Targeting arbitrary regions of intrinsically disordered proteins

Kejia Wu

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

Doctor of Philosophy

University of Washington
2024

Reading Committee:

David Baker, Chair

Neil King

Philip Bradley

Program Authorized to Offer Degree:

Biological Physics, Structure and Design

University of Washington

Abstract

Targeting arbitrary regions of intrinsically disordered proteins

Kejia Wu

Chair of the Supervisory Committee:

David Baker

Biochemistry

A general, robust approach to design proteins that bind tightly and specifically to intrinsically disordered regions (IDRs) of proteins and flexible peptides with minimal cost would have wide applications in biological research, therapeutics, and diagnosis. However, the lack of defined structures and the high variability in sequence and conformational preferences has complicated such efforts. Herein, we have built layers of work to solve this problem with two main computational methods, i.e., bottom-up and top-down. As for the bottom-up approach, we built a landscape to first develop components of geometric matching and amino acid sidechain-oriented pocket recognition for regular peptides; we then generalized it to the broad non-regular landscapes combined with deep learning tools under a rule of optimal binding geometric constraints, achieving specific recognition toward arbitrary unstructured protein sequence space. As for the top-down approach, we leveraged the power of deep learning, trained neural networks to predict and co-fold a disordered target and a designed binding protein to it all together. Using these computational methods, we have designed binders to more than 50 broadly diverse unstructured targets, including highly polar targets. Experimental testing

of dozens to hundreds of designs per target yielded binders with affinities better than 100 nM in most cases, and in the pM range straight out of the computer in five cases. Co-crystal structures of designed binder-target complexes as well as NMR structures with isotope labeled peptide targets were closely consistent with the design models. All-by-all in vitro binding crosstalk experiments for representative designs binding diverse targets show they are highly specific for the intended targets, with no crosstalk even for the closely related peptides. Designs were shown functional in a number of downstream assays indicating the therapeutic, diagnosis, intracellular tracking potentials. These methods were applied in the biologically relevant cancer target RAS to distinguish the four distinct isoforms in cells to the degree antibodies have never achieved. Our approach thus could provide a general solution to the intrinsically disordered protein and peptide recognition problem, while paving a road to site-specific recognition of post-translational modifications (PTMs) and enzymatic functional designs as well.

Acknowledgements

It has been an amazingly interesting journey in science, life, friendship, self-growth, and trying to understand the world and ourselves better. I had as much fun as I could have expected, fought through all the confusions that helped me to be a better person and scientist, and learned almost as much as I ever desired from everything around me. It was full of fun and no regrets. By all means, I hope this adventure continues far beyond. Therefore, I sincerely thank all the people, mentors, colleagues, collaborators, friends, families, strangers, pets, computers and experiments who ever shared a moment with me. May the adventure continue.

Chapter 1 – Introduction

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) within otherwise structured proteins play pivotal roles in various biological processes, including signaling, transcription regulation, and cellular homeostasis. Unlike structured proteins, IDPs lack a fixed three-dimensional conformation, which allows them to engage in highly dynamic interactions with multiple partners. This conformational flexibility makes targeting these regions therapeutically challenging, as traditional small-molecule and biologic approaches often rely on stable binding pockets that IDPs do not typically possess. The significance of this challenge lies in the involvement of IDPs and IDRs in critical cellular functions and their frequent dysregulation in diseases such as cancer, neurodegenerative disorders, and viral pathogenesis. Current methods in targeting these elusive regions are limited to scaffolded or rationally evolved molecules that often fail to achieve specificity and stability in binding. The field of protein engineering has seen considerable advancements, yet the de novo design of protein binders tailored to target IDRs with high precision and functional relevance offers a promising solution. The ability to create novel binders that can specifically interact with defined, albeit flexible, regions of IDPs could revolutionize therapeutic approaches and provide essential tools for probing the biology of disordered proteins. This dissertation explores the development and application of de novo designed protein binders, highlighting their advantages in overcoming the limitations of traditional approaches through rationally computational design, high experimental success rate, and iterative optimization.

1.1 – Dissertation overview

As I began my dissertation work, I set out with a rationale of rationally defining novel rules to bind any arbitrary intrinsically disordered protein sequence with minimal computation and experimental cost. In this work, I combined biophysical principles with deep learning tools to build a platform to target 39 broadly diverse bioactive disordered targets with ~36 designs tested per target of high specificity (Chapter 2). I also developed RFdiffusion based deep learning methods to co-fold a disordered target with a designed binding protein at the same time and used it to target 6 bioactive disordered targets with high selectivity (Chapter 3). I developed parametric matching algorithms to build orthogonal synthetic protein-peptide heterodimers for uses in synthetic biology, as the first design work in binding intrinsically disordered regions (Chapter 4). I supervised the work of combining and applying the above methods to develop the first novel Ras-isoform specific binding proteins in cells for cancer therapeutics and research, achieving better selectivity than any known developed antibodies (Chapter 5). I also developed the ultra-potent de novo trimeric SARS-Cov-2 RBD mini-binders, neutralizing all tested disease variants in mice; and functionalized de novo designed buttressed loop binding proteins in a number of cases (Chapter 6). May the audience find as much fun as I do.

Chapter 2 – Sequence-specific targeting of intrinsically disordered protein regions

2.0 – preface, authors, and abstract

Preface:

Note: this chapter is borrowed directly from the manuscript under revision at the same time, which is the original copy. I, Kejia Wu, am co-first and co-corresponding author for this work.

Authors:

Kejia Wu[1,2,3,‡*], Hanlun Jiang[1,2,4,‡], Derrick R. Hicks[1,2‡], Caixuan Liu[1,2], Edin Muratspahić[1,2], Theresa A. Ramelot[5], Yuexuan Liu[6], Kerrie McNally[7], Amit Gaur[5], Brian Coventry[1,2], Wei Chen[1,2], Asim K. Bera[1,2], Alex Kang[1,2], Stacey Gerben[1,2], Mila Ya-Lan Lamb[1,2], Analisa Murray[1,2], Xinting Li[1,2], Madison A. Kennedy[1,2], Wei Yang[1,2], Gudrun Schober[8,9], Stuart M. Brierley[8,9], Michael H. Gelb[1,6], Gaetano T. Montelione[5], Emmanuel Derivery[7], David Baker[1,2,10*]

1. Department of Biochemistry, University of Washington, Seattle, WA, USA.

2. Institute for Protein Design, University of Washington, Seattle, WA, USA.

3. Biological Physics, Structure and Design Graduate Program, University of Washington, Seattle, WA, USA

4. Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA

5. Department of Chemistry and Chemical Biology, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA

6. Department of Chemistry, University of Washington, Seattle, WA, USA.

7. MRC Laboratory of Molecular Biology, Cambridge, CB2 0QH, UK.

8. Visceral Pain Research Group, Hopwood Centre for Neurobiology, Lifelong Health Theme, South Australian Health and Medical Research Institute (SAHMRI), North Terrace, Adelaide, South Australia 5000, Australia

9. Faculty of Health and Medical Sciences, University of Adelaide, North Terrace, Adelaide, South Australia 5000, Australia

10. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

‡Contributed Equally

*To whom correspondence should be addressed: kejiawu@uw.edu, dabaker@uw.edu

Abstract:

A general approach to design proteins that bind tightly and specifically to intrinsically disordered regions (IDRs) of proteins and flexible peptides would have wide application in biological research, therapeutics, and diagnosis. However, the lack of defined structures and the high variability in sequence and conformational preferences has complicated such efforts. We sought to develop a method combining biophysical principles with deep learning to readily generate binders for any disordered sequence. Instead of assuming a fixed regular structure for the target, general recognition is achieved by threading the query sequence through diverse extended binding modes in hundreds of templates with varying pocket depths and spacings, followed by RFdiffusion refinement to optimize the binder-target fit. We tested the method by designing binders to 39 highly diverse unstructured targets, including polar targets. Experimental testing of ~36 designs per target yielded binders with affinities better than 100 nM in 34 cases, and in the pM range in four cases. The co-crystal structure of a designed binder in complex with dynorphin A is closely consistent with the design model. All by all binding experiments for 20 designs binding diverse targets show they are highly specific for the

intended targets, with no crosstalk even for the closely related dynorphin A and dynorphin B. Our approach thus could provide a general solution to the intrinsically disordered protein and peptide recognition problem.

2.1 – Main

Nature has evolved a variety of mechanisms to bind unstructured regions of peptides and intrinsically disordered proteins (IDPs)[1-6], including antibodies[1-3], the major histocompatibility complexes (MHCs)[4], Armadillo repeat proteins (ArmRPs)[5] and tetratricopeptide repeats (TPRs)[6]. Despite the diversity observed in natural systems, engineering general peptide recognition remains challenging. *De novo* design has succeeded in creating binders for peptides in beta-strand, alpha-helical, and polyproline II conformations[7-9]. However, more general recognition of disordered proteins and peptide regions requires the ability to bind more varied conformations as an arbitrary disordered sequence may not have any propensity for the same secondary structure throughout, or present suitable interfaces for binding in any regular secondary structures. For example, amphipathic helices or strands can be recognized using designs with grooves that bind primarily to the non-polar side of the helix or strand, but if charged residues are distributed around the target this binding mode becomes challenging. A method to achieve general peptide recognition across arbitrary sequence space would greatly enable applications in proteomics, targeting, sensing, and sequencing.

We aimed to develop a general method for designing binders to arbitrary intrinsically disordered sequences. We hypothesized that a family of designed templates with diverse amino acid recognition pockets spaced along an extended binding groove could enable general recognition of sequences with widely varying conformational preferences (Fig. 1A-B), provided that two conditions were satisfied. First, each template structure should “wrap” around

extended peptide conformations with numerous opportunities for the hydrogen bonding interactions required for high specificity. Second, the structural variation in the template family should be sufficient that for any disordered target at least one of the templates can induce it into a defined binding competent conformation. We set out to devise a protocol for designing binders to any arbitrary unstructured sequence in extended conformations with each amino acid residue fitting into a custom pocket (Extended Data Figure 1).

Our approach has four steps. In the first step, “pocket generation,” we construct binding pockets that are specific for single amino acids or dipeptides in extended conformations (Fig. 1C). We require that these pockets have side chains that not only interact with the target side chains, but also make hydrogen bonds with the target backbone to provide structural specificity and compensate for the cost of desolvation. We further require that the pockets are structurally compatible with assembly into larger structures that bind extended peptides (i.e., with N and C termini on opposite sides, and regular secondary structure elements for pocket-to-pocket interactions); to do this, we follow our previous work[7,10] and begin by explicitly designing pockets for recognition of repeating sequences, which yields tandem arrays of pockets which interact with the tandem repeats on the peptide. In the second step, “pocket assembly,” we go beyond the limitations of repeating structures, which are optimal for repeating sequences but not more general sequence targets, by recombining different pockets, using RFdiffusion[13] to generate interfaces between them where necessary to generate overall rigid structures (Fig. 1D). This generates a set of templates with the pockets arranged in different orders and geometries. In the third “threading” step, the target sequence is threaded through the backbone of each template to search for the best binding modes, and in the fourth “refinement” step, the best matches are refined to increase the fit between the designed binder and target peptide (Fig. 1B). Once a template library has been constructed with steps one and two, it can be reused, such that the design of binders for new IDR targets requires only steps three and four, greatly reducing the computational cost.

2.2 – Pocket generation

To construct a library of binding pockets specific for particular amino acids and dipeptides compatible with downstream fusion into single coherent structures (with extended peptide binding pockets), we adopted our previous superhelical matching approach[7] to designing repeat proteins to bind repeating peptides as well as a broader range of peptide conformations . In this approach, repeating conformations of the peptide target sequence were generated, the superhelical parameters were determined, repeat proteins with matching superhelical parameters were generated, and the two were docked together such that each repeat in the peptide makes identical interactions with a matching repeat in the designed protein. To achieve more general peptide recognition, we began by focusing on extended conformations, which are likely more highly populated for most protein sequences than the polyproline II conformation. As every other residue points in roughly the same direction in extended conformations, dipeptide repeats are more natural to target than tripeptide repeats (adjacent protein repeats do not need to have a large twist around the superhelical axis). We sampled the torsion angles of each of the two distinct residues in the dipeptide repeat from the extended region of the Ramachandran map (phi ranges [-150, -70], psi [-30, 150]), generated six repeat (twelve-residue) peptide conformations, computed their superhelical parameters, and generated helical repeat proteins with matching parameters. We then threaded the peptides through the matching repeat proteins, selecting for peptide-protein docks with bidentate hydrogen bonding interactions between protein sidechains and the peptide backbone. The remainder of the repeat protein sequence was then optimized using proteinMPNN[43], and designed complexes with favorable Rosetta binding energies (DDG)[37,38] and AlphaFold2[42] (AF2)

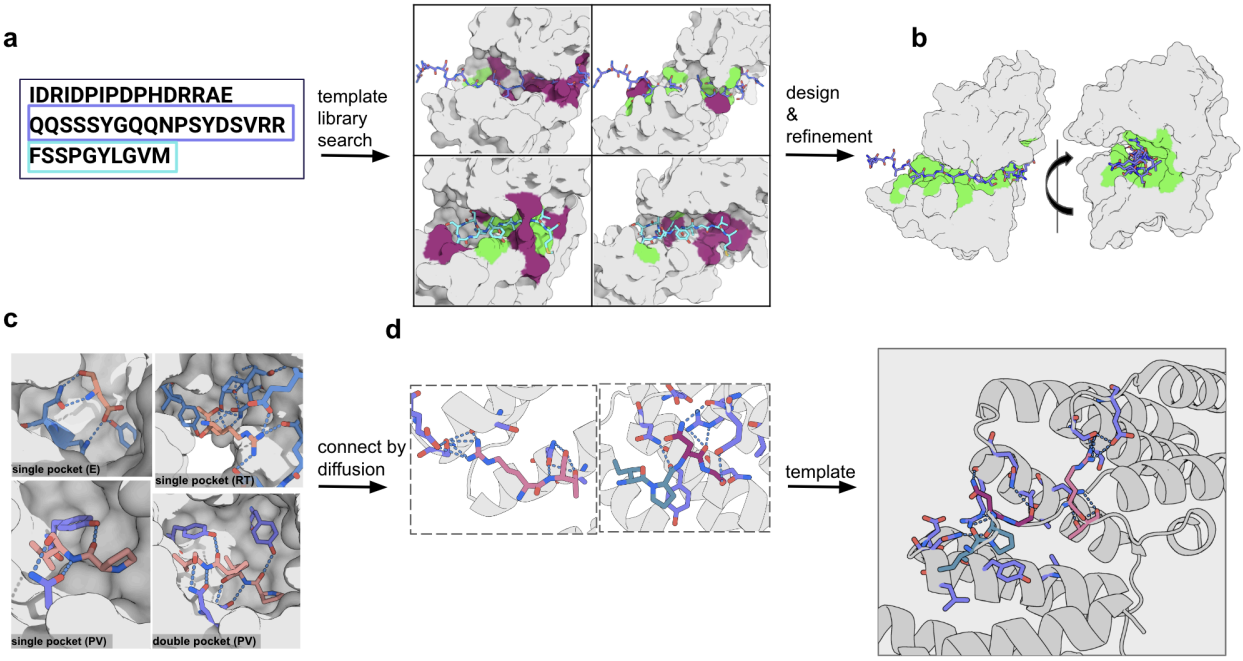
structure predictions matching the models were selected for experimental characterization (see Methods).

We chose dipeptide repeats with sequences LK, RT, YD, PV, and GA (single-letter amino acid codes) as initial targets. Experimental characterization by fluorescence polarization of four-repeat versions of the designs revealed nanomolar binding for the LK and PV repeat peptides, low micromolar binding for YD, but little binding for the more polar RT and highly flexible GA with small or no side chains (Extended Data Figure 3; all the repeat peptides have two flanking repeats for experimental characterization[7]). We hypothesized that these four-repeat proteins might not have sufficient binding energy for polar, flexible targets, and that additional repeats and more perfectly fitting pockets with higher side-chain placement accuracy might be necessary to fully compensate for the high associated entropy loss. To address these issues, we generated a second set of five- and six-repeat designs with more diverse binding pockets by backbone remodeling using parametric, Rosetta, and RFdiffusion approaches (Extended Data Figure 4 and Methods).

Experimental characterization showed that the most effective method for pocket backbone resampling was to use RFdiffusion, keeping the four to nine “critical” amino acids surrounding each sidechain bidentate hydrogen bond from the repeat protein to the peptide backbone fixed (Extended Data Figure 4D; we refer to this as “motif diffusion”). Following motif diffusion, the best designs had picomolar affinities for (LK)x7, and low nanomolar for (RT)x7 and (GA)x7 (Extended Data Figure 5). Keeping the interactions between polar amino acids on the binder and backbone N-H and C=O groups on the target peptide fixed while extensively diversifying hydrophobic interactions between designed binder and target likely succeeded because polar interactions have considerably more geometric requirements than the latter and hence are more efficiently templated than repeatedly sampled from scratch. Observation of tradeoffs between affinity and selectivity (Extended Data Figure 6) led us to more stringently filter on hydrogen bonds between design and target in subsequent designs (see Methods).

To provide coverage of the entire 20 amino acid alphabet, we split the dipeptide pockets into single amino acid binding pockets, yielding a total of twelve derived single amino acid pockets (see Extended Data Table 7).

Fig. 1. Overview of IDR binder design protocol.



a, Intrinsically disordered regions (IDRs) are threaded through a designed template library to identify matches or near matches between the possible dissected amino acid windows (examples shown as purple, cyan) on the peptide and pockets provided on the template. Individual pockets of matches were indicated as green and mismatches as red on the protein surface in the middle panel. **b,** The matches and near matches are used to further design and refine both the overall backbone geometry (gray surface) and individual sidechain pocket fit (green surface matches) between the target and the binder. A front view of an example

“peptide-wrapping” geometry is shown on the right. **c-d,** Template library construction. **c,** Examples of binding pockets for mono-peptide and di-peptide sequences. Clockwise from the top left single pocket for mono- (E), single pocket for di- (RT), single pocket for di- (PV), double pockets for di- (PV), with protein backbone context shown in gray cartoon and sphere. **d,** The binding pockets are connected using RFdiffusion from compatible pockets (each peptide window colored differently) into coherent templates for general sequence recognition, generating the library shown in the middle panel in **a**.

2.3 – Template generation by pocket assembly

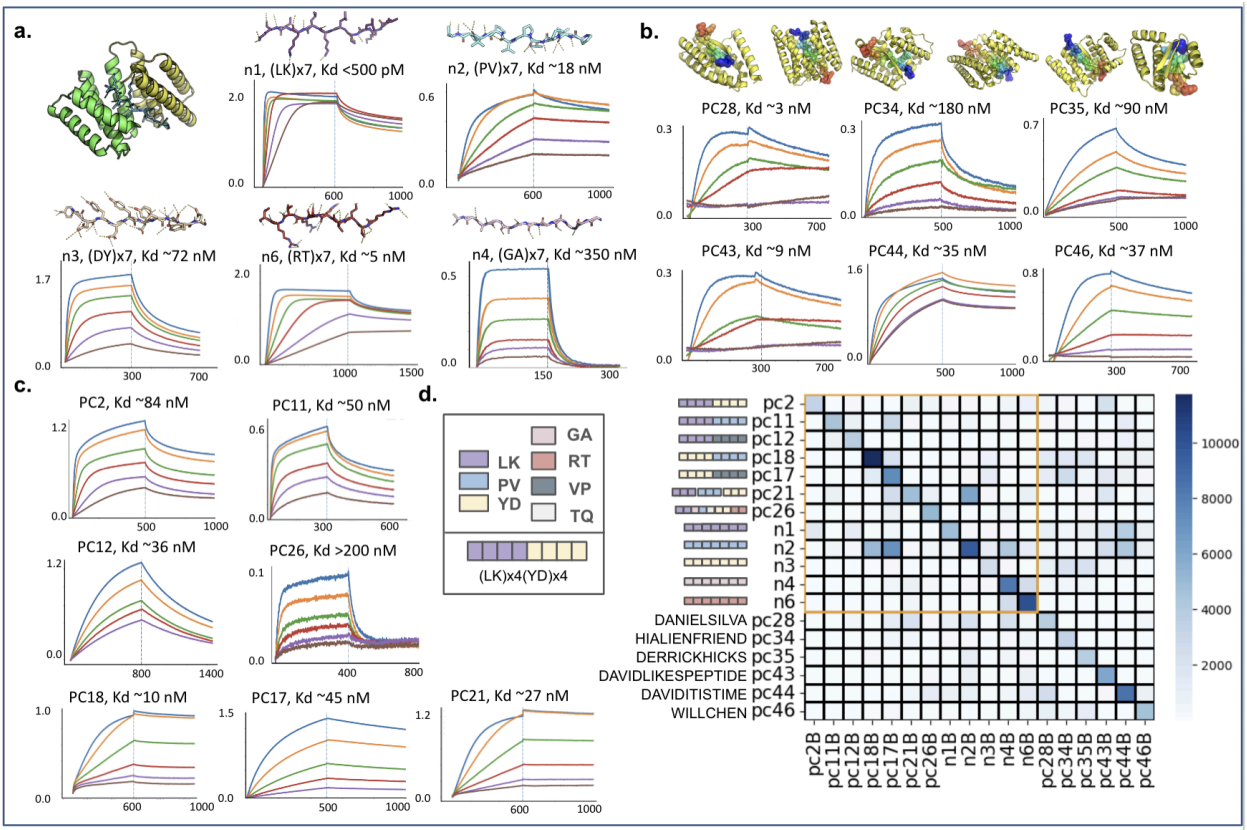
The designs described in the previous section were each custom-built for a specific repeating sequence target, and the binding pockets have different geometries customized to the target being recognized. We reasoned that more general recognition of non-repeating sequences could be achieved by assembling the binding pockets into new backbones, keeping them positioned to interact with peptide targets in continuously extended conformations. We assembled combinations of two to six binding pockets (see below and Methods) generating chimeric-designed protein and chimeric peptide target pairs. We refer to each binding unit (comprising a single amino acid or dipeptide and corresponding designed protein pocket) with a letter; thus AAABBB is a chimera of two designs from the previous section, while ABCDEF combines six different pockets. To do this, we positioned pockets parametrically (see Methods), and connected them with RFdiffusion. We used this approach to generate 70 designs against seven chimeric targets. Experimental characterization using nanoBiT split luciferase reconstitution[15] and biolayer interferometry (BLI) showed double-digit nanomolar binding for six out of seven of the targets, out of only ten designs tested per target on average (Fig. 2C).
We next targeted amino acid sequences spelling out six arbitrarily selected English words and names. Since the new targets differ considerably from the original repeating peptide targets, we used RFdiffusion as described above to optimize both the pockets and the overall spacing after design (see Methods). We tested 45 designs against six targets; the best binders for two out of the six targets had single-digit nanomolar affinities (K_d = 3 nM, 9 nM), three had double-digit affinities (K_d = 35 nM, 37 nM, 90 nM), and one a K_d = 180 nM (Fig. 2B).

We investigated the selectivity of the designs for their peptide targets by carrying out all-by-all (18 by 18) nanoBiT interaction measurements. While there was some cross-talk

between designs and targets with related sequences (for example, designs targeting four PV repeats also bound peptides with eight PV repeats), for the more diverse targets, such as the arbitrary words, the designs were quite specific (Fig. 2D).

To approach the general peptide recognition problem, we used pocket assembly to build 36 chimeric backbones containing pockets recognizing polar residues, and further diversified both binder and peptide target by two-sided sequence design (in the designs described above, the peptide sequence was always held constant). Together, this yielded a library of 340 templates used for the calculations described below. To provide general users with the best starting point for the library search, we subsequently expanded the library to 1,000 members. Each template consists of both a designed binding protein and a corresponding peptide backbone positioned such that the amino acids in the peptide fit into pockets in the design.

Fig. 2. Design of 18 synthetic peptide binding proteins.



a, Dipeptide repeat binder designs. Top left: A representative helical repeat protein scaffold with geometry optimized for binding peptides in an extended conformation, with hydrogen bonds to the peptide backbone and sidechains. Other panels: bound peptide conformation (top), with hydrogen bonds indicated by dashed lines (designed binder is not shown for simplicity, as overall topologies are all very similar), identity of target and binding affinity (middle), and bio-layer interferometry (BLI) for the intended target (bottom). Biotinylated target peptides were loaded onto streptavidin biosensors, and incubated with designed binders in solution to measure association and dissociation. Threefold serial dilutions were tested for each binder starting from 1000 nM (dark blue curve) to 4 nM (purple); the y-axis is the binding signal and the x-axis time (s) throughout the figure. **b,** Design of binders to six polypeptide sequences corresponding to arbitrary English words and names (as shown on the y-axis of panel c). Top row, design models with peptide shown in rainbow spheres from N to C terminus. **c,** Chimeric repeat peptide binder design. In each left panel: the identity of the peptide target and the

measured affinity (top), and the corresponding BLI data (below). Target sequences are color-coded on the y-axis of the heatmap (right) with each square representing one dipeptide motif using the color scheme shown in the legend (middle). Right panel: All by all binding measurements using the nanoBiT assay for 18 designed binder-synthetic peptide pairs, with lgBiT-binders at 100 nM and smBiT-targets at 0.5 nM. Heatmap indicates the average from N=3 experiments. Pairs within the orange square are composed of similar dipeptide repeats and hence have some cross-talk. Between 7 and 36 binder designs were experimentally tested per target.

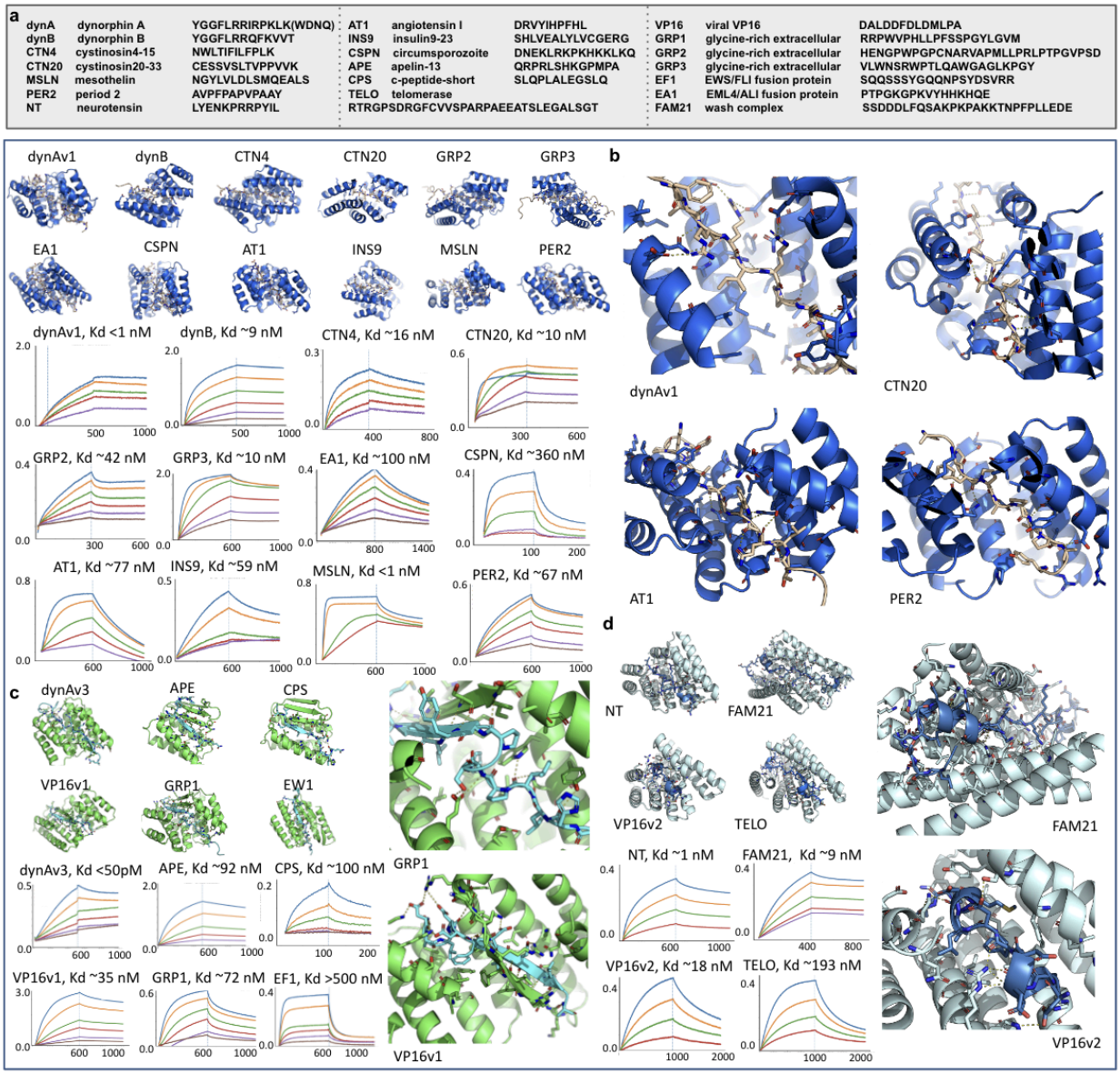
2.4 – Threading intrinsically disordered regions onto template library

For an intrinsically disordered protein (IDP) or region (IDR), there are, in general, a large number of possible peptide subsequences that can be targeted. To identify the most targetable peptide subsequences within an IDR, we map each sequence segment of 4-30 amino acids onto the target backbone with local backbone resampling for each library member, optimize the sequence of the binder using ProteinMPNN, and evaluate the designs based on the fit between designed binder and disordered target and the agreement between the AF2 prediction and the design model. This approach maps target segments with multiple polar residues into templates compatible with extended hydrogen bonding networks, which is likely important for achieving general recognition. In cases where AF2 metrics were suboptimal, we used RFdiffusion (see Methods) to customize the design backbone and certain pockets for the specific target.

