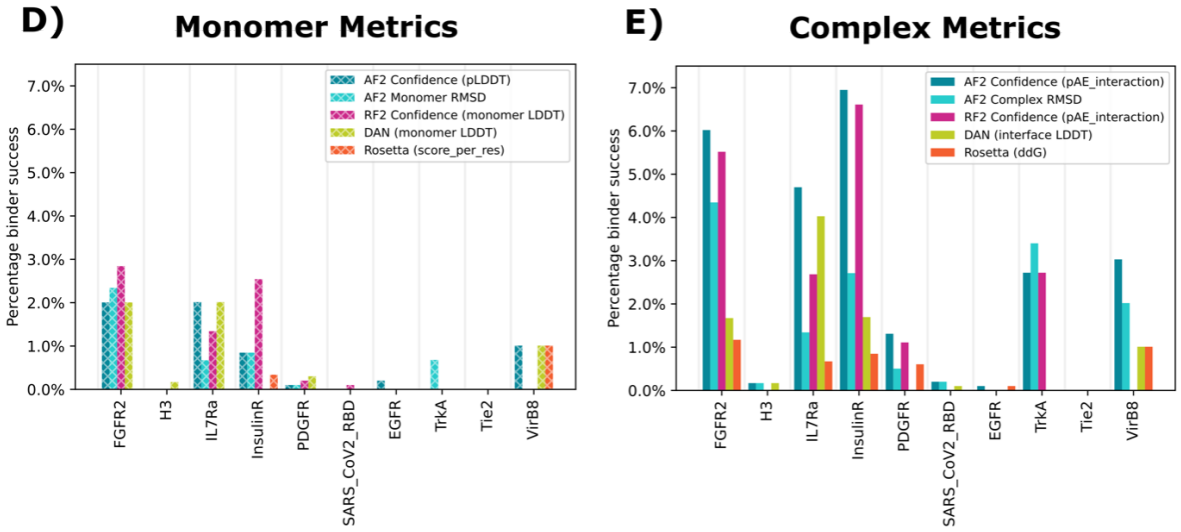

D)
Monomer Metrics
AF2 Confidence (pLDDT)
AF2 Monomer RMSD
RF2 Confidence (monomer LDDT)
DAN (monomer LDDT)
Rosetta (score_per_res)
Percentage binder success
7.0%
6.0%
5.0%
4.0%
3.0%
2.0%
1.0%
0.0%
FGFR2
H3
IL7Ra
InsulinR
PDGFR
SARS_CoV2_RBD
EGFR
TrkA
Tie2
VirB8
E)
Complex Metrics
AF2 Confidence (pAE_interaction)
AF2 Complex RMSD
RF2 Confidence (pAE_interaction)
DAN (interface LDDT)
Rosetta (ddG)
Percentage binder success
7.0%
6.0%
5.0%
4.0%
3.0%
2.0%
1.0%
0.0%
FGFR2
H3
IL7Ra
InsulinR
PDGFR
SARS_CoV2_RBD
EGFR
TrkA
Tie2
VirB8

Figure 1.1 **Monomer and protein complex structure prediction metrics distinguish previously designed binders from non-binders. A** For binder design to be successful, the designed sequence must fold to the designed binder monomer structure (left), and this structure must form the designed interface with the target protein (right). **B-C** Design failure modes. **B** Type-1 Failures. The designed sequence does not fold to the designed monomer structure. **C** Type-2 Failures. The designed sequence folds to the designed monomer structure but does not form the designed interface. **D,E** The retrospective experimental success rate (YSD $SC_{50} < 4$ μM) for the top 1% of designs selected according to different monomer (**D**) or protein complex (**E**) based metrics over 10 targets from Cao et al. Source data are provided as a Source Data file.

While DAN is very fast, taking ~0.5 GPU seconds per monomer structure, AF2 structure predictions are relatively slow (~5 GPU seconds). As an initial test of the utility of AF2 for monomer structure modeling, we evaluated the ability of AF2 to predict the structures of the binder monomers for the five minibinder structures from Cao et al. for which structures have been solved experimentally (for designs in complex with TrkA, FGFR2, IL-7Rα, and the SARS-CoV-2 Spike protein). Given only the single sequence for the designed binder, AF2 predicted the monomer structure with binder Cα accuracy between 0.2 Å-0.8 Å for all binders except for LCB1 which was predicted with 1.5 Å accuracy (Fig. 1.2). An updated version of RoseTTAFold (RF2[23,24]) was also found to predict all monomer structures with binder Cα accuracy between 0.2 Å-0.8 Å, except for TrkA which was predicted with 1.8 Å accuracy (Fig. 1.2).

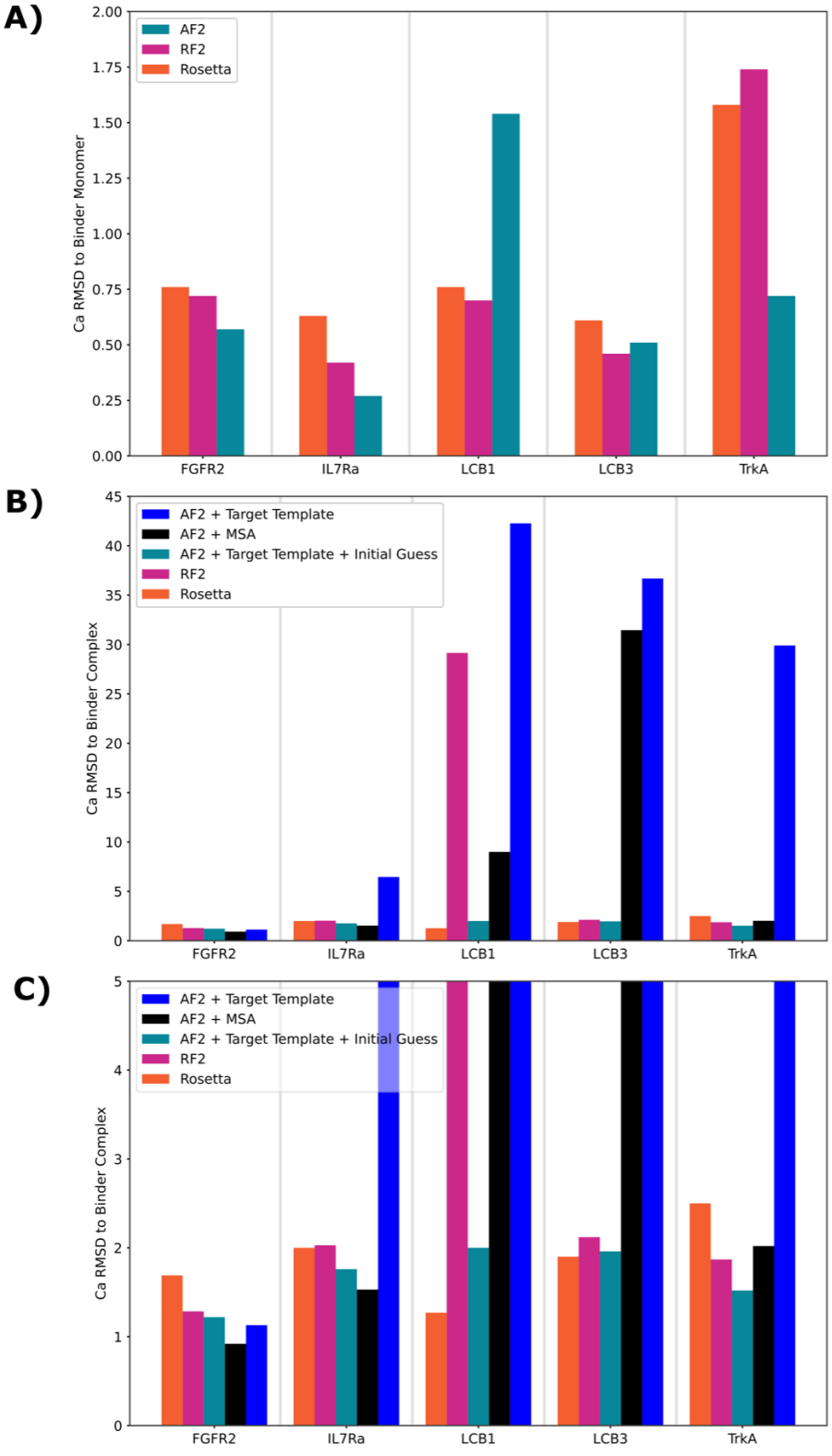
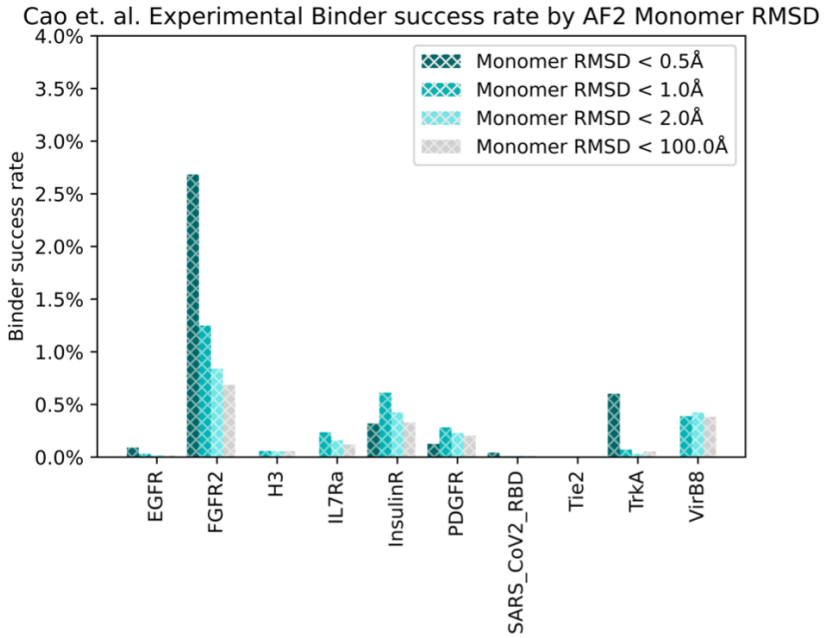

A)
AF2
RF2
Rosetta
Ca RMSD to Binder Monomer
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
FGFR2
IL7Ra
LCB1
LCB3
TrkA
B)
AF2 + Target Template
AF2 + MSA
AF2 + Target Template + Initial Guess
RF2
Rosetta
Ca RMSD to Binder Complex
45
40
35
30
25
20
15
10
5
0
FGFR2
IL7Ra
LCB1
LCB3
TrkA
C)
AF2 + Target Template
AF2 + MSA
AF2 + Target Template + Initial Guess
RF2
Rosetta
Ca RMSD to Binder Complex
5
4
3
2
1
0
FGFR2
IL7Ra
LCB1
LCB3
TrkA

Figure 1.2 **AF2 Prediction of Minibinder Crystal Structures. A** The accuracy in Cα RMSD of AF2 and RF2 predictions of binder monomer structure for the five minibinder complex structures reported in Cao et al. **B** The accuracy in Cα RMSD of AF2 and RF2 predictions of binder complex structure for the five minibinder complex structures reported in Cao et al. **C** A close-up view of **B**

Encouraged by this accuracy, we set out to filter the entire set of Cao et al. designs based on the similarity of the AF2 or RF2 predicted monomer structure to the designed structure (disagreement is an indication of a possible Type I failure). For each designed sequence for each target, using AF2 or RF2 with a single sequence as input, we predicted the structure of the binder monomer. We found that the closer the prediction of the binder structure was to the Rosetta-designed structure in Cα RMSD, the more likely a binder was to be successful (Fig. 1.3). We also found that the prediction confidence metric pLDDT was predictive of success (Fig 1.1d); the two metrics are quite correlated (Fig. 1.4; the pLDDT of AF2 and RF2 were equally discriminative). These results suggest that Type I failures contribute to the low success rate of binder design, and that such failures can, in part, be identified by discrepancies between design models and AF2 or RF2 structure predictions.

A)


Cao et. al. Experimental Binder success rate by AF2 Monomer RMSD
Binder success rate
4.0%
3.5%
3.0%
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
Monomer RMSD < 0.5Å
Monomer RMSD < 1.0Å
Monomer RMSD < 2.0Å
Monomer RMSD < 100.0Å
EGFR
FGFR2
H3
IL7Ra
InsulinR
PDGFR
SARS_CoV2_RBD
Tie2
TrkA
VirB8

B)

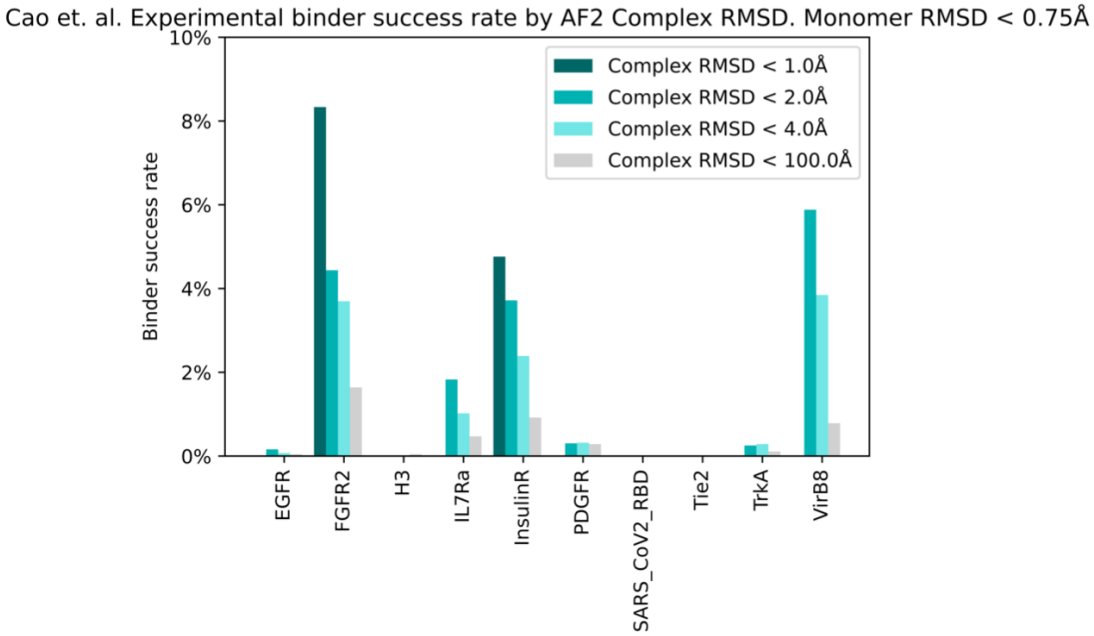

Cao et. al. Experimental binder success rate by AF2 Complex RMSD. Monomer RMSD < 0.75Å
Binder success rate
10%
8%
6%
4%
2%
0%
Complex RMSD < 1.0Å
Complex RMSD < 2.0Å
Complex RMSD < 4.0Å
Complex RMSD < 100.0Å
EGFR
FGFR2
H3
IL7Ra
InsulinR
PDGFR
SARS_CoV2_RBD
Tie2
TrkA
VirB8

Figure 1.3 **Retrospective Analysis of RMSD Metrics. A** The retrospective experimental success rate (YSD SC_{50} < 4μM) for designs passing 4 thresholds of the RMSD difference between the Rosetta monomer structure and the AF2 monomer structure. **B** The retrospective experimental success rate (YSD SC_{50} < 4μM) for designs passing 4 thresholds of the RMSD difference between the Rosetta complex structure and the AF2 complex structure.

Retrospective Analysis of Type II Failures

To estimate the likelihood of the designed binder structure forming an interface with the intended target, Cao et al. primarily used the difference in energy of the bound complex and the unbound monomers allowing sidechain repacking as computed by Rosetta (Rosetta ddg), and despite the extensive use of this metric during the original calculations, Rosetta ddg remains an effective filter (Fig 1.1e). We investigated the efficacy of DAN in supplementing Rosetta in assessing the accuracy of the designed complex structure. We found DAN's complex accuracy metric to be approximately as predictive of binder success as Rosetta ddg (Fig 1.1e).

We next investigated whether AF2 and RF2 complex prediction could be used to discriminate designs that form the intended complex structure from those that do not. We again began by evaluating the ability of AF2 and RF2 to reproduce the five experimentally determined minibinder structures from Cao et al. Given an MSA for the target protein and the single sequence of the designed binder, AF2 predicted the complex structure with binder Cα accuracy between 1.0Å-2.0Å for three of five, and RF2 for four of five. The two structures that were not correctly predicted by AF2 were LCB1 and LCB3 which both target the SARS-CoV-2 Spike

protein; AF2 was not able to correctly model a long loop in the Spike protein which caused the binders to be predicted as unbound. RF2 also predicted LCB1 as unbound (Fig. 1.2). To enable AF2 to be used for binding prediction in cases where the target is incorrectly modeled, we investigated providing the target structure to the model as a template. We found this allowed AF2 to predict the correct COVID spike structures but caused all of the interfaces except FGFR2 to be predicted incorrectly (Fig. 1.2). We next investigated initializing the AF2 pair representation with an encoding of the Rosetta binder structure; we call this protocol "AF2 initial guess" (see AF2 Initial Guess in Supplemental Information). Using AF2 with target template and an initial guess, AF2 is able to recapitulate the experimentally determined structures for all 5 minibinder interfaces with binder Cα accuracy between 1.0Å-2.0Å RMSD (Fig. 1.2). Notably, for all structures except LCB1 and LCB3, the AF2-predicted structures are closer to the experimentally determined structure than the original design models, even after extensive relaxation using Rosetta.

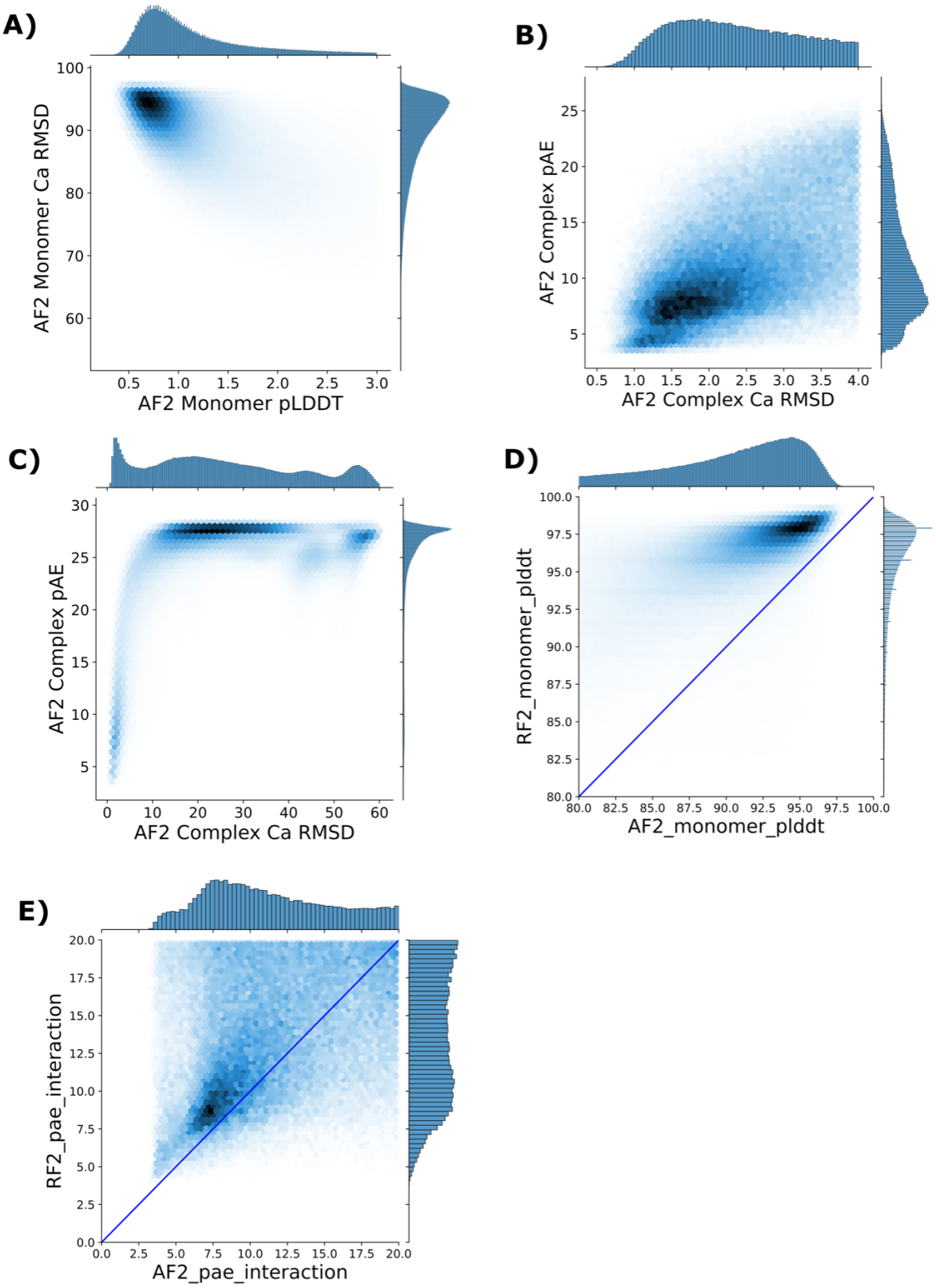

A)
AF2 Monomer Ca RMSD
AF2 Monomer pLDDT
B)
AF2 Complex pAE
AF2 Complex Ca RMSD
C)
AF2 Complex pAE
AF2 Complex Ca RMSD
D)
RF2_monomer_plddt
AF2_monomer_plddt
E)
RF2_pae_interaction
AF2_pae_interaction

Figure 1.4 **Correlation Between AF2 Confidence Metrics and AF2 RMSD. A** For all binders in the dataset from Cao et al. the AF2 monomer confidence (pLDDT) and the RMSD difference between the Rosetta binder structure and the AF2 binder structure is plotted. **B** For all binders in the dataset from Cao et al. the AF2 complex confidence (pAE_interaction) and the RMSD difference between the Rosetta complex structure and the AF2 complex structure is plotted. **C** Zoomed-out version of B. **D** For all binders in the dataset from Cao et al. the AF2 monomer confidence (AF2_monomer_plddt) and the RF2 monomer confidence (RF2_monomer_plddt) is plotted. The line x=y is plotted in solid blue. **E** For all binders in the dataset from Cao et al. the AF2 complex confidence (AF2_pAE_interaction) and the RF2 complex confidence (RF2_pAE_interaction) is plotted. The line x=y is plotted in solid blue.

We used the AF2 initial guess approach and RF2 without a starting model to generate complex models for each designed sequence for each target, and compared the predicted structure of the complex to the designed complex structure. The Cα RMSD of the predicted complex to the Rosetta designed complex model was predictive of design success in both cases (Fig 1.1e). We obtained the best discrimination of binders from non-binders using the pAE prediction confidence metrics produced by the two methods (Fig 1.1e). For the IL7Ra, TrkA, FGFR2, InsulinR, and PDGFR datasets from Cao et al, the average pAE of interchain residue pairs (pAE_interaction) was extremely effective in identifying the experimentally confirmed binders (Fig 1.1e); confident predictions had very high success frequencies (see the Receiver Operator Characteristic (ROC) curves in Fig. 1.5) with sharp increases in success rates for designs with pAE_interaction < 10. AF2 had slightly better performance than RF2 (Fig 1.1e), and we used this in the new design campaigns described in the following section. The excellent performance

of both AF2 and RF2 on the binder discrimination task strongly suggest that Type II errors are primarily responsible for the low success rates of Cao et al. We expect future, more accurate, networks will further improve the discrimination of binding proteins from non-binding proteins.

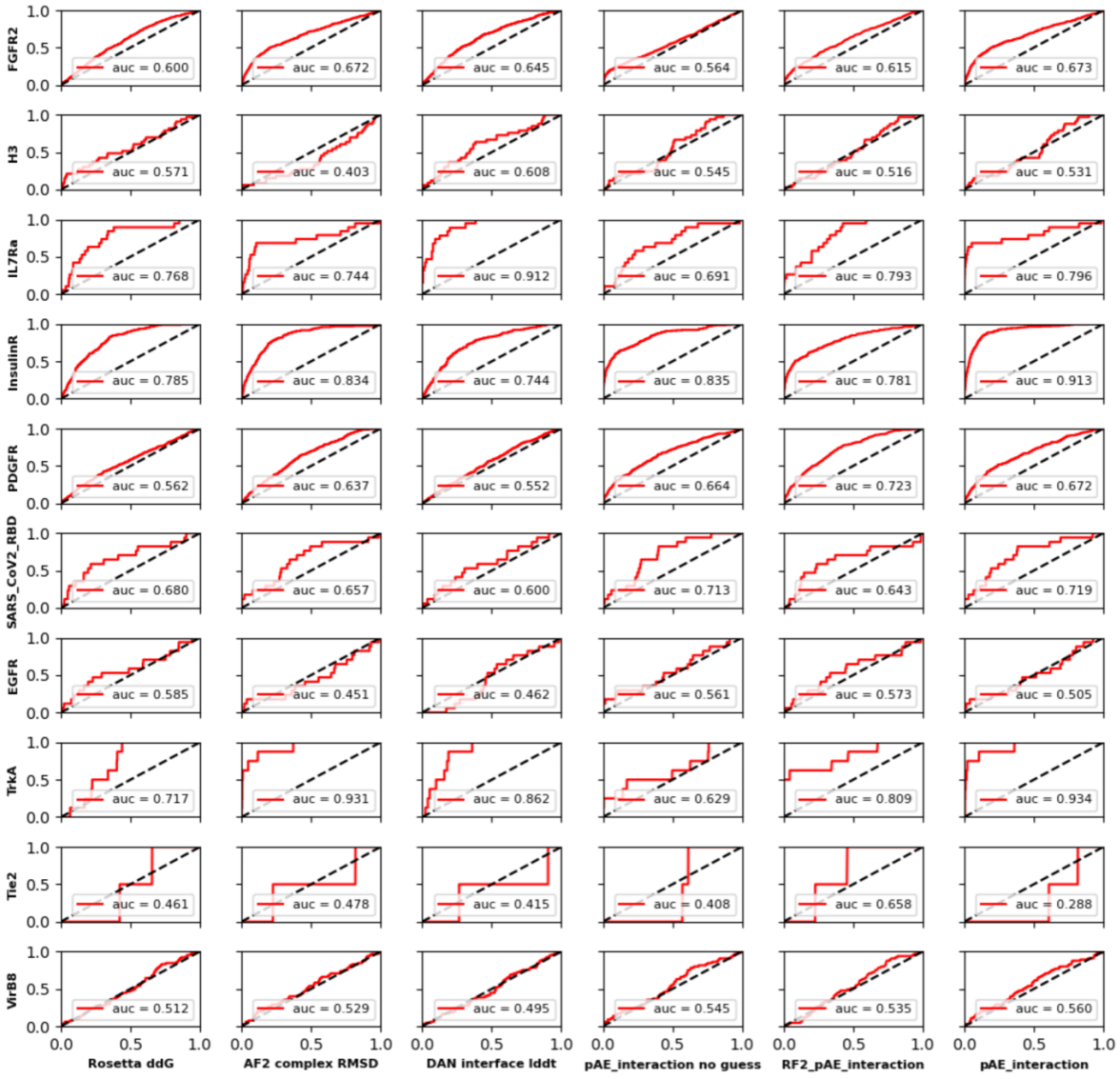
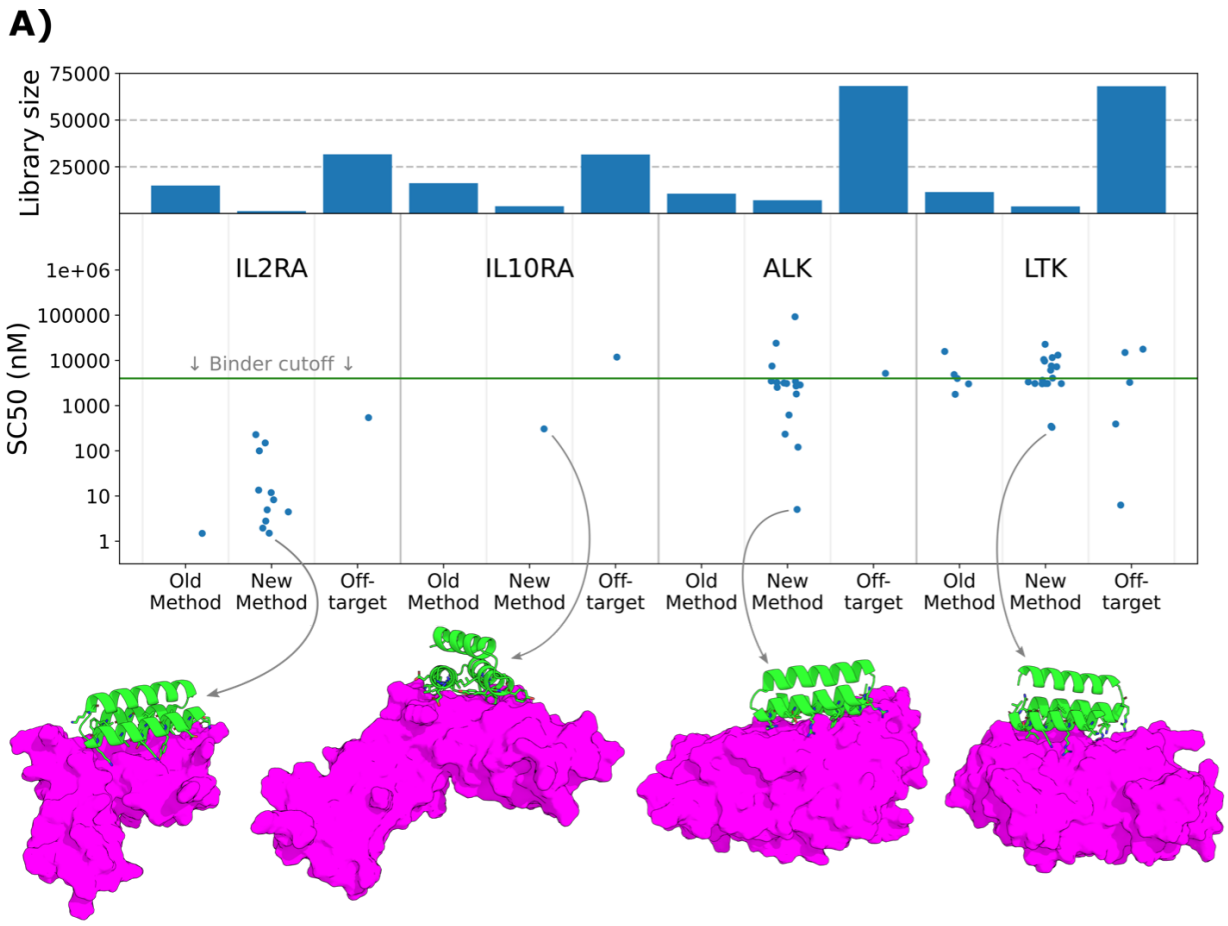

FGFR2
auc = 0.600
auc = 0.672
auc = 0.645
auc = 0.564
auc = 0.615
auc = 0.673
H3
auc = 0.571
auc = 0.403
auc = 0.608
auc = 0.545
auc = 0.516
auc = 0.531
IL7Ra
auc = 0.768
auc = 0.744
auc = 0.912
auc = 0.691
auc = 0.793
auc = 0.796
InsulinR
auc = 0.785
auc = 0.834
auc = 0.744
auc = 0.835
auc = 0.781
auc = 0.913
PDGFR
auc = 0.562
auc = 0.637
auc = 0.552
auc = 0.664
auc = 0.723
auc = 0.672
SARS_CoV2_RBD
auc = 0.680
auc = 0.657
auc = 0.600
auc = 0.713
auc = 0.643
auc = 0.719
EGFR
auc = 0.585
auc = 0.451
auc = 0.462
auc = 0.561
auc = 0.573
auc = 0.505
TrkA
auc = 0.717
auc = 0.931
auc = 0.862
auc = 0.629
auc = 0.809
auc = 0.934
Tie2
auc = 0.461
auc = 0.478
auc = 0.415
auc = 0.408
auc = 0.658
auc = 0.288
VirB8
auc = 0.512
auc = 0.529
auc = 0.495
auc = 0.545
auc = 0.535
auc = 0.560
0.0
0.5
1.0
Rosetta ddG
AF2 complex RMSD
DAN interface lddt
pAE_interaction no guess
RF2_pAE_interaction
pAE_interaction

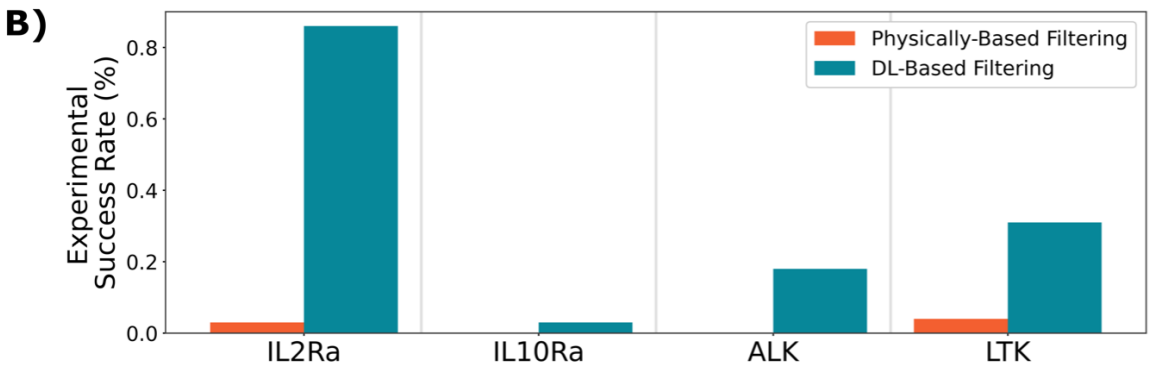
Figure 1.5 **Retrospective Analysis of Complex Metrics.** For each combination of 10 targets and 6 interface metrics a Receiver-Operator Characteristic curve quantifying the discriminatory power of the metric to separate successful designs (SC_{50} < 4 μM) from unsuccessful designs (SC_{50} >= 4 μM) is plotted. The True Positive Rate is plotted on the y-axis and the False Positive Rate on the x-axis. The Area Under the Curve (AUC) is included for each plot.