We used this approach to obtain binders for 21 diverse therapeutically relevant targets, including eight GPCR ligands, two insulin-related ligands, two tryptic peptides, two disease detection-related disordered regions, four IDRs from cancer-related receptors, and three human scaffolding complexes for which there are no good monoclonal antibodies. For each target, 3-70 designs (on average 40) were generated as described above. These proteins were expressed and purified, and target binding was measured by BLI. The best hits bound to their intended targets with nanomolar to picomolar affinity (Fig. 3A-C). Binders were also obtained against highly polar targets such as the fusion fragment EF1 of EWS/FLI onco-fusion protein for Ewing sarcoma[22-24] (84% polar residues), and the N-terminal fusion fragment CSP-N of Circumsporozoite protein (CSP) for malaria[25] (80% polar residues). Together with the 18 synthetic targets, we obtained binders of 39 of 43 targets attempted, testing on average 36

designs for each target. Twenty targets had greater than 50% polarity. (Extended Data Figure 9 for polarity distribution; Extended Data Table 2 for a summary of steps applied to all 39 targets). To explore the optimization potential of these experimentally characterized hits, we optimized a designed binder with Kd < 1nM for dynorphin A, DYNA_1b1, a kappa opioid receptor (KOR) peptide ligand implicated in chronic pain[26,27] using RFdiffusion as described above for the synthetic targets (see Methods). Forty-five out of 48 designs showed strong binding in the BLI screening assay, six had K_d <= 100 pM by BLI; fluorescence polarization measurements for two of these indicated K_ds <60 pM and <200 pM (Fig. 4B; Extended Data Figure 10). The dynorphin A binder and dynorphin B binder are completely orthogonal and bind only to their intended targets (Fig. 5D), despite having 62% sequence homology.

Fig. 3. Design of binders to 21 native protein disordered regions and peptides.



a, Identities and amino-acid sequences of target native bioactive peptides and IDRs. **b,** Binders against IDRs in random coil conformations. **c,** Binders against IDR conformations containing some beta strands. **d,** Binders against IDRs in partial helical conformations. DynA (dynorphin A) and VP16 (viral protein VP16) were each targeted in two different conformations; tight binding was obtained in both cases. In each group, zoom-in cartoons are shown for representative examples. All the binding measurements were made with BLI against the corresponding peptide targets; threefold serial dilutions starting from 1000 nM except for AT1 (333 nM), INS9 (333 nM);

CTN4 (111 nM), CTN20 (111 nM), MSLN (111 nM), dynAv3 (111 nM), NT (111 nM), FAM21 (111 nM), VP16v2 (111 nM). Between 3 and 70 binder designs were experimentally tested per target.

2.5 – Structural validation

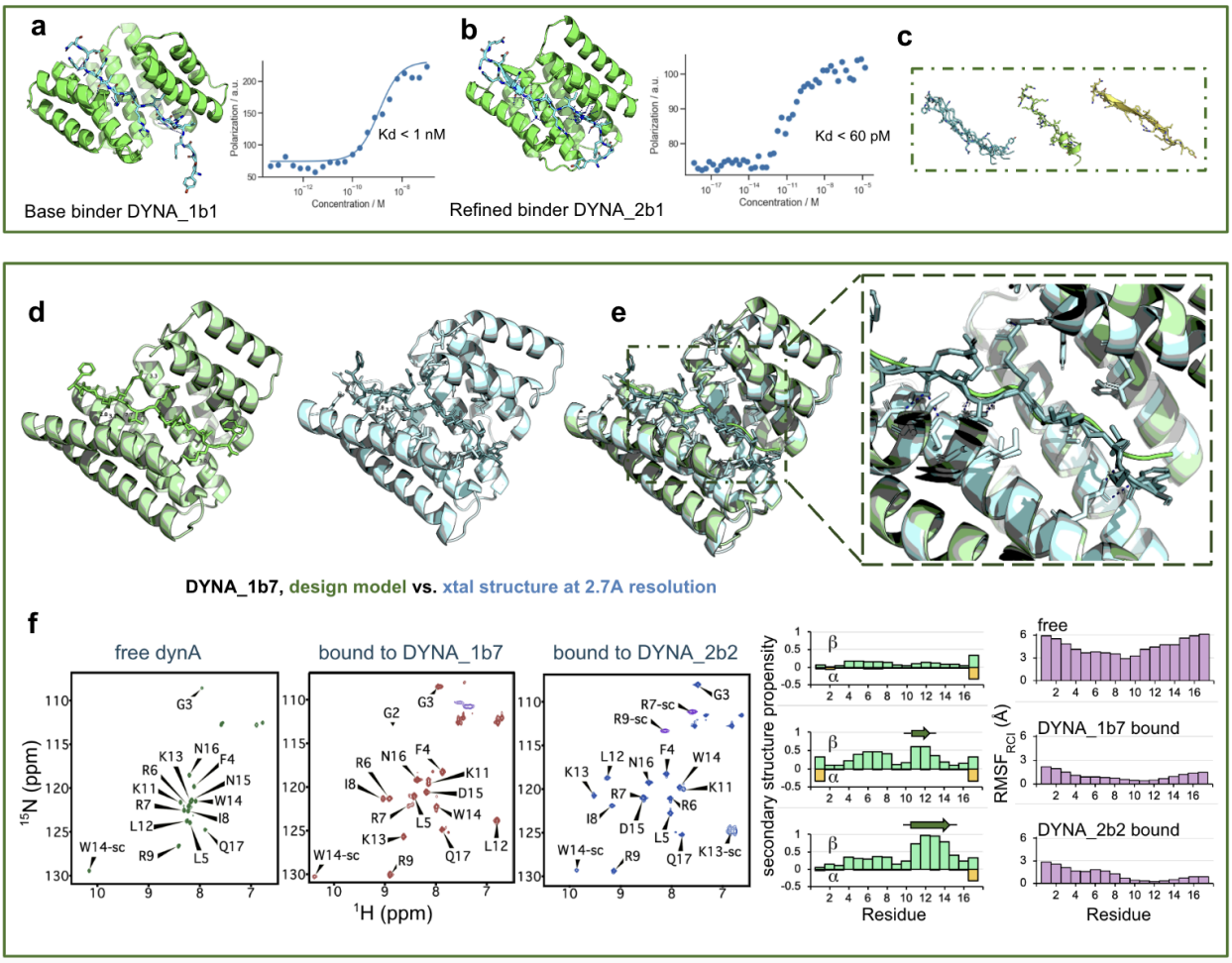
We succeeded in solving a co-crystal structure of a 7 nM K_d dynorphin A binding design, DYNA_1b7, in complex with dynorphin A (residues 1-17) at 2.7 Å resolution. The backbones of both the protein and peptide in the crystal structure match the design model well, with an interface backbone RMSD of 1.2 Å for the complex and interface sidechain RMSD of 2.9 Å (Fig. 4D-E). The key interactions are in the central region of the peptide: during design, we excluded the N-terminal YGGF sequence which is shared between dynorphin A and B and other neuropeptides[32], aiming to distinguish closely related peptides in the family where antibodies often fail, and the C-terminal region (-WDNQ) which is missing in some species (Fig. 4E). This design underwent two-side partial diffusion during optimization, which resulted in retention of two pairs of the original asparagine-peptide backbone bidentate interactions (Extended Data Figure 11). The crystal structure confirmed that all the designed hydrogen bonds to the peptide backbone (ASN19, ASN69, ASN70 on binder) were present as designed; the corresponding peptide region (from LEU5 to ARG9) also aligned precisely to the design with Cα RMSD=0.6 Å. There were minor shifts of side chains in hydrophobic grooves observed, and density was missing for the excluded termini (YGG- and -DNQ).

To investigate changes in dynorphin structure upon binding, we also examined the NMR spectra of isotope-labeled dynorphin A unbound in solution, bound to the first round design solved by X-ray crystallography, DYNA_1b7 (K_d = 7 nM), and bound to the refined round two binder, DYNA_2b2 (K_d <200 pM; Extended Data Figure 10C, a variant of binder DYNA_2b1) (Fig. 4F). NMR confirmed that free dynorphin A is intrinsically disordered and becomes ordered upon binding except for the intentionally excluded regions (Fig. 4F). For both DYNA peptides the NMR data clearly support an extended bound-state conformation, as expected from the design models (Fig. 4F and Extended Data Figure 12A). Consistent with the design model of the more extended binding site with hydrogen bonding networks on both sides, the extent of

ordering upon binding to the optimized DYNA_2b2 binder was increased compared to the original binder DYNA_1b7 in both the C-terminal region and around the W14-ASN137 bidentate interaction (as indicated by the $RMSF_{rci}$ and secondary structure propensity data of Fig. 4F and Extended Data Figure 12A).

The extended conformation of the dynorphin peptide in the designed complexes, confirmed by the X-ray and NMR data, is considerably different from previously solved cryoEM or NMR structures of dynorphin with native KOR (Extended Data Figure 12), where it binds in a compact, partial helix conformation (PDB ID 2n2f)[47]. These data highlight the power of computational design for targeting disordered proteins and peptides in a wide variety of conformations, including non-native extended conformations.

Fig. 4. Structural characterization of dynorphin A binder designs.



a, Design model of first-round design DYNA_1b1 bound to dynorphin A in an extended backbone conformation, with five pairs of peptide backbone-protein sidechain bidentate hydrogen bonds (left). Fluorescence polarization (FP) binding with 1nM TAMRA labeled peptide indicates a K_d <1nM. **b,** Diffusion refined binder DYNA_2b1 with estimated K_d < 60 pM by fluorescence polarization with 100 pM peptide (K_d cannot be accurately measured below the concentration of peptide used in the FP assay). **c,** During diffusion-based refinement, the target peptide backbone conformation is sampled as well as that of the binder. **d,** Computational design model (green, left) and the 2.7 Å co-crystal structure (cyan, right) of dynorphin A bound with design DYNA_1b7. **e,** Superposition of the design model and the crystal structure. The inset on the right shows zoom-in on the central portion of the dynorphin A sequence

LRRIRPKLKW. **f,** Assigned NMR ^{1}H-^{15}N HSQC spectra (left) of $^{15}N^{13}C$-labeled dynorphin A unbound (free), bound to unlabeled DYNA_1b7, and bound to unlabeled DYNA_2b2 (a variant of DYNA_2b1) in solution, and (right) secondary structure propensity and root-mean-squared fluctuation of C positions ($RMSF_{RCl}$) based on backbone chemical shift data. Sidechain amide resonance peaks in these HSQC spectra are not labeled.

2.6 – Applications of Designed Binders

The WASH complex is a pentameric complex consisting of WASH (WASHC1), FAM21 (WASHC2), CCDC53 (WASHC3), SWIP (WASHC4), and Strumpellin (WASHC5) and is responsible for the nucleation of branched actin on endosomes[28,29]. FAM21 contains a C-terminal disordered region of ~1,000 residues in length involved in multiple protein:protein interactions (Fig. 5A). We designed two binders of a 27-amino acid disordered region in FAM21 (Fig. 3A, D). Immunoprecipitation studies with these binders retrieved the entire WASH complex from cell lysate (Fig. 5A). Designed binders for less well characterized complexes involving disordered proteins could considerably enhance our understanding of the roles played by this important class of proteins.

To date, no antibodies, peptides, or small molecules have been developed to inhibit dynorphin A; existing ligands instead modulate KOR signaling by engaging the deep binding pocket of the receptor[27,30]. To explore the potential of our binders to block KOR signaling mediated by dynorphin A (Extended Data Figure 13), we performed an *in vitro* cAMP assay using mammalian cells stably expressing the human KOR. The binder DYNA_2b2 inhibited dynorphin A-dependent KOR signaling with an IC_{50} of 50 nM (Fig. 5B). As noted above, to increase specificity, during design we excluded the N-terminal YGGF sequence during design to distinguish between dynorphin A and B; this region is critical for opioid receptor activation[16] (Fig. 5B) and extension to include the YGGF- motif would likely increase potency for potential therapeutic use.

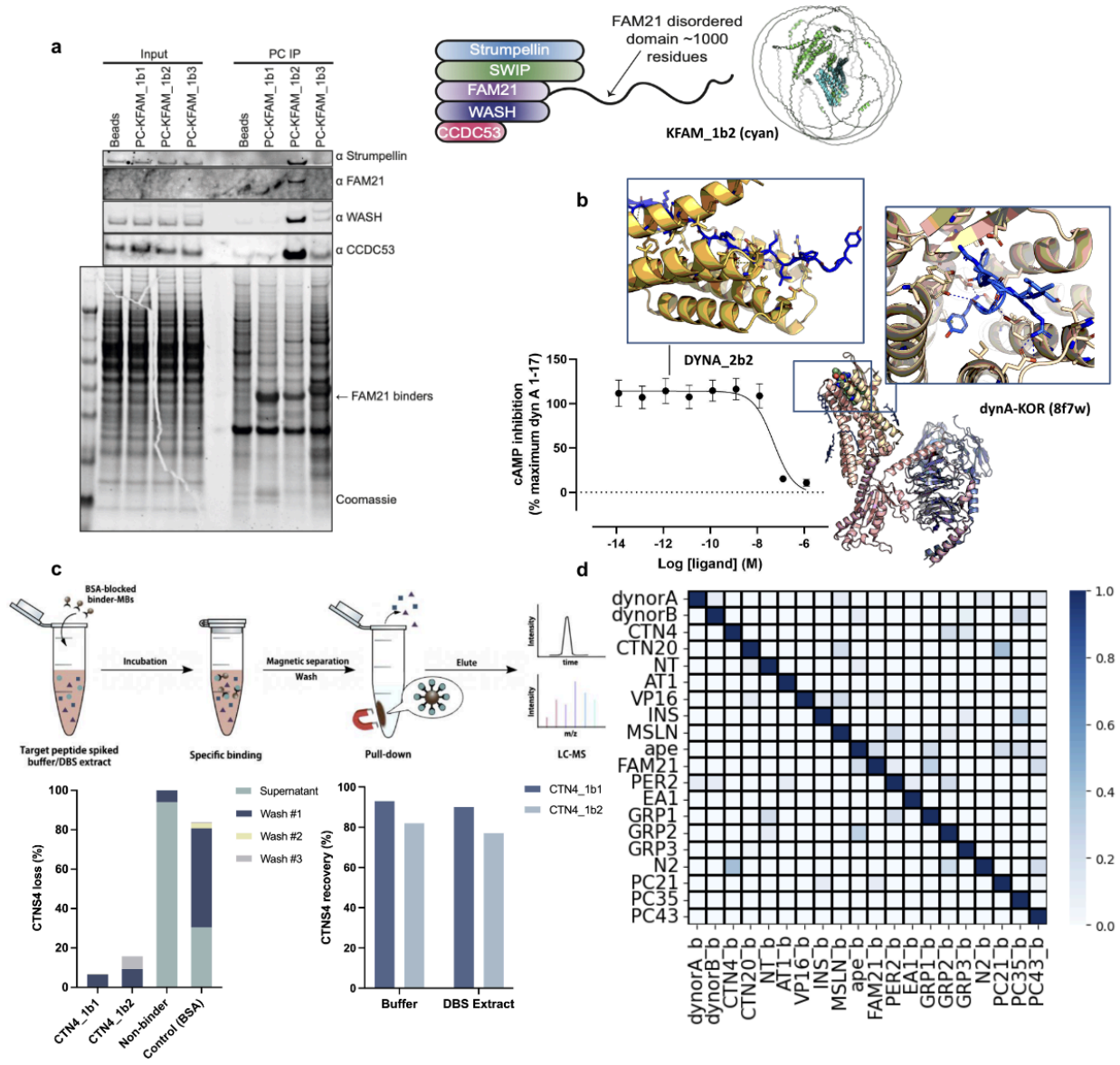
We also explored the use of the designed binders in affinity enrichment coupled with LC-MS for detecting low-abundance proteotypic peptides (Fig. 5C). Binders CTN4_1b1 and CTN4_1b2 targeting CTN4, a 12-amino acid tryptic peptide of the lysosomal cystine transporter cystinosin protein were coupled to magnetic beads, and incubated with buffer and blood samples to which CTN4 had been added. LC-MS showed that CTN4 was captured by the

binder-conjugated magnetic beads (MBs) but not the control unconjugated or BSA-conjugated beads. CTN4_1b1 enriched and recovered 90% of the CTN4 from both buffer and blood samples, much higher than achieved with previously described helical peptide binders[8].

2.6 – Binder orthogonality

We next investigated the specificity of the designed binders for 16 native targets and four representative synthetic targets with the most unique target amino acid sequences. The affinities of these designs for their targets are all tighter than 100 nM. We measured the binding affinity for each design against all 20 targets using BLI. We observed remarkably little crossreactivity at concentrations up to 1 uM. Hence, even though all the targets are disordered, each design only binds well to the target it was designed to bind (Fig. 5D; Extended Data Figure 14).

Fig. 5. Designed binders are functional and orthogonal.



A, FAM21, a subunit of the pentameric WASH complex, contains a large disordered C terminal domain, to which a binder was designed (shown in cartoon: green, FAM21; cyan, designed binder). This binder is sufficient to immunoprecipitate the WASH complex from HEK293 cells. **b,** antagonism of dynorphin A-stimulated KOR signaling by DYNA_2b2 binder. The inhibition was measured in a cAMP assay in CHO cells. Data are shown as mean ± SEM (n=4). The IC_{50} of DYNA_2b2 binder was 50.3 ± 0.7 nM. The structural mechanism of dynorphin A activating KOR is shown in cryoEM complex structure (PDB ID 8f7w); zoom-in interaction shows how the interaction from YGGF- motif differs in our design model (left) and native cryoEM structure

(right). **c,** BSA-blocked binder-conjugated magnetic beads (MBs) were used to capture spiked target peptides. The amount of peptide recovered from elution was quantified by LC-MS. (Left) Followed by affinity pull-down with binder-MBs, the supernatant and three subsequent wash fractions were collected, and the amount of unbound peptide from each step was measured by LC-MS. Percentages were normalized to the peak area of peptide standards. Non-binder MBs and BSA-blocked unfunctionalized MBs were used as negative controls. (Right) The percentages of peptide recovery were measured by LC-MS and normalized to the peak area of peptide standards; **d,** 20x20 orthogonality binding matrix determined using BLI. Biotinylated target peptide was loaded onto streptavidin biosensors, and incubated with the designed cognate binder as well as the other 19 non-cognate binders at concentration of 1000 nM in solution. The heat map shows the maximum response signal for each binder–target pair normalized by the maximum response signal of the cognate designed binder–target pair.

2.7 – Discussion

We described a general approach for designing proteins that bind to IDRs with high affinity and specificity. We use the approach to design binders for 39 out of 43 broadly diverse unstructured sequences that were targeted, with widely varying polarity (Extended Data Figure 15). These binders were obtained with only, on average, 36 designs tested per target, with the majority of affinities better than 100 nM. For entirely polar targets, testing larger numbers of designs would likely increase success rates and improve binding affinities. Compared to previously designed proline-rich tri-peptide repeat binders[7], the hydrogen bonding density is increased by ~210% and contact molecular surface from 360 Å to 620 Å (Extended Data Figure 16), likely accounting for the improved affinity, specificity, and success rate despite the greater target polarity and flexibility. The diversity in binding modes makes it possible to target essentially any unstructured sequence. As highlighted in the case of dynorphin A, our designs induce the target disordered region into the bound structure; such 'induced fit' is a general feature of disordered protein binding interactions in nature[11,35,36]. Diffusion methods can generate binders to targets that can adopt regular secondary structures (helix[8], strand, and combinations of helix and strand) as the PDB training set is rich in examples of interacting helices and beta strands; our approach covers highly disordered regions with weak secondary structure propensity and where interactions with nearly every residue are important as in the case of polar and charged targets.

Our designed binders should be broadly useful in therapeutic, diagnosis, synthetic biology, and cell biology, especially in cases lacking good monoclonal antibodies. Cancer and disease-related cell receptors such as Mesothelin (MSLN)[44], onco-fusion proteins of childhood cancer such as EWS/FLI for Ewing sarcoma and EML4-ALK for lung cancer[45], and

Circumsporozoite protein (CSP) for malaria[25] can potentially be targeted and localized through their unique unstructured regions for delivery or degradation. Many transcription factors, epigenetic regulation, and viral-host protein-protein interactions involve intrinsically disordered regions that could be modulated by IDR-binding proteins[46]. Soluble proteins and peptides from neuropeptides in the brain to poorly studied noncanonical open reading frames (ORFs)[20,21] which are often intrinsically disordered, such as GREP1[21] implicated in breast cancer, could become accessible for imaging and sensing. The synthetic heterodimer pairs could be useful in synthetic biology, for example construction of new transcriptional regulation networks. The potential of our designs for diagnostics is illustrated by the detection of tryptic peptide CTN4 of cystinosin, a newborn disease-related membrane protein variant. More generally, the ability to design binding proteins for short peptide sequences could open the door to next-generation low-cost proteomics platforms based on arrays of specifically designed short peptide binding proteins.

There are a number of exciting directions for extending our design approach. First, it should be possible to construct sites appropriate for post-translational modifications such as phosphorylation to enable specific recognition of modified peptides. Second, as catalytic site design methods improve, it should be possible to incorporate proteolytic or covalent modification sites into the designs; the extended conformation of the peptide bond and the pocket-by-pocket sidechain recognition make them both very accessible as substrates. The binding pockets and conformations of peptides in most natural proteases resemble that of our designs, but completely redesigning natural enzyme specificity has proven challenging—instead, designing binders to the target of interest as described here, and then incorporating catalytic sites could provide a more generally customizable approach.

References

1. Chiu, M. L., Goulet, D. R., Teplyakov, A. & Gilliland, G. L. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies* **8**, 55 (2019).

2. Lee, J. H., Yin, R., Ofek, G. & Pierce, B. G. Structural Features of Antibody-Peptide Recognition. *Front. Immunol.* **13**, (2022).

3. Nelson, A. L., Dhimolea, E. & Reichert, J. M. Development trends for human monoclonal antibody therapeutics. *Nat. Rev. Drug Discov.* **9**, 767–774 (2010).

4. Perez, M. A. S., Cuendet, M. A., Röhrig, U. F., Michielin, O. & Zoete, V. Structural Prediction of Peptide–MHC Binding Modes. in *Computational Peptide Science: Methods and Protocols* (ed. Simonson, T.) 245–282 (Springer US, New York, NY, 2022). doi:10.1007/978-1-0716-1855-4_13.

5. Ernst, P. & Plückthun, A. Advances in the design and engineering of peptide-binding repeat proteins. *Biol. Chem.* **398**, 23–29 (2017).

6. Zeytuni, N. & Zarivach, R. Structural and Functional Discussion of the Tetra-Trico-Peptide Repeat, a Protein Interaction Module. *Structure* **20**, 397–405 (2012).

7. Wu, K. *et al.* De novo design of modular peptide-binding proteins by superhelical matching. *Nature* **616**, 581–589 (2023).

8. Vázquez Torres, S. *et al.* De novo design of high-affinity binders of bioactive helical peptides. *Nature* **626**, 435–442 (2024).

9. Sahtoe, D. D. *et al.* Design of amyloidogenic peptide traps. *Nat. Chem. Biol.* 1–10 (2024) doi:10.1038/s41589-024-01578-5.

10. Jiang, H. *et al.* De novo design of buttressed loops for sculpting protein functions. *Nat. Chem. Biol.* 1–7 (2024) doi:10.1038/s41589-024-01632-2.

11. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).

12. Liu, J., Tan, H. & Rost, B. Loopy Proteins Appear Conserved in Evolution. *J. Mol. Biol.* **322**, 53–64 (2002).

13. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

14. Hansen, S. *et al.* Structure and Energetic Contributions of a Designed Modular Peptide-Binding Protein with Picomolar Affinity. *J. Am. Chem. Soc.* **138**, 3526–3532 (2016).

15. Rozbeh, R. & Forchhammer, K. Split NanoLuc technology allows quantitation of interactions between PII protein and its receptors with unprecedented sensitivity and reveals transient interactions. *Sci. Rep.* **11**, 12535 (2021).

16. Ernst, P. *et al.* Structure-Guided Design of a Peptide Lock for Modular Peptide Binders. *ACS Chem. Biol.* **15**, 457–468 (2020).

17. Liu, J. K. H. The history of monoclonal antibody development – Progress, remaining challenges and future innovations. *Ann. Med. Surg.* **3**, 113–116 (2014).

18. Lu, R.-M. *et al.* Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* **27**, 1 (2020).

19. Pantaleo, G., Correia, B., Fenwick, C., Joo, V. S. & Perez, L. Antibodies to combat viral infections: development strategies and progress. *Nat. Rev. Drug Discov.* **21**, 676–696 (2022).

20. Erady, C. *et al.* Novel open reading frames in human accelerated regions and transposable elements reveal new leads to understand schizophrenia and bipolar disorder. *Mol. Psychiatry* **27**, 1455–1468 (2022).

21. Varabyou, A., Erdogdu, B., Salzberg, S. L. & Pertea, M. Investigating open reading frames in known and novel transcripts using ORFanage. *Nat. Comput. Sci.* **3**, 700–708 (2023).

22. Li, X., McGee-Lawrence, M. E., Decker, M. & Westendorf, J. J. The Ewing's Sarcoma Fusion Protein, EWS-FLI, Binds Runx2 and Blocks Osteoblast Differentiation. *J. Cell. Biochem.* **111**, 10.1002/jcb.22782 (2010).

23. Selvanathan, S. P. *et al.* Oncogenic fusion protein EWS-FLI1 is a network hub that regulates alternative splicing. *Proc. Natl. Acad. Sci.* **112**, E1307–E1316 (2015).

24. One oncogene, several vulnerabilities: EWS/FLI targeted therapies for Ewing sarcoma - ClinicalKey. https://www.clinicalkey.com/#!/content/playContent/1-s2.0-S2212137421000592?returnurl=https:%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2212137421000592%3Fshowall%3Dtrue&referrer=.

25. Tan, J. *et al.* Functional human IgA targets a conserved site on malaria sporozoites. *Sci. Transl. Med.* **13**, eabg2344 (2021).

26. Hang, A., Wang, Y., He, L. & Liu, J. The role of the dynorphin/κ opioid receptor system in anxiety. *Acta Pharmacol. Sin.* **36**, 783–790 (2015).

27. Podvin, S., Yaksh, T. & Hook, V. The Emerging Role of Spinal Dynorphin in Chronic Pain: A Therapeutic Perspective. *Annu. Rev. Pharmacol. Toxicol.* **56**, 511–533 (2016).

28. Derivery, E. *et al.* The Arp2/3 Activator WASH Controls the Fission of Endosomes through a Large Multiprotein Complex. *Dev. Cell* **17**, 712–723 (2009).

29. Gomez, T. S. & Billadeau, D. D. A FAM21-Containing WASH Complex Regulates Retromer-Dependent Sorting. *Dev. Cell* **17**, 699–711 (2009).

30. O'Connor, C. *et al.* NMR structure and dynamics of the agonist dynorphin peptide bound to the human kappa opioid receptor. *Proc. Natl. Acad. Sci.* **112**, 11852–11857 (2015).

31. Wang, Y. *et al.* Structures of the entire human opioid receptor family. *Cell* **186**, 413-427.e17 (2023).

32. Margolis, E. B., Moulton, M. G., Lambeth, P. S. & O'Meara, M. J. The life and times of endogenous opioid peptides: Updated understanding of synthesis, spatiotemporal dynamics, and the clinical impact in alcohol use disorder. *Neuropharmacology* **225**, 109376 (2023).

33. Plückthun, A. Designed Ankyrin Repeat Proteins (DARPins): Binding Proteins for Research, Diagnostics, and Therapy. *Annu. Rev. Pharmacol. Toxicol.* **55**, 489–511 (2015).

34. Stumpp, M. T., Binz, H. K. & Amstutz, P. DARPins: A new generation of protein therapeutics. *Drug Discov. Today* **13**, 695–701 (2008).

35. Fujiwara, K., Toda, H. & Ikeguchi, M. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC Struct. Biol.* **12**, 18 (2012).

36. Sipeki, S. *et al.* Novel Roles of SH2 and SH3 Domains in Lipid Binding. *Cells* **10**, 1191 (2021).

37. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

38. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).

39. Ferré, G., Czaplicki, G., Demange, P. & Milon, A. Chapter Two - Structure and dynamics of dynorphin peptide and its receptor. in *Vitamins and Hormones* (ed. Litwack, G.) vol. 111 17–47 (Academic Press, 2019).

40. Macias, M. J., Wiesner, S. & Sudol, M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Lett.* **513**, 30–37 (2002).

41. Koch, C. A., Anderson, D., Moran, M. F., Ellis, C. & Pawson, T. SH2 and SH3 Domains: Elements that Control Interactions of Cytoplasmic Signaling Proteins. *Science* **252**, 668–674 (1991).

42. Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).

43. J. Dauparas *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022).

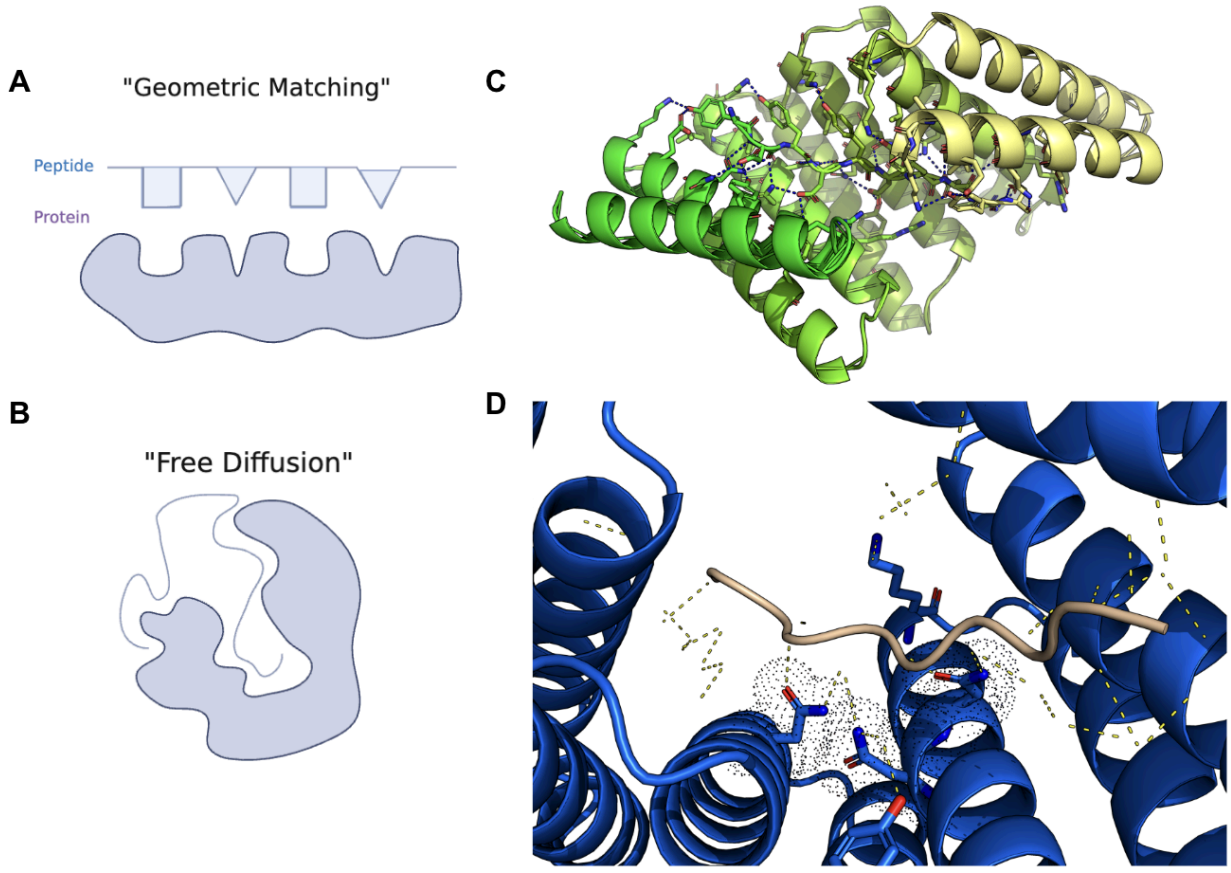
44. Lv, J., Li, P. Mesothelin as a biomarker for targeted therapy. *Biomark Res* 7, 18 (2019).

45. Sasaki T, Rodig SJ, Chirieac LR, Jänne PA. The biology and treatment of EML4-ALK non-small cell lung cancer. Eur J Cancer. 2010 Jul;46(10):1773-80.

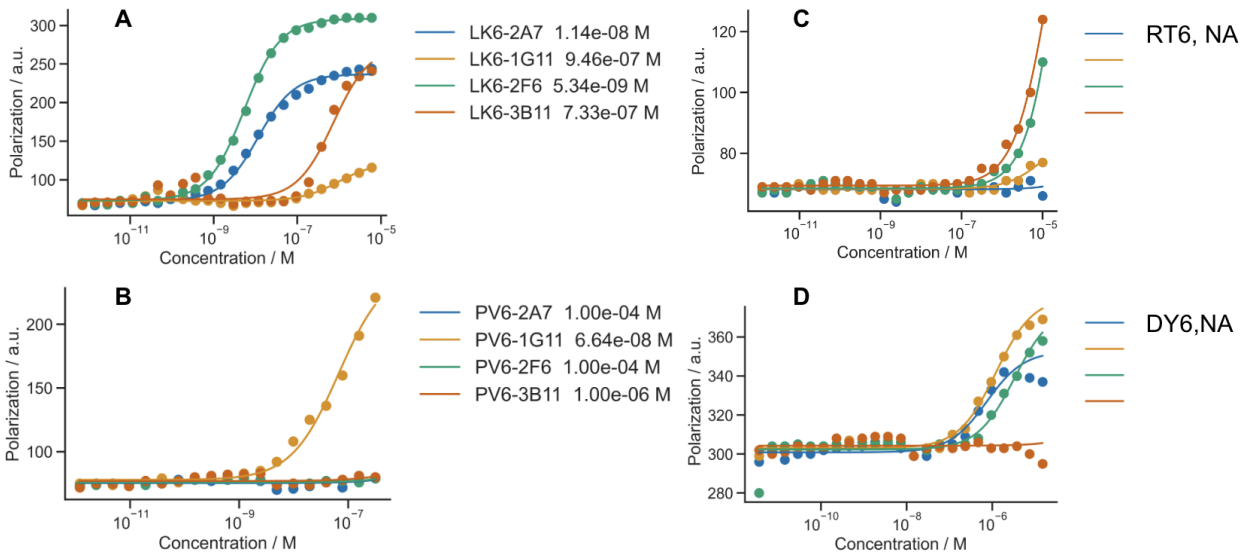
46. Aiyer S. *et al.* A common binding motif in the ET domain of BRD 3 forms polymorphic

structural interfaces with host and viral proteins, *Structure*, **29-8**, 886-898 (2021).

47. O'Connor, C.; White, K.L.; Doncescu, N.; Didenko, T.; Roth, B.L.; Czaplicki, G.; Stevens, R.C.; Wüthrich, K.; Milon, A. NMR structure and dynamics of the agonist dynorphin peptide bound to the human kappa opioid receptor. Proc. Natl. Acad. Sci. USA 2015, 112, 11852–11857.



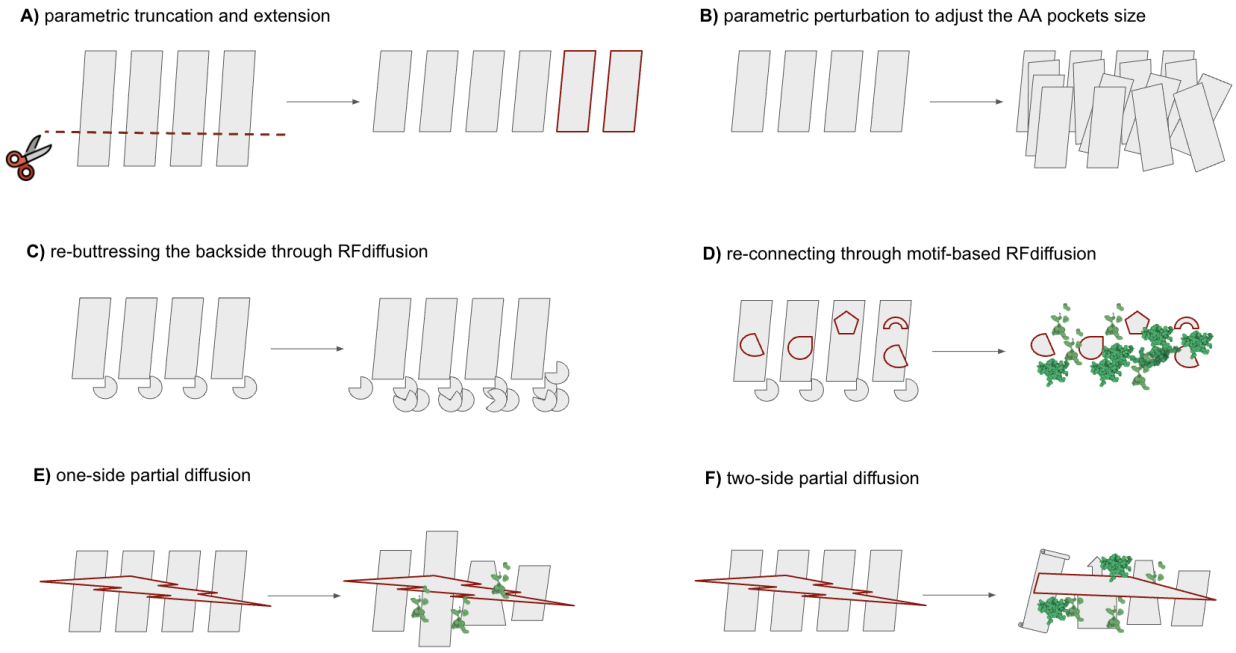
Extended Data Fig. 1: Template repeat di-peptide binding protein platform. A) Previous geometric matching approaches required a definite number of pre-built pockets as well as one-to-one parametric matching. B) Free diffusion approach folds relatively large disordered targets (30-40 amino acids) into conformational combinations of alpha-helical or beta-strands, or bind shorter targets in the alpha-helical or beta-strand conformation; C) In the new protein family, a single five-repeat DHR structurally fully surrounds 10 AA peptide target, with adjacent residues on the peptide pointing to the opposite directions to make residue-to-residue interaction (the first residue in each repeat points in one direction, and the second in the other direction to the protein); D) The pre-inserted protein side chain—peptide backbone bidentate hydrogen bonding donors (shown in stick and dot) lock the peptide in more general extended conformations, meanwhile the geometric constraints enable massive protein side chain - peptide side chain hydrogen bonding networks (shown in yellow dotted lines) for affinity and specificity.



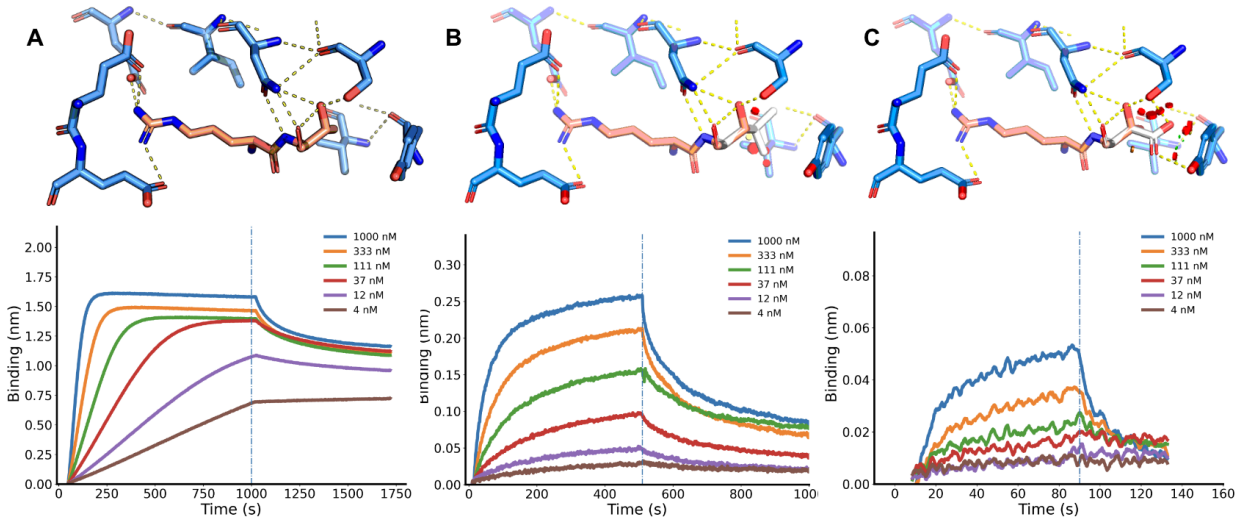
Extended Data Fig. 2: Initial binding characterizations of the round one di-peptide repeat binders by fluorescence polarization (FP). FP characterization of the first-round four-repeat designed binders titrating against TAMRA-labeled six-repeat di-peptides: **A)** (LK)x6, **B)** (PV)x6, **C)** (RT)x6, **D)** (DY)x6. TAMRA-peptides were maintained at 1 nM concentration, while designed binders started from 40 uM (DY), 10 uM (LK, RT), and 1 uM (PV) with two-folded titration.

Number of Targets	Optimization	Initial computation		Computation-experiment cycle	
		Base	Diffusion refinement	one-time	two-time
25	×	✓		✓	
2	×	✓			✓
11	×		✓	✓	
1	✓	✓	✓		✓

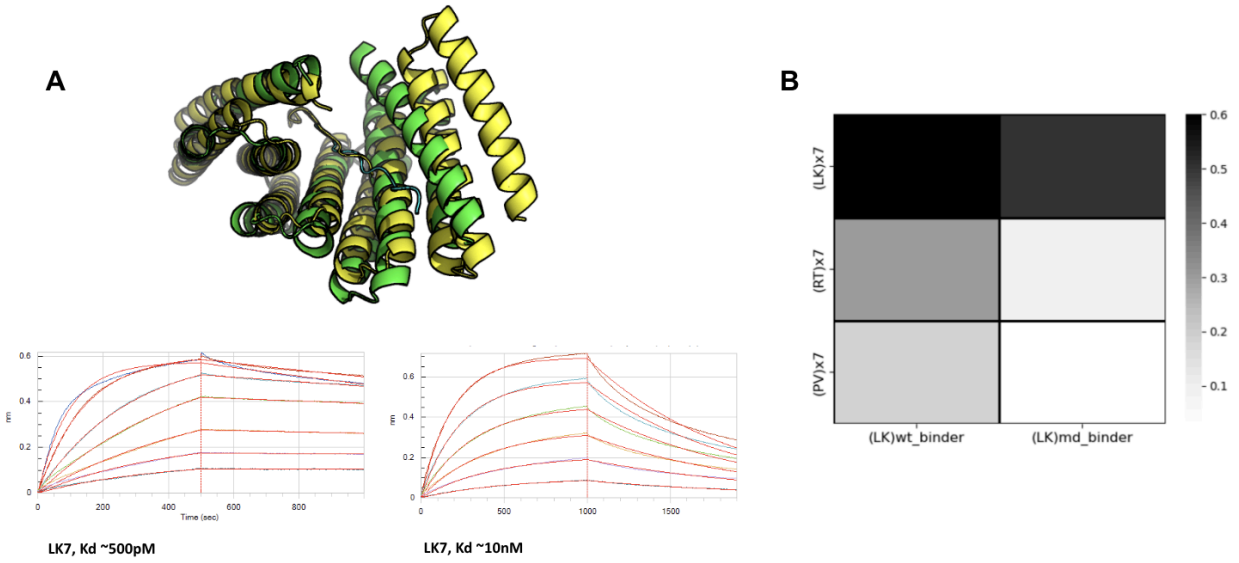
Extended Data Table. 3: Summary of computational and experimental protocols. To summarize, binders to 25/39 targets have not been experimentally or computationally optimized, and produced straight out of the computer through the base design pipeline. Binders to 2/39 targets have not been experimentally or computationally optimized, and produced straight out of the computer through the base design pipeline, but were experimentally tested in two batches to slightly increase the tested number. Binders to 11/39 targets have not been experimentally or computationally optimized, and produced straight out of the computer through the base design pipeline followed by the diffusion refinement. Binders to 1/39 target (dynorphin A) went through experiment-guided computationally optimized. In the first batch, they were produced straight out of the computer through either the base design pipeline or diffusion-refined pipeline; in the second batch, optimized designs were produced by diffusion refinement on the best experimental hits from batch one.



Extended Data Fig. 4: Template backbone optimization approaches. Multiple approaches were applied to generate the library templates during pocket design. **A)** Original four-repeat proteins with long helices were truncated by helical length and extended by repeating geometry parametrically to generate five- and six-repeat proteins with short helices for a suitable peptide binding interface. **B)** Parametric perturbations between the repeat-to-repeat transition were applied both symmetrically and asymmetrically to adjust the pocket size. **C)** Each of the original repeat proteins was made of two-helix bundles. To re-buttress and re-pack the core of the protein, all the helices interacting with the peptide were fixed, while the back helices were masked, denoise, and regenerated with RFdiffusion with the context of the target peptide. **D)** Similar to C), per repeat, only the four to nine interacting residues (around the bidentate hydrogen bond donor) to the peptide were fixed. The rest of the protein was masked, denoised, and regenerated with RFdiffusion in the context of the target peptide. **E)** One-side (chain A, which is the designed protein) partial diffusion was conducted as published with PT ranging from [10, 12, 15, 18]. **F)** Two-side (chain A and chain B, which are the designed protein and the peptide target) partial diffusion was conducted with PT ranging from [12, 15, 18, 22].



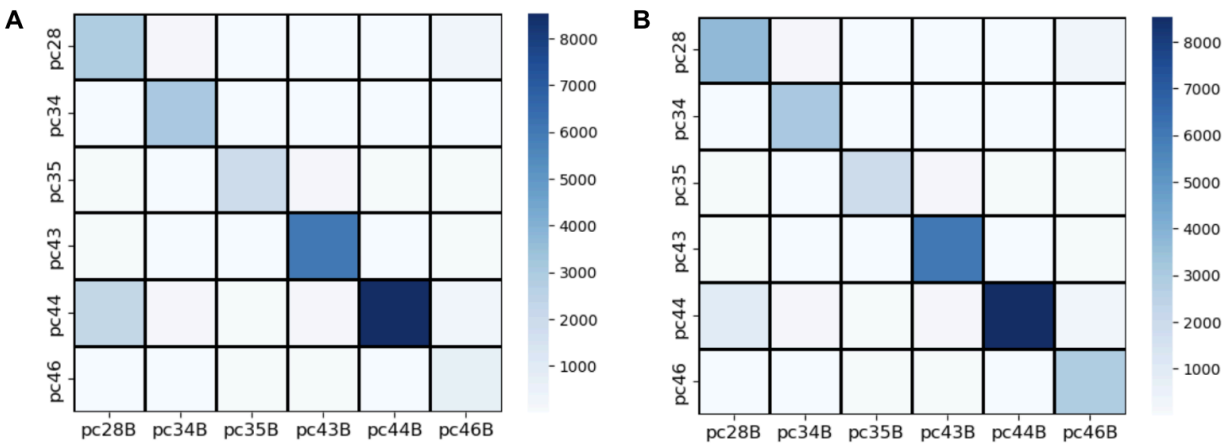
Extended Data Fig. 5: Repeat di-peptide (RT) binder showed selectivity between threonine, aspartate, and leucine by BLI. Bio-layer interferometry characterization for the designed (RT) binder titrating against three closely related peptide targets, **A)** (RT)x7, **B)** (RL)x7, **C)** (RD)x7. Three-fold serial dilutions were tested for each binder, and the full tested concentration is labeled. The biotinylated targets were loaded onto the streptavidin (SA) biosensors, and incubated with designed binders in solution to measure association and dissociation.



Extended Data Fig. 6: Refined binder (refinement from motif diffusion) showed increased selectivity while decreased affinity in the case of (LK) binder. A) WT (LK) binder (cartoon in yellow, BLI plot below, left) showed a better K_d ~500 pM against its biotinylated target (LK)x7, while motif-diffused (LK) binder (cartoon in green, BLI plot below, right) showed a weaker K_d ~10 nM. **B)** Motif-diffused ((LK)md) binder showed better selectivity against (PV)x7 and (RT)x7 off-targets than WT ((LK)wt) binder in BLI assay. The biotinylated targets were loaded onto the streptavidin (SA) biosensors, and incubated with designed binders in solution to measure association and dissociation. Binders were measured starting at 100 nM concentration, followed by two-fold titrations. Heatmap was plotted with the highest binding signals.

Base pocket	Related pockets
LK	IK,VK, LR, IR, VR, LH, IH, VH
RT	KT, HT, RS, KS, HS
DY	EY, DW, DF, EW, EF, EY
PV	PI, PL
GA	GC
VP	IP, LP
TQ	SQ, SN, TN
L	I, V
K	R, H
R	K, H
T	S
D	E
Y	W, F
P	
V	I, L
G	
A	C
Q	N
M	

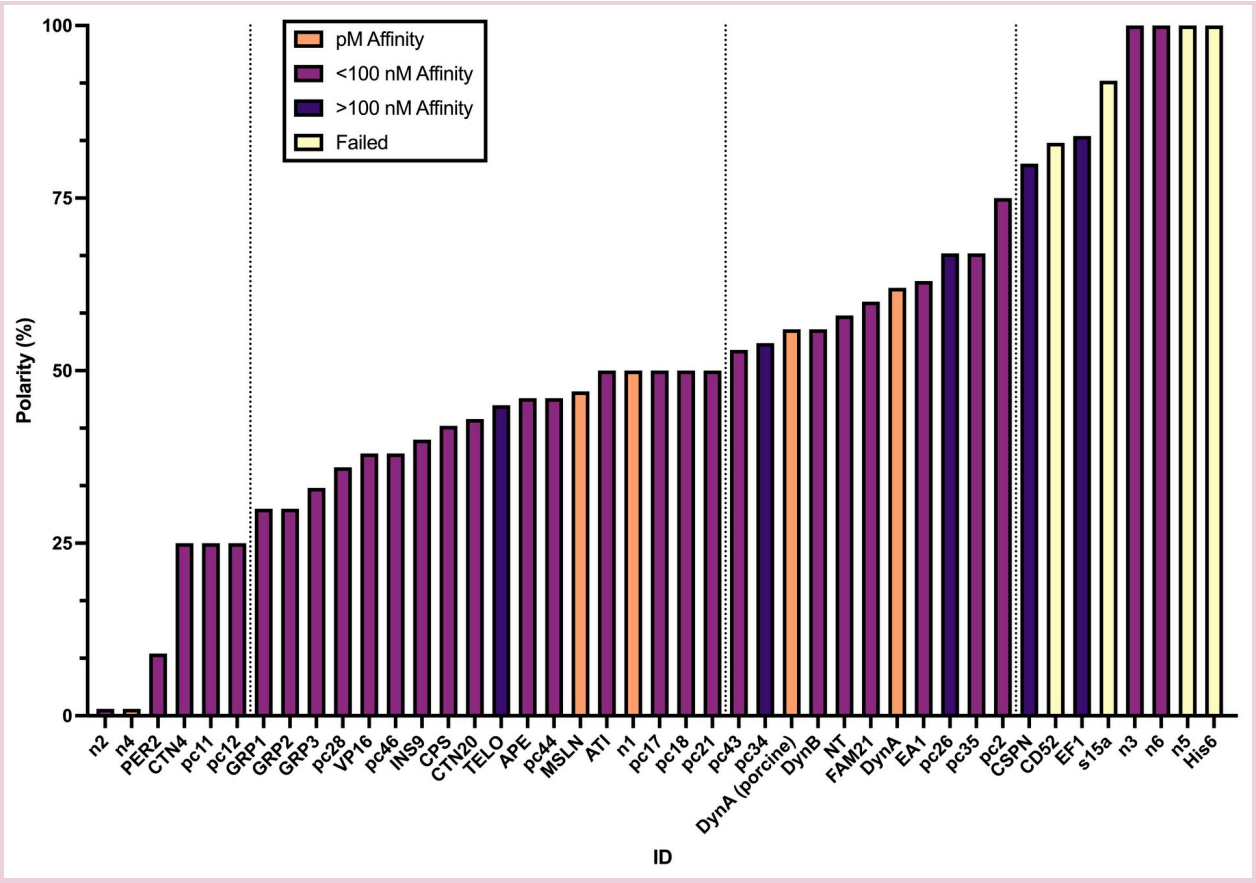
Extended Data Table 7: Recognition pockets (base and related) generated in this work. Collection of all pockets generated in this work, with either experimentally validated or computationally predicted confidence in context of disordered targets binding. The related pockets were structurally grouped by their side chain properties, such as charge, polarity, hydrophobicity, aromatics, and size.



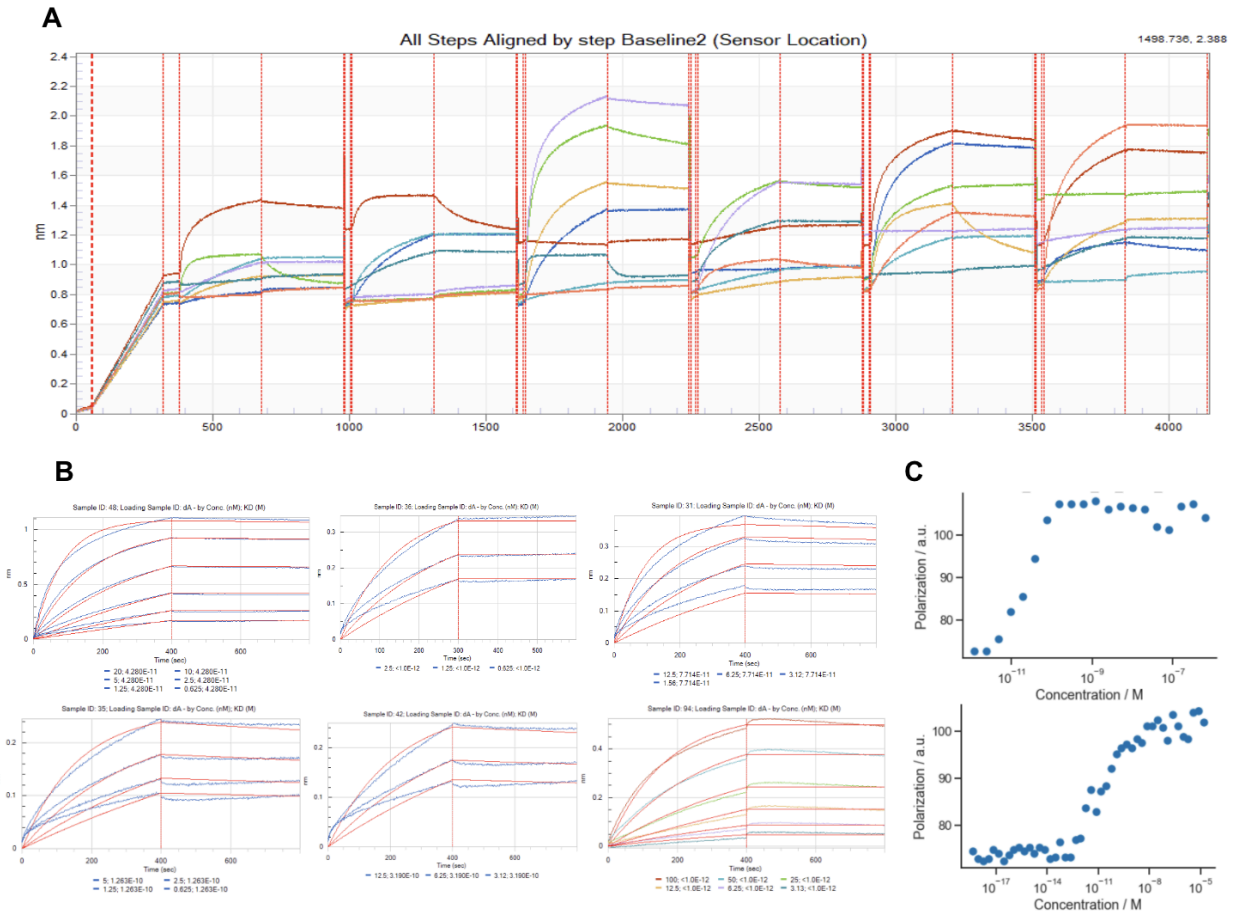
C

Target	Batch 1 tested designs	Batch 2 tested designs
pc28	10	20
pc46	7	20

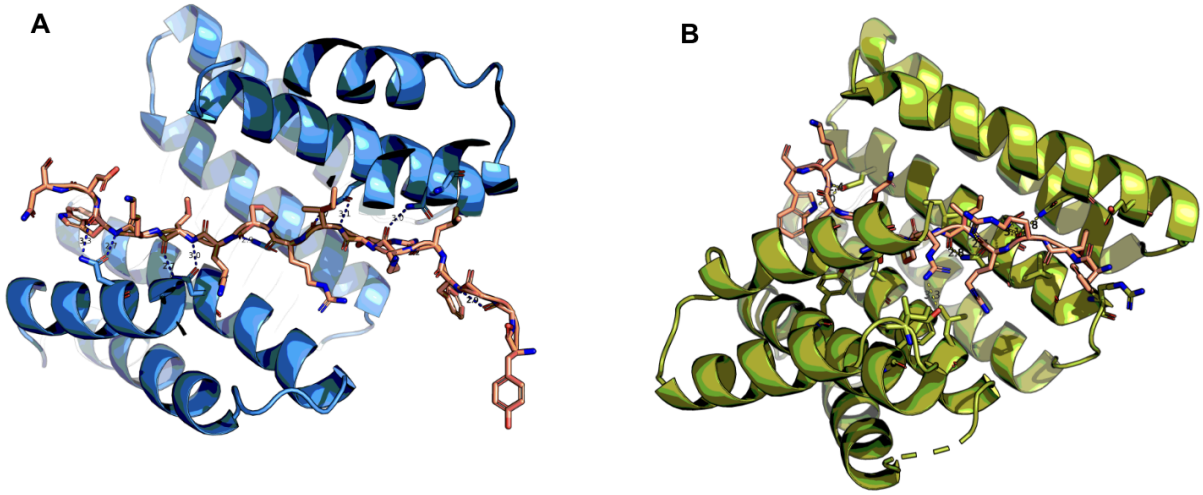
Extended Data Fig. 8: Orthogonality improved between targets of arbitrary English words through testing a slightly bigger set of binders (from 8 on average to 20 on average). A) orthogonality matrix before; **B)** orthogonality matrix after; **C)** number of designs tested in these two batches respectively. Spells out of the English words are as follows: DANIELSILVA (pc28), HIALIENFRIEND (pc34), DERRICKHICKS (pc35), DAVIDLIKESPEPTIDE (pc43), DAVIDITISTIME (pc44), WILLCHEN (pc46). Orthogonality screening was performed by the nanoBiT assay for the six designed binder-synthetic peptide pairs, with lgBiT-binders at 100 nM and smBiT-targets at 0.5 nM. Heatmap indicates the average from N=3 experiments.