Prospective Analysis

The retrospective analysis in Fig 1.1 suggests incorporation of AF2 or RF2 into the design pipeline as a final evaluation filter could considerably increase the design success rate. To directly test this hypothesis, we carried out binder design campaigns on four new targets of considerable biological importance: ALK[25], LTK[25], IL10 receptor-α (IL-10Rα)[26], and IL2 receptor-α (IL-2Rα)[27–30]. As is clear from the retrospective analysis of the Cao et al. data (Figures 1D and 1E), binder success rate and the predictivity of metrics varies between targets: generating designs for new targets (where there is no *a priori* knowledge of which filters would be predictive) is the most unbiased approach for comparing different design protocols. For IL-2Rα, two separate sites were targeted with independent campaigns. Using the Rosetta-based design protocol of Cao et al., we generated computational libraries of ~2 million designs for each target and filtered these down to ~20,000 designs to be experimentally tested for each target: ~15,000 designs using the physically-based filters of Cao et al. and ~5,000 designs with AF2 pAE_interaction < 10 (these designs were also filtered by additional metrics as described in the Supplement). Synthetic genes were obtained for the ~80,000 designs, transformed into yeast, and the resulting library sorted for display of the proteins on yeast cells, followed by sorts at 1uM

target with avidity, and sorts at decreasing concentrations of target. The frequency of each design at each sort was determined by deep sequencing, and SC_{50} values (the concentration where half of the expressing yeast-cells are collected) estimated as described in Cao et al. Designs with SC_{50} values better than 4 μM were considered successes; the number of successes for the four targets ranged from 1 to 17. For each target, several designs which were measured to bind by YSD were expressed in E. coli and their binding was verified by single-concentration Biolayer Inferometry (BLI). All designs which showed binding by YSD also showed binding by BLI (Fig. 1.7; for IL-10Rα where only one binder was identified, only this single design was screened by BLI). For all four targets, there was a considerably higher success rate (number of successes / number of designs tested) in the AF2 filtered design set than in the Rosetta set (Fig. 1.6). Physically-based filtering yielded successful binders for two targets: LTK and Site 1 of IL-2Rα; for these the AF2 filtered libraries had 8- and 30-fold higher success rates, respectively. AF2 filtered libraries also yielded successful binders to both ALK and IL-10Rα; physically-based filtering yielded no successful binders to either of these targets (Neither filtering method was able to generate successful binders to Site 2 on IL-2Rα). Thus, AF2 filtering performs as expected in prospective tests, increasing success rates (for targets where physically-based filtering is successful) and expanding the set of targets for which successful minibinders can be generated.


A)
Library size
75000
50000
25000
SC50 (nM)
1e+06
100000
10000
1000
100
10
1
↓ Binder cutoff ↓
IL2RA
IL10RA
ALK
LTK
Old Method
New Method
Off-target


B)
Experimental Success Rate (%)
0.8
0.6
0.4
0.2
0.0
Physically-Based Filtering
DL-Based Filtering
IL2Ra
IL10Ra
ALK
LTK

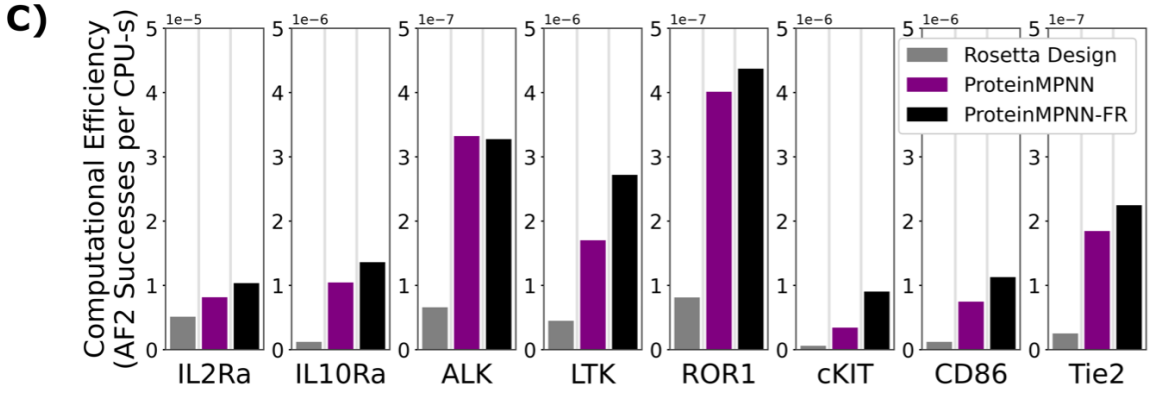

C)
Computational Efficiency (AF2 Successes per CPU-s)
1e−5
1e−6
1e−7
Rosetta Design
ProteinMPNN
ProteinMPNN-FR
IL2Ra
IL10Ra
ALK
LTK
ROR1
cKIT
CD86
Tie2

Figure 1.6 **Incorporation of structure prediction metrics increases design success rate on new targets. A** Results of Prospective Campaigns. For each target the SC_{50} from YSD is shown for all designs which showed binding by YSD (like Kd's, lower values are better). The number of designs included in each library for each target is indicated by the bars in the top panel. The AF2-predicted structure of the top scoring on-target design is shown as a cartoon. No binders were identified to Site 2 of IL2 receptor-α so this campaign is not included here or in panel C. **B** The experimental success rate for libraries filtered by DL-based filtering versus Physically-based filtering for the four prospective targets. **C** The computational efficiency (the number of designs with pAE_interaction < 10 per CPU-s) for the ProteinMPNN sequence design plus Rosetta relax protocol outperforms that of the original Rosetta sequence design protocol. Source data are provided as a Source Data file.

Increasing binder design pipeline compute efficiency with ProteinMPNN

While an effective predictor of binder success, the AF2 filter is computationally expensive (~30 GPU-seconds per design) and only ~2.3% of designs pass, so large numbers of prediction calculations must be run. To enable the testing of large (~5,000) pools of designs, it is desirable to decrease the computational demand of the design pipeline, in particular to maximize the number of designs passing the AF2 filter a method can generate per unit compute time (the time to generate all designs and run AF2; we use a conversion factor of 100 CPU-s to 1 GPU-s because of the relative scarcity of GPU resources).

$$Efficiency = Success\ Rate * Throughput = \frac{Number\ of\ Designs\ with\ pAE\_interaction < 10}{Total\ Number\ of\ Generated\ Designs} * \frac{1}{Compute\ Time\ to\ Generate\ \ One\ Design}$$

Using this metric, we find that Rosetta-design has an efficiency of about 7.6×10^{-7} successful designs per CPU-s equivalent.

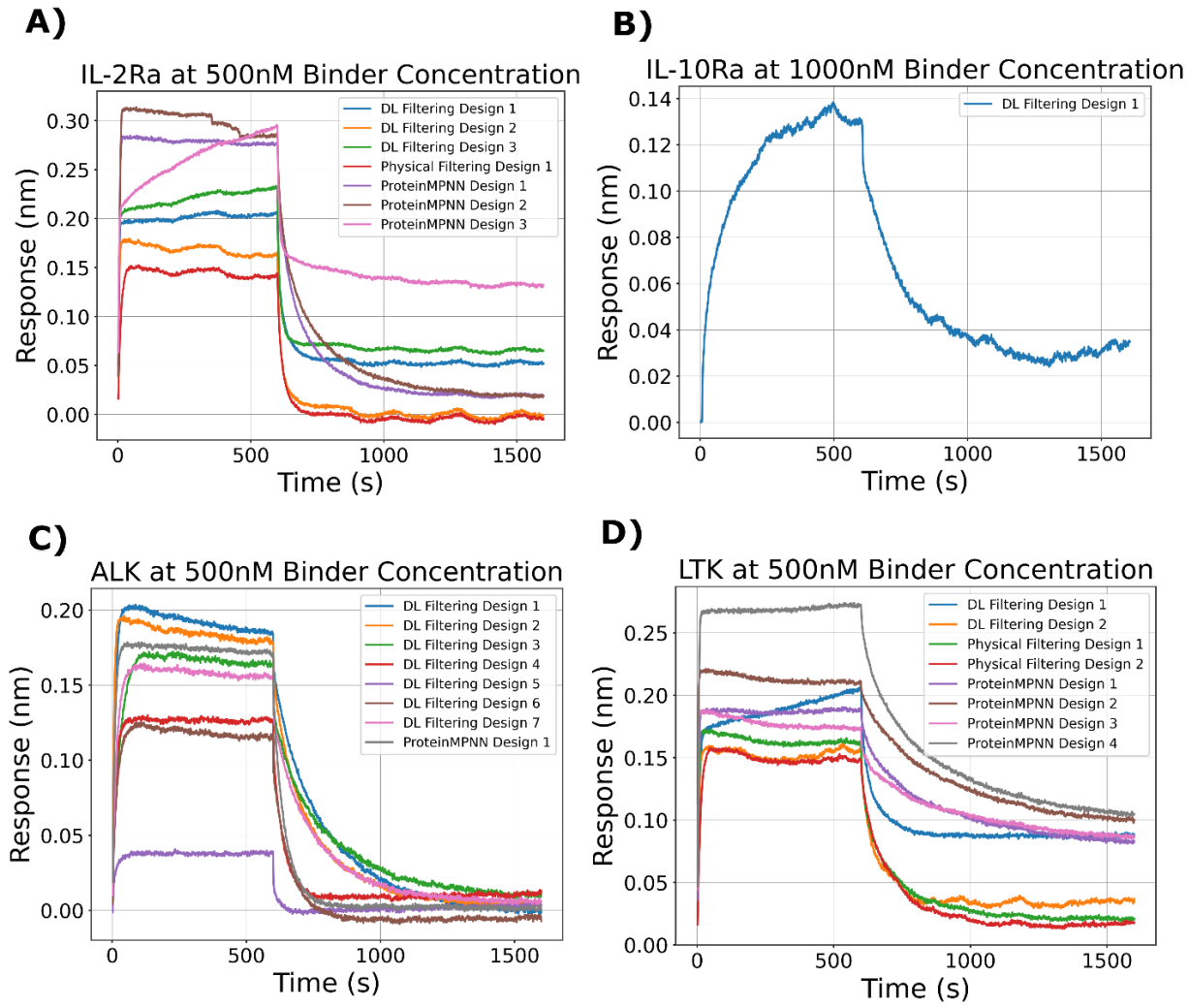


Figure 1.7 **Biophysical Characterization of Select Designs.** For each target, 7-8 designs to that target, which showed binding by Yeast Surface Display, were screened for binding by single-concentration Biolayer Interferometry (BLI). For IL-10Rα only one design was identified

and that is the only design which was screened by BLI. All designs tested showed binding signal by BLI. No successful ProteinMPNN designs against IL-10Rα were identified and only one successful ProteinMPNN design was identified to bind ALK.

We investigated whether the recently developed deep learning graphical model based sequence design method ProteinMPNN[3] could be used to increase the efficiency of the design pipeline. ProteinMPNN is very fast, generating a sequence for a minibinder backbone in ~2 CPU-s compared to ~350 CPU-s for Rosetta-design. We first compared the experimental success rate of ProteinMPNN designs to Rosetta designs by generating sequences for backbones generated by AF2 for Rosetta designs to the four new targets that had low complex Cα RMSD to the AF2 prediction (~10^4 designs in total). Genes encoding designs with AF2 pAE_interaction < 10 (~10^3 per method) were synthesized, and the binding evaluated by FACS followed by deep sequencing as described above. For each target, several designs from ProteinMPNN were expressed in E. coli and their binding was verified with BLI, we again found that all designs which bound by YSD showed binding by BLI (Fig. 1.7). We found that the design success rate of ProteinMPNN and Rosetta-design were similar (Fig. 1.8), thus the considerable increase in speed comes with no decrease in performance.

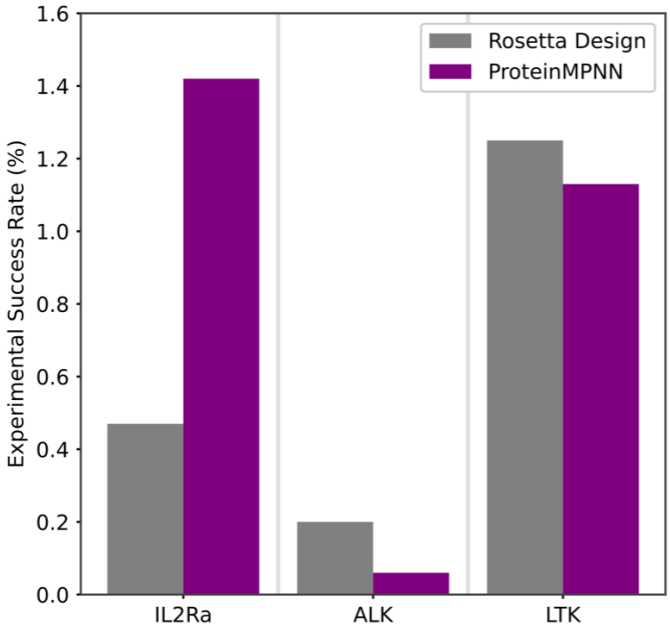


Figure 1.8 **Experimental Success Rate of Binders Generated by ProteinMPNN versus Rosetta Design**

The experimental success rate for libraries generated by redesign of DL filtered designs with Rosetta FastDesign versus ProteinMPNN and filtered by DL-based filtering for three prospective targets.

Encouraged by the speed and performance of ProteinMPNN design, we next evaluated its efficiency in generating sequences passing the AF2 cutoffs. ProteinMPNN design alone had an efficiency of 1.6×10^{-6} successful designs per CPU-s equivalent. When calculating the fold efficiency improvement for each target and taking the average, ProteinMPNN has ~5-fold higher efficiency compared to Rosetta-design (Fig. 1.6c). Since unlike Rosetta, ProteinMPNN keeps the protein backbone fixed, it is sensitive to the input backbone structure quality. Inspired by the very efficient alternation between sequence optimization and structure refinement in Rosetta flexible backbone design[31], we evaluated similar cycling between ProteinMPNN and Rosetta structure refinement (FastRelax), hoping to converge on a high-quality backbone that would then allow ProteinMPNN to generate a high-quality sequence. This hybrid ProteinMPNN/Rosetta sequence design protocol (henceforth referred to as ProteinMPNN-FR) generated AF2 pAE_interaction < 10 structures at a rate of ~6.6% with a throughput of 1 design per 120 CPU-s for an efficiency of 2.2×10^{-6}. The average per-target efficiency improvement of ProteinMPNN-FR over Rosetta-design is ~8-fold (Fig. 1.6c).

Discussion

These experiments show that by complementing physically-based methods with deep learning-based approaches trained on large numbers of protein structures, significant improvements to the one-sided protein-interface design challenge can be achieved. Our retrospective and prospective studies suggest an increase in design success rate of ten fold. In contrast to Rosetta energy calculations and DAN structure accuracy measures, which operate on single protein structures (or with Rosetta relax calculations, structures very close to the query),

structure prediction calculations implicitly assess the fit of the sequence with the desired target structure compared to all others. As observed previously[32], such consideration of the overall folding landscape enables considerably more accurate assessment of the likelihood a design will fold and bind as intended compared to evaluation of only the depth of the designed energy well. Although the protocol reported here is an order-of-magnitude improvement over the previous state-of-the-art, it is clear that much about interface energetics remains poorly understood; success rates among the targets remain low (<1%) and no binders were identified to Site 2 of IL2 receptor-α. There is also considerable room for improvement in designing high affinity; as with the original pipeline the initially generated binders are in the high nM affinity range. Given the rate of progress in the field, we anticipate further increases in design success rates and affinities in the near future, which will make computational protein design methods even more powerful compared to empirical selection methods for generating affinity reagents and therapeutic candidates. While continued progress is nearly certain, an open question is whether this will come from integration of deep learning and physically based methods, or from deep learning alone–there are exciting times ahead!