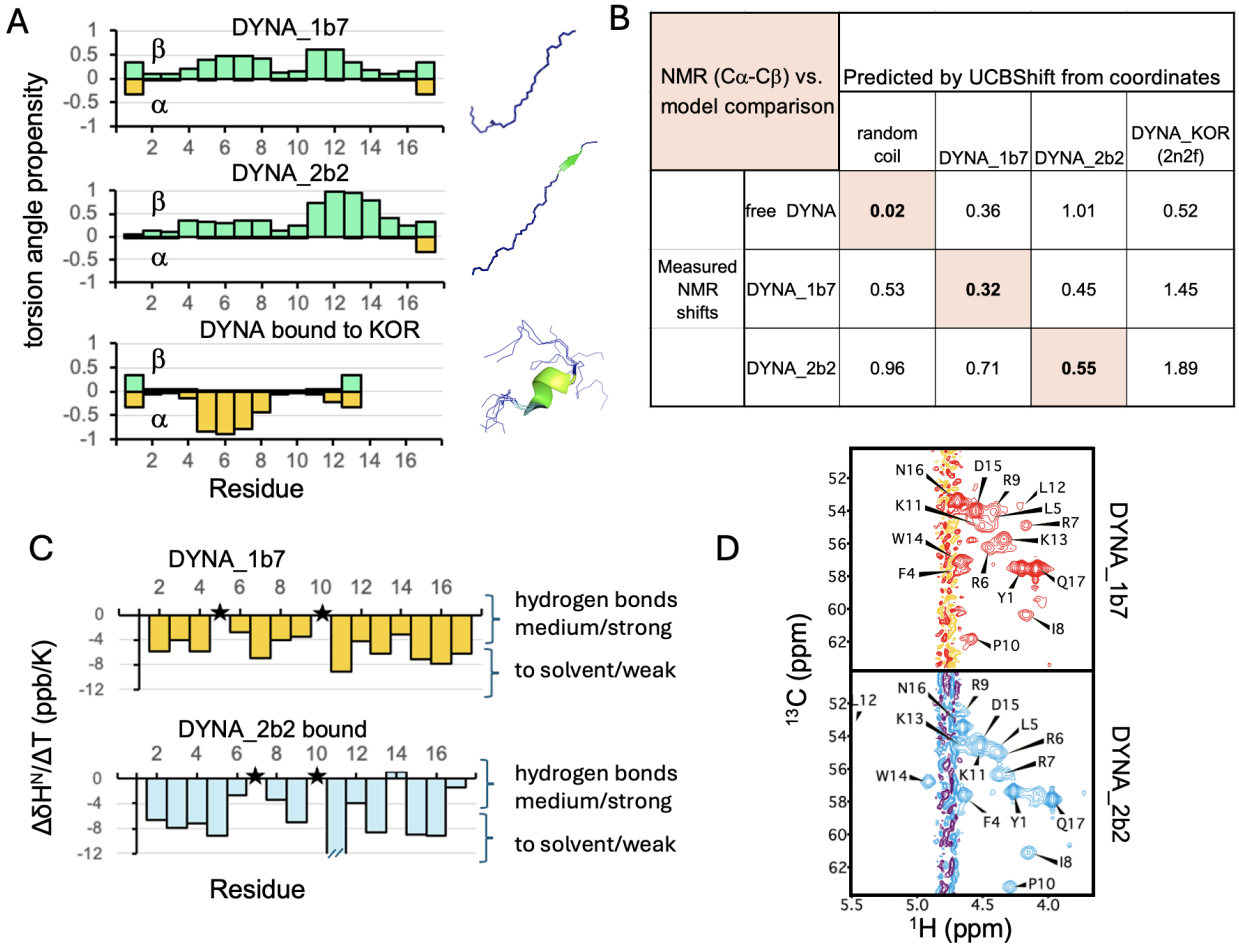
Extended Data Fig. 9: Target sequence polarity distribution on the 21 native targets and 18 synthetic targets. Polarity was calculated based on the percentage of polar and charged amino acids (D, E, H, K, N, Q, R, S, T, Y) among the targeted window of each target.



Extended Data Fig. 10: Tight binding of designed dynorphin A binders. In the second design round, the optimized dynorphin A binders showed **A)** exceptional screening success rate at 5 nM binder concentration, **B)** ultra-tight binding for six binders' titration by Octet, BLI (tested concentrations and kinetic Kd fittings were labeled under each plot), estimated Kd ranging from <1pM ~ 300pM, and **C)** ultra-tight binding for two binders' thermodynamic fitting by fluorescence polarization (FP) with 100pM TAMRA labeled peptide, estimated as Kd <200 pM, Kd <60 pM (Kd cannot be accurately measured below the concentration of peptide used in the FP assay).

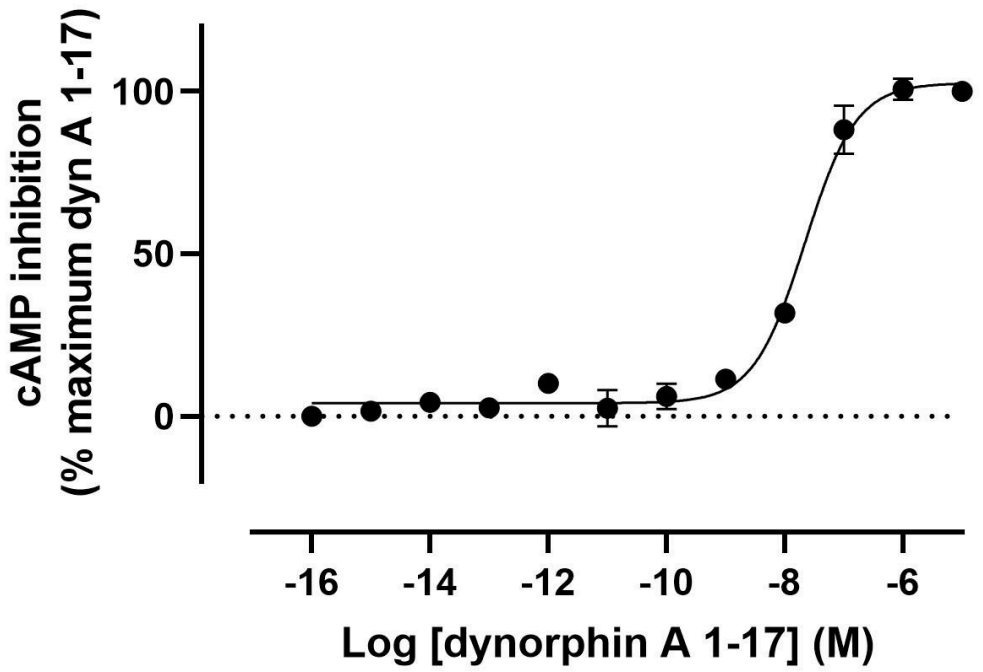


Extended Data Fig. 11: Comparison of the structure of A) initial design DYNA_1b1 to B) two-sided partial diffusion refined design DYNA_1b7. During partial diffusion, the binding mode was further changed and diversified on both the binder and the target.

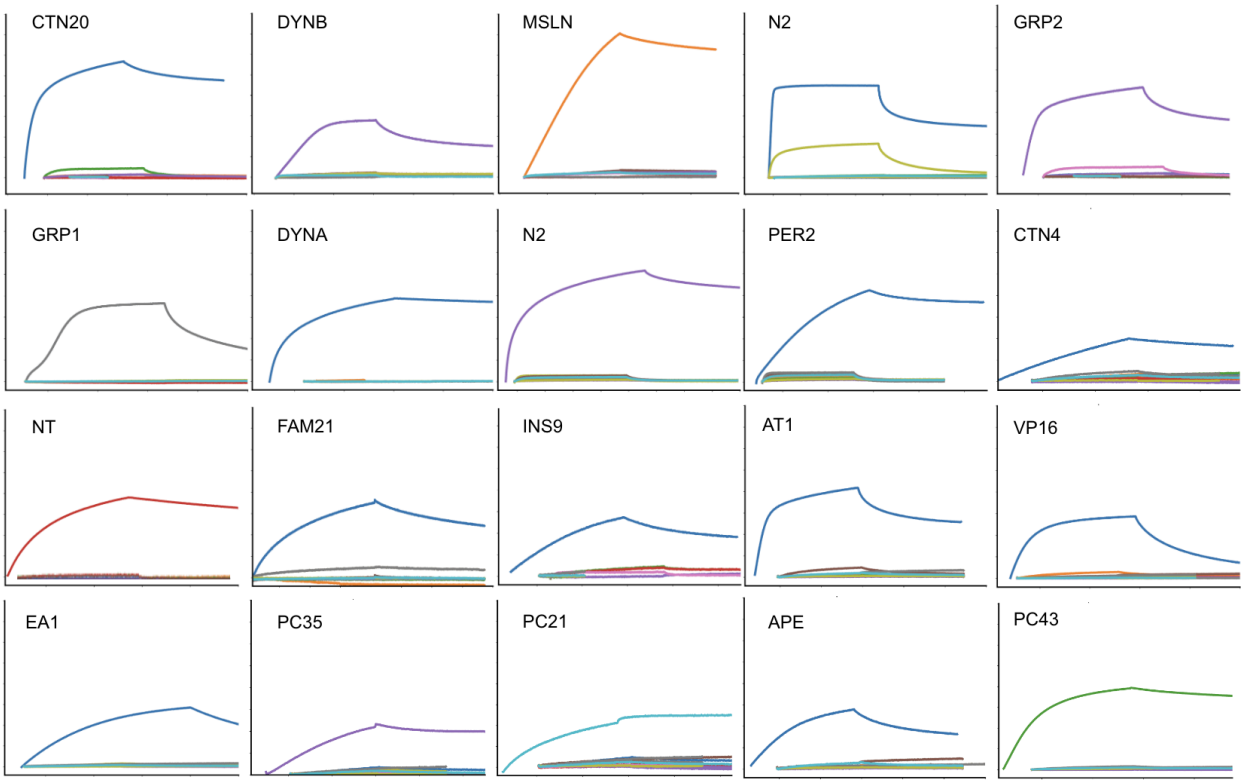


Extended Data Fig. 12: NMR analysis on dynorphin A complexes. Dynorphin A binds as extended and partial beta strands as designed. **A)** NMR chemical shift-based analysis using Talos-N of isotope-enriched dynorphins bound to DYNA_1b7 and DYNA_2b2 distinguish helical (α) from extended (β) conformations for each residue along its sequence. Its binding mode is different from the short helix formed when bound to the native human KOR (PDB 2n2f with chemical shifts predicted by UCBShift). **B)** Validation of design DYNA_2b2 by comparison of experimental ^{13}Cα and ^{13}Cβ chemical shifts with back-calculated chemical shifts (scaled RMSE analysis) shows that each set of measured NMR chemical shift values agrees best with those predicted from the corresponding atomic coordinates (DYNA_1b7 (PDB 9cce), DYNA_2b2, and PDB 2n2f) by UCBShift (*ref*), rather than with chemical shifts predicted from the other models. **C)** Amide temperature coefficients indicating hydrogen-bond donor amide protons; the backbone amide of W14 of DYNA_2b2 has the strongest hydrogen bond, corresponding to the bidentate hydrogen bonds between W14 backbone atoms and the polar Asn180 side chain of the binder. On either side of W14, there are predicted dynA to binder backbone-backbone hydrogen bonds for K13 and D15 to receptor S5 and K3, in two short antiparallel beta strands; two between K13 NH/CO and receptor S5 CO/NH, and two between D15 NH/CO and receptor K3 CO/NH forming the hydrogen bonded network. The star symbol indicates amide sites not

accessible to measurement. **D)** ^{1}H-^{13}C HSQC spectra of Hα – Cα region. The downfield $^{1}H\alpha$ shifts of L12, K13, and W14 in 2b2 versus 1b7 indicate enhanced binding interactions.



Extended Data Fig. 13: cAMP assay of dynorphin A (1-17). Gi-protein mediated cAMP inhibition was measured following KOR activation by dynorphin A (1-17). Dynorphin A (1-17) activates KOR in the cAMP assay with an EC_{50} of 23.1 ± 1.04 nM (mean ± SEM, n=2).



Extended Data Table 14: BLI specificity test of 20 binder-target pairs with Kd under 100nM. In each panel, one biotinylated disordered target (as labeled in upper left corner) was loaded onto the streptavidin (SA) biosensors, and incubated with its designed cognate binder and other 19 non-cognate binders all at 1000 nM in solution to measure association and dissociation. Traces were pulled together from multiple individual runs, which have varied association and dissociation times on x-axis. All binding signals were plotted using the same y-axis range (0.0, 2.0 ABU). Among all 400 individual runs, only the cognate pairs were showing obvious strong binding signals; with the exception of target N2 also binds to designed binder CTN4b1 (light green curve), though the non-cognate binding signal showed much smaller with a faster dissociation.

A

ID	Native	Amino acid sequence	Best affinity attained
1	DynA	(YGGF)LRRIRPKLKWDNQ	Low pM
2	DynA (porcine)	(YGGF)LRRIRPKLK	<1nM
3	DynB	(YGGF)LRRQFKVVT	9nM
4	NT	LYENKPRRPYIL	~1nM
5	ATI	DRVYIHPFHL	77nM
6	APE	QRPRLSHKGPMPA	92nM
7	INS9	SHLVEALYLVCGERG	59nM
8	CPS	(SLQP)LALEGSLQ	100nM
9	CTN4	NWLTIFILFPLK	16nM
10	CTN20	ESSVSLTVPPVVK	10nM
11	CSPN	DNEKLRKPKHKKLKQ	~360nM
12	VP16	DALDDFDLDML	18nM
13	EA1	PTPGKGPKVYRRKHQE	~100nM
14	EF1	SQQSSSYGQQNPSYDSVRR	>500nM
15	FAM21	SSDDDLFQSAKPKPAKKTNPFPLLEDE	9nM
16	PER2	AVPFPAPVPAAY	67nM
17	TELO	SWAHPGRTRGPSDRGFCVVSPARPAEEATSLEGALSGT	193nM
18	GRP1	RRPWVPHLLPFSSPGYLG	72nM
19	GRP2	HENGPWPGPCNARVAPMLLPRLPTPGVPSD	42nM
20	GRP3	VLWNSRWPTLQAWGAGLKPGY	~10nM
21	MSLN	NGYLVLDLSMQEALS	<1nM
22	s15a	TSTSPSSSSSSPLSSSSSSSSSSS	NA
23	HIS6	HHHHHH	NA
24	CD52	GQNDTSQTSSPS	NA

B

ID	Synthetic	Amino acid sequence	Best affinity attained
25	n1	LKLKLKLKLKLKLKLK	<500pM
26	n2	PVPVPVPVPVPVPVPV	18nM
27	n3	YDYDYDYDYDYDYDYD	72nM
28	n4	GAGAGAGAGAGAGAGA	~350nM
29	n6	RTRTRTRTRTRTRTRT	5nM
30	pc2	LKLKLKLKYDYDYDYD	84nM
31	pc11	LKLKLKLKPVPVPVPV	50nM
32	pc12	LKLKLKLKVPVPVPVP	36nM
33	pc17	DYDYDYDYVPVPVPVP	45nM
34	pc18	DYDYDYDYPVPVPVPV	~10nM
35	pc21	LKLKLKPVPVPVDYDYDY	27nM
36	pc26	LKLKGAPVQQYDYDEHRD	>200nM
37	pc28	DANIELSILVA	3nM
38	pc34	HIALIENFRIEND	~180nM
39	pc35	DERRICKHICKS	~90nM
40	pc43	DAVIDLIKESPEPTIDE	9nM
41	pc44	DAVIDITISTIME	35nM
42	pc46	WILLCHEN	37nM
43	n5	RDRDRDRDRDRDRD	NA/10+ uM

Extended Data Table 15: Amino acid sequences of all 43 targets we have attempted in this work. A) the group of native targets; **B)** the group of synthetic targets. The highest affinity attained was noted in the last column each. The four failed targets (affinity labeled as “NA”) are included.

Computation method	HBD (hydrogen bonding density / 12AA)	CMS (contact molecular surface / 12AA)	DDG (Rosetta binding energy/ 12AA)	BUNS (buried unsatisfied hydrogen bonds/ 12AA)
logos	18	620	-57	1.5
PPII	8	360	-36	1.4

Extended Data Table 16: Comparison of the Rosetta metrics of binders designed in this work (*logos*) and the previous poly-proline II repeat peptide binders averaging against the ordered designs. Calculations were done and averaged over all the experimentally tested designs. Due to the variation on target length, all the calculations were normalized to a base complex unit of per 12 amino acids on the target.

Supplementary outline

➢ Computational Methods

I. IDR threading
1) Threading
2) Cycle of sequence design
3) Predictions
4) Illustration of the customized packages (MPNN, AF2)
5) Instructions and step-by-step guide

II. Refinement - Scaffold-refined Diffusion
1) Partial diffusion (one-sided)
2) Partial diffusion (two-sided for library construction)
3) Motif diffusion
4) RFdiffusion

III. Pocket design
1) Base scaffold generation
2) Hashing provides initial high-quality complex docks for ProteinMPNN
3) Base binders for di-peptide repeat recognition
4) Backbone library through truncation and extension
5) Backbone library through parametric perturbation
6) Backbone library through motif diffusion

IV. Pocket assembly
1) Backbone pockets connecting through RFdiffusion
2) Assemble to target arbitrary sequences beyond pockets
3) Two-sided interface design

➢ Experimental Methods

1) Gene Construction of Designed Binders
2) Protein production and purification (small scale and medium scale)
3) BLI assays for initial screening and affinity measurement
4) BLI all by all orthogonality matrix
5) NanoBiT assays for initial binding screen and all by all orthogonality matrix
6) Protein Purification for Crystallography

7) TAMRA Peptide Synthesis
8) Biotinylated Peptide Synthesis
9) Fluorescence Polarization

➢ **Individual Design Characterization**

1) Purification of KFAM binders
2) Immunoprecipitation with KFAM binders
3) Dynorphin A inhibition cAMP assay
4) Affinity enrichment of CTN4 analyzed by LC-MS
5) NMR sample preparation and data collection of dynorphin-binder complex data
6) X-ray crystallography

Computational Methods

Due to the flexible nature of disordered regions, we considered them as polymer chains of arbitrary combinations of amino acids. We broke the problem into two main components: 1) providing binding pockets to all 20 amino acids (pocket design), and 2) assembling pockets flexibly to induce any unstructured backbones to fit (pocket assembly).

Here we present the final, simplified pipeline for general users first, i.e., logos and scaffold-refinement diffusion in sections I and II (detailed instructions and github repo follows). Then we describe the pipeline construction in sections III and IV.

I. Logos - IDR threading

Supplementing the representative experimentally validated design complexes, we assembled an initial library of 340 templates to start with. We reasoned that the current set was broad enough to take most arbitrary disordered sequences, assuming that 1) at least one of the templates could map to at least one 8 AA window with a BLOSUM score > -1.0; and 2) a minimum of 70% of the 8 AA window was mapped (especially the polar amino acids), then the rest of the pocket and backbone could be easily optimized with deep learning. The whole pipeline is as mostly automated as three steps, 1) threading, 2) sequence design, and 3) prediction.

We tested this idea by arbitrarily targeting 21 broadly diverse native bioactive targets (small peptides and intrinsically disordered protein regions), where one can not choose or change any sequence identities on the target side. During this benchmark, we discovered our hypothesis held mostly true, with two new observations. First, the later tested targets performed better in general, both in silico and in vitro—If the target had a low BLOSUM score (<1.0) or the complex went through dramatic diffusion refinement (RMSD > 3.5) we recycled and supplemented the new “representative” validated binders back into the library. Second, adding the iteration of MPNN-AF2-MPNN-AF2 cycle with a customized AF2 package (see below) further improved the in silico metrics, possibly due to the local RMSD diversification guided by AF2 during design.

To further extend the capability of this pipeline, we implemented these findings into the base logos pipeline, intending to keep accepting new, diverse, challenging targets with help from AF2 and RFdiffusion. The final library presented in this paper contained 1,000 templates, picked from the experimentally tested designs with diverse binding modes, diverse unstructured targets, and diverse binder RMSD and sequences.

Threading

To input the target sequence information of any unstructured target as a string of amino acids a target fasta file was required. The sequence of interest was then threaded through our ‘template library’ (made of 1,000 diverse docks, each with a peptide backbone). The library contained four sub-libraries: walle (binds targets in mostly extended conformations with the massive amount of bidentate hydrogen bonds; extensively tested against dozens of targets throughput this work), walle2 (supplemented experimentally validated designs throughput this work; similar to walle despite being less heavily tested), strand (binds targets in strand pairing conformations), and mini (shorter binders designed to bind short targets; this is in a beta testing stage). For general users, we recommend prioritizing with the order of walle, walle2, strand, mini when limited numbers of designs can be experimentally tested.

This threading process followed the register of the target backbone and through each individual template to search for the potential binding windows (for the target fragment) and feasible binding modes (for the target-binder complex; later judged by AF2). A flag called --keep_bidentates could have been enabled to keep the bidentate hydrogen bonding donors fixed. After threading, the new complex dock was repacked and minimized[1]. FlexPepDock Rosetta package[2,3] was then used to enable local resampling of the binder and target backbone and sidechains, outputting 10 different complex docks per run. With the flag of --max_farep, only complexes with acceptably low energy (i.e., farep(protein + peptide_seq) - farep(protein + peptide_backbone) < 300.0 EU) were saved.

Cycle of Sequence Design

All accepted docks were sent to ProteinMPNN with a customized script, which contains empirical weights on certain amino acids. Two iterations of ProteinMPNN and FastRelax (to delicately adjust chain B backbones further), and binder SAP score[1] optimization were performed. By default, the sequence with the lowest ProteinMPNN score from iteration one and five sequences with the lowest ProteinMPNN score from iteration two were collected and sent to our customized AF2 package (see below). This filtering, in general, varies depending on the target difficulty level. Designs with PAE <= 5, RMSD <= 3, iPTM >= 0.83 (for easy targets, when > 7,000 designs would pass this criteria) or designs with PAE <= 10, RMSD <= 4, iPTM >= 0.79 (for hard targets, when differing from above) were selected to undergo one more round of the customized pipeline. Criteria can be further adjusted to individual targets and design campaigns as needed.

Predictions

After two cycles of sequence design and AF2 (which we call prediction guided sequence design), passing designs are collected to go through AlphaFold-multimer. Empirically, we found in the cases of highly polar targets (polarity >= 80%) and proline-rich targets (proline >= 30%, and/or more than one proline appears in the center of the sequence window), AlphaFold-multimer was more predictive than AlphaFold initial. These filtering criteria were used for the final complex predictions: PAE <= 4-7, PLDDT >= 90 for AlphaFold-initial, and iPTM >= 0.8-0.9, PLDDT >= 90 for AlphaFold-multimer. Note that we found AF2 favored binding complexes with the disordered targets in secondary structural conformations (i.e., alpha-helical or beta-strand) and tended to give better scores. Therefore, we scored each sub-library separately with a more stringent criteria toward ‘strand’ (the sublibrary containing strand pairing binding modes; PAE <= 5; iPTM >= 0.88-0.9) than ‘walle’ (the sublibrary with diverse local extended binding modes, but less regular secondary structures; PAE

<=5-9; iPTM >= 0.8-0.9). At last, AlphaFold-monomer was applied on the remaining designs on chain A only (binder) to select the final designs to order, with PLDDT >= 90. The designs with the best scores were selected to be experimentally tested in this work.

Illustration of the customized packages (MPNN, AF2)

MPNN

For sequence design, ProteinMPNN[4] was set up with a customized weight on certain amino acids: {"A": -0.15, "G": -0.15, "P": 0.15, "E": 0.1}. For polar targets (polarity > 80%), this set of weights has been experimented: {"A": -0.15, "G": -0.15, "M": -0.35, "P": 0.15, "E": 0.1, "Q": 0.3, "H": 0.3, "W": 0.1, "Y": 0.25}. Each round of MPNN was followed by Rosetta FastRelax[1] where one sequence with the lowest score from round one and five to ten (differ by target difficulty level) lowest score from round two were generated for each design.

Templated AF2-initial-multimer

For templated prediction, we developed a package called 'AF2-initial-multimer.' The AF2 initial guess (and target templating) method of Bennett et al.[5] was implemented into a local install of colabfold such that the initial guess could easily be utilized with the AF2-multimer models[6]. We also added pyrosetta to our local install of colabfold to allow input and output of silent files which helps with file management when working with large numbers of designs. For each design from ProteinMPNN, we predicted whether the binder would bind its target peptide with the initial guess for each of the five multimer models using single sequence prediction. We averaged the metrics iPTM, binder_plddt, and interface_rmsd, from all five predictions, and utilized the averaged metrics for filtering purposes (see below).

MPNN-AF2 cycle

ProteinMPNN and AF2-multimer's initial guess filtering were iterated two times such that ProteinMPNN could assign a new sequence to the initial AF2 predictions when the predictions were close to the designs. AF2 predictions are able to improve the accuracy of the binder structure and local binding interface (including peptide structure) as has been demonstrated for previous de novo designs. By allowing ProteinMPNN to assign a new sequence to the more accurate backbone and docking configuration, we were able to improve the sequence structure agreement. After this, AF2 filtering was applied again, but with more stringent thresholds. The AF2 filtering metrics (averages of all five multimer models) for the first round were iPTM > 0.8, binder_plddt > 90 and interface_rmsd < 2.7. The AF2 filtering metrics (averages of all five multimer models) for the second round were iPTM > 0.88, binder_plddt > 92, and interface_rmsd < 1.0. For

peptide targets that were particularly easy or difficult, the stringency of these filtering metrics were increased or decreased to control the number of designs moving to subsequent design or filtering steps. We grouped the passing designs with a clustering method based on binders' amino acid sequence identities, with a dynamic AF2 metrics threshold, aiming to select the most structurally diverse set, with the top 10 scored designs from each cluster for AF2-multimer prediction.

Final AF2 prediction

Designs passing the previous AF2 initial guess filtering were now subjected to AF2 multimer single sequence predictions without initial guess using a local install of colabfold. The top <100 designs were selected from these predictions based on iPTM and subjected to visual inspection to select 3-70 designs for gene synthesis and experimental characterization.

Instructions and step-by-step guide:

The final pipeline of arbitrary IDR targeting consists of these steps (the code is attached below and through https://github.com/drhicks/Kejia_peptide_binders).

A computational pipeline to target arbitrary unstructured sequence fragments (4-30 amino acids) of intrinsically disordered proteins and peptides, with de novo designed binding proteins.
Prerequisites

- A Python environment with PyRosetta.
- Make sure silent_tools is in your PATH
- Download DL weights for AlphaFold, ProteinMPNN, and RF_Diffusion.
- Correct paths in MPNN scripts for your local installs.
- Update paths in `path_to/threading/make_jobs.py` to use your Python environment.
- Ensure `path_to/threading/make_jobs.py` has the correct path for `path_to/threading/thread_peptide_sequence_new.py`

Step-by-Step Guide

from your working directory for this binder project.

1. Threading

1. Make a directory for threading:

mkdir 1_threading

2. cd 1_threading
3. Make all jobs: make your target fasta file make your template list file
4. python path_to/threading/make_jobs.py path_to/peptide.fasta path_to/templates.list | sort -R > all_jobs
5. Split jobs into smaller sets:
6. split -l 3 all_jobs
7. Add all job sets to SLURM array list:
8. for i in x* ; do echo "bash $i" ; done > jobs
9. Submit array file jobs to SLURM:
10. path_to/threading/submit_jobs.sh
11. After jobs finish, collect all PDB outputs into a silent file:
12. silentfrompdbs path_to_pdbs/*pdb > threading.silent

2. MPNN

1. Back in your original working directory:

mkdir 2_mpnn

2. cd 2_mpnn
3. Run the initial MPNN (without relax; preferred):

path_to/job_creation/dev_mpnn_design_job_create -prefix mpnn -script path_to/mpnn_git_repo/design_scripts/killer_mpnn_interface_design.py -p cpu -t 12:00:00 -mem 5 -cpus 1 -conda path_to/env/mpnn_pyro -structs_per_job 100 -silent path_to/threading.silent -args "--num_seq_per_target 5 --max_out 5 --sampling_temp 0.1"

4. ./run_submit.sh
5. or (with apptainer)

path_to/job_creation/mpnn_design_job_create -prefix mpnn -script path_to/mpnn_git_repo/design_scripts/killer_mpnn_interface_design.py -p cpu -t 12:00:00 -mem 5 -cpus 1 -apptainer path_to/your_apptainer -structs_per_job 100 -silent path_to/threading.silent -args "--num_seq_per_target 5 --max_out 5 --sampling_temp 0.1"

6. ./run_submit.sh
7. In the paper, we have been routinely doing 2-rounds sequence design (i.e., MPNN-relax-MPNN-relax) for many targets in an earlier time. To do this, one can simply run the above script twice with the flag --relax on. Set --num_seq_per_target 1 in the first run and --num_seq_per_target 5 in the second.
8. Concatenate all the silent files together:
9. cat mpnn_runs/*/*silent > mpnn_out.silent

3. AlphaFold Filtering and Refinement (AF2-initial-multimer)

1. Back in your original working directory:

mkdir 3_af2_im

2. cd 3_af2_im
3. Make array jobs:

path_to/job_creation/interfaceaf2create -prefix af2 -script path_to/colabfold_initial_guess/AlphaFold2_initial_guess_multimer.py -silent ../2_mpnn/mpnn_out.silent -gres "gpu:1" -apptainer /home/drhicks1/scripts/Kejia_peptide_binders/colabfold_initial_guess/make_apptainer/colab_fold_ig.sif -structs_per_job 300 -p gpu-bf -t 06:00:00

4. ./run_submit.sh
5. Concatenate all the silent files together:
6. cat af2_runs/*/*silent > af2_out.silent
7. Create a scorefile:
8. silent_tools/silentscorefile af2_out.silent

9. Filter with sequence clustering and picking the top AlphaFold output/s per cluster after averaging 5 models:

```
python path_to/af2_filtering/average_af2_model_scores.py af2_out.sc > af2_out_averaged.sc
python path_to/af2_filtering/dynamic_filtering_by_group.py af2_out_averaged.sc af2_out.silent > cluster.log
```

10. column_number=$(head -1 cluster.log | tr '\t' '\n' | grep -n 'description' | cut -d: -f1); awk -v col=$column_number 'NR > 1 {print $col}' cluster.log | grep -oE '[a-zA-Z0-9_]+_af2mv3_[0-9]+' > tags

4. MPNN/AF2 cycle

1. Repeat MPNN step on filtered silent file.
2. Repeat AlphaFold IM and filtering.

5. Sequence Only AlphaFold Filtering and Refinement (potentially optional)

1. Make fasta file.
2. silentsequence path_to/af2_out_filtered.silent | awk '{print ">"$3"\n"$1":"$2}' > colabfold_input.fasta
3. Run colabfold.
4. /home/drhicks1/scripts/Kejia_peptide_binders/colabfold_initial_guess/make_apptainer/colab_fold_ig.sif AlphaFold2_jupyter-batch_hack_new_v2.py --fasta colabfold_input.fasta --num_recycles 10
5. Concatenate all the silent files together:
6. cat af2_runs/*/*silent > af2_out.silent
7. Create a scorefile:
8. silent_tools/silentscorefile af2_out.silent
9. Filter with sequence clustering and picking the top AlphaFold output/s per cluster after averaging 5 models:

```
python path_to/af2_filtering/average_af2_model_scores.py af2_out.sc > af2_out_averaged.sc
```

10. python path_to/af2_filtering/dynamic_filtering_by_group.py af2_out_averaged.sc af2_out.silent --not_initial_guess

Additional Steps

- Other filtering steps as desired such as af2_filtering/rosetta_min_ddg.py and visual inspection to order.
- Incorporate motif diffusion or partial diffusion as needed

Notes

- You may choose to run MPNN with Rosetta relax, but this is slow and questionably useful. If you do, add the flag `--relax`.

- If you run relax or minimization methods that can perturb the rigid body and/or binder/target backbones, you could run a second MPNN on the output of the first to potentially design a better sequence. However, the current preference is to go straight to AlphaFold filtering/refinement.
- Diffusion can be run at various steps such as:

1. After 1 or 2 rounds of mpnn/af2 , if not enough designs (< 70) passing the final filtering criteria, in which case you will want to repeat the two cycles of mpnn/af2 on the output from diffusion.
2. On the final designs before ordering, if enough designs (>= 70) passing the final filtering criteria, but one may want to order on chips and/or include arbitrary refined designs in initial test, in which case you can repeat either the one cycle or two cycles of mpnn/af2 on the output from diffusion. Depending on available computation resources and chip quota.
3. On the initial hits after experimental screening and characterization, in which case you will want to repeat the two cycles of mpnn/af2 on the output from diffusion.

- In general the pipeline works most times without the use of diffusion, however, intelligent use of diffusion can increase in silico success rates for difficult targets and potentially improve the affinity and specificity of characterized binders.
- There are many knobs that can be tuned and variations of the pipeline that can be run depending on the ease or difficulty of individual targets.
- Diverse + good in silico designs generated by the pipeline (regardless of experimental characterization) can be added into the templates to continue building new diverse binding modes for the future.

Motif Diffusion
bash make_motifd_jobs
bash motif_d_jobs
Submit jobs

Partial DIffusion
If partial diffusion, refer to the normal published partial diffusion with small partial_T.

II. Refinement

Pocket recognition and pocket assembly enabled refinement of precise interactions with individual amino acids in a sequence-specific manner. This was critical for challenging polar or charged targets, or targets that would not adopt regular secondary structures when bound.

To avoid running into a theoretical modular assembly limit introduced from a finite number of assemblies and an infinite number of possible targets, we developed a refinement package based on RFdiffusion[7] called 'scaffold-refined diffusion.' This aimed to provide adequate high-resolution refinement and perturbations to the 'mismatches' between pockets, spacers, and protein backbones even when the overall match was a coarse-grained fit, i.e., when accepting a new enough target in terms of sequence identities. This way, we maintained the high-efficiency engineering for targeting disordered regions with low computation resources (1-3 computation days, 1,500 CPU hours, 500 GPU hours per target) and minimum experimental cost for most arbitrary disordered targets (3-70, on average 36 tested designs per target). Meanwhile, we provided an additional option to surpass the theoretical limitation (potentially with a slightly higher computational/experimental cost).

Scaffold-refined Diffusion

If the design did not pass the above complex prediction criteria, or if a larger number of tested designs were preferred, scaffold-refined diffusion was developed to optimize the customized fits between the less ideal interacting pockets and the less ideal binder backbone. To do this, a fair amount of the original binding interaction and/or binding motifs were kept to maintain the advantage of our general platform. Therefore, partial diffusion[8] with small steps (unless for the purposes of new library construction) and multi-motif-constrained diffusion (motif diffusion) were developed for this purpose. For the work in this paper, a customized build of RFdiffusion was implemented.

Partial diffusion (one-sided)

As described in other cases, RFdiffusion was modified to allow the input structure to be noised only to a user-specified time step rather than completing the full noising schedule. Consequently, the starting point of the denoising trajectory retained information about the input distribution, leading to denoised structures that are structurally similar to the original input. In our cases, the time step was set to be lower to maintain as much structural similarity as possible as the high design resolution of hydrogen bonding interactions (especially to the polar amino acids) were reasoned critical. Therefore, 10-18 noising timesteps (specifically, 10, 12, 15, 18) out of a total of 50 in the noising/denoising scheme were chosen for the work in this paper. Two

hundred to three thousand partially diffused designs were generated for each target parameter.

The new backbones went through ProteinMPNN sequence design and AF2 predictions as described above. The designs were then filtered using the same criteria. It was noticed that partial diffusion tended to increase the number of overall passing designs, while sometimes the original hydrogen bonding networks (especially the bidentate hydrogen bonds to the target backbone) were broken without being caught by AF2. Therefore, we implemented another filter ‘buns’ (buried unsatisfied hydrogen bonds)[9] <=1 to explicitly count the unsatisfied heavy atoms. In some cases, overall correlated with target polarity, this step would filter out 50-80% of designs. Partial diffusion seemed to increase the wet lab success rate while being inconclusive as to the affinity increase unless tested on a much larger number of designs (>1,000 designs, unpublished data).

Partial diffusion (two-sided for library construction)

Unlike one-sided directional partial diffusion, which only diversifies the conformation of the binder while keeping the target unchanged, two-sided partial diffusion enabled simultaneous conformational changes in both the target and the binder.

For the input designs, we applied 10-25 noising timesteps out of a total of 50 in the noising scheme followed by denoising. This process generated approximately 2,000-10,000 partially diffused designs for each target. The new backbones went through ProteinMPNN sequence design and AF2 predictions as described above. The designs were then filtered using the same criteria. It was noticed that two-sided partial diffusion tended to increase the number of complex passing designs and slightly decrease the number of monomer passing designs. We did not see a direct correlation with the following wet lab success rates as targets varied.

Motif diffusion

We utilized a custom pyrosetta script to identify binding motifs; defined as any residue making atomic contacts (heavy atoms < 4.5 Å) to atoms in the target peptide. We input these binder motifs into diffusion, allowing variable lengths (from 0.75x to 1.25x, normally with bigger variety on the N/C-terminal) of protein backbone (fully noised atoms) to connect the motif residues, such that diffusion could reform and create new binding proteins scaffolded around the input binding motifs with variety. This improved and customized the shape complementarity between each binder and the individual target, while maintaining the key input motif residues, especially the ones that are hydrogen bonded; improved and diversified the core packing of the binder (if needed); and sometimes increased the total number of binder residues contacting the peptide

target. Motif diffusion has seen consistent improvement in both in silico and in vitro binding metrics in most of the cases we tested.

RFdiffusion

In this work, we were using accelerated denoising schedules in RFDiffusion that make fewer calls to the underlying RosettaFold2 (RF2) module. We implemented twpst which is partial diffusion with only one frame. We additionally used a technique to skip even more calls to RF2: Instead of adjusting the denoising schedule (which seems to have a limit of no fewer than 40 RF2 calls while maintaining quality), we repeatedly seeded the Px0 (re-noised to the current t) as the current xt structure allowing us to achieve schedules without only 15 RF2 calls.

III. Pocket design

Ideally, a *de novo* designed binding protein would recognize each individual AA along an extended polymer chain with specific AA "pockets." However, even assembling pockets for a short 10 AA peptide leads to an exponential number of $20^{10} = 1.024^{13}$ designs. Based on the previous work[10], specificity to the unstructured peptides could be achieved by computationally redesigning pockets to recognize, say, 2-3 AA out of an 18 AA extended peptide target. Additionally, RFdiffusion showed extraordinary performance in adjusting the binding interface to achieve better affinity on helical peptide targets[8] and specificity on TNF protein targets [unpublished]. By combining both we established a possible solution: *de novo* design of a set of recognition pockets covering the alphabet of a representative 10 out of 20 AAs with the option to fine-tune the pocket's sensitivity later to increase the specificity. Twenty amino acids (AA) were grouped by their side-chain properties, such as charge (R/K/H, E/D), polarity (Q/N, T/S), torsion angle space (P, G), aromatics (W/Y/F), and hydrophobicity and size (M, I/L/V, C/A).

Base scaffold generation

To make the pocket templates programmable and experimental characterization modular, designed helical repeat (DHR) was chosen as base scaffolds to generate recognition pockets.

Each scaffold was constructed using a helix-loop-helix-loop pattern repeated at least four times. The helices typically consisted of 18 to 30 AAs, while the loops contained 3 to 5 AAs. The scaffold design process involved backbone design, sequence design, and structural prediction. The designs exhibited a wide range of twist (omega) between 0.5 and 1.2 radians, a radius ranging from 3 to 15 Å, and a rise between 0 and 20 Å to align with the peptides. The geometry of a repeat protein can be described by the super-helix radius, axial displacement, and twist.

The backbone design was achieved using Rosetta fragment assembly, guided by parameters and motifs. This involved 9,600 Monte Carlo fragment assembly steps, utilizing fragments from a non-redundant set of Protein Data Bank (PDB) structures. After inserting each fragment, the rigid-body transformation was extended to subsequent repeats. The scoring for fragment assembly included Van der Waals interactions, packing, backbone dihedral angles, and residue-pair-transform (RPX) motifs. RPX motifs provided a rapid evaluation of full-atom hydrophobic packability of the backbone prior to side chain assignment. Post-design, backbones were screened

for native-like features. Loops must have been within 0.4 Å of a naturally occurring loop or be rebuilt. Structures with helices deviating more than 0.14 Å were considered bent or kinked and were discarded. Structures with fewer than eight helices in contact were also filtered out. To select base proteins more curved or twisted than Armadillo repeat proteins (ArmRPs)[11,12] and tetratricopeptide repeats (TPRs)[13], the distance between the first and the last helix was required, as D_sml <= 15A, D_avg <= 25A (measured from Cα to Cα) to close up and surround the target.

Sequence design was performed on each filtered backbone using ProteinMPNN, with a customized weight on certain AAs: {"A": -0.15, "G": -0.15, "M": -0.35, "P": 0.15, "E": 0.1}. Sequence-level internal repeats were not forced in this round of backbone generation, as we reasoned that the sequence diversity would potentially help with gene synthesis and protein expression for repeat proteins. At last, structural prediction was conducted by AlphaFold2 (AF2)[14] with PLDDT > 90, Cα RMSD < 2 Å.

Hashing provides initial high-quality complex docks for ProteinMPNN

Following our previous work[10], a hash table was considered to maintain the highest resolution to store the privileged protein-peptide motif pairs of hydrogen bonding. To create hash tables that store pre-computed privileged side-chain–backbone bidentate interactions (to place the peptide backbone into the right ‘docks’), we employed the previous hash database created for the polyproline II tri-peptide binders, but only focused on bidentate interactions.

To sample repeat peptides that align with the superhelical parameters of the designed helical repeats, we randomly generated sets of backbone torsion angles φ and ψ (e.g., φ ranges from [-150, -70], ψ ranges from [-30, 150] for mono-peptide, di-peptide, tri-peptide, tetra-peptide repeats). If any pair of φ and ψ angles yielded a Rosetta Ramachandran score above the −0.5 threshold, indicating potential steric clashes, we regenerated new pairs of angles until they met the Rosetta score criteria. Di-peptide repeats were a focus of the initial templates, as less enumeration of sequence searching space was needed to cover amino acid alphabet while mono-peptide could introduce torsion angle limitation for future assembly. These torsion angles were then repetitively applied across the six repeats of the repeat peptide, and the superhelical parameters were calculated using the 3D coordinates of adjacent repeat units. Repeat peptides matching the superhelical parameters of any curated designed helical repeats were retained for docking.

For docking cognate repeat proteins and peptides with matching superhelical parameters, both were initially aligned to the z-axis based on their superhelical axes. A 2D grid search (involving rotation around and translation along the z-axis) was then performed to identify compatible positions of the repeat peptide within the binding groove of the repeat protein. Upon generating a reasonable dock without steric clashes, the relevant hash function iterated through potential peptide–protein interacting residue sets to compute hash keys. If a hash key was found in the hash table, the corresponding interacting side-chain identities and torsion angles were retrieved and applied to all equivalent positions in the docking conformation. Docked peptide–designed helical repeat pairs were saved for the interface design step if the peptide–designed helical repeat hydrogen-bond interactions were satisfied. Once accepted, this docked pair was stored in our initial base library logos-di-0, which produced a derivative library logos-di-1 with peptide docks perturbed through FlexPepDock (10 low-energy docks per template). We reasoned that the hash table approved initial peptide-designed helical repeat docks would enable the highest percentage of side-chain–backbone hydrogen bonding recovery followed by ProteinMPNN.

Base binders for di-peptide repeat recognition

For the first round of di-peptide repeat binder design, target candidates were generated through the enumeration of 19 x 18 AAs, excluding Cys (to avoid disulfide bond formation) and itself. Only combos containing at least one polar AA (i.e., Asp, Glu, His, Lys, Asn, Gln, Arg, Ser, Thr, Tyr, or Pro for its special Ramachadran constraint) were selected. Target (GA)n was intentionally added as an extreme case of a highly flexible target with the smallest side-chain combinations—which would then be forced to make the smallest pockets and maximal backbone interactions. ProteinMPNN was applied with the same AA weights as above. All the designed complexes were validated by AF2 with a cutoff of PAE_interaction ≤10, PLDDT > 92 for the complex and PLDDT > 92, Cα RMSD < 2 Å for the binder monomer before experimental characterization. For the fully polar targets, AF2 was only used to judge the monomers, combined with a Rosetta scoring matrix as DDG < -50, contact_molecular_surface (CMS) > 500, BUNS (buried_unsatisfied_penalty) < 1.

In this work, the polar AAs were prioritized for a few reasons. 1) Polar targets can be particularly challenging due to the increased need for precise side-chain placements and hydrogen bonding networks; meanwhile, since it is significantly underrepresented in the PDB database used for deep learning training, it becomes an excellent option for

rational computation design. 2) Polar interactions are presumably the driving force of target specificity due to the accurate side-chain placement from reason 1. 3) In order to reduce the vast search space of flexible backbone assemblies, we reasoned that designing a set of assemblies capable of precise polar targeting could be engineered to effectively target non-polar targets using deep learning resampling capabilities (due to reasons 1 and 2).

Therefore, we chose to first model a number of di-peptides (each one of the two AAs in an extended backbone conformation, occupying an individual pocket) in a repeat binding manner. This modeling provided experimental convenience (as we could easily enhance or diminish the binding energy contribution of any single pocket by changing its repeat number) and ensured the compatibility of later pocket assembly (all pockets being inserted on coherent helical bundles).

Backbone library through truncation and extension

After inspecting the first hits from round one of di-peptide repeat binders, helical fragments outside of the binding interface (with distance between Cα > 10 Å while Cβ > 8 Å) on the same helix were considered unnecessary. For each repeat, truncations were made symmetrically for each interface helix on the same spot, together with the matching sides on the buttressing helices. Loops were reconstituted with the exact same loop residues plus or minus (G) or (GS). After modification, ProteinMPNN was applied to produce 10 sequences for each base designed helical repeat. All 11 sequences per base designed helical repeat were sent to AF2. Designs with PLDDT > 92, Cα RMSD < 2 Å were selected to parametrically extend to five and six repeats. Another round of MPNN-AF2 was applied to create the new di-peptide binding library.

Backbone library through parametric perturbation

To diversify the shapes of the binding pockets, we adapted the previously developed parametric repeat protein generation method for both symmetric and asymmetric perturbation with specified geometric parameters[15]. For each helix in a base designed helical repeat, we performed a grid search of its six rigid-body degrees of freedom. The translation along any axis was limited to 2 Å, and the rotation around any axis was limited to 10 degrees. For asymmetric perturbation, only the designated helices were subjected to this protocol, and for symmetric perturbation, each of the sampled geometric transforms was propagated to the rest of the repeats in the designed helical

repeat to maintain symmetry. We filtered the diversified designed helical repeats by removing the backbones that have high steric clashes (fa_rep>100), high intra-repeat inter-helix distance (distance between the centers of mass of the two helices in a repeat unit>12 Å), or less than 28% of the residues in a buried core.

Backbone library through motif diffusion

We manually picked the interacting residues surrounding the bidentate hydrogen bond donors to the target peptide, often four to nine residues per repeat. This is chosen with the hypothesis that the backbone hydrogen bonding is the prerequisite to lock the disordered target in an extended, random coil conformation. We input these binder motifs to diffusion, allowing variable lengths (from 0.75x to 1.25x compared to the parent backbone, normally with bigger variety on the N/C-terminal) of the protein backbone (fully noised atoms) to connect the motif residues, such that diffusion could reform and create new binding proteins scaffolded around the input binding motifs with variety. This improved and customized the shape complementarity between each binder and the individual target (while maintaining the key input motif residues, especially the ones needed for hydrogen bonding), improved and diversified the core packing of the binder (if needed), and sometimes increased the total number of binder residues contacting the peptide target.

IV. Pocket assembly

Once we collected five di-peptides and ten amino acid binding pockets, we considered the next two milestones toward general sequence recognition to be, 1) efficiently assemble and shuffle the existing pockets; and 2) expand this assembly manner to arbitrary sequences beyond the pockets.

To do 1, we explored parametric assembly taking advantage of the modular nature of repeat proteins together with RFdiffusion to insert varied ‘spacers’ between the regular modules to create backbone diversity on both binders and targets. To do 2, we tested ‘derivative pockets’ (meaning, this pocket targets slightly newer amino acids) on sites of the existing pockets with deep learning-based protein design (i.e., ProteinMPNN, AF2, RFdiffusion).

Backbone pockets connecting through RFdiffusion

Based on the five di-peptide binders of representative amino acid motifs (LK), (RT), (YD), (PV), and (GA), we reasoned arbitrarily assembling them into 12-18 AA sequences would allow us to expand and cover a huge space of random target residue identities. This is judged in terms of side chain group size, overall charge distribution, hydrophobicity extent, aromatic properties, and backbone torsion angles.

To do this, we first enumerated the origami combos of target pattern generation, i.e., AAAAAA, AAABBB, AABBCC, AABCDD, ABCDEF, where each letter represented a unique existing di-peptide motif. Note that, to insert a single di-peptide motif into a continuous disordered fragment caused the recognition pocket before and after the pocket to fit into different backbone hydrogen bonding patterns. In other words, the assembled final interface distinguished (XX)n(PV)n and (XX)n(VP)n with a minor one-amino-acid pocket out of phase. Pockets with mutations on target (ABCDEF) were generated on the fly while assembling the target backbone with varied spacers as well to satisfy the backbone phi-psi angles better, such as (TR) to (TQ) in pc26.

We generated 36 synthetic assembled target sequences following the above pattern. Twenty target candidates with more diverse side-chain properties were chosen for binder design. For each target, we first generated the peptide conformation by copying and pasting their phi-psi angles from the individual amino acid motif pockets, with a tolerance of 30° to connect and smooth the backbone transition between motifs. With exceptions, the ‘spacer’(s) between motif A and B were intentionally constructed with 0.5, 1.5, and 2.0 repeat units by Monte Carlo sampling a di-peptide backbone to fit in,

instead of using 1 unit spacer to create peptide backbone diversity. Rosetta FastRelax was performed to maintain their lower energy states.

With the 20 target sequences in place with the calculated backbone ensembles, the interacting motifs from the di-peptide pockets were also placed accordingly using two strategies. 1) A small continuous helical chunk on the protein (with loopy regions occasionally included); or 2) individual interacting amino acids distributed on the protein. RFdiffusion was applied to connect these motifs to be a single chain A protein, while in the context of chain B target being present. As for the gap regions between the motifs (for RFdiffusion to connect and fill in), AA length ranged from 0.75x to 1.25x of the original AA length from the parent repeat-motif binder.

A thousand trajectories were carried out for each strategy of each design task. Two sequences per backbone were generated by ProteinMPNN. All sequences were then validated through complex predictions (AF2 with initial guess[5], PAE_interaction ≤10, PLDDT > 92 followed by AF multimer[6]) and monomer predictions (AF2 on chain A only).

Assemble to target arbitrary sequences beyond pockets

To go beyond existing pockets and motifs, we tested a set of eight arbitrary English words and human names ranging from eight to eighteen amino acids, i.e., DANIELSILVA, HIALIENFRIEND, DERRICKHICKS, DAVIDLIKESPEPTIDES, DAVIDITISTIME, WILLCHEN. To target arbitrary targets, we dissected them as individual amino acids instead of di-peptide motifs, and evaluated the assembled binders as a string of individual pockets. To do this, we first analyzed and compared the sequence identity of ‘the random words’ and ‘motif combos’ by BLOSUM62[16], mapping each new target to the closest set of existing combos. This way, for two ‘unfamiliar targets,' HIALIENFRIEND and WILLCHEN (BLOSUM scores < -2.0), indicating there were no good matches in the current small set of libraries (~70 template pairs). We then assembled a set of 20 new templates following the above pocket assembly, with the purpose of targeting 70% of the target sequence (as we didn’t have full amino acid pockets by then).

For all the identified and newly assembled templates, we threaded the target into the template chain B backbone, performed chain B local backbone resampling through FlexPepDock[2], filtered out the backbone clashed docks and the docks losing more than 30% of the backbone hydrogen bonding. Survivors were then sent to customized ProteinMPNN for sequence design and customized AF2, as described below in detail in the ‘IDR threading’ section. Since this set of targets was new, we used several strategies of diffusion refinement on most of the designs as described below in detail in the ‘Refinement’ section.

Two-sided interface design

To increase the diversity of target sequences in fusion protein and multi-motif scaffolding binder generation, two-sided interface design was employed at the last stage. AF2 passing complexes were taken back, with both chain A and chain B redesigned through ProteinMPNN using the exact same protocol above of two iterations. Designs with DDG < -50 were sent to the same AF2 prediction. Out of this round at chain B sequence diversification (without intentionally modifying the backbones), 267 AF2 passing new targets were generated in silico, which could suggest the potential of this platform in general sequence recognition.

Experimental Methods

Gene Construction of Designed Binders

The designed protein sequences were optimized for efficient expression in *E. coli*. linear DNA fragments encoding these design sequences were obtained (eBlocks, Integrated DNA Technologies) and included overhangs compatible with Golden Gate cloning into the LM670 vector (Addgene #191552) for *E. coli* protein expression. The LM670 vector is a modified expression system featuring a Kanamycin resistance gene, a ccdB lethal gene flanked by BsaI cut sites, and a C-terminal hexahistidine (6xHis) tag. Peptide genes were purchased as fusion proteins to either the C terminus of sfGFP or the N terminus of a GB1−AviTag−His6x construct separated by (PAS) linker or (GGSGSG) linker.

Protein production and purification (small scale and medium scale)

Linear gene fragments encoding the binder design sequences were cloned into the LM670 vector using Golden Gate assembly. These subcloning reactions were performed in 96-well PCR plates with a 4 µL reaction volume. Following this, 1 µL of the reaction mixture was transformed into chemically competent *E. coli* BL21 (DE3) cells.

After a 1-hour recovery period in 100 µL of SOC medium, the transformed cell suspensions were transferred directly into a 96-deep well plate containing 900 µL of LB media supplemented with Kanamycin. After overnight incubation at 37°C, 100 µL of the growth culture was inoculated into 96-deep well plates containing 900 µL of auto-induction media (autoclaved TBII media with Kanamycin, 2 mM $MgSO_4$, 1X 5052). The cultures were then expressed overnight at room temperature.

Cells were harvested by centrifugation at 4000 x g for 15 minutes. The bacterial pellets were lysed in 100 µL of lysis buffer (1X BugBuster (Millipore #70921-4), 0.01 mg/mL DNase, 1 Pierce protease inhibitor tablet per 50 mL lysis) for 30 minutes on shaker, 220 RPM. The lysates were spun down by centrifugation at 4000 x g for 10 minutes, followed by purification using Ni-charged MagBeads (GenScript #L00295). The wash buffer contained 25 mM Tris pH 8.0, 300 mM NaCl, and 10 mM Imidazole, while the elution buffer contained 25 mM Tris pH 8.0, 300 mM NaCl, and 500 mM Imidazole. 250 µL of wash buffer was used to wash twice while 120 µL of elution buffer was used to collect eluted proteins. Filtered elutions were then submitted to HPLC, S200. For samples showing major monomeric peaks, protein concentrations were determined by measuring absorbance at 280 nm with a NanoDrop spectrophotometer (Thermo

Scientific), using extinction coefficients and molecular weights calculated from their amino acid sequences.

For further validation, proteins were expressed at a 50 mL scale using autoinduction for approximately 24 hours. During the first 6 hours, cultures were grown at 37°C, followed by incubation at 22°C for the remaining time. Cultures were harvested by centrifugation at 4000 x g for 10 minutes and resuspended in approximately 20 mL of lysis buffer (25 mM Tris-HCl, 150 mM NaCl, 0.1 mg/mL lysozyme, 0.01 mg/mL DNase, 1 mM PMSF, and 1 Pierce protease inhibitor tablet per 50 mL culture). Sonication was performed with a four-prong head for a total of 5 minutes, with 10-second on/off pulses at 70% amplitude. The resulting lysate was clarified by centrifugation at 14000 x g for 30 minutes. The lysate supernatants were applied directly to a 1 mL bed of Ni-NTA agarose resin equilibrated with a binding buffer. After sample application and flow-through, the resin was thoroughly washed two times, and samples were eluted with an elution buffer containing 500 mM imidazole. Post-elution, protein samples were filtered and injected into an Akta Pure system equipped with an autosampler, using a Superdex S75 Increase 10/300 GL column at room temperature. The SEC running buffer consisted of either 25 mM Tris-HCl, 150 mM NaCl, pH 8.0 or 1x PBS. Protein concentrations were determined by measuring absorbance at 280 nm with a NanoDrop spectrophotometer (Thermo Scientific), using extinction coefficients and molecular weights calculated from their amino acid sequences.

BLI assays for initial screening and affinity measurement

As for quantitatively screening all designs, Bio-Layer Interferometry (BLI) was employed (detailed method described below). These samples were screened by BLI with concentrations ranging from 200 nM to 2 μM. Designs exhibiting obvious binding signals (>= 0.2 AU) were further analyzed by BLI through titration (with double controls) and 50 mL expression.

BLI experiments were conducted using an Octet Red96 (ForteBio) instrument with streptavidin-coated tips (Sartorius Item no. 18-5019). The assay buffer was 1X HBS-EP+ (Cytiva BR100669) supplemented with 0.1-0.2% w/v bovine serum albumin and 0.1% sucrose. Each design was initially tested for non-specific binding against unloaded tips. Biotinylated target peptides (50-200 nM) were loaded onto the tips for 50-300 seconds to reach their 1/3 to 2/3 maximal loading signals (conducted by a loading test in advance), followed by a 60-second baseline measurement. After loading, all runs underwent a 60-second baseline, 100-800 seconds association, and 100-700 seconds dissociation. Baseline measurements from unloaded tips were subtracted from their corresponding loaded tip measurements. Twofold or Threefold serial dilutions were

performed during the titration, with double controls in each run (i.e., one as Octet buffer to the loaded tip for signal subtraction, one as the highest tested protein concentration to the non-loaded tip to check for non-specific binding). For some runs toward the same target, streptavidin tips were regenerated five times in HCl pH 1.0 buffer for three 5-second exposures after each single run. For each individual titration run, the target and binder concentrations, association and dissociation times were adjusted as needed. Steady-state and global kinetic fits were carried out using the manufacturer's software, Data Analysis 9.1, based on the assumption of a 1:1 binding model.

BLI all by all orthogonality matrix

BLI experiments were performed as described above for individual binding measurements. Here, each of the 20 biotinylated-targets was loaded onto the streptavidin-coated tips again to reach their 1/3 to 2/3 maximal loading signals. Each target was tested against its cognate designed binder and other 19 noncognate binder individually at titration ranges from 1 uM to 4 nM. For the same target, streptavidin tips were regenerated in HCl pH 1.0 buffer for three 5-second exposures after each single run. Titration experiments were conducted at 25°C with continuous rotation at 1,000g. Association and dissociation with various binders were allowed for 200-1,000 seconds for each step. All binding signals at 1 uM were normalized by their cognate pairs and converted for heatmap plotting.

NanoBiT assays for initial binding screen and all by all orthogonality matrix

To qualitatively screen designs, the split-luciferase assay (nanoBiT) was carried out using the Nano-Glo Luciferase Assay System (Promega). The coding sequence for the small-BiT was fused to the gene encoding the peptide binders, while the large-BiT coding sequence was fused to the gene encoding the target peptide. The BiT-fused proteins and peptides were expressed and purified following the same protocol as above.

All assays were conducted in a buffer containing 20 mM sodium phosphate, 100 mM NaCl, pH 7.4, and 0.05% v/v Tween 20. Reactions were set up in 96-well plates (Corning, cat. no. 3686) with Nano-Glo substrate (Promega, cat. no. N1130), diluted 500× for endpoint measurements. Luminescence signals were recorded using a Synergy Neo2 plate reader (BioTek).

Equilibrium binding assays were performed with one component (i.e., the target peptide) held constant at 1 nM or 0.5 nM while the other protein (i.e., the designed binder) was titrated. Serial dilutions were prepared over four points with a one-quarter dilution factor between each step. The highest concentration point of the curve was provided by the

purified protein, normally around 500 nM to 2 uM. The plates were incubated overnight at room temperature before adding the substrate and immediately measuring luminescence. Constructs showing obvious binding (high luminescence signals) at relatively lower concentrations were identified as initial hits for Octet characterization to further determine the binding affinity.

During the all-by-all (18x18) orthogonality screen, all the target peptides were held constant at 0.5 nM, while the designed binders were at 100 nM. The plates were incubated overnight at room temperature before adding the substrate and immediately measuring luminescence. Each experiment was repeated three times, the averaged luminescence signals were used for plotting.