METHODS

AF2 Initial Guess

The protein structure provided to the model as an initial guess is first converted to AF2 atom positions. These positions are then provided, along with the standard model inputs into the AlphaFold Model Runner. In the AlphaFold class of the AlphaFold code, on the first recycle, the

prev_pos variable is initialized to the input AlphaFold atom positions as opposed to the standard initialization of all zeros. A script to run AF2 with an initial guess and the modified source code is provided here: https://github.com/nrbennet/dl_binder_design [33]. The AlphaFold model used in the script and in this work is configured to run with a reduced number of extra MSA sequences which speeds the inference of the network dramatically, as described in previous work[34].

ProteinMPNN FastRelax

This protocol takes as input a protein complex structure. ProteinMPNN is then provided the complex structure with the binder sequence masked and asked to assign the binder a sequence. The new sequence is then threaded back onto the binder structure in the complex and the complex structure is relaxed using Rosetta FastRelax. The relaxed complex structure can then be used as the input to ProteinMPNN to continue the cycle. A python script to perform this design technique is provided here: https://github.com/nrbennet/dl_binder_design [33].

Design and filtering procedure for Prospective Study

The prospective study was performed at a time of rapid protocol discovery with a tight deadline for placing the gene-order. As such, not every experiment that could have been performed was performed. However, the comparison of Rosetta filtering to AF2 filtering was the main goal and the data required for this comparison was plentiful.

The standard procedure from Cao et. al. was followed for the 4 targets starting with the following pdbs: IL2RA (1Z92, 2B5I , 3NFP , 2ERJ), IL10RA (1LQS), ALK (Privately communicated structure. Now 7NWZ), LTK (Privately communicated structure. Now 7NX0). The "recommended_scaffolds.list" from Cao et. al. were used and on the order of 10M RifDock[35] outputs were generated for each target with about 500K FastDesigned. ~6K motifs were extracted, grafted up to 10M docks, and 500K FastDesigned again. The resulting 1M designs for each target were predicted by AF2.

From this set of 1M designs, 3 overlapping subsets were selected. The first subset was the Rosetta-control group where the AF2 predictions were ignored and the top ~18K per target were selected by the pareto-front method from Cao et. al. looking at target_delta_sap, ddg, contact_patch, and contact_molec_sq5_apap_target. The second subset was the AF2-filtered group where all designs passing pae_interaction < 10 and af2_complex_rmsd < 5Å were included. This set was typically around 8K per target. The third subset was all predictions with af2_complex_rmsd < 5Å. These designs were designated to be redesigned and were typically about 12K in scale.

These AF2-predicted interfaces were then designed either with Rosetta or ProteinMPNN. Here, ProteinMPNN was used to generate a protein sequence from the input coordinates and no further optimization was performed. The Rosetta-redesigned and ProteinMPNN-redesigned pools were predicted again by AF2 and were filtered either with the Rosetta filters mentioned above or the AF2 filters mentioned above resulting in pools of sizes 9K (Rosetta-Rosetta), 2K (Rosetta-AF2),

and 2K (ProteinMPNN-AF2). The Rosetta filters weren't used to filter ProteinMPNN designs because Rosetta models of ProteinMPNN outputs didn't exist.

DNA Library Preparation

DNA libraries were prepared in the manner described in Cao et al., we review this protocol here:

The sequences of protein designs were padded to 65 amino acids through addition of a (S)n linker at the C-terminus. The protein sequences were reverse translated and codon optimized for *Saccharomyces cerevisiae* using DNAworks2.0[36]. After reverse translation, DNA adapter sequences are added to the N (GGTGGATCAGGAGGTTCG) and C (GGAAGCGGTGGAAGTGG) terminus. Designs were purchased as oligonucleotide libraries from Agilent Technologies.

Oligonucleotide libraries were amplified using Kapa HiFi polymerase (Kapa Biosystems) with a qPCR machine (Bio-Rad, CFX96). The PCR product was run on a DNA agarose gel, the band with the correct size was cut out of the gel and cleaned (Qiagen QIAquick Clean up kit). The extracted DNA products were then re-amplified and purified following the above protocol. The resulting DNA inserts and linearized pETcon3 vector were transformed into EBY100 yeast following an established protocol[37].

To prepare libraries for deep sequencing, yeast plasmids were isolated from $5x10^7$ to $1x10^8$ yeast cells by Zymoprep (Zymo Research). Two qPCR amplifications were then performed following

the protocol in the above paragraph. Illumina adapters and 6-bp pool-specific barcodes were added in the second amplification. The final DNA product was purified by gel extraction. The libraries were sequenced using Illumina NextSeq sequencing.

Yeast Surface Display

Yeast surface display experiments were performed in the manner described in Cao et al., we review this protocol here:

EBY100 yeast were grown in C-Trp-Ura media supplemented with 2% (w/v) glucose. Yeast cells were centrifuged and resuspended in SGCAA media supplemented with 0.2% (w/v) glucose. Cells were resuspended to a concentration of $1x10^7$ cells per ml and induced at 30℃ for 16-24 hours. Cells were washed with PBSF (PBS with 1% (w/v) BSA) and then labeled with biotinylated target. To allow for the identification of low affinity binders, an initial sort with target avidity was performed for all libraries. In the avidity sort, the cells are incubated with biotinylated target, anti-c-Myc fluorescein isothiocyanate (FITC, Miltenyi Biotech) and steptavidin-phycoerythrin (SAPE, ThermoFisher). To allow all SAPE molecules to display four biotinylated target molecules, the biotinylated target is provided at a 4x excess over the concentration of SAPE. When sorting without avidity, the cells are incubated first with biotinylated target alone, then washed in PBSF and subsequently incubated with SAPE and FITC. Each library was sorted against a titration of target concentrations. Sorts were performed using a Sony SH800S cell sorter with software version 2.1.5.

Protein Expression

Proteins were expressed and purified in the manner described in Cao et al., we review this protocol here:

Genes encoding the designed protein sequences were purchased from Integrated DNA Technologies (IDT). All genes included an N-terminal 8-His tag followed by a TEV cleavage site. The genes were cloned into modified pET-29b(+) *E. coli* plasmid expression vectors. Plasmids were transformed into chemically competent *E. coli* BL21(DE3) cells (NEB). Cells were either grown overnight in Studier autoinduction media supplemented with antibiotics or induced using the IPTG expression system and then grown overnight. Cells were then lysed by sonication and the protein samples were purified by immobilized metal affinity chromatography (Qiagen) followed by size-exclusion fast protein liquid chromatography (Superdex 75 10/300 GL, GE Healthcare).

Target Protein Preparation

Expression and purification of biotinylated ALK and LTK ectodomains

DNA encoding for the cytokine binding domains of ALK ($ALK_{TG\text{-}EGFL}$, residues 648-1030) and LTK ($LTK_{TG\text{-}EGFL}$, residues 63-420) were cloned in the pHLsec vector in frame with a N-terminal chicken RTPμ-like signal peptide sequence and a C-terminal Avi-tag followed by a caspase-3-cleavable Fc-His_{x6} tag (ref[38]).

Proteins were produced in HEK293S suspension cells maintained in growth medium consisting of 50% Freestyle (Thermofisher) and 50% Ex-Cell (Sigma-Aldrich). Transient transfection was performed using linear 25kDA polyethyleneimine (Polysciences) as transfection reagent. To allow specific in vivo biotinylation of the Avi-tag, both constructs were co-transfected with the pDisplay-BirA-ER plasmid in a 4:1 pHLsec:pDisplay stoichiometric ratio (ref[39]). The growth medium was supplemented with D-biotin to a final concentration of 100 μM to ensure complete biotinylation of the recognition sequence. After 4 days of expression, conditioned medium was clarified by centrifugation and filtered through a 0.22 μm filter prior to chromatographic steps.

Proteins were captured via their Fc tag on a protein A column (HiTrap Protein A HP, Cytiva) and eluted in HBS (20mM HEPES, pH 7.4, 150 mM NaCl) after an on-column digestion with caspase-3 for 1 h at 37 °C and an additional 2-h incubation at room temperature. As a final polishing step, recombinant proteins were concentrated and injected onto a Superdex 200 increase 10/300 GL (Cytiva) size-exclusion chromatography column pre-equilibrated with HBS. Purified biotinylated proteins were flash frozen in liquid nitrogen and stored at -80°C until further use.

Biotinylated IL-10Rα was purchased from R&D Systems (AVI9044). Biotinylated IL-2Rα was purchased from Acro Biosystems (ILA-H82E6).

Biolayer Interferometry Binding Experiments

Biolayer interferometry (BLI) measurements were performed on an Octet Red96 (ForteBio) or Octet R8 (Sartorius) instrument with Octet BLI Discovery 12.2.1.18 software, with streptavidin coated tips (Sartorius Item no. 18-5019). The binding buffer consisted of 1X HBS-EP+ buffer (Cytiva BR100669) supplemented with 1.0% w/v bovine serum albumin. 30-50nM (depending on target availability) of target protein was loaded onto the tips. After target loading, a baseline measurement was performed in binding buffer alone for 120s. The tips were then dipped in a solution of 500nM (1000nM for the IL-10Rα design) protein analyte in binding buffer for 600s (association phase). The tips were then dipped back into binding buffer alone for 1000s (dissociation phase).

Chapter 3

Broadly Applicable Protein Design with RF*diffusion*

Introduction

The protocol described in the previous section presents a robust path to protein binders. The previous protocol still relies upon docking members of the Scaffold Set against the target protein using RifDock[35] and PatchDock[40] followed by a Predictor step to generate the initial candidate docks which are then fed to ProteinMPNN-FastRelax[33] for flexible backbone design. This protocol to identify designable docks is complicated, requiring the installation of RifDock and Rosetta on the users' system and requiring expert knowledge to pick suitable hyperparameters; the protocol is extremely computationally expensive, requiring on-the-order of 50,000 CPU-hours per target to execute, well beyond the resources available to many institutions interested in binder design; finally, the protocol is limited to generating binders with a scaffold contained in the Scaffold Set, adding new scaffolds to the Scaffold Set is slow and labor-intensive, as a result the Scaffold Set remains relatively small and fails to contain many protein folds-of-interest.

Generative models are trained to learn the distribution of a dataset, sampling from this distribution yields novel examples of the same type as those contained in the dataset. These models can also be trained to perform conditional sampling which can yield novel examples of a user-specified type. Denoising Diffusion Probabilistic Models[41–43] (DDPMs) have emerged as a powerful class of generative models and have shown performance surpassing that of Generative Adversarial Networks[44] (GANs) while being significantly simpler to train. Recent work has demonstrated the feasibility of using DDPMs to sample protein structures that appear native-like[45,46]. Prior to this work, generation of an experimentally verified protein from a DDPM had not yet been shown, nor had the fully-DL design of a protein binder been shown, we set out to demonstrate the applicability of DDPMs at both of these tasks.

This chapter presents work with Joseph L. Watson, David Juergens, Brian L. Trippe, Jason Yim, Helen E, Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakola, Frank DiMaio, Minkyung Baek, and David Baker. It was published on BioRxiv[19] in 2022. J.L.W., D.J., N.R.B, B.L.T., J.Y., and D.B. conceived the study. J.L.W., D.J., N.R.B, W.A., B.L.T., and J.Y. trained RF*diffusion*. B.L.T. and J.Y. extended diffusion to residue orientations with assistance from V.D.B., and E.M. H.E.E. D.J., J.L.W., N.R.B., N.H., W.S., P.V., and I.S. generated experimentally characterized designs. W.A., B.L.T., J.Y., D.J., J.L.W., and N.R.B. generated computational designs. H.E.E., A.J.B., R.J.R., L.F.M., B.I.M.W., S.J.P., N.H., A.C., S.V.T., J.L.W., and B.L.T. experimentally characterized designs. J.W., A.L., and W.S contributed additional code. S.O. implemented RF*diffusion* on Google Colab. M.B., and F.D. trained RoseTTAFold. D.B., T.S.J., and R.B. offered supervision throughout the project. J.L.W., D.J., B.L.T., N.R.B., J.Y., H.E., and D.B. wrote the manuscript. All authors read and contributed to the manuscript. J.L.W. and D.J. agree that the order of their respective names may be changed to best suit their own interests for personal pursuits.

De novo protein design seeks to generate proteins with specified structural and/or functional properties, for example making a binding interaction with a given target[1], folding into a particular topology[47], or stabilizing a desired functional "motif" (geometries and amino acid identities that produce a desired activity)[48]. Denoising diffusion probabilistic models (DDPMs), a

powerful class of machine learning models recently demonstrated to generate novel photorealistic images in response to text prompts[49,50], have several properties well-suited to protein design. First, DDPMs generate highly diverse outputs, as they are trained to denoise data (for instance images or text) that have been corrupted with Gaussian noise. By learning to stochastically reverse this corruption, diverse outputs closely resembling the training data are generated. Second, DDPMs can be guided at each step of the iterative generation process towards specific design objectives through provision of conditioning information. Third, for almost all protein design applications it is necessary to explicitly model 3D structure; rotationally-equivariant DDPMs are able to do this in a global representation frame independent manner. Recent work has adapted DDPMs for protein monomer design by conditioning on small protein "motifs"[51] or on secondary structure and block-adjacency ("fold") information[45]. While promising, these attempts have shown limited success in generating sequences that fold to the intended structures *in silico*[51,52], likely due to the limited ability of the denoising networks to generate realistic protein backbones, and have not been tested experimentally.

We reasoned that improved diffusion models for protein design could be developed by taking advantage of the deep understanding of protein structure implicit in powerful structure prediction methods like AlphaFold2 (AF2) and RoseTTAFold (RF). RF has properties well suited for use in a protein design DDPM (Fig. 2.1A). First, RF can generate protein structures with very high precision, and in our previous work we demonstrated considerable success in accurately scaffolding motifs following fine tuning of RF for protein design ("RF_{joint} Inpainting")[48]. Second, RF operates on a rigid-frame representation of residues with rotational equivariance. Third, the RF architecture enables conditioning on design specifications at three different levels: individual

residue properties, pairwise distances and orientations between residues, and 3D coordinates. In RF_{joint} Inpainting, we fine-tuned RF to design protein scaffolds in a *single* step. Experimental characterization showed that the method can scaffold a wide range of protein functional motifs with atomic accuracy[48], but the approach fails on minimalist site descriptions that do not sufficiently constrain the overall fold, and because it is deterministic, can produce only a limited diversity of designs for a given problem. We reasoned that by instead fine-tuning (an updated version of) RoseTTAFold[1]* as the denoising network in a generative diffusion model, we could overcome both problems: because the starting point is random noise, each denoising trajectory yields a different solution, and because structure is built up progressively through many denoising iterations, little to no starting structural information should be required. While in this study we use RoseTTAFold as the basis for the denoising network architecture, other equivariant structure prediction networks (AF2[2], OmegaFold[53], ESMFold[54]) could in principle be substituted into an analogous DDPM.

[1]* This updated version of RoseTTAFold is described in Methods 1, and will be fully described in a future publication.

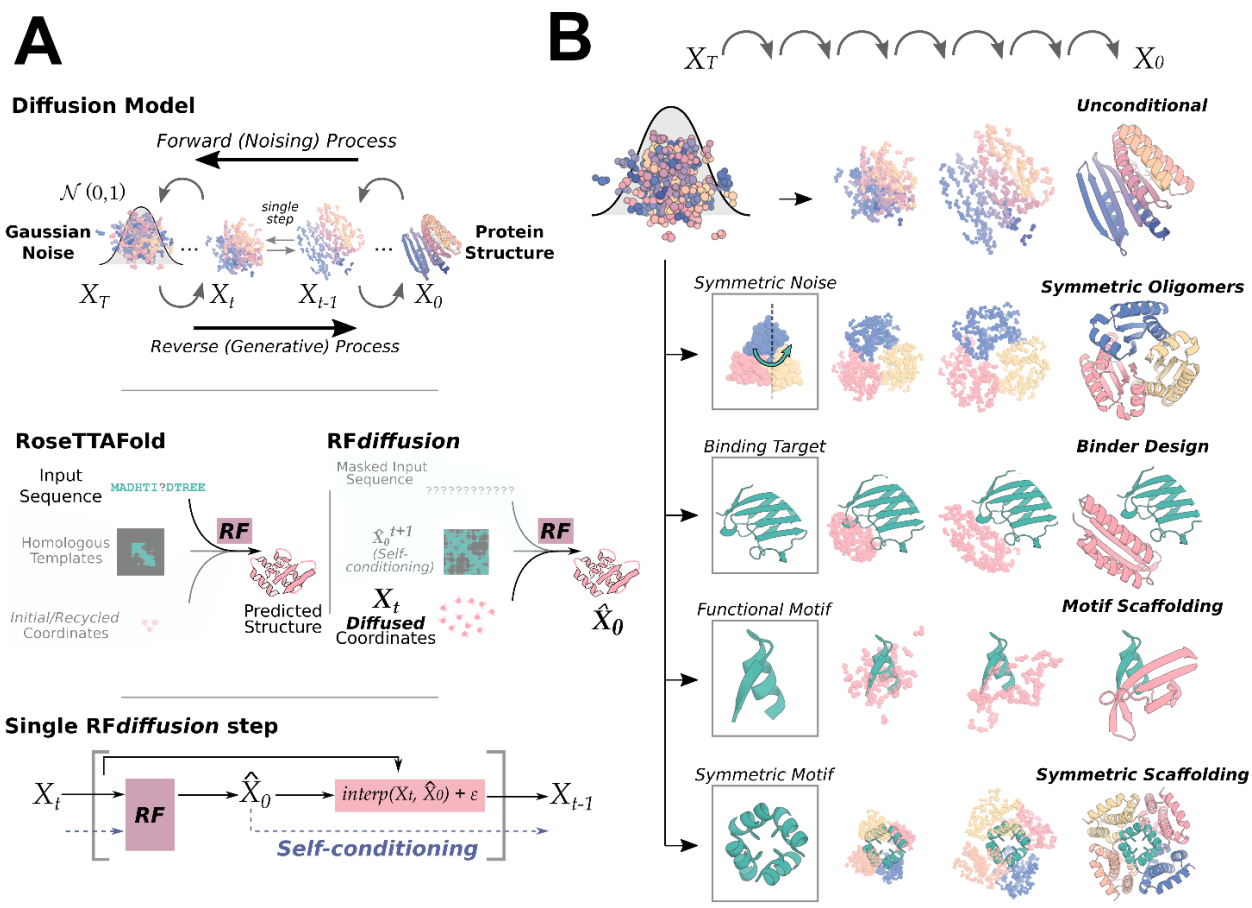


Figure 2.1. **RF*diffusion* is a denoising diffusion probabilistic model with RoseTTAFold fine-tuned as the denoising network. A** Top panel: Diffusion models for proteins are trained to recover structures of proteins corrupted with noise, and generate new structures by reversing the corruption process 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. RF*diffusion* (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 RF*diffusion*, the primary input is diffused residue frames. In both cases, the model predicts final 3D coordinates directly (denoted $\hat{X}_0$ in RF*diffusion*). In RF*diffusion*, 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), RF*diffusion* 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 $\hat{X}_0^{t}$. **B** RF*diffusion* is of broad applicability to protein design. RF*diffusion* 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 oligomer scaffolds (bottom row). In each case random noise, along with conditioning information, is input to RF*diffusion*, which iteratively refines that noise until a final protein structure is designed.

We construct a RoseTTAFold-based diffusion model, RF*diffusion*, using the RF frame representation which comprises a Cα coordinate and N-Cα-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[55] (PDB). For translations, we perturb Cα coordinates with 3D Gaussian noise. For residue orientations, we use Brownian motion on the manifold of rotation matrices (building on refs [56,57]). To enable RF*diffusion* to learn to reverse each step of the noising process, we train the model by minimizing a mean squared error (MSE) loss between frame predictions and the *true* protein structure (without alignment), averaged across all residues. This loss drives denoising trajectories to match the data distribution at each timestep and hence to converge on structures of designable protein backbones (Fig. S1A). 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.

To generate a new protein backbone, we first initialize random residue frames and RF*diffusion* makes a denoised prediction. Each residue frame is updated by taking a step in the direction of this prediction with some noise added to generate the input to the next step. The nature of the noise added and the size of this reverse step is chosen such that the denoising process matches the distribution of the noising process. RF*diffusion* 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. 2.2A left). However, through many such steps, the breadth of possible protein structures from which the input could have arisen narrows, and RF*diffusion* predictions come to closely resemble protein structures (Fig. 2.2A right). We use the ProteinMPNN network[3] to subsequently design sequences encoding these structures, typically sampling 8 sequences per design, in line with previous work[51,52]. We also considered simultaneously designing structure and sequence within RF*diffusion*, but given the excellent performance of combining ProteinMPNN with the diffusion of structure alone, we did not extensively explore this possibility.

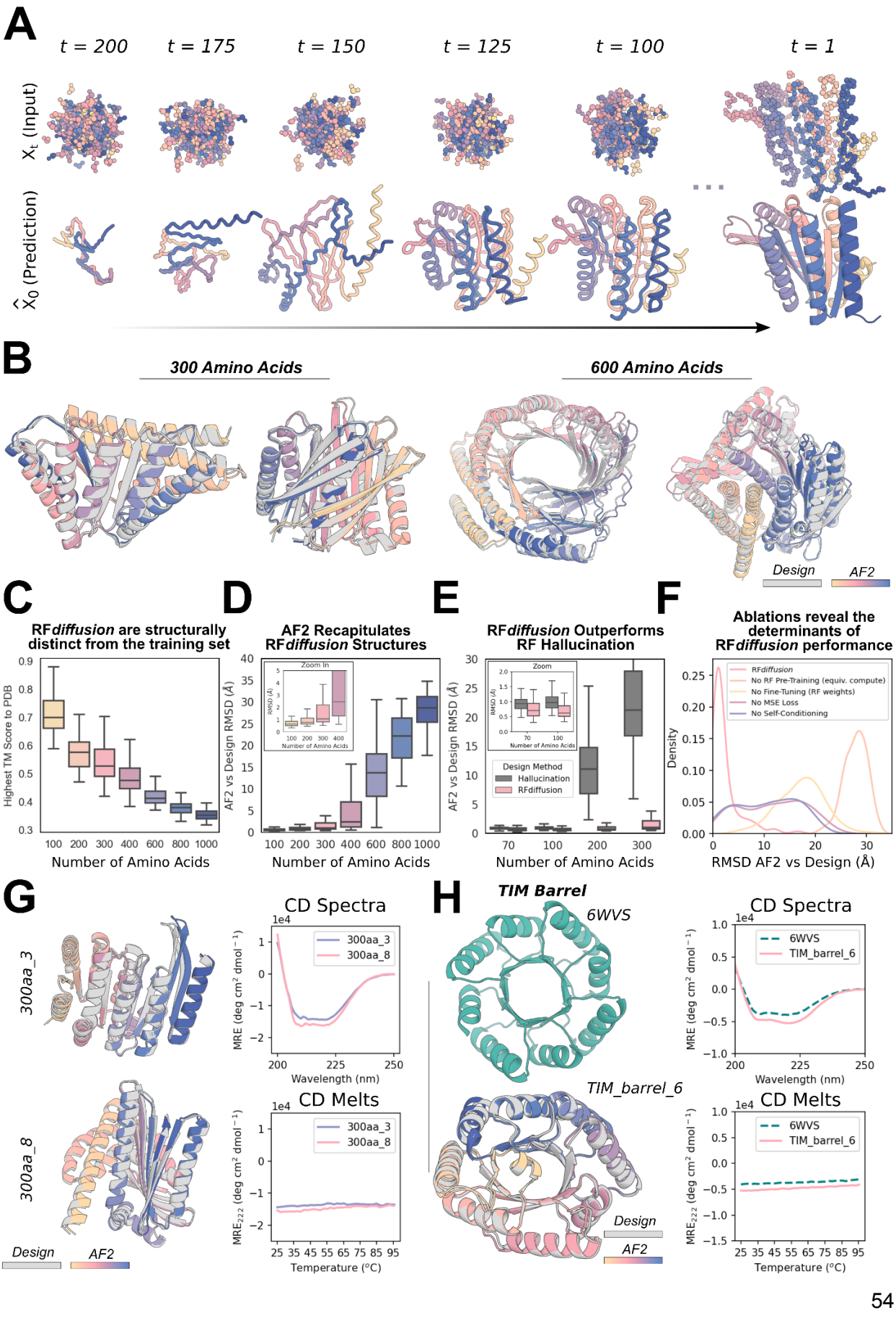

A
t = 200
t = 175
t = 150
t = 125
t = 100
t = 1
X_t (Input)
$\hat{X}_0$ (Prediction)
B
300 Amino Acids
600 Amino Acids
Design
AF2
C
RFdiffusion are structurally distinct from the training set
Highest TM Score to PDB
Number of Amino Acids
D
AF2 Recapitulates RFdiffusion Structures
AF2 vs Design RMSD (Å)
Zoom In
Number of Amino Acids
E
RFdiffusion Outperforms RF Hallucination
AF2 vs Design RMSD (Å)
Zoom
Design Method
Hallucination
RFdiffusion
Number of Amino Acids
F
Ablations reveal the determinants of RFdiffusion performance
RFdiffusion
No RF Pre-Training (equiv. compute)
No Fine-Tuning (RF weights)
No MSE Loss
No Self-Conditioning
Density
RMSD AF2 vs Design (Å)
G
300aa_3
300aa_8
CD Spectra
300aa_3
300aa_8
MRE (deg cm² dmol⁻¹)
Wavelength (nm)
CD Melts
MRE222 (deg cm² dmol⁻¹)
Temperature (°C)
Design
AF2
H
TIM Barrel
6WVS
TIM_barrel_6
CD Spectra
6WVS
TIM_barrel_6
MRE (deg cm² dmol⁻¹)
Wavelength (nm)
CD Melts
MRE222 (deg cm² dmol⁻¹)
Temperature (°C)
Design
AF2

Figure 2.2. **Outstanding performance of RF*diffusion* for monomer generation. A** An example trajectory of an unconditional 300 amino acid design, depicting the input to the model (X_t) and the corresponding $\hat{X}_0$ prediction. At early timesteps (high t), $\hat{X}_0$ predictions bear little resemblance to a protein, but are gradually refined into a protein structure. **B** RF*diffusion* 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 RF*diffusion* 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** RF*diffusion* 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, *in silico* success rates rapidly deteriorate beyond this length. **F** Ablating pre-training (by starting from untrained RF), RF*diffusion* fine-tuning (i.e., using original RF structure prediction weights as the denoiser), self-conditioning, or MSE losses (by training with FAPE) each dramatically decrease the performance of RF*diffusion*. RMSD between design and AF2 is shown, for the unconditional generation of 300 amino acid proteins (see Methods 5.8). **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** RF*diffusion* can condition on fold information. An example TIM barrel is shown (bottom left), conditioned on the secondary structure and

block-adjacency of a previously designed TIM barrel, PDB: 6WVS (top left). Designs have very similar CD spectra to 6WVS (top right), and are highly thermostable (bottom right). See also Fig. S9 for additional traces.

Fig. 2.1A highlights the similarities between RoseTTAFold structure prediction and an RF*diffusion* denoising step: in both cases, the networks transform coordinates into a predicted structure, conditioned on inputs to the model. In RoseTTAFold, sequence is the primary input, with additional structural information provided as templates and initial coordinates to the model. In RF*diffusion*, the primary input is the noised coordinates from the previous step. For design tasks, we optionally provide a range of auxiliary conditioning information, including partial sequence, fold information, or fixed functional motif coordinates.

We explored two different strategies for training RF*diffusion*: 1) in a manner akin to "canonical" diffusion models, with predictions at each timestep independent of predictions at previous timesteps (as in previous work[45,46,51,52]), and 2) with self-conditioning[58], where the model can condition on previous predictions between timesteps (Fig. 2.1A 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 RF*diffusion* dramatically improved performance on *in silico* benchmarks encompassing both conditional and unconditional protein design tasks (Fig. 2.2F). Increased coherence of predictions within self-conditioned trajectories may, at least in part, explain these performance increases. Fine-tuning RF*diffusion* from pre-trained RF weights was far more successful than training for an equivalent length of time from untrained weights and the MSE loss was also crucial for unconditional generation. For all *in silico* benchmarks in

this paper, we use the AF2 structure prediction network[2] for validation and define an *in silico* "success" as an RF*diffusion* 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 2 Å backbone-RMSD of the designed structure, and (3) within 1 Å backbone-RMSD on any scaffolded functional-site. This measure of *in silico* success has been found to correlate with experimental success[33,48,59] and is significantly more stringent than TM-score based metrics used elsewhere (refs [45,51,52,60,61]).

Unconditional protein monomer generation

As illustrated in Fig. 2.2B-D starting from random noise, RF*diffusion* can readily generate elaborate protein structures with little overall structural similarity to structures seen during training, indicating considerable generalization beyond the PDB (see Extended Data 1 for comparison of all designs in the paper to the PDB). The designs are diverse (Fig. S6A), spanning a wide range of alpha-, beta- and mixed alpha-beta- topologies, with AF2 and ESMFold (Fig. 2.2D) predictions very close to the design structure models for *de novo* designs with as many as 600 residues. RF*diffusion* 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[45]). 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. S7).

Physics-based protein design methodologies have struggled in unconstrained generation of diverse protein monomers due to the difficulty of sampling on the very large and rugged conformational landscape[62], and overcoming this limitation has been a primary test of deep learning based protein design approaches[19,45,51,52,63,64]. RF*diffusion* strongly outperforms Hallucination with RoseTTAFold, an experimentally validated method using Monte Carlo search or gradient descent to identify sequences predicted to fold into stable structures (Fig. 2.2E). RF*diffusion* generation is also more compute efficient than unconstrained Hallucination with RoseTTAFold, requiring ~2.5 minutes on an NVIDIA RTX A4000 GPU to generate a 100 residue structure compared to ~8.5 minutes for Hallucination. The computational efficiency of RF*diffusion* is also dramatically improved by taking larger steps at inference time, and by truncating trajectories early - the latter of which is an advantage of predicting the *final* structure at each timestep (e.g. a 100 residue protein can be generated in around 11s).

It is often desirable to be able to specify a particular protein fold during designed (such as TIM barrels or cavity-containing NTF2s for small molecule binder and enzyme design[65,66]), and thus we further fine-tuned RF*diffusion* to condition on secondary structure and/or fold information, enabling rapid and accurate generation of diverse designs with the desired topologies (Fig. 2.2H). *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. 2.2H).

Design of higher order oligomers

There is considerable interest in designing symmetric oligomers, which can serve as vaccine platforms[67], delivery vehicles[68], and catalysts[69]. 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[59].

We set out to generalize RF*diffusion* 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 RF*diffusion* 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, RF*diffusion* is able to generate symmetric oligomers with high *in silico* success rates, particularly when guided by an auxiliary inter- and

intra-chain contact potential. As illustrated in Fig. 2.3, RF*diffusion* 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. 2.3A, C10 symmetry) and complex mixed alpha/beta topologies (Fig. 2.3A, C8 symmetry; closest TM align in PDB: 6BRP, 0.47; 6BRO, 0.43 respectively).