Protein Purification for Crystallography

Constructs were transformed into LEMO21 or NEB BL21(DE3) *E. coli* and expressed as 0.5 L cultures in 2L flasks. Proteins were produced using Studier's M2 autoinduction media containing 50 µg/mL kanamycin. Pre-cultures were grown at 37°C for 4 hours, then shifted to 22°C for 14 hours before being used to inoculate the main cultures with 10 mL of pre-culture. After growth, cells were harvested by centrifugation at 4000 x g for 10 minutes, and the supernatant was discarded. The cell pellets were resuspended in 40 mL of lysis buffer (100 mM Tris-HCl pH 8, 100 mM NaCl, 400 mM imidazole, 1 mM PMSF, 1 mM DNase). Cells were lysed using a Microfluidics M-100P microfluidizer at 18,000 psi, and the lysate was clarified by centrifugation at 14,000 x g for 30 minutes.

His-tagged proteins were bound to 8 mL of Ni-NTA resin (Qiagen) using gravity flow. The bound proteins were washed with 10 mL of lysis buffer followed by 30 mL of high salt wash buffer (25 mM Tris-HCl pH 8, 1 M NaCl, 20 mM imidazole), and then with 10 mL of SNAC cleavage buffer (100 mM CHES, 100 mM acetone oxime, 150 mM NaCl, 500 mM GnCl, pH 8.6). To initiate cleavage, 40 mL of SNAC cleavage buffer and 80 µL of 1 M NiCl2 were added, and the columns were sealed and shaken on a nutator for 12 hours. After cleavage, the flowthrough was collected and concentrated prior to further purification using SEC/FPLC on a HiLoad 20/600 Superdex 75 pg column in TBS (20 mM Tris pH 8.0, 150 mM NaCl), with 14 mL fractions collected between 100 and 290 mL.

TAMRA Peptide Synthesis

Peptides were synthesized in-house using a CEM Liberty Blue microwave synthesizer. All amino acids were sourced from P3 Biosystems. Oxyma Pure was obtained from CEM, DIC from Oakwood Chemical, and diisopropyl ethylamine (DIEA) and piperidine from Sigma-Aldrich. Dimethylformamide (DMF) was procured from Fisher Scientific and

pre-treated with an Aldraamine trapping pack. The 5(6)-carboxytetramethylrhodamine carboxylic acid (5(6)-TAMRA) was purchased from Novabiochem.

The synthesis was conducted on a 0.1 mmol scale using CEM Cl-MPA resin. Each amino acid, at a concentration of five equivalents, was activated with 0.1 M Oxyma and 2% (v/v) DIEA in DMF, combined with 15.4% (v/v) DIC. Coupling was performed twice on the resin for 2 minutes per coupling with microwave irradiation. For TAMRA-labeled peptides, the peptides were washed with DMF post-synthesis, then incubated for 3 hours with 5(6)-TAMRA carboxylic acid (3 equivalents), HATU (3 equivalents), and DIEA (5 equivalents) in DMF. After incubation, the peptides were washed with DMF (three times) and DCM (three times) in preparation for global deprotection.

Global deprotection was achieved using a mixture of TFA/water/TIPS/2,2'-(ethylenedioxy) diethanethiol (92.5:2.5:2.5:2.5) for 3 hours. This deprotection mixture was concentrated in vacuo to 2-3 mL, then precipitated in 30 mL of ice-cold ethyl ether, centrifuged, and decanted. The peptide was washed twice more with fresh ether and dried under nitrogen to yield the crude peptide, which was then purified by high-pressure liquid chromatography (HPLC). The crude peptide was dissolved in a minimal amount of ACN and water to ensure complete solubility. Purification was performed on a Zorbax Stablebond C18 (9.4 x 250 mm, 5 µm) column using an Agilent 1260 Infinity HPLC with a linear gradient of water (0.1% TFA) and increasing ACN (0.1% TFA). UV signals were monitored at 214 nm and all peaks were collected.

Peak masses were verified using an Agilent G6230B LC-MS, and purity was assessed with a C18 column (Higgins Analytical PROTO 300 C18, 10 µm, 10 x 250 mm) on an analytical Agilent 1260 Infinity II HPLC.

Biotinylated Peptide Synthesis

All Fmoc-protected amino acids were sourced from P3 Bio. The synthesized biotinylated peptides were modified at the N terminus with biotin-Ahx-GGGS, using biotin-Ahx building blocks also obtained from P3 Bio. Oxyma was procured from CEM, and DIC was acquired from Oakwood Chemicals. Dimethylformamide (DMF) was purchased from Fisher Scientific and treated with an AldraAmine trapping pack from Sigma-Aldrich prior to use. Piperidine was also purchased from Sigma-Aldrich. Cl-TCP(Cl) resins were obtained from CEM.

Peptides were synthesized on a 0.1 mmol scale using microwave-assisted solid-phase peptide synthesis on a CEM Liberty Blue system. After synthesis, the peptides were cleaved using a cocktail of trifluoroacetic acid (TFA), TIPS, water, and DODT in a 92.5:2.5:2.5:2.5 ratio. The cleavage solution was concentrated under vacuum,

precipitated into cold ether, and centrifuged. The resulting pellet was washed and centrifuged again with ether (twice), dried under nitrogen, resuspended in water and acetonitrile (ACN), and purified by reverse-phase high-performance liquid chromatography (RP-HPLC) using an Agilent 1260 Infinity semi-preparative system with a gradient from 20% to 70% over 15 minutes (A: H2O with 0.1% TFA; B: ACN with 0.1% TFA). The purified peptide fractions were combined, lyophilized, and weighed in a tared scintillation vial.

Depending on their isoelectric points, lyophilized peptides were solubilized in buffers containing either 100 mM Tris pH 8.0 or 100 mM MES pH 6.5 and stored at −20°C.

Fluorescence Polarization

All fluorescence polarization (FP) measurements were conducted at 25°C in 96-well plates (Corning 3686) using a Synergy Neo2 plate reader equipped with a 530/590 nm filter cube. Peptide targets were synthesized with N-terminal tetramethylrhodamine labels. The buffer for all FP measurements consisted of 20 mM Tris-HCl, 100 mM NaCl, and 0.05% v/v TWEEN20, adjusted to pH 8. Titrations were performed in a 96-well format with four replicates per plate, comprising either 24 or 48 data points per titration (including 23 or 47 steps of two-fold serial dilution of designed binders in the presence of a TAMRA-labeled peptide at a constant concentration between 0.1 nM and 1 nM), with a final sample volume of 80 µL per well. To ensure complete equilibration, the titration plates were incubated from 3 hours to overnight at room temperature before measurements were taken.

The polarization signal SSS (as calculated by the Neo2 software) was fitted to the equation: $$S = S_0 + S_1 \cdot f_{AB}$$

where:

$$f_{AB} = \frac{1}{2B_{\text{tot}}} \left(A_{\text{tot}} + B_{\text{tot}} + K_D - \sqrt{(A_{\text{tot}} + B_{\text{tot}} + K_D)^2 - 4 \cdot A_{\text{tot}} \cdot B_{\text{tot}}} \right)$$

Here, f_{AB} represents the fraction of the peptide that is bound, A_{tot} is the absolute concentration of the hinge, B_{tot} is the absolute concentration of the peptide, S_0 is the baseline polarization of the free peptide, and S_1 is the change in polarization upon complex formation.

The fitting process was performed using the `scipy.optimize.curve_fit` function in Python. The uncertainties for K_D values were calculated as standard deviation errors derived from the covariance matrix of the fits. In instances where the fitted K_D was lower than the concentration of the labeled peptide B_{tot}, the K_D was reported as $K_D < B_{tot}$.

Individual design characterizations

Purification of KFAM binders

Plasmids encoding FAM21 binders tagged with an N-terminal PC tag and C-terminal 10x his tag were transformed into BL21(DE3) *E. coli*. Bacteria were cultured at 37°C in 2xTY medium to an optical density at 600 nm of 0.8 and then induced using 0.5 mM isopropylthiogalactoside (IPTG) at 20°C for at least 16 h. Cells were pelleted using centrifugation. Unless stated otherwise, all purification steps were performed at 4°C. Cell pellets were resuspended in lysis buffer (25 mM HEPES pH 7.2, 500 mM NaCl, 10% glycerol, 1 mM CaCl2, 100 ug/ml lysozyme, 1 mM DTT, and cOmplete protease inhibitor cocktail tablets (Roche)) and lysed by sonication. Lysates were spanned for 1 hour at 18,000 rpm and the supernatant containing soluble proteins was incubated with pre-equilibrated Protein C affinity resin for 2 hours. The resin was then packed into an empty column (Bio-Rad), washed with high salt wash buffer (25 mM HEPES pH 7.2, 500 mM NaCl, 10% glycerol) and then with low salt wash buffer (25 mM HEPES pH7.2, 150 mM NaCl, 10% glycerol). Binders were eluted with an elution buffer (25 mM HEPES pH 7.2, 150 mM NaCl, 10% glycerol, 5 mM EGTA). Binders were then concentrated and buffer exchanged into storage buffer (25 mM HEPES pH 7.2, 150 mM NaCl, 10% glycerol, 1 mM CaCl2) using an Amicon Ultra 4 ml 10kDa cutoff centrifugal filter unit.

Immunoprecipitation with KFAM binders

Immunoprecipitation steps were performed at 4°C unless otherwise stated. PC tagged FAM21 binders in storage buffer (or buffer only as a control) were incubated with pre-equilibrated Protein C affinity resin for 1 hour. Unbound proteins were removed by washing the resin three times with a storage buffer. HEK293 cells were lysed in 25 mM HEPES pH 7.2, 150 mM NaCl, 10% glycerol, 0.5% NP40, 1 mM DTT and cOmplete protease inhibitor cocktail tablets (Roche). Lysate was spun on a tabletop centrifuge at 14,000 rpm for 10 minutes at 4°C. Equal volumes of supernatant were added to the binder-Protein C resin, an input was taken for analysis and then binder:Protein C resin:HEK293 lysate was incubated for 16 hours. Resin was then gently pelleted (2,000 rpm, 1 min) and washed three times with 25 mM HEPES pH7.2, 150 mM NaCl, 10% glycerol, 1 mM DTT. After removing the final wash buffer, beads were resuspended in

2x Nupage LDS Sample Buffer + 2.5% beta-mercaptoethanol and beads were boiled at 98°C for 10 minutes.

Inputs and elution samples were run on 4–12% Bis-Tris gels (Invitrogen NP0323BOX) and transferred to a nitrocellulose membrane using the iBlot system (Thermo Fisher Scientific). Membranes were blocked in 5% (w/v) milk in TBS-TWEEN (10 mM Tris-HCl, 120 mM NaCl and 1% (w/v) TWEEN20, pH 7.4) for 30 min at room temperature with gentle shaking. Mouse anti-Strumpellin (SantaCruz, sc-377146), Rabbit anti-FAM21 (Gift from Dan Billadeau), Rabbit anti-WASH (Gift from Dan Billadeau), Rabbit anti-CCDC53 (Merck, ABT69) were diluted in 1% (w/v) BSA in TBS-TWEEN and incubated with the membrane overnight at 4°C with gentle shaking. The membrane was washed three times in TBS-TWEEN then incubated with goat anti-rabbit Alexa 647 (Invitrogen, A-21246, 1:2,000) or donkey anti-Mouse Alexa 488 (Invitrogen, A-21202, 1:2,000)for 1 hour at room temperature with gentle shaking. The membrane was washed twice with TBS-TWEEN, followed by a final wash with TBS-TWEEN with 0.001% SDS. Membranes were imaged using a ChemiDoc system (BioRad). Alternatively, the same samples were analyzed using 4–12% Bis-Tris gels (Invitrogen NP0323BOX) and stained with InstantBlue Coomassie stain (Sigma ISB1L) for total protein staining.

Dynorphin A inhibition cAMP assay

Functional cAMP assay was performed according to previously described protocols[3–5](https://www.zotero.org/google-docs/?RsHTSR) and using Chinese hamster ovarian (CHO) cells stably expressing human KOR. Briefly, 3,000 cells per 5 µL per well were seeded into a white 384-well plate and incubated with 5 µL of varying concentrations of binders prepared (4×) in a 1× stimulation buffer. The reaction mixture was incubated at 37°C for 45 min. Following, 5 uL of an EC80 concentration of dynorphin A 1-17 and 5 uL of forskolin (1 µM final) were added and incubated at 37°C for an additional 45 min. After adding 10 uL of Europium cryptate-labeled cAMP and cAMP d2-labeled antibody, respectively, and incubating the reaction mixture for 1 hour at room temperature, cAMP quantification was determined by measuring homogeneous time-resolved fluorescence resonance energy transfer on Neo2 plate reader using a ratio of 665/620 nm. The concentration response curve for dynorphin A 1-17 was generated in analogy to measuring binder antagonism with the exception of using 10 uL of cells per well.

Affinity enrichment of CTN4 analyzed by LC-MS

Designed 6xHis-tagged protein binders were conjugated to Dynabeads™ His-Tag Isolation and Pulldown (Invitrogen, 10103D) according to the manufacturer's instructions. Functionalized magnetic beads (MBs) were subsequently blocked with 1% (w/v) bovine serum albumin.

Synthesized peptide CTN4 (NWLTIFILFPLK) was custom ordered from GenScript (NJ, USA). For affinity enrichment, per 2 nmol protein binders in 30 uL beads slurry were used for capturing per 0.2 nmol CTN4. Binder-MBs were added to either 200 µL spiked pull-down buffer (50 mM ammonium bicarbonate pH 8.0, 1 mM DTT, 0.1% Triton X-100) or 200 µL spiked dried blood spot (DBS) extract in same buffer and incubated overnight with rotation at 4°C. After incubation, the supernatant were removed and the beads were washed with 200 µL phosphate-buffered saline for three times. The beads were then suspended with 30 µL elution buffer (30% acetonitrile, 0.1% formic acid) and incubated for 10 minutes at room temperature with shaking at 250 rpm to elute the peptides, and the peptides were then removed by magnetic separation.

0.5 nmol CTN4 were added to the pull-down/elution buffer in the same volume as the experiment setting to make peptide standards representing 100% loss/recovery. In the peptide wash & loss experiment, BSA-blocked non-functionalized beads were used as the negative control to quantify non-specific binding. Aliquots of both peptide standards and samples were diluted in LC mobile phases prior to injection.

Table 1. MS source parameters for the Waters Xevo-TQS Micro instrument.

Parameter (units)	Value
Source Polarity	ESI+
Capillary Voltage (V)	3500
Cone Voltage (V)	35
Source temperature (°C)	150
Desolvation temperature (°C)	500
Cone gas flow (L/hr)	20
Desolvation gas flow (L/hr)	1000

Table 2. Liquid chromatography conditions

Parameter (units)	Value
Mobile phase	Phase A: water with 0.1% formic acid Phase B: acetonitrile with 0.1% formic acid
Column	ACQUITY UPLC BEH C18 Column, 130Å, 1.7 µm, 2.1 mm x 50 mm

Temperature (°C)	22
Flow rate (mL/min)	0.4
Injection volume (µL)	10
Gradient	The gradient started with 5% B, increased linearly to 60% B at 2.0 min, then linearly to 100% B at 2.30 min, then switched back to 5% B at 2.50 min for re-equilibration, with a flow rate of 0.4 mL/min.

NMR sample preparation and data collection of dynorphin-binder complex data

The plasmid encoding for 6xHis-SUMO-dynorphin was purchased from Genscript (based on the pET15_SUMO2_NESG vector) and expressed in *E. coli* BL21 (DE3) grown in MJ9 media with $^{15}NH_4SO4$ and ^{13}C-glucose[17] and 50 ug/ml Carbencillin. Briefly, 1 L of MJ9 media was inoculated with 25 ml of overnight grown starter culture in an incubator shaker set at 37^{o}C, 200 RPM. Bacterial culture was induced with 0.25 mM of IPTG at 18^{o}C and left on shaking for 18 hrs. Next day, cells were pelleted down by centrifugation and resuspended in Tris buffer (50 mM Tris-HCl pH 7.5, 500 mM NaCl) with 5 mM imidazole, followed by mild sonication. The lysate was clarified by centrifugation and loaded onto an IMAC column (Cytiva 5 ml HisTrap HP column). After washing and eluting with Tris buffer containing 250 mM imidazole, the 6xHis-SUMO-dynorphin was further purified using SEC (Cytiva HiLoad 16/600 Superdex 75 pg) on an ÄKTA pure chromatography system (Cytiva) followed by overnight cleavage with SUMO protease at room temperature to remove the His_6-SUMO tag. Solution containing Sumo Protease, sumo tag, and cleaved dynorphin was lyophilized, resuspended in cold methanol, and centrifuged at high speed. Further methanol containing dynorphin was vacuum dried, similar to a previously described peptide purification protocol[18]. Final purification was done using reverse-phase chromatography with an acetonitrile/water gradient (Resource RPC, 3 mL, Cytiva) on an ÄKTA pure system as described previously for dynorphin A[19], and the molecular weight was confirmed using MALDI-TOF mass spectrometry. Eluted samples were lyophilized and stored at -80^{o}C, and resuspended to 1-4 mM (300 ul) in 20 mM phosphate buffer pH 6.5, 150 mM NaCl prior to complex formation.

His_6-tagged protein binders DYNA_1b7 and DYNA_2b2 were cultured in LB broth, and purified using IMAC and SEC, similarly as for dynorphin A. After elution into 20 mM phosphate buffer pH 6.5, 150 mM NaCl, samples were concentrated to 20-100 uM using Vivaspin centrifugal concentrators (Sartorius, 5 kDa MWCO). Binder-dynorphin complexes were prepared by mixing $^{15}N^{13}C$-labeled dynorphin to a solution of purified binders. After incubation overnight at 4°C, the complex was further purified by SEC (Superdex 75) with the same buffer, and the monomer elution fraction was concentrated

to 100-250 uM by centrifugal filter concentration. The sizes and stability of dynorphin and its binder were confirmed by MALDI-TOF MS. NMR samples were prepared with 5% D_2O, and 300 uL was loaded into 5 mm Shigemi tubes.

NMR spectra were collected at 25 or 35°C with a 5 mm TCI CryoProbe on a Bruker Avance III 600 MHz or Avance Neo 800 MHz spectrometer. Chemical shift assignments for dynorphin complexes were obtained from standard spectra including 1D ^{1}H spectra, 2D ^{1}H-^{15}N SOFAST HMQC and ^{1}H-^{13}C HSQC (aliphatic and aromatic) spectra, and 2D (free) or 3D (bound to DA7 or 1C10) HNCO, HNCA, HNcoCA, HNCACB, ^{15}N-edited NOESY (100 ms), and ^{1}H-^{13}C hCCH TOCSY (11 ms) spectra. 3D 15N-edited NOESY data was collected at 800MHz with a non-uniform sampling rate of 50%. Chemical shift assignments for free dynorphin (4 mM) were obtained from 2D HC faces of standard 3D HNCA, HNcoCA, HNCACB, HNcoCACB, HNCO, ^{15}N-edited HccONH and CccoNH spectra and HH faces of ^{15}N-edited TOCSY (60 ms) and ^{1}H-^{13}C hCCH TOCSY (11 ms) as 3D data was not required. All spectra were referenced to internal DSS.
Spectra were processed using TopSpin 4.0 (Bruker) or NMRPipe software[20] and visualized using NMRFAM-SPARKY[21]. Non-uniform sampling data were processed using the SMILE package[22] package integrated with NMRPipe. Secondary structure propensities were calculated using TALOS-N from NMR backbone (and CB) chemical shifts[23] for unbound dynorphin, DA7-, and 1C10-bound. Positive (green) bars are beta-strand, and negative (gold) indicate helical propensity, with predicted beta-strand secondary structure shown as green arrows above this data. Chemical-shift based RMSF (Å) was calculated by scaling the random coil index by 12.7[24,25].

X-ray crystallography

Crystallization experiments were conducted using the sitting drop vapor diffusion method. Initial crystallization trials were set up in 200 nL drops using the 96-well plate format at 20˚C. Crystallization plates were set up using a Mosquito LCP from SPT Labtech, then imaged using UVEX microscopes and UVEX PS-256 from JAN Scientific. Diffraction quality crystals formed in 0.1 M Phosphate/citrate pH 4.2 and 40 % v/v PEG 300 for DNYA_1b7-1. Diffraction quality crystals formed in 0.1 M Sodium acetate pH 4.6 and 25% (v/v) PEG 550 MME for DA7-2.

Diffraction data were collected at the National Synchrotron Light Source II on beamline 17-ID-1 (AMF) for DYNA_1b7-1 and at the Advanced Light Source beamline 821 for DA7-2. X-ray intensities and data reduction were evaluated and integrated using XDS[31] and merged/scaled using Pointless/Aimless in the CCP4 program suite[27]. Structure determination and refinement starting phases were obtained by molecular replacement

using Phaser[27] using the designed model for the structures. Following molecular replacement, the models were improved using phenix.autobuild[28]; with rebuild-in-place to false, and using simulated annealing. Structures were refined in Phenix[28]. Model building was performed using COOT[29]. The final model was evaluated using MolProbity[30]. Data collection and refinement statistics are recorded in Table 3. Data deposition, atomic coordinates, and structure factors reported in this paper have been deposited in the Protein Data Bank (PDB), http://www.rcsb.org/ with accession code 9CCE and 9CCF.

Supplementary table 3. Data collection and refinement statistics.

	DYNA_1b7 (PDB Code: 9CCE)	DA7-2 (PDB Code: 9CCF)
Resolution range	28.84 - 3.15 (3.23 - 3.15)	47.76 - 4.00 (4.47 - 4.0)
Space group	*P* 2_1	*P* 2_1
Unit cell	44.43, 67.78, 68.36; 90.00, 98.03, 90.00	43.42, 67.21, 68.36; 90.00, 96.75, 90.00
Unique reflections	6972 (517)	3361 (941)
Multiplicity	3.1 (3.1)	5.0 (5.1)
Completeness (%)	98.3 (96.8)	99.5 (99.8)
Mean I/sigma(I)	5.6 (1.3)	11.4 (6.5)
Wilson B-factor	84.85	129.92
R-merge	0.138 (0.658)	0.047 (0.118)
R-pim	0.102 (0.498)	0.047 (0.119)
$CC_{1/2}$	0.989 (0.790)	0.997 (0.979)
Reflections used in refinement	6951 (511)	3343 (900)

R-work	0.2626 (0.3187)	0.24.23 (0.2526)
R-free	0.3080 (0.3800)	0.29.63 (0.3408)
Number of non-hydrogen atoms	3355	3221
macromolecules	3355	3221
Protein residues	410	394
RMS(bonds)	0.002	0.002
RMS(angles)	0.42	0.429
Ramachandran favored (%)	96.97	96.83
Ramachandran allowed (%)	3.03	3.17
Ramachandran outliers (%)	0.00	0.00
Average B-factor	81	126
macromolecules	81	126

The highest-resolution shells are shown in parentheses.

Acknowledgements:

This research used resources (FMX/AMX) of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. The Center for BioMolecular Structure (CBMS) is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011).

Reference

1. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

2. London, N., Raveh, B., Cohen, E., Fathi, G. & Schueler-Furman, O. Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions. *Nucleic Acids Res.* **39**, W249–W253 (2011).

3. Marcu, O. *et al.* FlexPepDock lessons from CAPRI peptide–protein rounds and suggested new criteria for assessment of model quality and utility. *Proteins* **85**, 445–462 (2017).

4. Dauparas, J. *et al.* Robust deep learning based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

5. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).

6. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. 2021.10.04.463034 Preprint at https://doi.org/10.1101/2021.10.04.463034 (2022).

7. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

8. Vázquez Torres, S. *et al.* De novo design of high-affinity binders of bioactive helical peptides. *Nature* **626**, 435–442 (2024).

9. Coventry, B. & Baker, D. Protein sequence optimization with a pairwise decomposable penalty for buried unsatisfied hydrogen bonds. *PLOS Comput. Biol.* **17**, e1008061 (2021).

10. Wu, K. *et al.* De novo design of modular peptide-binding proteins by superhelical matching. *Nature* **616**, 581–589 (2023).

11. Reichen, C. *et al.* Computationally Designed Armadillo Repeat Proteins for Modular

Peptide Recognition. *J. Mol. Biol.* **428**, 4467–4489 (2016).

12. Reichen, C., Hansen, S. & Plückthun, A. Modular peptide binding: From a comparison of natural binders to designed armadillo repeat proteins. *J. Struct. Biol.* **185**, 147–162 (2014).

13. Zeytuni, N. & Zarivach, R. Structural and Functional Discussion of the Tetra-Trico-Peptide Repeat, a Protein Interaction Module. *Structure* **20**, 397–405 (2012).

14. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

15. Jiang, H. *et al.* De novo design of buttressed loops for sculpting protein functions. *Nat. Chem. Biol.* 1–7 (2024) doi:10.1038/s41589-024-01632-2.

16. Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.* **22**, 1035–1036 (2004).

17. Jansson, M. *et al.* High-level production of uniformly 15N-and 13C-enriched fusion proteins in Escherichia coli. *J. Biomol. NMR* **7**, 131–141 (1996).

18. Kumar, P. & Aradhyam, G. K. Easy and efficient protocol for purification of recombinant peptides. *Protein Expr. Purif.* **95**, 129–135 (2014).

19. O'Connor, C. *et al.* NMR structure and dynamics of the agonist dynorphin peptide bound to the human kappa opioid receptor. *Proc. Natl. Acad. Sci.* **112**, 11852–11857 (2015).

20. Delaglio, F. *et al.* NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).

21. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).

22. Ying, J., Delaglio, F., Torchia, D. A. & Bax, A. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR* **68**, 101–118 (2017).

23. Shen, Y. & Bax, A. Protein Structural Information Derived from NMR Chemical Shift with the Neural Network Program TALOS-N. *Methods Mol. Biol. Clifton NJ* **1260**, 17–32 (2015).

24. Berjanskii, M. V. & Wishart, D. S. A Simple Method To Predict Protein Flexibility Using Secondary Chemical Shifts. *J. Am. Chem. Soc.* **127**, 14970–14971 (2005).

25. Berjanskii, M. V. & Wishart, D. S. The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. *Nucleic Acids Res.* **35**, W531–W537 (2007).

26. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

27. (IUCr) Phaser crystallographic software. https://journals.iucr.org/j/issues/2007/04/00/he5368/index.html.

28. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

29. (IUCr) Coot: model-building tools for molecular graphics. https://journals.iucr.org/d/issues/2004/12/01/ba5070/index.html.

30. MolProbity: More and better reference data for improved all‐atom structure validation - Williams - 2018 - Protein Science - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/pro.3330.

31 . Kabsch, W. XDS. Acta Crystallogr. D Biol. Crystallogr. 66, 125–132 (2010).

Chapter 3 – Diffusing protein binders to intrinsically disordered proteins

3.0 – preface, authors, and abstract

Preface:

Note: this chapter is borrowed directly from the manuscript under review of the same name, which is the original copy. I, Kejia Wu, am co-first and co-corresponding author for this work.

Authors:

Caixuan Liu†,1,2, Kejia Wu†,*,1,2,3, Hojun Choi†,1,2, Hannah Han‡,1,2, Xulie Zhang‡,6,7, Joseph L.Watson‡,1,2, Sara Shijo‡,5, Asim K. Bera1,2, Alex Kang1,2, Evans Brackenbrough1,2, Brian Coventry1,2, Derrick R. Hick1,2, Andrew N. Hoofnagle5, Ping Zhu6,7, Xingting Li1,2, Justin Decarreau1,2, Stacey R. Gerben1,2, Wei Yang1,2, Xinru Wang1,2, Mila Lamp1,2, Analisa Murray1,2, Magnus Bauer1,2, David Baker*,1,2

†Equal contribution

‡ Equal contribution

*To whom correspondence should be addressed: kejiawu@uw.edu, dabaker@uw.edu

1.Department of Biochemistry, University of Washington, Seattle, WA, USA.

2.Institute for Protein Design, University of Washington, Seattle, WA, USA.

3. Biological Physics, Structure and Design Graduate Program, University of Washington, Seattle,

WA, USA

4.Graduate Program in Molecular Engineering, University of Washington, Seattle, WA 98105, USA

5. Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, 98105, USA.

6.Key Laboratory of Epigenetic Regulation and Intervention, Institute of Biophysics, Chinese Academy of Sciences Beijing, 100101, China
7.University of Chinese Academy of Sciences, Beijing 100049, China
Keywords
Intrinsically disordered protein, intrinsically disordered region, Amyloid fibril dissociation, diagnostics, protein design, RFdiffusion, Rosetta, deep learning

Abstract

Proteins which bind intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) with high affinity and specificity could have considerable utility for therapeutic and diagnostic applications. However, a general methodology for targeting IDPs/IDRs has yet to be developed. Here, we show that starting only from the target sequence of the input, and freely sampling both target and binding protein conformation, RFdiffusion can generate binders to IDPs and IDRs in a wide range of conformations. We use this approach to generate binders to the IDPs Amylin, C-peptide and VP48 in a range of conformations with Kds in the 3 -100nM range. The Amylin binder inhibits amyloid fibril formation and dissociates existing fibers, and enables enrichment of amylin for mass spectrometry-based detection. For the IDRs G3bp1, common gamma chain (IL2RG) and prion, we diffused binders to beta strand conformations of the targets, obtaining 10 to 100 nM affinity. The IL2RG binder colocalizes with the receptor in cells, enabling new approaches to modulating IL2 signaling. Our approach should be widely useful for creating binders to flexible IDPs/IDRs spanning a wide range of intrinsic conformational preferences.

3.1 – Main

IDPs and IDPRs (structured proteins with intrinsically disordered regions) are abundant in nature and carry out important biological functions without adopting a single well-defined structure, and hence are well established biomarkers in clinical care and biomedical research (Fig. 1a). Designing binders specific for disordered regions could be valuable for clinical diagnosis, therapeutic development, and scientific research1-4 . Current methods largely rely on antibodies, which have limitations such as high production costs, reproducibility, and complex engineering requirements5,6; the dynamic nature of disordered proteins can also complicate the elicitation of antibodies7,8. Computational protein design has created binders of peptides in extended beta strand9,10, helical11, and polyproline II conformations12. While powerful, these methods require prespecification of the target peptide geometry, which can be limiting because the optimal conformation given both the intrinsic sequence biases of the peptide, and the opportunities for making high affinity interactions, may be quite irregular.