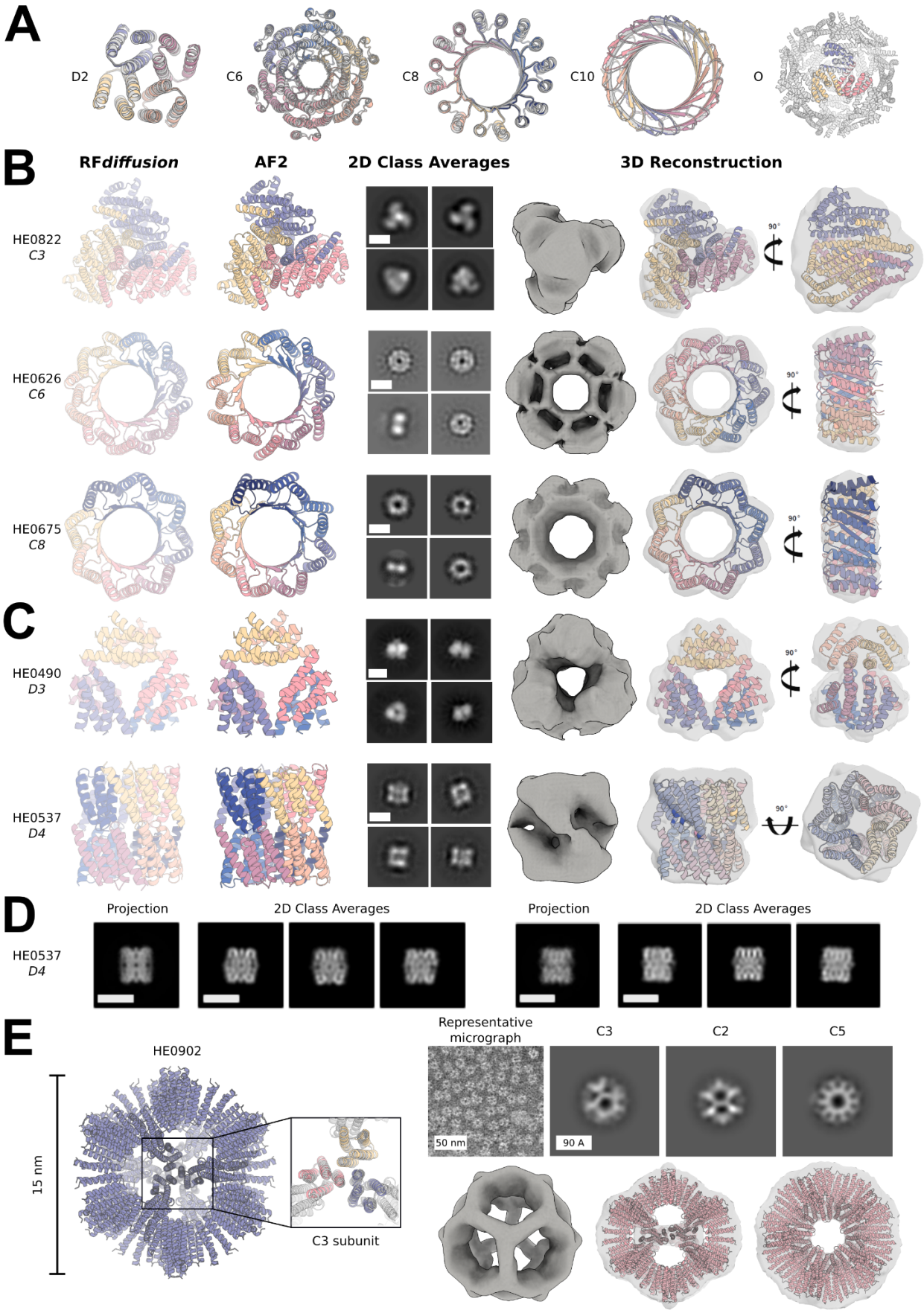

A
D2
C6
C8
C10
O
B
RFdiffusion
AF2
2D Class Averages
3D Reconstruction
HE0822
C3
HE0626
C6
HE0675
C8
C
HE0490
D3
HE0537
D4
D
Projection
2D Class Averages
Projection
2D Class Averages
HE0537
D4
E
HE0902
15 nm
C3 subunit
Representative micrograph
C3
C2
C5
50 nm
90 A

Figure 2.3 **Design and experimental characterization of high-order symmetric oligomers. A** RF*diffusion*-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 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.

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. 2.3). We describe these designs in the following paragraphs.

Electron microscopy characterization of a C3 design (HE0822) with 350 residue subunits (1050 residues in total) suggests that the actual structure is very close to the design, both over the 350 residue subunits and the overall C3 architecture. 2D class averages are clearly consistent with both top- and side-views of the design model, and a 3D reconstruction of the density has key features consistent with the design, including the distinctive pinwheel shape (Fig. 2.3B, top row). Electron microscopy 2D class averages of C5 and C6 designs with greater than 750 residues (HE0795, HE0789, HE0841) were also consistent with the respective design models.

RF*diffusion* also generated cyclic oligomers with alpha/beta barrel structures that resemble expanded TIM barrels and provide an interesting comparison between innovation during natural evolution and innovation through deep learning. The TIM barrel fold, with 8 strands and 8

helices, is one of the most abundant folds in nature[70]. Electron microscopy characterization confirmed the structure of two RF*diffusion* designed cyclic oligomers which considerably extend beyond this fold (Fig. 2.3B, bottom rows). HE0626 is a C6 alpha/beta barrel composed of 18 strands and 18 helices, and HE0675 is a C8 octamer composed of an inner ring of 16 strands and an outer ring of 16 helices arranged locally in a very similar repeating pattern to the TIM barrel (1:1 helix:strand). By nsEM, we observed 2D class averages for HE0626 that resemble this two ring organization, and for both HE0626 and HE0675 we obtained 3D reconstructions that are in agreement with the computational design models. The HE0600 design is also an alpha-beta barrel, but has two strands for every helix (24 strands and 12 helices in total) and is hence locally quite different from a TIM barrel. Whereas natural evolution has extensively explored structural variations of the classic 8-strand/8-helix TIM barrel fold, RF*diffusion* can more readily explore global changes in barrel curvature, enabling discovery of TIM barrel-like structures with many more helices and strands.

RF*diffusion* readily generated structures with dihedral and tetrahedral symmetries (Fig. 2.3C). SEC characterization indicated that 38 D2, 7 D3, and 3 D4 designs had the expected molecular weights (these have 4, 6, and 8 chains, respectively). While the D2 dihedrals are too small for nsEM, 2D class averages–and for some, 3D reconstructions– of D3 and D4 designs were congruent with the overall topologies of the design models (Fig. 2.3C). The reconstruction for the D3 HE0490 shows the characteristic triangular shape of the design. Similarly, the 3D reconstruction of the D4 HE0537 closely matches the design model, recapitulating the approximate 45° offset between tetramic subunits. Cryogenic electron microscopy (cryo-EM) 2D class averages for HE0537 are in very close agreement with 2D projections of the design model

(Fig. 2.3D) with similar placements of alpha helices; 3D reconstruction was complicated by preferred orientation effects. 2D nsEM class averages for a 12 chain tetrahedron (HE0964) were consistent with the design model, but we were unable to generate a 3D reconstruction of high confidence due to a lack of clear discernable design features at the resolution range provided by nsEM.

Icosahedra have 60 subunits arrayed around 2-fold, 3-fold and 5-fold symmetry axes. Of the 48 icosahedra selected for experimental validation, one was confirmed by nsEM to form the intended assembly. As shown in Fig. 2.3E on the left, HE0902 is a 15nm (diameter) highly-porous icosahedron composed of alpha helical subunits. The nsEM micrographs reveal highly homogeneous particles, and the corresponding 2D class averages and 3D reconstruction nearly perfectly match the design model (Fig. 2.3E), with triangular hubs arrayed around the empty C5 axes. Designs such as HE0902 (and future similar large assemblies) should be useful as new nanomaterials and vaccine scaffolds, with robust assembly and (in the case of HE0902) the outward facing N- and C-termini offering multiple possibilities for antigen display.

Functional motif scaffolding

We next investigated the use of RF*diffusion* 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 RF*diffusion*, we input motifs as 3D coordinates (including sequence and sidechains) both during conditional training and inference, and build scaffolds that hold the motif atomic coordinates in place. A number of deep learning methods

have been developed recently to address this problem, including RF_{joint} Inpainting[48], constrained Hallucination[48], and other DDPMs[45,51,61]. To rigorously evaluate the performance of these methods in comparison to RF*diffusion* 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[48,51,61,71–73]. 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.

RF*diffusion* solves 23 of the 25 benchmark problems, compared to 15 for Hallucination and 19 for RF_{joint} Inpainting (Fig. 2.4A-B). For 19/23 of the problems solved by RF*diffusion,* the fraction of successful designs is higher than either Hallucination or RF_{joint} Inpainting. The excellent performance of RF*diffusion* 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, RF*diffusion* generated successful solutions with higher *in silico* success rates when noise was not added during the reverse diffusion trajectories. Furthermore, the ability of RF*diffusion* to scaffold functional motifs is not related to their presence in the RF*diffusion* training set.

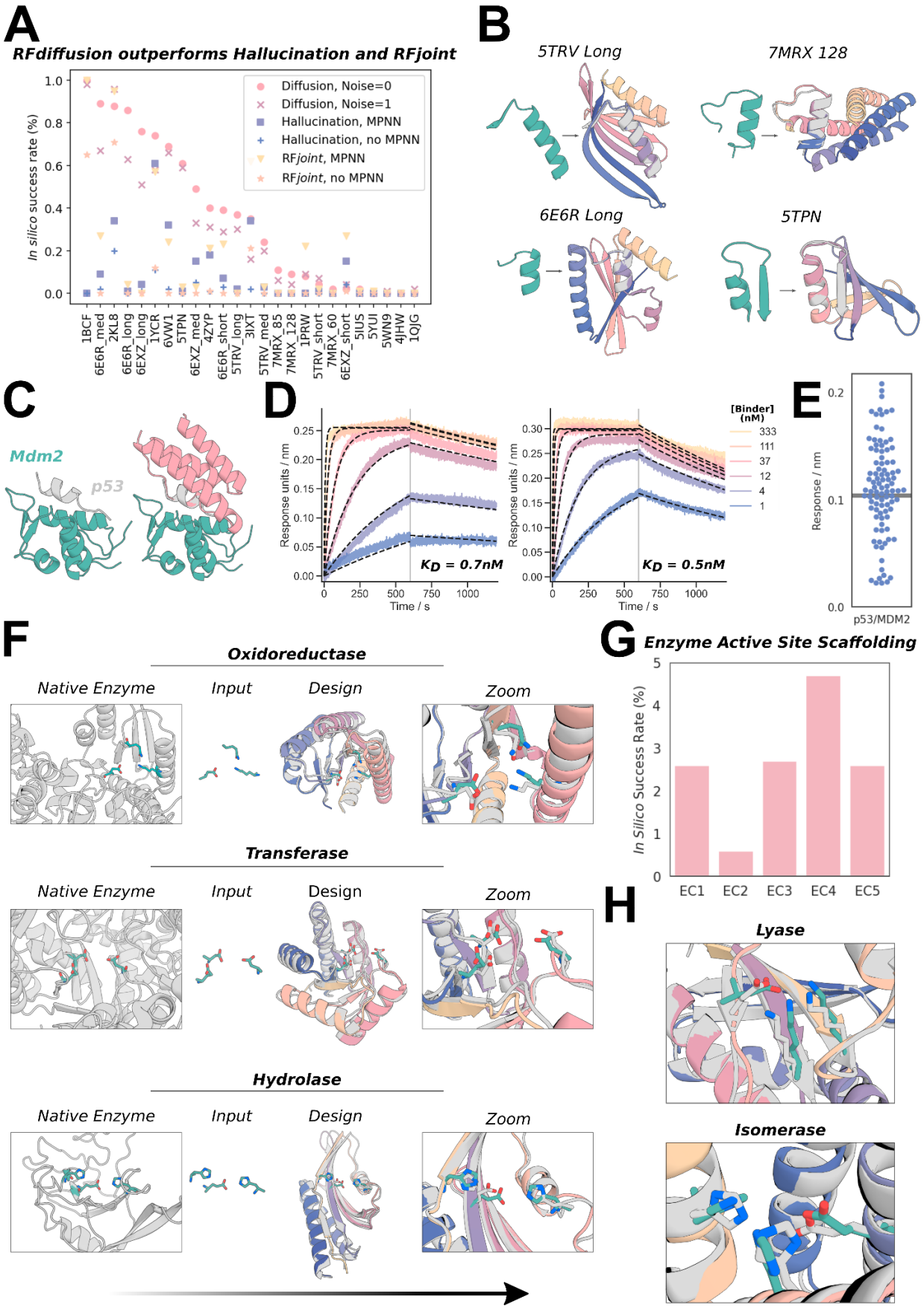

A
RFdiffusion outperforms Hallucination and RFjoint
In silico success rate (%)
Diffusion, Noise=0
Diffusion, Noise=1
Hallucination, MPNN
Hallucination, no MPNN
RFjoint, MPNN
RFjoint, no MPNN
1BCF
6E6R_med
2KL8
6E6R_long
6EXZ_long
1YCR
6VW1
5TPN
6EXZ_med
4ZYP
6E6R_short
5TRV_long
3IXT
5TRV_med
7MRX_85
7MRX_128
1PRW
5TRV_short
7MRX_60
6EXZ_short
5IUS
5YUI
5WN9
4JHW
1QJG
B
5TRV Long
7MRX 128
6E6R Long
5TPN
C
Mdm2
p53
D
Response units / nm
Time / s
KD = 0.7nM
KD = 0.5nM
[Binder] (nM)
333
111
37
12
4
1
E
Response / nm
p53/MDM2
F
Oxidoreductase
Transferase
Hydrolase
Native Enzyme
Input
Design
Zoom
G
Enzyme Active Site Scaffolding
In Silico Success Rate (%)
EC1
EC2
EC3
EC4
EC5
H
Lyase
Isomerase

Figure 2.4 **Scaffolding of diverse functional-sites with RF*diffusion*. A** RF*diffusion* has state of the art performance across 25 benchmark motif scaffolding problems collected from six recent publications, encompassing a broad range of motifs. *In silico* success was defined as AF2 RMSD to design model < 2 Å, AF2 RMSD to the native functional site (the "motif") < 1 Å, and AF2 predicted alignment error (pAE) < 5, and the examples are ordered by *in silico* success rate with RF*diffusion* (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). Note that while RF*diffusion* robustly outperforms both RF hallucination and RF_{joint} Inpainting, *in silico* success rates on the problems are correlated between the three methods, and RF*diffusion* 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 RF*diffusion* significantly outperforms existing methods. Teal: native motif; colors: AF2 prediction of an RF*diffusion* 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** RF*diffusion* 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 RF*diffusion* 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[41]. **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, Fig. S20H. **F** After fine-tuning on a task that mimics active-site scaffolding, RF*diffusion* can scaffold a broad range of enzyme active sites. Three examples are shown (Enzyme Classes, EC, 1-3; ref [54]). Left to

right: native enzyme (PDB: 1A4I, 1CWY, 1DE3); catalytic site (teal); RF*diffusion* output (gray: model, colors: AF2 prediction); zoom of active site. **G** *In silico* success rates on active sites derived from EC1-5 (AF2 Motif RMSD vs native: backbone < 1 Å, backbone and sidechain atoms < 1.5 Å, RMSD AF2 vs design < 2 Å, AF2 pAE < 5). H) Zoom in views of two further *in silico* successful designs, for EC4 and EC5 (active sites from PDB: 1P1X, 1SNZ). Metrics for examples in (**F**) and (**H**) (AF2 vs design backbone RMSD, AF2 vs design motif backbone RMSD, AF2 vs design motif full-atom RMSD, AF2 pAE): EC2: 0.93 Å, 0.50 Å, 1.29 Å, 3.51; EC3: 0.92 Å, 0.60 Å, 1.07 Å, 4.59; EC4: 0.93 Å, 0.80 Å, 1.03 Å, 4.41; EC5: 0.78 Å, 0.44 Å, 1.14 Å, 3.32.

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[74]. *In silico* success has been described elsewhere[48], but experimental success has not been reported. We used an RF*diffusion* 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 RF*diffusion*, and experimentally identified 0.5nM and 0.7nM binders (Fig. 2.4C-D), three orders of magnitude higher affinity than the reported 600nM affinity of the p53 peptide alone[75]. The experimental success rate for this problem was particularly striking with 55/95 designs showing some detectable binding at 10μM (Fig. 2.4E) and multiple designs with affinities in the low- to sub- nanomolar range (Fig. 2.4D).

Scaffolding enzyme active sites

A grand challenge in protein design is to scaffold minimal descriptions of enzyme active sites comprising a few single amino acids. While some *in silico* success has been reported previously[48], a general solution that can readily produce high-quality, orthogonally-validated outputs remains elusive. Following fine-tuning on a task mimicking this problem, RF*diffusion* was able to scaffold enzyme active sites comprising multiple sidechain and backbone functional groups with high accuracy and *in silico* success rates across a range of enzyme classes (Fig. 2.4F-H; *in silico* successes were not present without fine-tuning). While RF*diffusion* is currently unable to *explicitly* model bound small molecules (see conclusion), the substrate can be *implicitly* modeled using an external potential to guide the generation of "pockets" around the active site. As a demonstration, we scaffold a retroaldolase active site triad while implicitly modeling its substrate.

Symmetric functional-motif scaffolding for metal coordinating assemblies and antiviral therapeutics and vaccines

A number of important design challenges involve the scaffolding of multiple copies of a functional motif in symmetric arrangements. For example, many viral glycoproteins are trimeric, and symmetry matched arrangements of inhibitory domains can be extremely potent[13,76–78]. Conversely, symmetric presentation of viral epitopes in an arrangement that mimics the virus could induce new classes of neutralizing antibodies[79,80]. To explore this general direction, we sought to design trimeric multivalent binders to the SARS-CoV-2 spike protein. In previous

work, flexible linkage of a binder to the ACE2 binding site (on the spike protein receptor binding domain) to a trimerization domain yielded a high-affinity inhibitor that had potent and broadly neutralizing antiviral activity in animal models[76]. Ideally, however, symmetric fusions to binders would be rigid, so as to reduce the entropic cost of binding while maintaining the avidity benefits from multivalency. We used RF*diffusion* to design C3 symmetric trimers which rigidly hold three binding domains (the functional motif in this case) such that they exactly match the ACE2 binding sites on the SARS-CoV-2 spike protein trimer. Design models were confidently predicted by AF2 to both assemble as C3-symmetric oligomers, and to scaffold the AHB2 SARS-CoV-2 binder interface with high accuracy (Fig. 2.5A).

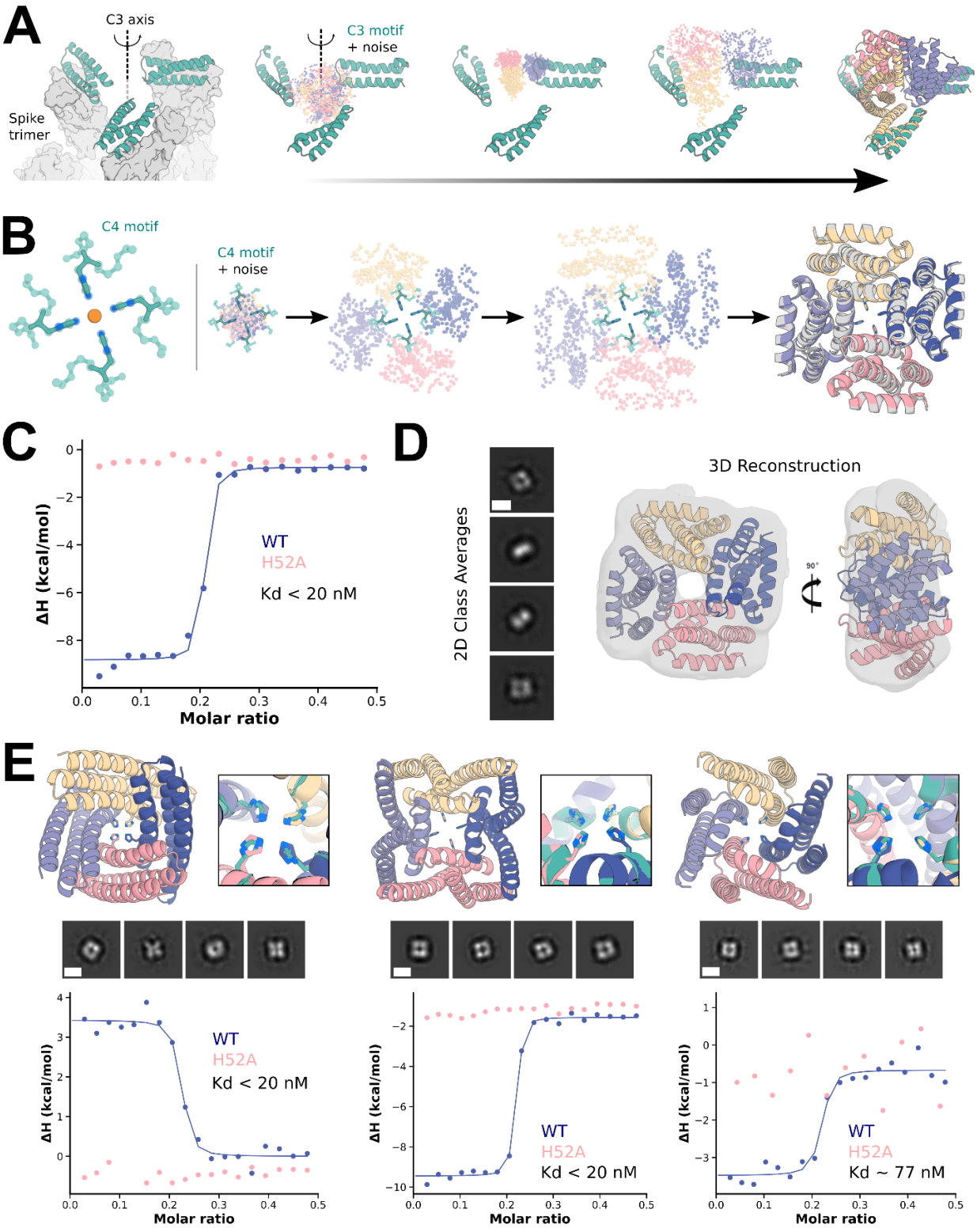

A
C3 axis
Spike trimer
C3 motif + noise
B
C4 motif
C4 motif + noise
C
ΔH (kcal/mol)
Molar ratio
WT
H52A
Kd < 20 nM
D
2D Class Averages
3D Reconstruction
90°
E
ΔH (kcal/mol)
Molar ratio
WT
H52A
Kd < 20 nM
Kd < 20 nM
Kd ~ 77 nM

Figure 2.5 **Symmetric motif scaffolding with RF*diffusion*.** **A** Design of C3-symmetric oligomers to scaffold the binding interface of the designed ACE2 mimic, AHB2 (left, teal), against the SARS-CoV-2 spike trimer (left, gray). Starting from AHB2 bound to each of the three ACE2 binding sites on the spike trimer, RF*diffusion* was used to generate C3-symmetric oligomers that hold the three AHB2 exactly in place to simultaneously engage the binding sites on all three spike subunits.The first 55 amino-acids of each minibinder copy are used as the symmetric motif input to RF*diffusion* (middle). The method produces designs whose AF2 predictions (right) recapitulate the mini-binder motif with high accuracy on the asymmetric unit (0.6 Å RMSD) and good accuracy the symmetric motif (2.9 Å RMSD). **B** Design of C4-symmetric oligomers to scaffold a theoretical Ni^{2+} binding motif (left). Starting from a square-planar set of histidine rotamers within three-residue helical fragments (Methods 5.9) and C4-symmetric noise, an RF*diffusion* trajectory iteratively builds a symmetric oligomer scaffolding the theoretical Ni^{2+} binding domain (middle). AF2 predictions (color) overlaid with the RF*diffusion* design model (gray) agree closely, with backbone RMSD for the particular example < 1.0 Å (right). **C** 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. **D** 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). **E** Additional experimentally

characterized Ni^{2+} binding oligomers NiB2.15 (left), NiB1.12 (middle), and NiB1.20 (right) from RF*diffusion* 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 Ni^{2+} binding mediated directly by the scaffolded histidines at the designed 1:4 stoichiometry.

The ability to scaffold functional sites with any desired symmetry opens up new approaches to designing metal-coordinating protein assemblies[81,82]. 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. RF*diffusion* provides a general route to building up symmetric protein assemblies around such sites, with the symmetry of the assembly matching the symmetry of the coordination geometry. As a first test, we sought to design square planar Ni^{2+} binding sites. We designed C4 protein assemblies with four central histidine imidazoles arranged in an ideal Ni^{2+}-binding site with square planar coordination geometry. Diverse designs starting from various different C4-symmetric histidine square planar sites (Fig. 2.5B) had good *in silico* success, with the histidine residues in near ideal geometries for coordinating metal in the AF2 predicted structures (Fig. 2.5E).

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 Ni^{2+} coordination by isothermal titration calorimetry. 18 designs bound Ni^{2+} with dissociation constants ranging from low nanomolar to low micromolar (Fig. 2.5C,E). 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. 2.5E), suggesting that Ni^{2+} coordination is entropically driven in these assemblies. To confirm that Ni^{2+} binding was indeed mediated by the scaffolded histidine 52, we mutated this residue to alanine, which abolished or dramatically reduced binding in 17/17 cases with successful expression (Fig. 2.5C,E; 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. 2.5D,E), 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. 2.5D).

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[83]. A general method to de novo design binders from target structure information alone using the physically-based Rosetta method was recently described[1]. Subsequently, utilizing ProteinMPNN for sequence design and AF2 for design filtering was found to improve design success rates[33]. However, experimental success rates were low, still requiring many thousands of designs to be screened for each design campaign[1], 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[1]. To our knowledge, no deep-learning method has yet demonstrated experimental general success in designing completely *de novo* binders.

We reasoned that RF*diffusion* 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 RF*diffusion* 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. 2.6A). 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.

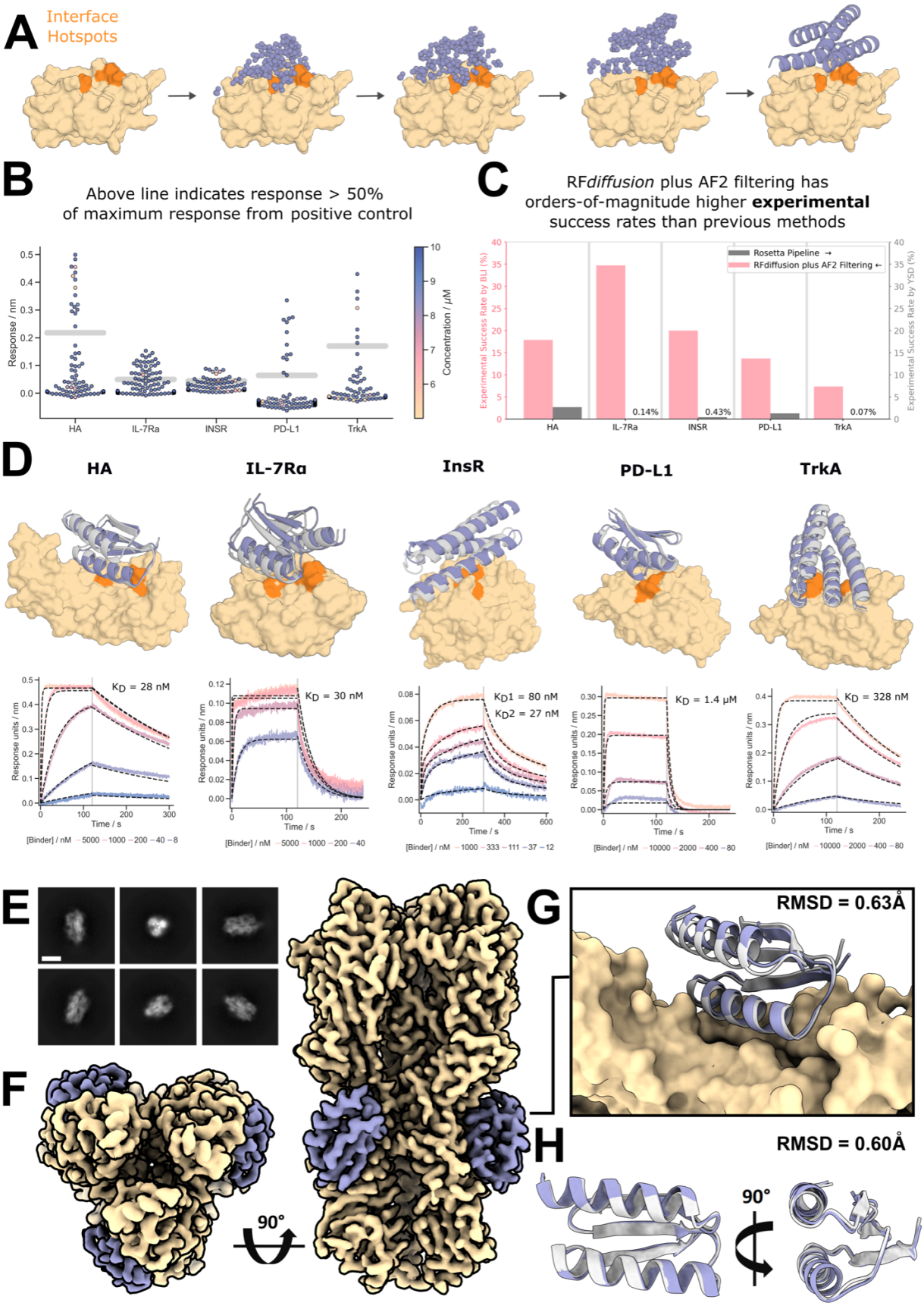

A
Interface
Hotspots
B
Above line indicates response > 50%
of maximum response from positive control
Response / nm
Concentration / μM
HA
IL-7Ra
INSR
PD-L1
TrkA
C
RFdiffusion plus AF2 filtering has
orders-of-magnitude higher experimental
success rates than previous methods
Rosetta Pipeline →
RFdiffusion plus AF2 Filtering ←
Experimental Success Rate by BLI (%)
Experimental Success Rate by YSD (%)
0.14%
0.43%
0.07%
D
HA
IL-7Rα
InsR
PD-L1
TrkA
KD = 28 nM
KD = 30 nM
KD1 = 80 nM
KD2 = 27 nM
KD = 1.4 μM
KD = 328 nM
Response units / nm
Time / s
[Binder] / nM 5000 1000 200 40 8
[Binder] / nM 5000 1000 200 40
[Binder] / nM 1000 333 111 37 12
[Binder] / nM 10000 2000 400 80
[Binder] / nM 10000 2000 400 80
E
F
90°
G
RMSD = 0.63Å
H
RMSD = 0.60Å
90°

Figure 2.6 **Design of *de novo* protein-binding proteins. A-B** *De novo* binders were designed to five protein targets; Influenza A H1 Hemagglutinin (HA), Interleukin-7 Receptor-α (IL-7Rα), Insulin Receptor (InsR), Programmed Death-Ligand 1 (PD-L1), and Tropomyosin Receptor Kinase A (TrkA.) RF*diffusion* generates protein binders by conditioning on interface hotspot residues. Additionally, the general topology of the binders generated by RF*diffusion* can be controlled using fold-conditioning. **B** *De novo* protein binders were identified for all five of the targets. Designs that bound at 10μM 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 10μM and thus may be false negatives. **C** RF*diffusion* designed binders have very high experimental success rates compared to the previous design campaigns against the same targets. For IL-7Rα, Insulin Receptor, PD-L1, and TrkA, RF*diffusion* has success rates ~2 orders-of-magnitude higher than the original design campaigns. We attribute one order-of-magnitude in success rate to RF*diffusion*, 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-7Rα: 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 K_Ds 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: 0.7 Å, 0.9 Å, 1.2 Å, 0.7 Å). **E** Cryo-EM 2D class averages of the RF*diffusion* binder, *HA_20*, bound to Influenza Hemagglutinin, strain A/USA:Iowa/1943 H1N1 (scale bar = 10 nm). **F** 2.9 Å cryo-EM 3D reconstruction of the corresponding complex viewed along two orthogonal axes. The RF*diffusion* 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 RF*diffusion* design: 0.63 Å). **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 RF*diffusion* to previous binder design methods, we performed binder design campaigns against 5 targets: Influenza A H1 Hemagglutinin (HA)[11], Interleukin-7 Receptor-α (IL-7Rα)[1], Programmed Death-Ligand 1 (PD-L1)[1], Insulin Receptor[1], and Tropomyosin Receptor Kinase A (TrkA)[1]. 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[33], and 95 were selected for each target for experimental characterization.

The designed binders were expressed in *E. coli* and purified, and binding was assessed through single point biolayer interferometry (BLI) screening at 10μM binder concentration (Fig. 2.6B). In each case, a positive control was included that binds to the site targeted by the designs on the target protein[1]. The overall experimental success rate, defined as binding at or above 50% of the maximal response for the positive control, was 19% (this is a conservative estimate as some designs which showed binding had insufficient material to permit screening at 10μM (Fig. 2.6B)); an increase of approximately 2 orders-of-magnitude over our previous Rosetta-based method on the same targets (Fig. 2.6C). 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-7Rα binders with affinities of approximately 30nM (Fig. 2.6D). To assess binder specificity, 6 of the highest affinity IL-7Rα binders were assessed via competition BLI, and all 6 competed for binding with the structurally validated positive control.

We solved the structure of a designed Influenza binder, *HA_20*, in complex with Iowa43 HA using cryo electron microscopy. Raw electron micrographs revealed a well-folded HA glycoprotein with clearly discernible side, top, and tilted view orientations suspended in a thin layer of vitreous ice. 2D class averages further show clear secondary structure elements corresponding to both Iowa43 HA, as well as the *HA_20* binder bound to the stem (Fig 6E). 3D heterogenous refinement without symmetry revealed full occupancy of all three HA stem epitopes by the *HA_20* binder. A final non-uniform 3D refinement reconstruction with C3 symmetry yielded a 2.9 Å map of the HA/*HA_20* protein-protein complex (Fig 6F) and corresponding 3D structure which nearly perfectly matches the computational design model (0.63 Å, Fig 6F,G; the sidechain interactions at the interface are very different from the closest structure in the PDB). Over the binder alone, the experimental structure deviates from the RF*diffusion* design by only 0.60 Å (Fig 6H). These results demonstrate the ability of RF*diffusion* to generate new proteins with atomic level accuracy, and to precisely target functionally relevant sites on therapeutically important proteins.