We sought to develop a general approach to design high-affinity binders for intrinsically disordered proteins that starts from the target sequence alone and does not require prespecification of the target geometry (Fig. 1b ①). We reasoned that a version of RFdiffusion trained on two chain systems from the PDB, noising the structure on one and providing only the sequence on the second, could have such capability. This was used previously to generate binders to bioactive peptide hormones restricted to helical conformations11; here we begin by investigating the application of the approach to IDPs in a much broader range of conformations (the sequences of many targets are not compatible with uninterrupted helical conformations). To target shorter IDRs, we reasoned that strand pairing, as employed by Sahtoe et al using Rosetta13, coupled with RFdiffusion14 to sample the many different possible variations of strand conformation, could provide a general approach to maximizing interactions over a short

region since backbone backbone hydrogen bonds contribute to binding energy in addition to sidechain-sidechain interactions (Fig. 1b ②).

We first experimented with designing binders to the human islet amyloid polypeptide (hIAPP), also known as amylin, a 37-residue hormone co-secreted with insulin by pancreatic islet β-cells to modulate glucose levels15,16. Cysteine residues 2 and 7 form disulfide bridge which is critical for the full biological activity of amylin15 . NMR studies conducted in lipid environments or under SDS micelle binding conditions have indicated helical propensity in Amylin fragments17,18; the overall structure appears to be intrinsically disordered19,20.
We employed the flexible target fine-tuned RFdiffusion to design binders against Amylin using only the Amylin sequence as input – the structure of the binding protein, the Amylin conformation, and the binding mode are entirely unspecified. Starting from the amino acid sequence of Amylin, RFdiffusion generated complexes encompassing a variety of conformations for both peptides and binders. Representative design trajectories are shown in Supplementary Video 1; starting from a random distribution of residues of both Amylin and binder; in sequential denoising steps, the Amylin adopts different conformations while the binder residue distribution shifts to surround Amylin and progressively organizes into a folded structure which cradles nearly the entire surface of the peptide (Fig. 1b①). The resulting library of backbones were sequence designed using ProteinMPNN21, and filtered using AlphaFold2 (AF2)22 for the monomer conformation and AF2 initial guess for the complex23.

We obtained synthetic genes encoding 96 designs binding amylin in a variety of conformations, expressed the proteins in E.Coli, and purified them using immobilized metal ion affinity chromatography (IMAC). Amylin binding affinities determined using bio-layer interferometry (BLI) ranged from 100 nM to 454 nM (Supplementary Fig. 1a). Since binders to peptides in entirely helical conformations have been studied11, here we focused on other geometries. To optimize the binding affinity of initial hits to αβ, αβL, and αα conformations, we implemented a two sided partial diffusion approach (see Methods; in contrast to one sided

partial diffusion which only diversifies the binder conformation and keeps the target fixed, two sided partial diffusion allows simultaneous conformation changes of both target and binder which leads to broader sampling (Fig. 1c, Supplementary Fig. 2a)). We carried out 5,000 two sided diffusion trajectories from initial designs noised over 5 to 20 steps (complete randomization corresponds to 50 steps), and found that this yielded designs with generally better metrics than one sided diffusion likely because the peptide conformation can adapt to that of binder resulting in greater shape complementarity and more extensive interactions (Supplementary Fig. 2). We obtained synthetic genes encoding the 174 resulting designs with the best metrics that span amylin conformations in the αβ, αβL, and αα conformations. 107 out of 174 refined designs bound Amylin; the highest affinity binders (Amylin-68nαβ, Amylin-36αβ, Amylin-75αα and Amylin-22αβL) which bind Amylin in different conformations, have affinities of 3.8 nM, 10 nM, 15 nM and 100 nM, respectively (Fig. 2a-d). While the Amylin adopts very different conformations in different designs, the diffusion process was able to maintain the disulfide bond, key to amylin function, in all designs15 (Fig. 2ad). Circular dichroism studies showed that all four binders were largely helical as designed and thermostable up to 95 °C (Supplementary Fig. 1b).

C-peptide is a 31 residue peptide secreted by islet β cells that is made from the same precursor – proinsulin – as insulin24. Measurement of plasma C-peptide levels is important for accurate classification and diagnosis of type I and type II diabetes25. We carried out sequence-input diffusion with C-peptide allowed to sample diverse conformations (Supplementary Fig. 3a). Of 96 designs tested, one in which the C-peptide forms a long strand, followed by a long dynamic loop and a small strand paired with the long strand had weak binding affinity (Supplementary Fig. 3bc). This design had more hydrogen bonds between target and binder (13) than all but 5 of the 96 designs (Supplementary Fig. 3d), and we hypothesized that this was important for binding. To optimize the initial hit to improve binding affinity, we again used two sided partial diffusion and included the number of hydrogen bonds in filtering.

Screening with BLI revealed a much higher success rate, with six designs binding C peptide with better than 100nM binding affinity; the highest affinity binder (CP-35) had a Kd of 28nM (Fig. 2e). Circular dichroism studies showed that CP-35 was largely helical, consistent with the design model, and thermostable up to 95 °C (Supplementary Fig.3e).

We next chose to target VP48 (39 amino acid), a potent activator of transcription26. In a first round of 30,000 unconstrained RFdiffusion trajectories, the most enriched conformations after filtering contained substantial secondary structure as in the above cases. To explore binding to more loop-containing conformations, we filtered these designs based on target backbone conformation and a relatively loose PAE cutoff (PAE <16); within this pool, 20 designs were manually selected and further optimized by iterative partial diffusion and backbone extension (see Methods). Of 95 designs tested, 2 showed binding at 2 μM by BLI with the highest affinity 750nM for a design with the VP48 in a conformation with three short helical fragments connected with relatively long loops. Further partial diffusion optimization yielded a design with a Kd of 39nM (Fig. 2f), that again was thermostable up to 95C (Supplementary Fig.

3f).

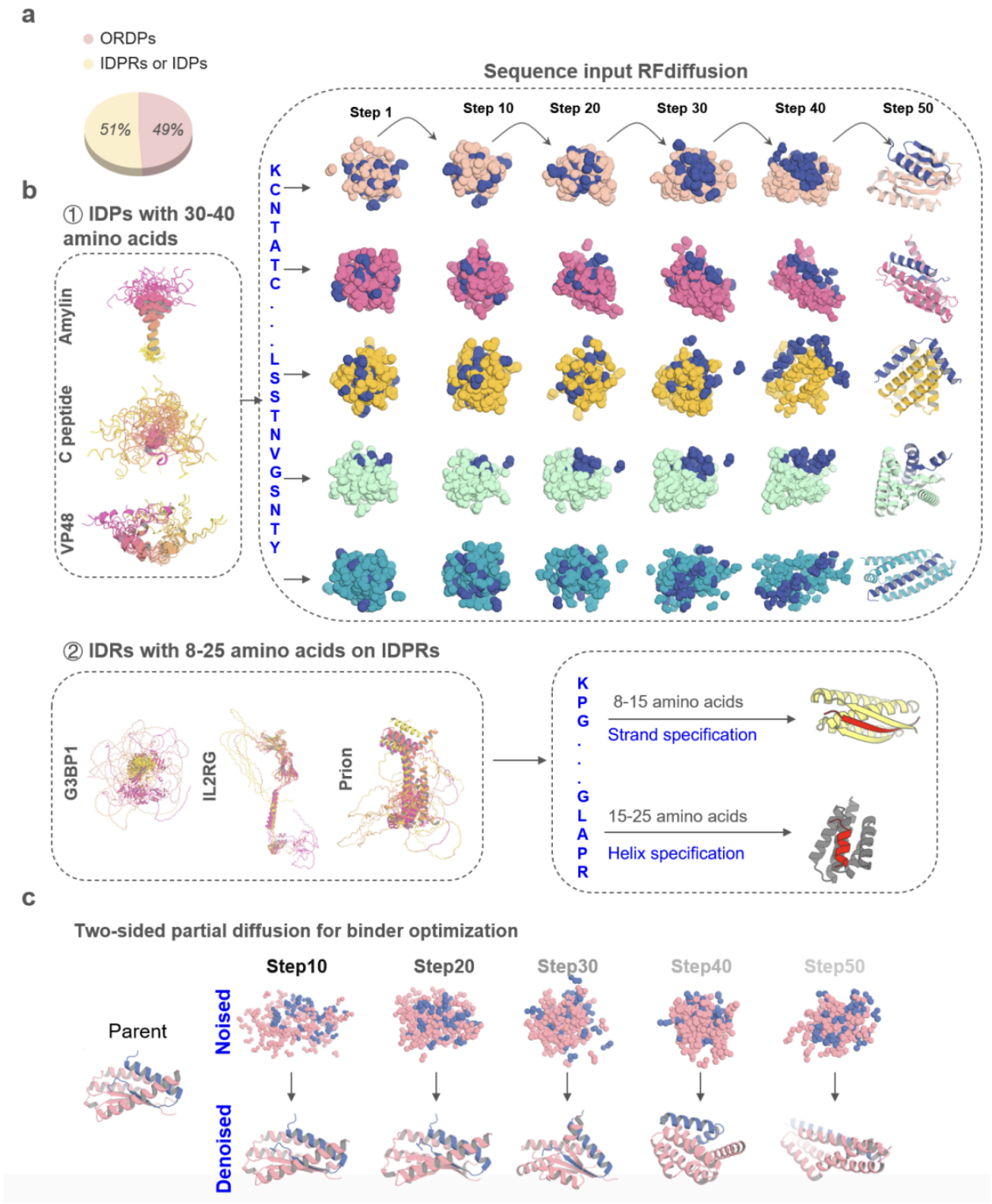

a
ORDPs
IDPRs or IDPs
51%
49%
b
Sequence input RFdiffusion
Step 1
Step 10
Step 20
Step 30
Step 40
Step 50
① IDPs with 30-40 amino acids
Amylin
C peptide
VP48
② IDRs with 8-25 amino acids on IDPRs
G3BP1
IL2RG
Prion
8-15 amino acids
Strand specification
15-25 amino acids
Helix specification
c
Two-sided partial diffusion for binder optimization
Step10
Step20
Step30
Step40
Step50
Parent
Noised
Denoised

Figure. 1 Design strategies for binding conformational flexible peptides. a, Frequency of ORDPs (ordered proteins), IDPRs /IDPs (intrinsically disordered proteins) in the human proteome41 . b, ① Left, the NMR structure of Amylin (PDBID: 2KB8), C peptide (PDBID: 1T0C) , the predicted structures of VP48 by five AlphaFold models22. The 5 predicted structures of VP48 are aligned together, revealing the flexibility of the intrinsically disordered protein. Right, Diffusion models for proteins are trained to recover noised protein structures and to generate new structures by reversing the corruption process through iterative denoising of initially random noise into a realistic structure. Here, A modified version of RFdiffusion was trained on two chain systems from the PDB to permit the design of protein binders to targets, for which only the sequence of the target was specified. The fine-tuned was found to generate binders to peptides in finely varying helix conformationsWith solely sequence input. ② Left, the predicted structures of G3BP1, IL2RG and prion by five AlphaFold models22 . Right, A modified version of RFdiffusion was trained, allowing for specification of the secondary structure of a region, along with its sequence (See Method). When provided with the same target sequence input but different secondary structure specifications (helix or strand), the resulting conformations of the target could vary. c, Top: two sided partial diffusion. RFdiffusion is used to denoise a randomly noised starting parent design for both target and binder ; varying the extent by different noised step of initial noising (top row) enables control over the extent of introduced structural variation (bottom row; colours, new designs; grey, parent design).

3.2 – Targeting shorter IDRs using beta strand interactions

Consistent with the observations of Sahtoe et al using the non-deep learning Rosetta method13, we found that for targeting shorter segments, the RFdiffusion generated designs with the best metrics often made extensive beta strand interactions to targets adopting beta strand conformations. To increase the efficiency of generating such designs, we incorporated into the

RFdiffusion sequence input approach the ability to define the secondary structure of the target (See Methods), to enable the specification of either the entire or a portion of the target sequence in helical, strand, or loop conformation. This is particularly important for strand conformations which can vary considerably in actual 3D coordinates; the coordinate specifying approach used by Vasquez et al11 for helical peptides would be less efficient for targeting strands as many trajectories would have to be carried out for beta strand conformations with different twists, etc. To explore the power of this approach, we used it to design binders to three IDR containing targets.

G3BP1 is a central node within the core stress granule (SG) network27 and plays a crucial role in RNA metabolism and stress response, with a disordered RNA-binding domain (abbreviated as RBD; KPGFGVGRGLAPR, 13 amino acid) mediating interactions with RNA molecules, regulating RNA metabolism, and contributing to the assembly and disassembly of stress granules. A first round of 10,000 RFdiffusion trajectories with sequence only specification of the RBD domain of G3BP1, abbreviated as G3bp1RBD yielded designs with the peptide adopting a roughly 5.7 :3.8 :0.5 ratio for helix:strand:loop, respectively (Supplementary Fig. 4a), but only the 23 strand containing designs had AF2 pae_interaction < 10 and plddt_binder > 90 (Supplementary Fig. 4a-b). Based on these observations, we specified the secondary structure as a strand and conducted 10,000 trajectories. The resulting ratio of G3bp1RBD conformation in the complex was 0.54:8.9:0.6 for helix:strand:loop, respectively, with 1,192 designs meeting the same filtering criteria, a ~51 fold improvement; in all passing designs the target had a strand conformation. We narrowed these down to 78 designs by filtering on structure prediction and Rosetta interaction metrics (monomer plddt, hbonds_count, monomer RMSD, sap_score, ddg, and contact_molecular_surface). The 78 designs were subsequently expressed in E. coli and subjected to initial screening using BLI. 5 out of 78 designs were found to bind to G3bp1RBD, with the tightest exhibiting a binding affinity at 18 nM. Through two-side partial diffusion, we further optimized 4 of the binders (G3bp1-4, G3bp1-45, G3bp1-53, and G3bp1-77;

Supplementary Fig. 4c); 40 of the 95 refined designs bound G3bp1RBD , with the tightest G3bp1-11 having an affinity of 11nM.

We next sought to make binders of the prion protein which is primarily found in neuronal cells in mammals. Aggregated forms of this protein are linked to prion diseases, a group of transmissible neurodegenerative disorders28,29 . The pathological hallmark of prion diseases is the conformational conversion of the native, monomeric cellular prion protein (PrPC) into a misfolded and aggregated form (PrPSc) characterized by a cross-β structure30-33. To target the amyloid core region of the prion protein, we targeted the amino acid sequence VNITIKQH (positions 180-187), specifying its secondary structure as a β-strand and conducted 20,000 trajectories. Using in silico filtering strategies similar to those employed for G3bp1RBD, we selected 48 designs for further validation via BLI. Among these, the tightest binder, PRI28, had a binding affinity of 14 nM (Fig. 2h) with high stability up to 95 °C (Supplementary Fig. 5a), higher affinity and specificity than generally achieved with our earlier Rosetta based β-strand targeting method13 (Supplementary Fig. 5b). Moreover, we found that specifying the secondary structure of the target region as a β-strand resulted in binders with higher affinity than using the target sequence information alone (14 nM from secondary structure specification (PRI28) vs 1.88 μM sequence input (PRI22), Fig. 2h and Supplementary Fig. 5c-d). After refinement through two-sided partial diffusion, the affinity of PRI22 improved to 80 nM, still weaker than PRI28 (Supplementary Fig. 5c-d).

Signal transduction via cell surface receptors is mediated by their intracellular domains, which contain long disordered regions34,35. Developing binders targeted at these domains would be broadly useful for co-localization imaging applications and for the modulation of receptor activation. The common cytokine receptor γ chain (common gamma chain, IL2RG) is a receptor subunit shared among the interleukin (IL) receptors for IL-2, IL-4, IL-7, IL-9, IL-15 and IL-21. Each receptor within the γc family uniquely contributes to the adaptive immune system, influencing the development of T, B, natural killer, and innate lymphoid cells36. To target the

intracellular domain of IL2RG, we selected the amino acid sequence ERLCLVSEIP (positions 327-336) as the target region, specifying its secondary structure as a strand and conducted 40,000 trajectories. Employing in silico filtering strategies similar to those used for G3bp1, we selected 94 designs for further validation via BLI. Among these, one design had a binding affinity of 493 nM. Through two sided partial diffusion, we increased the binding affinity to 97 nM, and we named it IL2RG-30 (Fig. 2i); this optimized design again had high thermal stability (Supplementary Fig. 4e).

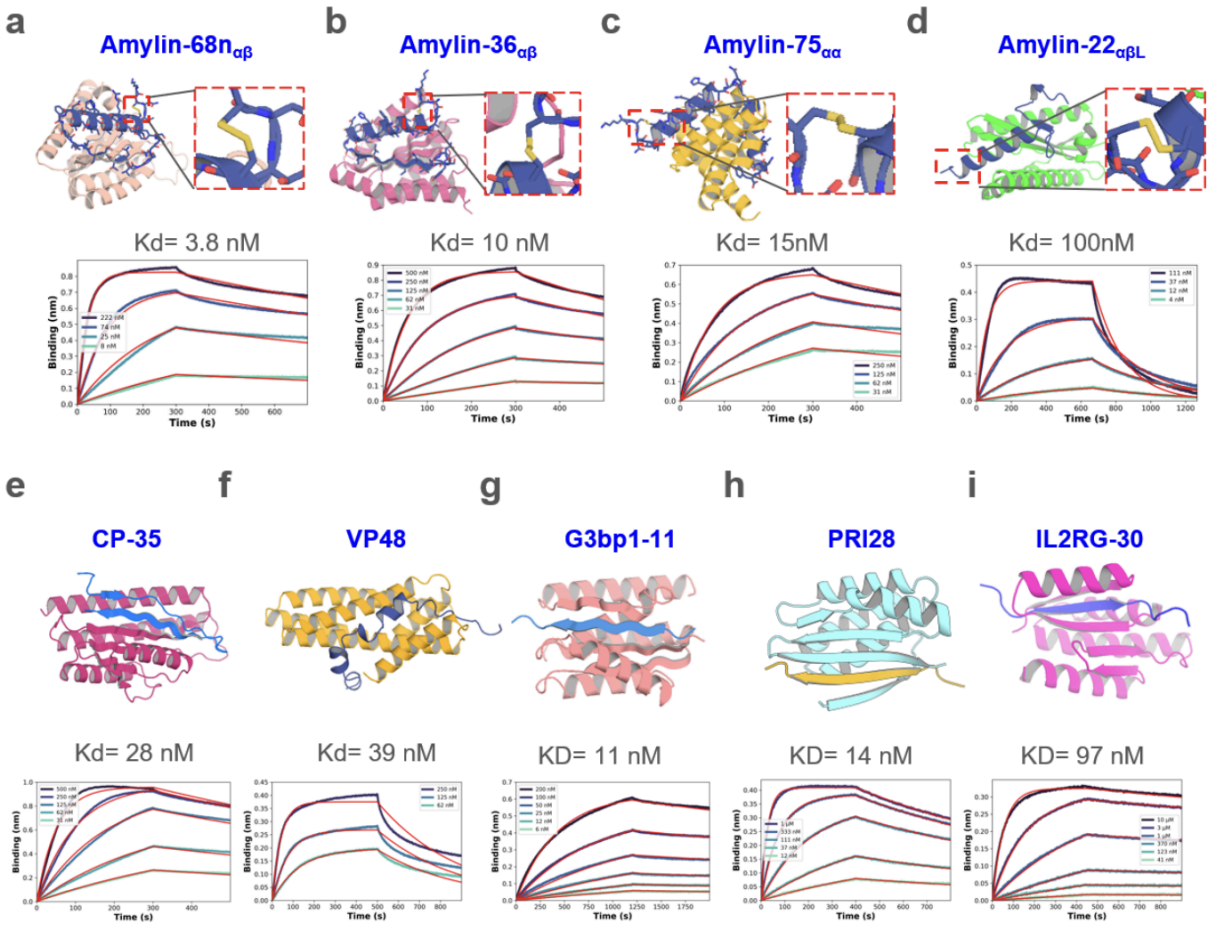


Figure.2 Design of disordered region binder a-d, Binder design of Amylin using sequence input diffusion. Top, from left to right, design model of Amylin and its binder Amylin-68nαβ, Amylin-36αβ, Amylin-75αα and Amylin-22αβL, respectively. The secondary structure of Amylin is indicated in the subscript of the binder's name. For each of the designs, the Amylin disulfide bonds between 2nd Cysteine and 7th Cysteine were retained well. Bottom, from left to right, the BLI measurement indicated that the binding affinity between Amylin-68nαβ, Amylin-36αβ, Amylin-75αα, Amylin-22αβL and Amylin are 3.8, 10, 15, 100 nM respectively. e-f, Binder design of CP and VP48 using sequence input diffusion, the binder affinity of CP and VP48 are 28 and 39 nM, respectively. g-i, Binder design using strand specification. Top, from left to right, design model of G3BP1RBD , prion and IL2RG and their binders G3bp1-11, PRI28 and IL2RG-30. Bottom, the BLI measurement indicated that the binding affinity of G3bp1-11, PRI28 and IL2RG-30 binders are 11, 14 and 97 nM, respectively.

3.3 – Structure analysis of designed complexes

We obtained crystal structures of Amylin-22αβL and G3bp1-11 in complexes with their target at 1.8-Å-resolution and 2.4-Å-resolution, respectively. For Amylin-22αβL, the designed conformation comprises a helix, a strand, and an unstructured loop (Fig. 3a, left). The Amylin helix is embedded within a groove formed by the helix and strand segments of the binder. Adjacent to this, the Amylin strand pairs with a corresponding strand of the binder. The Amylin loop is predicted to be disordered based on the low per-residue AF2 pLDDT (predicted Local Distance Difference Test) (Fig. 3a, left, Supplementary Fig. 1c) 22,37 . In the crystal structure, the main helix and strand are well resolved, and closely match the computational model; the

disordered loop is as anticipated not resolved (Fig. 3a-b). The Ca RMSD between the design model and the crystal structure over the backbone of the binder alone, and over the backbone of the full complex excluding the missing loop of Amylin, are 0.96 and 2.04, respectively. The backbone and sidechains at the designed binder-target interface are also in close agreement between crystal structure and design model (Fig. 3b, interface Ca and sidechain RMSD are 1.33 and 1.87, respectively).

In the G3bp1-11 design model, the peptide is in a β-strand conformation and lies within a cleft formed by two α/β structures, T1 and T2, in the designed binder, pairing with two adjacent strands (Fig. 3c). An additional helix in T2 also interacts with the target, potentially enhancing binding affinity and specificity (Fig. 3c, Supplementary Fig. 6a). The crystal structure of G3bp1-11 closely recapitulates the design model, with the peptide clamped in a β-strand conformation (Fig. 3c-d, Ca RMSD 0.8 Å for entire complex between design and crystal structure) with the interface residues nearly perfectly aligned with the design model structure (Fig. 3c-d, interface Ca and sidechain RMSD are 0.86 and 2.29, respectively).

We were unable to solve crystal structures of the CP binders, so we instead obtained a lower resolution structural footprint of the binding site by generating a site saturation mutagenesis library (SSMs) for CP-35 in which every residue was substituted with each of the 20 amino acids one at a time. Next generation sequencing before and after FACS sorting for CP binding revealed that residues at the binding interface and protein core were largely conserved (Fig. 3e-f and Supplementary Fig. 6b-c), supporting the design model.

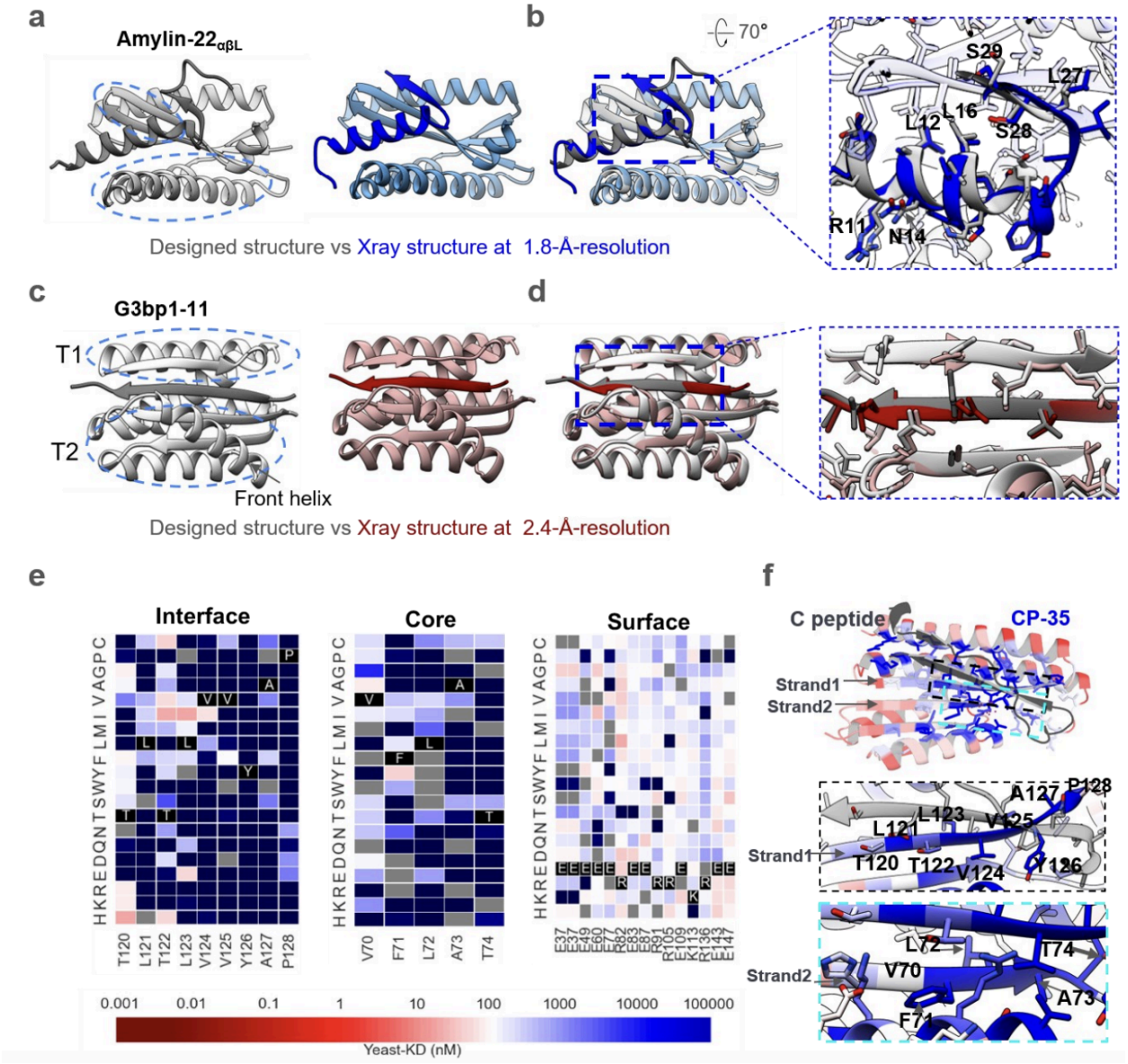


Figure. 3 Structural characterizations. a, Left, the designed model of Amylin-22αβL, with target and binder proteins rendered in dim gray and gray, respectively. The helical and strand segments that create the groove in the binder, docking the helical segment of Amylin, are highlighted with blue dashed ellipsoid. Right, the crystal structure of Amylin-22αβL at 1.8 Å-resolution, with target and binder proteins rendered in blue and cornflower blue, respectively. b, Left, the overlay of the design model and the crystal structure of Amylin-22αβL. Right, magnified views of the regions indicated with black dotted frames in the left panel are provided

to illustrate the detailed interface view of the design and crystal structure. The binder proteins are rendered with 90% transparency to enhance the visibility of the peptide target. The key residues on the Amylin are labeled to illustrate the good alignment of the key residues between designed protein and crystal structure. c, Left, the designed model of G3bp1-11, with target and binder proteins rendered in dim gray and gray, respectively. The two α/β topologies (T1 and T2) of the binders, forming the cleft where the target strand is positioned, are highlighted with blue dashed ellipses. The front helix of T2 is denoted by a black arrow. Right, the crystal structure of G3bp1-11 at 2.4 Å-resolution, with target and binder proteins rendered in dark red and rosy brown, respectively. d, Left, the overlay of the design model and the crystal structure of G3bp1-11. Right, magnified views of the regions indicated with black dotted frames in the left panel. The front helix of T2 has been surface capped to reveal the strand pairing interface. e, Heat maps representing C peptide-binding Kd (nM) values for single mutations in the designed interface (left), core (middle) and the surface (right). Substitutions that are heavily depleted are shown in blue, and beneficial mutations are shown in red, gray color indicates the lost yeast strains. For the interface region, we highlighted and showcased strand 1 (indicated by the arrow), which serves as the primary interaction secondary structure with the C peptide. For the core region, we showcased the right segment of strand 2 (indicated by the arrow), representing a main core region that does not form interactions with the C peptide. For the surface region, we selected the most exposed surface residues that don't form any connections with other residues (Supplementary Fig. 6c). Full SSM map over all positions for CP35 is provided in Supplementary Fig. 6b. f, Top, designed binding proteins are colored by positional Shannon entropy from site saturation mutagenesis, with blue indicating positions of low entropy (conserved) and red those of high entropy (not conserved). Bottom, zoomed-in views of central regions of the design interface and core with the C peptide.

3.4 – Specificity of designed binders

We investigated the specificity of the binders by carrying out all by all binding experiments (Fig. 4). BLI binding characterization of 9 binders against 6 targets showed that the designs had high specificity for their intended peptide targets. Very weak off target binding was observed at high concentrations in two cases: VP48 weakly bound Amylin above 800 nM, perhaps reflecting the ~50% helical content of both peptides (specificity could potentially be further improved through another round of partial diffusion, or decreasing the helical percentage through secondary structure specification) and G3BP1-11 weakly bound IL2RG at 2 uM. Overall, the much higher on-target than off-target binding suggests the binders should be broadly usable as affinity reagents.