Discussion

RF*diffusion* is a major improvement over current physically-based and deep learning based protein design methods over a wide range of design challenges. For the classic unconstrained protein structure generation problem, RF*diffusion* readily generates diverse protein structures with as many as 600 residues that are accurately predicted by AF2 (and ESMFold). This exceeds the complexity and accuracy achieved by previously described methods, diffusion-based or otherwise (during review of this manuscript, a recent hallucination based approach was described that also generates complex structures[84]). Half of our tested (length 200 and 300) unconditional designs express solubly, exhibiting CD spectra consistent with the design models and high thermostability - despite the substantially increased complexity, the ideality and stability of RF*diffusion* designs is akin to that of previous *de novo* design methods. Success rates are even higher for conditional generation of specific folds, as illustrated by the design and characterization of 8 thermostable TIM barrels. The versatility and control provided by RF*diffusion* enables the extension of unconditional generation to higher order architectures with any desired symmetry - surpassing Hallucination methods, which have so far been limited to cyclic symmetries. Experimental characterization of a subset of these oligomers using electron microscopy revealed structures very similar to the design models, and in many cases without global similarity to known protein oligomers.

There has been recent progress in scaffolding protein functional motifs using deep learning methods (RF Hallucination, RF_{joint} Inpainting, and diffusion), but Hallucination becomes very

slow for large systems, inpainting fails when insufficient starting information is provided, and previous diffusion methods had quite low accuracy. Our benchmark tests show that RF*diffusion* considerably outperforms previous methods in the complexity of the motifs that can be scaffolded, the ability to precisely position sidechains (for catalysis and other functions), and the accuracy of motif recapitulation by AF2. The robust design of MDM2 binding proteins with three orders of magnitude higher affinities than the scaffolded P53 motif experimentally demonstrates the power of RF*diffusion* for scaffolding functional motifs. By combining accurate motif scaffolding with the design of symmetric assemblies, we demonstrate consistent and atomically precise positioning of sidechains to coordinate Ni^{2+} ions across large and diverse tetramers.

For binder design, substantial progress was recently made using Rosetta in designing binding proteins from target structural information alone[1], but this required testing tens of thousands of sequences. RF*diffusion* and improved filtering[33] now enable experimental success rates that are approximately two orders of magnitude higher, and consequently, high affinity binders can be identified through testing only dozens of designs (at least for the somewhat non-polar sites targeted here; further studies will be required to assess success rates on more polar target sites). A high resolution cryo-EM structure of one of these designs in complex with influenza hemagglutinin further shows that RF*diffusion* can design functional proteins with atomic accuracy. In the accompanying paper (Vázquez Torres *et al.*)[85], we demonstrate the ability of RF*diffusion* to design picomolar affinity binders to flexible helical peptides with x-ray crystallographically confirmed accuracy, further highlighting the utility of RF*diffusion* for *de novo* binder design. Vázquez Torres *et al.* also show how RF*diffusion* can be extended for

protein model refinement by partial noising and denoising, which enables tunable sampling around a given input structure. For peptide binder design, this enabled increases in affinity of nearly three orders of magnitude without high-throughput screening.

Overall, the complexity of the problems solvable with RF*diffusion* and the robustness and accuracy of the solutions (extensively validated both *in silico* and by experiment) far exceeds what has been achieved previously. In a manner reminiscent of the generation of images from text prompts, RF*diffusion* makes possible, with minimal specialist knowledge, the generation of functional proteins from very simple molecular specifications (for example, high affinity binders to a user-specified target protein, and diverse protein assemblies from user-specified symmetries).

The power and scope of RF*diffusion* can be extended in several directions. RF has recently been extended to nucleic acids and protein-nucleic acid complexes[23], which should enable RF*diffusion* to design nucleic acid binding proteins, and perhaps folded RNA structures. Extension of RF to incorporate ligands should similarly enable extension of RF*diffusion* to explicitly model ligand atoms, and allow the design of protein-ligand interactions. The ability to customize RF*diffusion* to specific design challenges by addition of external potentials and by fine-tuning (as illustrated here for catalytic site scaffolding, binder-targeting and fold-specification), along with continued improvements to the underlying methodology, should enable *de novo* protein design to achieve still higher levels of complexity, to approach and – in some cases – surpass what natural evolution has achieved.

Chapter 4

Conclusions

The work presented in this dissertation has transformed *de novo* protein binder design. Prior to this work, the design of protein binders required months of computational work and the ordering of tens of thousands of designs to identify experimentally successful binders. Now, the computational step takes ~1 week and designs can be ordered and screened in an additional week's work. It is also now routine to identify successful binders from ~100 designs. Protein binders may now be designed with arbitrarily large scaffolds which are designed at runtime as opposed to being limited to a constrained set of *a priori*-designed scaffolds. The design tools described in this work are easy to use and install and are available free and open source. Now anyone with access to a GPU or two can design protein binders.

This rapid and high success pipeline enables compelling applications in probing biological function and rapidly iterating on therapeutic target discovery. By shrinking the time from target structure and hypothesis to initial binding hit from 2-3 months for animal immunization and antibody identification to ~2 weeks for this pipeline, many more biological hypotheses may be explored. The combination of this rapid pipeline with a reliable disease model is quite compelling.

Finally, it is notable that the deep learning pipeline still contains all of the major steps from the previous pipeline: design of binder structure and dock, design of a sequence to encode the structure of the binder, and a filtering step to select those designs likely to work experimentally. Assessment of each step in this pipeline can be used to judge the feasibility of future, yet-unsolved, protein design tasks; steps in the pipeline that do not work well for specific tasks are excellent candidates for future methods development.

References

1. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* (2022) doi:10.1038/s41586-022-04654-9.
2. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
4. Nelson, A. L., Dhimolea, E. & Reichert, J. M. Development trends for human monoclonal antibody therapeutics. *Nat. Rev. Drug Discov.* **9**, 767–774 (2010).
5. Brennan, D. J., O'Connor, D. P., Rexhepaj, E., Ponten, F. & Gallagher, W. M. Antibody-based proteomics: fast-tracking molecular diagnostics in oncology. *Nat. Rev. Cancer* **10**, 605–617 (2010).
6. Stern, L. A., Case, B. A. & Hackel, B. J. Alternative non-antibody protein scaffolds for molecular imaging of cancer. *Curr. Opin. Chem. Eng.* **2**, 425–432 (2013).
7. Warram, J. M. *et al.* Antibody-based imaging strategies for cancer. *Cancer Metastasis Rev.* **33**, 809–822 (2014).
8. Gray, A. *et al.* Animal-free alternatives and the antibody iceberg. *Nat. Biotechnol.* **38**, 1234–1239 (2020).
9. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1**, 755–768 (2006).
10. Hackel, B. J., Kapila, A. & Dane Wittrup, K. Picomolar Affinity Fibronectin Domains

Engineered Utilizing Loop Length Diversity, Recursive Mutagenesis, and Loop Shuffling. *J. Mol. Biol.* **381**, 1238–1252 (2008).

11. Chevalier, A. *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
12. Silva, D.-A. *et al.* De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019).
13. Strauch, E.-M. *et al.* Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nat. Biotechnol.* **35**, 667–671 (2017).
14. Fleishman, S. J. *et al.* Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science* **332**, 816–821 (2011).
15. Baran, D. *et al.* Principles for computational design of binding antibodies. *Proc. Natl. Acad. Sci.* **114**, 10900–10905 (2017).
16. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
17. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
18. Baek, M. & Baker, D. Deep learning and protein structure modeling. *Nat. Methods* **19**, 13–14 (2022).
19. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
20. Wang, J. *et al. Deep learning methods for designing proteins scaffolding functional*

sites. http://biorxiv.org/lookup/doi/10.1101/2021.11.10.468128 (2021) doi:10.1101/2021.11.10.468128.

21. Hiranuma, N. *et al.* Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* **12**, 1340 (2021).
22. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis and testing. *Science* **357**, 168–175 (2017).
23. Baek, M., McHugh, R., Anishchenko, I., Baker, D. & DiMaio, F. Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. 2022.09.09.507333 Preprint at https://doi.org/10.1101/2022.09.09.507333 (2022).
24. Watson, J. L. *et al.* Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. 2022.12.09.519842 Preprint at https://doi.org/10.1101/2022.12.09.519842 (2022).
25. De Munck, S. *et al.* Structural basis of cytokine-mediated activation of ALK family receptors. *Nature* **600**, 143–147 (2021).
26. Jones, B. C. *et al.* Crystal structure of human cytomegalovirus IL-10 bound to soluble human IL-10R1. *Proc. Natl. Acad. Sci.* **99**, 9404–9409 (2002).
27. Stauber, D. J., Debler, E. W., Horton, P. A., Smith, K. A. & Wilson, I. A. Crystal structure of the IL-2 signaling complex: Paradigm for a heterotrimeric cytokine receptor. *Proc. Natl. Acad. Sci.* **103**, 2788–2793 (2006).
28. Yang, H. *et al.* Structural basis of immunosuppression by the therapeutic antibody daclizumab. *Cell Res.* **20**, 1361–1371 (2010).
29. Wang, X., Rickert, M. & Garcia, K. C. Structure of the Quaternary Complex of

Interleukin-2 with Its α, ß, and γ c Receptors. *Science* **310**, 1159–1163 (2005).

30. Rickert, M., Wang, X., Boulanger, M. J., Goriatcheva, N. & Garcia, K. C. The Structure of Interleukin-2 Complexed with Its Alpha Receptor. *Science* **308**, 1477–1480 (2005).
31. Leaver-Fay, A. *et al.* Rosetta3. in *Methods in Enzymology* vol. 487 545–574 (2011).
32. Norn, C. *et al.* Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci.* **118**, e2017228118 (2021).
33. Bennett, N. *et al.* Improving de novo Protein Binder Design with Deep Learning. 2022.06.15.495993 Preprint at https://doi.org/10.1101/2022.06.15.495993 (2022).
34. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
35. Dou, J. *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
36. Hoover, D. M. & Lubkowski, J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43 (2002).
37. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C.-M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel.* **23**, 155–159 (2010).
38. Aricescu, A. R., Lu, W. & Jones, E. Y. A time- and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 1243–1250 (2006).
39. Howarth, M. *et al.* Monovalent, reduced-size quantum dots for imaging receptors on

living cells. *Nat. Methods* **5**, 397–399 (2008).

40. Duhovny, D., Nussinov, R. & Wolfson, H. J. Efficient Unbound Docking of Rigid Molecules. in *Algorithms in Bioinformatics* (eds. Guigó, R. & Gusfield, D.) 185–200 (Springer, 2002). doi:10.1007/3-540-45784-4_14.
41. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. Preprint at http://arxiv.org/abs/2006.11239 (2020).
42. Dhariwal, P. & Nichol, A. Diffusion Models Beat GANs on Image Synthesis. Preprint at http://arxiv.org/abs/2105.05233 (2021).
43. Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. & Ganguli, S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. Preprint at http://arxiv.org/abs/1503.03585 (2015).
44. Goodfellow, I. *et al.* Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
45. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. Preprint at http://arxiv.org/abs/2205.15019 (2022).
46. Luo, S. *et al.* Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models. 2022.07.10.499510 Preprint at https://doi.org/10.1101/2022.07.10.499510 (2022).
47. Kuhlman, B. *et al.* Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **302**, 1364–1368 (2003).
48. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**,

387–394 (2022).

49. Ramesh, A. *et al.* Zero-Shot Text-to-Image Generation. Preprint at http://arxiv.org/abs/2102.12092 (2021).
50. Saharia, C. *et al.* Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. (2022) doi:10.48550/ARXIV.2205.11487.
51. Trippe, B. L. *et al.* Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. Preprint at http://arxiv.org/abs/2206.04119 (2022).
52. Wu, K. E. *et al.* Protein structure generation via folding diffusion. (2022) doi:10.48550/ARXIV.2209.15611.
53. Wu, R. *et al. High-resolution* de novo *structure prediction from primary sequence*. http://biorxiv.org/lookup/doi/10.1101/2022.07.21.500999 (2022) doi:10.1101/2022.07.21.500999.
54. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
55. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
56. De Bortoli, V. *et al.* Riemannian Score-Based Generative Modelling. (2022) doi:10.48550/ARXIV.2202.02763.
57. Leach, A., Schmon, S. M., Degiacomi, M. T. & Willcocks, C. G. DENOISING DIFFUSION PROBABILISTIC MODELS ON SO(3) FOR ROTATIONAL ALIGNMENT. 8 (2022).
58. Chen, T., Zhang, R. & Hinton, G. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. Preprint at http://arxiv.org/abs/2208.04202

(2022).

59. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

60. Ingraham, J. *et al. Illuminating protein space with a programmable generative model*. http://biorxiv.org/lookup/doi/10.1101/2022.12.01.518682 (2022) doi:10.1101/2022.12.01.518682.

61. Lee, J. S., Kim, J. & Kim, P. M. *ProteinSGM: Score-based generative modeling for* de novo *protein design*. http://biorxiv.org/lookup/doi/10.1101/2022.07.13.499967 (2022) doi:10.1101/2022.07.13.499967.

62. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).

63. Anand, N. & Huang, P. Generative modeling for protein structures. in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).

64. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. *AlphaDesign: A* de novo *protein design framework based on AlphaFold*. http://biorxiv.org/lookup/doi/10.1101/2021.10.11.463937 (2021) doi:10.1101/2021.10.11.463937.

65. Basanta, B. *et al.* An enumerative algorithm for de novo design of proteins with diverse pocket structures. *Proc. Natl. Acad. Sci.* **117**, 22135–22145 (2020).

66. Pan, X. *et al.* Expanding the space of protein geometries by computational design of de novo fold families. *Science* **369**, 1132–1136 (2020).

67. Marcandalli, J. *et al.* Induction of Potent Neutralizing Antibody Responses by a Designed Protein Nanoparticle Vaccine for Respiratory Syncytial Virus. *Cell* **176**, 1420-1431.e17 (2019).

68. Butterfield, G. L. *et al.* Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).

69. Goodsell, D. S. & Olson, A. J. Structural Symmetry and Protein Function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).

70. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the $(\beta\alpha)_8$ -Barrel Enzyme Fold. *Chem. Rev.* **105**, 4038–4055 (2005).

71. Sesterhenn, F. *et al.* De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science* **368**, eaay5051 (2020).

72. Yang, C. *et al.* Bottom-up de novo design of functional proteins with complex structural features. *Nat. Chem. Biol.* **17**, 492–500 (2021).

73. Glasgow, A. *et al.* Engineered ACE2 receptor traps potently neutralize SARS-CoV-2. *Proc. Natl. Acad. Sci.* **117**, 28046–28055 (2020).

74. Chène, P. Inhibiting the p53–MDM2 interaction: an important target for cancer therapy. *Nat. Rev. Cancer* **3**, 102–109 (2003).

75. Kussie, P. H. *et al.* Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* **274**, 948–953 (1996).

76. Hunt, A. C. *et al.* Multivalent designed proteins neutralize SARS-CoV-2 variants of concern and confer protection against infection in mice. *Sci. Transl. Med.* **14**, eabn1252 (2022).

77. Silverman, J. *et al.* Multivalent avimer proteins evolved by exon shuffling of a family of human receptor domains. *Nat. Biotechnol.* **23**, 1556–1561 (2005).

78. Detalle, L. *et al.* Generation and Characterization of ALX-0171, a Potent Novel Therapeutic Nanobody for the Treatment of Respiratory Syncytial Virus Infection. *Antimicrob. Agents Chemother.* **60**, 6–13 (2016).

79. Boyoglu-Barnum, S. *et al.* Quadrivalent influenza nanoparticle vaccines induce broad protection. *Nature* **592**, 623–628 (2021).

80. Walls, A. C. *et al.* Elicitation of Potent Neutralizing Antibody Responses by Designed Protein Nanoparticle Vaccines for SARS-CoV-2. *Cell* **183**, 1367-1382.e17 (2020).

81. Salgado, E. N., Lewis, R. A., Mossin, S., Rheingold, A. L. & Tezcan, F. A. Control of Protein Oligomerization Symmetry by Metal Coordination: C_2 and C_3 Symmetrical Assemblies through Cu^{II} and Ni^{II} Coordination. *Inorg. Chem.* **48**, 2726–2728 (2009).

82. Salgado, E. N. *et al.* Metal templated design of protein interfaces. *Proc. Natl. Acad. Sci.* **107**, 1827–1832 (2010).

83. Quijano-Rubio, A., Ulge, U. Y., Walkey, C. D. & Silva, D.-A. The advent of de novo proteins for cancer immunotherapy. *Curr. Opin. Chem. Biol.* **56**, 119–128 (2020).

84. Frank, C. *et al. Efficient and scalable* de novo *protein design using a relaxed sequence space*. http://biorxiv.org/lookup/doi/10.1101/2023.02.24.529906 (2023) doi:10.1101/2023.02.24.529906.

85. Vázquez Torres, S. *et al. De novo design of high-affinity protein binders to bioactive helical peptides*. http://biorxiv.org/lookup/doi/10.1101/2022.12.10.519862 (2022) doi:10.1101/2022.12.10.519862.