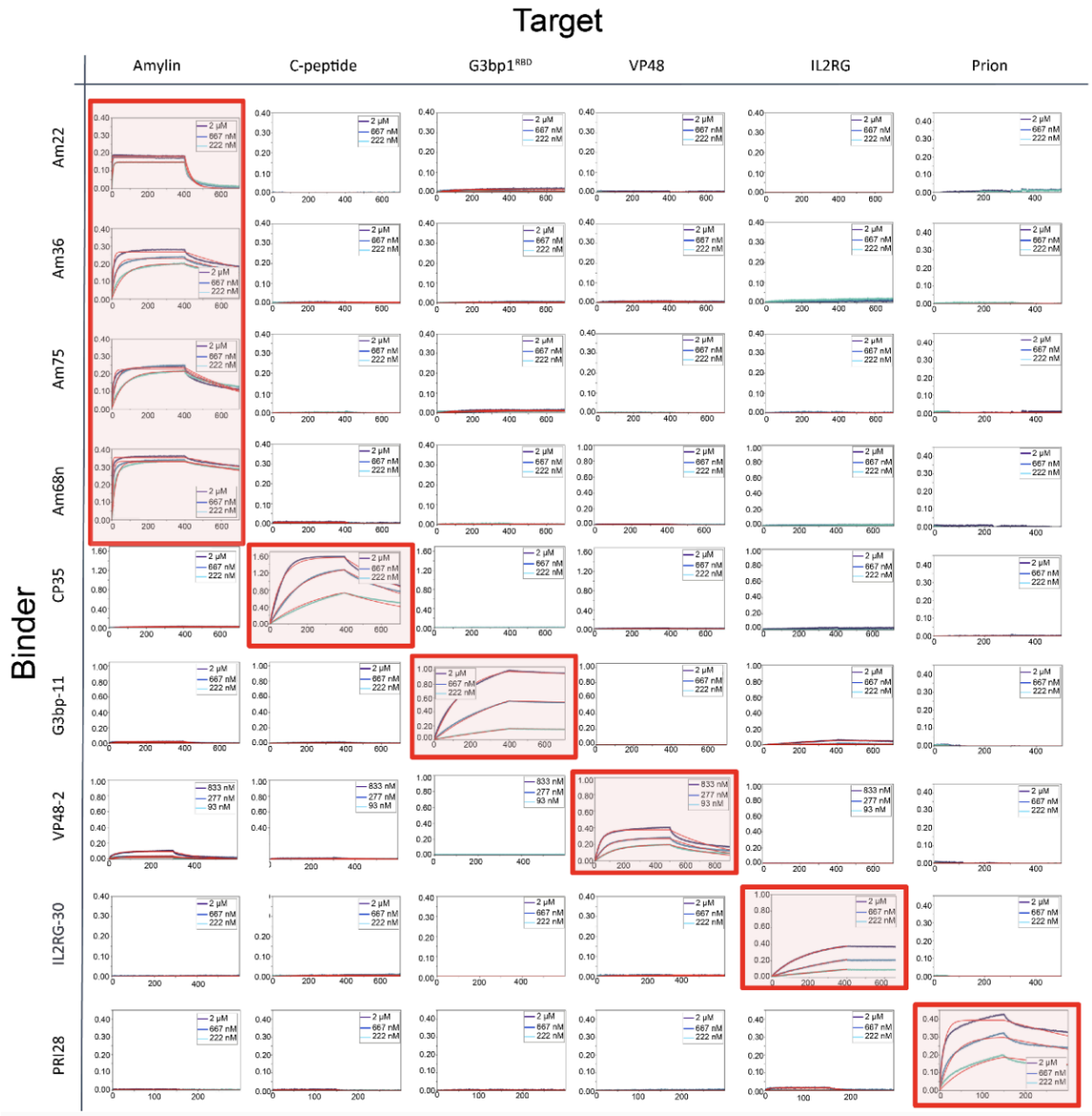


Figure. 4 Specificity profile of designed binders in BLI. Biotinylated peptides were immobilized onto octet streptavidin biosensors at equal densities and incubated with all binders in separate experiments at three concentrations (2, 0.667 and 0.222 μM except VP48 binder at 0.833, 0.277 and 0.093 μM). Amylin-68nαβ, Amylin-36αβ, Amylin75αα, Amylin-22αβL are

abbreviated as Am68n, Am36, Am75 and Am22, respectively. The designed on-target interactions are indicated with a light red background.

3.5 – Designed binders colocalize with their targets in mammalian cells

To examine whether the designs could fold properly and bind to the target proteins in mammalian cells, we knocked out the endogenous IL2RG in HeLa cells using CRISPR-Cas9, and then transfected the cells with a construct encoding IL2RG fused to EGFP. When cells were additionally transfected with mScarlet-labeled IL2RG binder IL2RG-30, colocalization of GFP and mScarlet was observed, indicating binding (Fig. 5a). In IL2RG knockout cells transfected only with IL2RG-30-mScarlet, no colocalization was observed (Fig. 5a, left), confirming that the interaction occurs through the designed interface.

3.6 – Enrichment for LC–MS/MS detection

We explored the use of amylin binder Amylin-68n as a capture agent for immunoaffinity enrichment combined with liquid chromatography–tandem mass spectrometry (LC–MS/MS), a general platform for detecting low-abundance protein biomarkers in human serum38. We prepared Amylin-binder-conjugated beads as described in the Methods. Amylin enrichment was calculated based on detection of intact, alkylated amylin in either human plasma or simplified PBS-CHAPS matrix39 (Methods). We found that the designed binder enabled capture of Amylin from buffer and human plasma supplemented with Amylin (the endogenous levels are too low for reliable detection) with recoveries of 62.2% and 53.5%, respectively (Figure. 5b).

3.7 – Designs inhibit Amylin fibril formation and dissociate existing fibrils

Amylin fibril formation is implicated in type 2 diabetes, where the aggregation of amylin into insoluble fibrils contributes to islet amyloid deposition and β-cell dysfunction40. We investigated the effect of four binders—Amylin-68nαβ, Amylin-36αβ, Amylin-75αα and Amylin-22αβL—on Amylin fibril formation. At a binder to Amylin molar ratio of 1:4, with concentrations of 40 μM for Amylin and 10 μM for binders, all binders completely inhibited fibril formation (Fig. 5e). Further tests with Amylin-22αβL and Amylin-36αβ at binder to Amylin molar ratios of 1:4, 1:40, and 1:400 revealed a concentration-dependent retardation of fibril formation (Supplementary Fig. 7a). Inhibition of fibril formation was also observed by negative stain electron microscopy (NSEM), with Amylin-22αβL and Amylin-36αβ at binder to Amylin molar ratios of 1:4. Addition of Amylin-36αβ blocked fiber formation at both 1 h and 18 h, whereas some short fibrils were observed 18 hours post-addition of Amylin-22αβL (Supplementary Fig. 7b-c).

We next investigated whether the amylin binders were able to disaggregate pre-formed amylin fibrils. We generated short Amylin fibrils by incubating the peptide at 40 μM for 3 hours at 37 °C, to reach the elongation phase, and then incubated with 10 μM Amylin-36αβ. NS-EM revealed no fibrillar structures after treatment with Amylin-36αβ at both 1 h and 18 h time points (Fig. 5c). Thioflavin T (ThT) assays with Amylin-36αβ added at the 3-hour Amylin fiber stage also showed fiber disassembly in a design concentration-dependent manner (Fig. 5f).

To test whether Amylin-36αβ could dissociate mature fibrils that had formed over 24 hours at 10 μM, we incubated them with 10 μM of the binder. Small oligomers were still observed at 1 hour, but were completely dissociated by 18 hours (Fig. 5d). Fibril ThT fluorescence again decreased in a designed binder concentration-dependent manner (Fig. 5g).

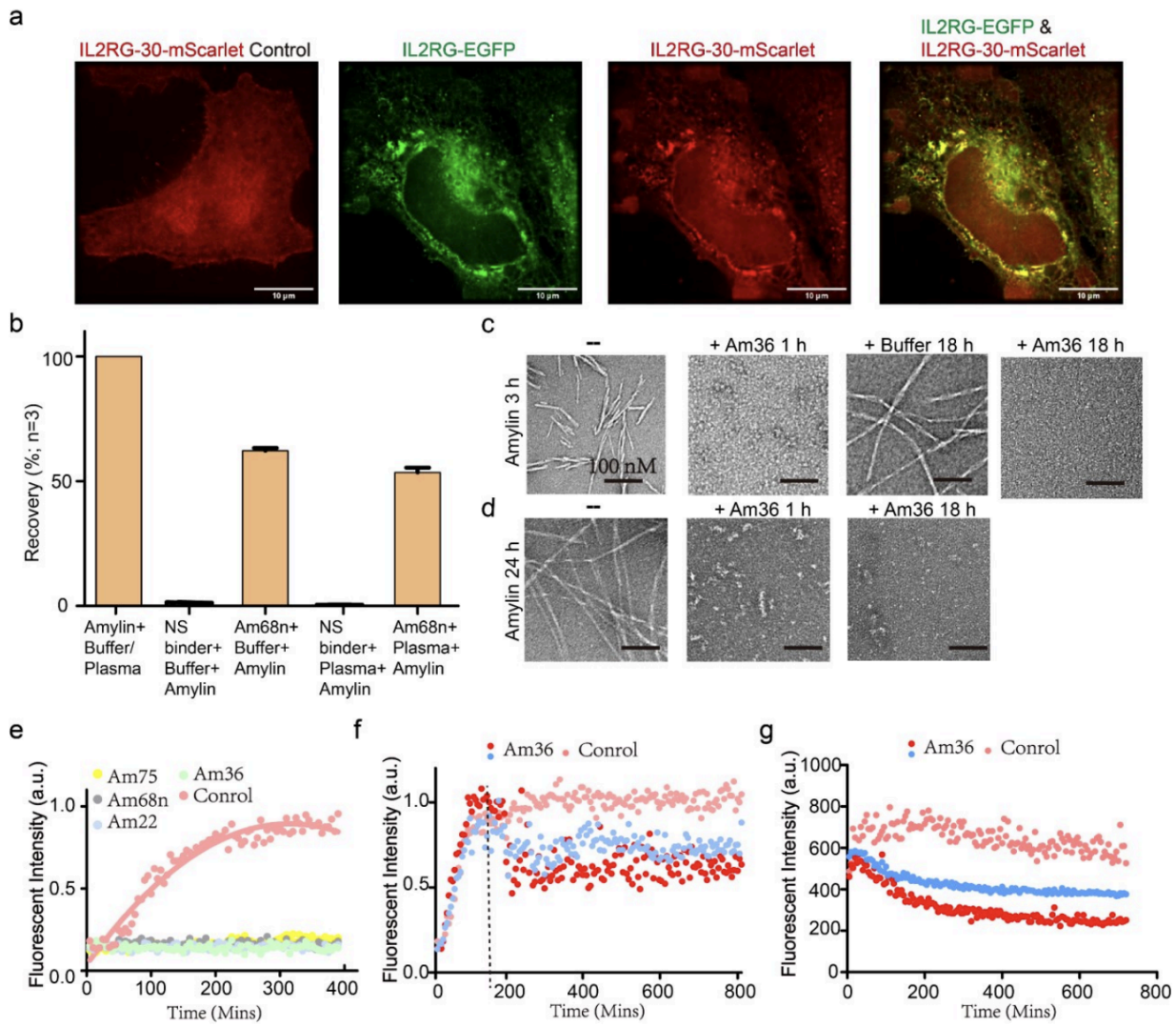


Figure. 5 Applications of designed binders a, Colocalization of binder IL2RG-30 and target membrane receptor IL2RG in HeLa Cells. Cells with endogenous IL2RG knocked out express only the red fluorescent mScarlet-tagged binder IL2RG-30, which is uniformly distributed throughout the cell (left). In contrast, cells co-expressing green EGFP-tagged IL2RG and red mScarlet-tagged IL2RG-30 show specific colocalization of both proteins. b, The LC–MS/MS recovery percent of Amylin from PBS-0.1% CHAPS buffer and EDTA-anticoagulated plasma was compared between BSA-blocked tosyl-activated bead, an offtarget binder, and amylin-targeted binders (Am68n). Percent recovery was calculated using the peak area of a sample of pure amylin peptide in elution solvent as the denominator (i.e.,

100% recovery of the peptide). Error bars represent SD (n=3). c-d, Visualization of fibril dissociation by Amylin-36αβ binder using negative staining electron microscopy. panels (c) and (d) demonstrate the dissociation of existing fibrils at elongation phase (c) and mature phase (d) following the addition of Amylin-36αβ. Scale bars, 100 nM. e, Thioflavin T (ThT) assay revealed that all 4 binders could strongly inhibit fibril formation at molar ratio of binder to Amylin 1:4. f, Amylin36αβ could dissociate fibrils at elongation phase in concentration-dependent manner. The Tht assay was performed since the Amylin monomer, Amylin-36αβ was added at 3h when Amylin fibrils were at elongation phase, marked with a dotted line. Red dot and blue dot indicate that Amylin36αβ to Amylin is 1:4 and 1:40, respectively. g, Tht assay was performed after the mature Amylin fibrils were formed for 24 h, at the same time, Amylin-36αβ was added, the data revealed that fibril fluorescence decreased in a concentration-dependent manner. Red dot and blue dot indicate that Amylin-36αβ to Amylin is 1:4 and 1:40, respectively.

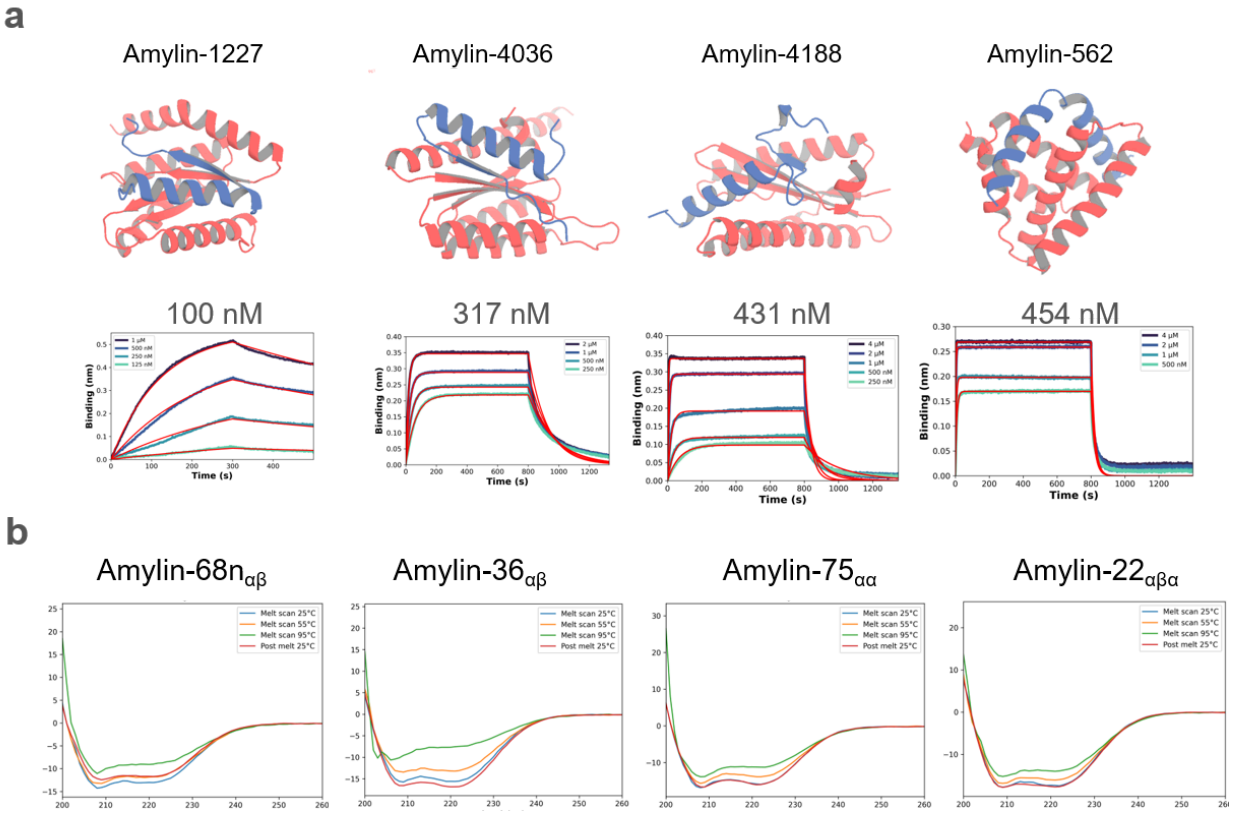
3.9 – Discussion

Our results demonstrate the utility of RFdiffusion in designing binders for IDPs ranging from 30- 40 amino acids in length in diverse conformations, expanding its applicability beyond helical peptides. The ability to target IDPs without specifying the target structure is important as such proteins have no single defined conformation. During the design process, the target protein samples a wide range of possible conformations as the designed binding protein diffuses around it; the co-folding of design and target effectively enables the selection of conformations particularly suitable for binding. The versatility of our approach is highlighted by the design binders for Amylin in diverse conformations while consistently forming the Amylin peptide disulfide.

For shorter peptides which can adopt beta strand like conformations, we show the introduction of a secondary structure type specification feature within the RFdiffusion model enables targeting of peptides in the beta strand conformation. The generated structures resemble previous strand targeting designs generated using Rosetta, but exhibit higher specificity and binding affinity.

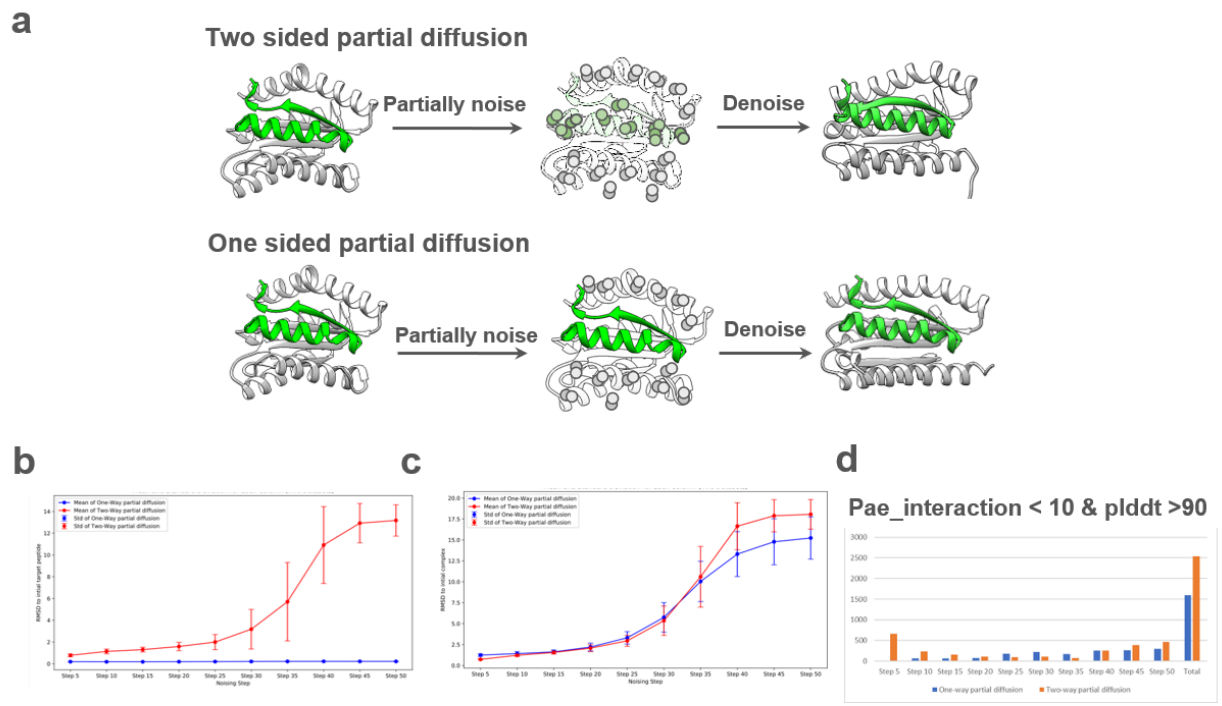
The binders and approaches described here could be broadly useful given the current difficulty in targeting IDPs and IDRs, and the important roles these play in both normal physiology and disease. For example, the Amylin binder both inhibits the formation of Amylin fibers and dissociating preexisting fibers, which could have therapeutic utility. Additionally, it facilitates the enrichment and detection of Amylin using mass spectrometry. The designed binders bind their targets in cells, as illustrated by the colocalization of PRI28 with the intracellular tail of the IL2 receptor gamma subunit, opening up new ways of modulating cytokine signaling in feedback loops for adoptive cell therapies and other applications.

Supplementary data



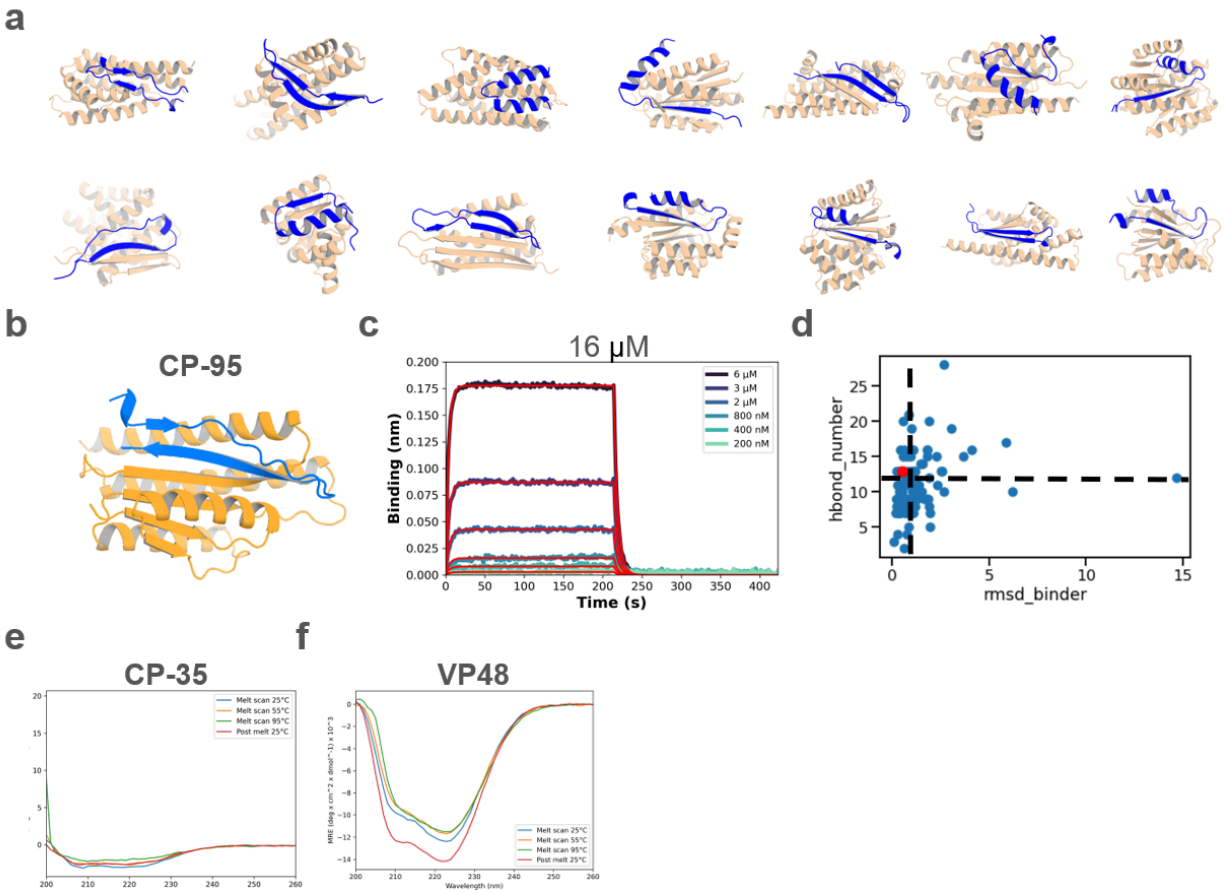
Supplementary figure 1. Diffusing de novo peptide binder design to Amylin.

a, Top, the designed structures of four initial hits, Amylin-1227, -4036, -4188, -562, which serve as starting point of two sided partial diffusion. Bottom, the BLI result of the four hits revealing the binding affinity of the 4 initial hits are 100, 317, 431, 454 nM, respectively. **b,** Circular dichroism data show that the optimized binders have helical secondary structure and is stable up to 95 °C (inset).



Supplementary figure 2. Two sided partial diffusion and the comparison with one sided partial diffusion

a, Top, two sided partial diffusion allows simultaneous conformational changes in both the target and the binder. Botomn, one sided partial diffusion solely diversifies the conformation of the binder while keeping the target fixed. **b,** Two sided partial diffusion (in red) diversifies the target while one sided partial diffusion (in blue) keeps the target fixed. **c,** The peptide-binder complex diverse magnitudes of two sided (in red) and one sided partial diffusion (in blue) remain comparable before nosing step 35, after step 35, the diverse magnitude of two sided parietal diffusion is larger than one sided one. **d,** Take the interface pAE <10, pLDDT >90 as cutoff criteron, two sided partial diffusion yielded designs with generally better metrics than one sided diffusion. At steps 25, 30, and 35 exclusively, one-sided partial diffusion exhibited superior performance. However, in practical cases, we typically operate within fewer than 20 steps to remain the main features of parent structure.

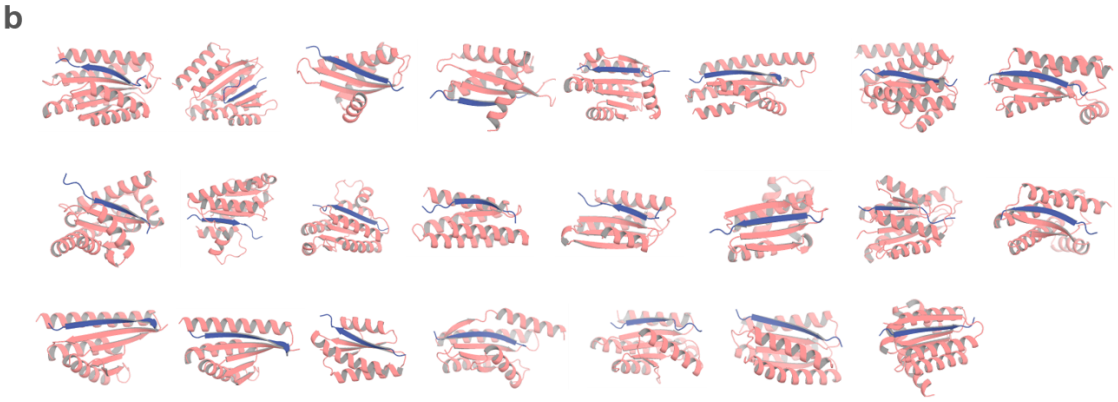


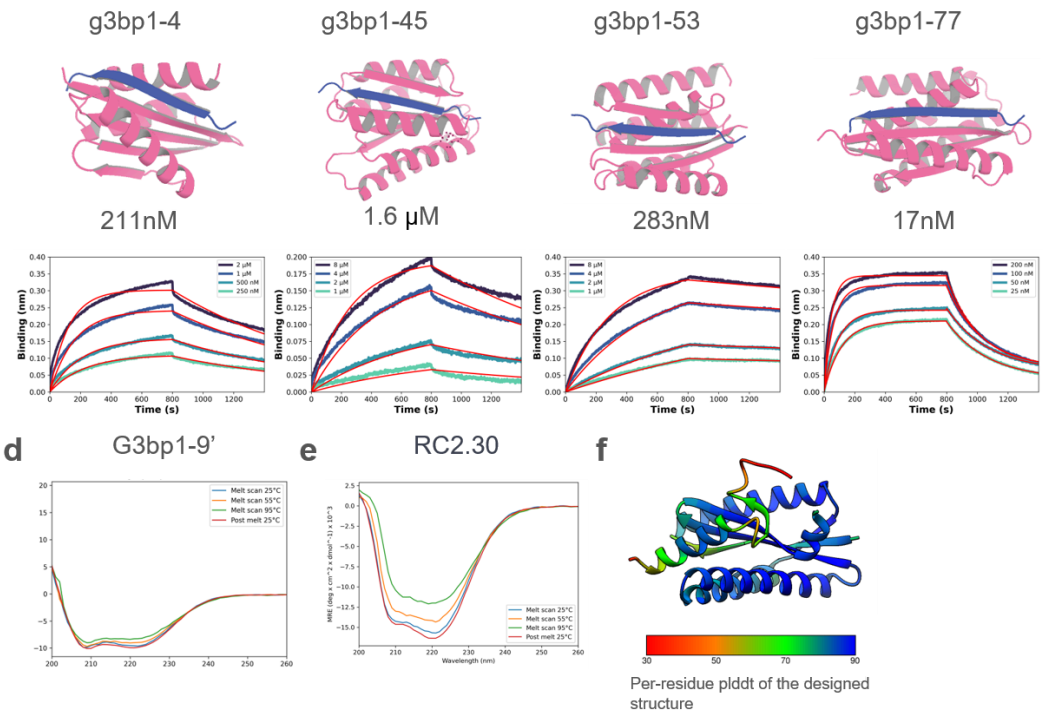
Supplementary figure 3. Diffusing de novo peptide binder design to C peptide.

a, Sequence-input diffusion was carried out, allowing C peptide to sample diverse conformations. The diverse conformations of C peptide and protein binder are rendered in blue and wheat color, respectively. **b,** Design model of the initial hit CP-95 which was also the starting point of two-sided partial diffusion. **c**, the BLI data revealed the binding affinity of the initial hit is 16 µm. **d,** Scatter plot showing the distribution of designs based on the number of hydrogen bonds (hbond_number) and the RMSD of the binder (rmsd_binder). Each blue dot represents a design, while the red dot marks a validated hit. The dashed black lines indicate the cutoff values based on the initial hit criteria (hbond_number = 13 and rmsd_binder = 0.545). Analysis revealed that only 6 out of 96 designs met these criteria (hbond_number > 13 and rmsd_binder < 0.545), indicating a low success rate. **e-f,** Circular dichroism data show that the binder CP35 (e) and VP48 (f) have helical secondary structure and is stable up to 95 °C (inset).

a

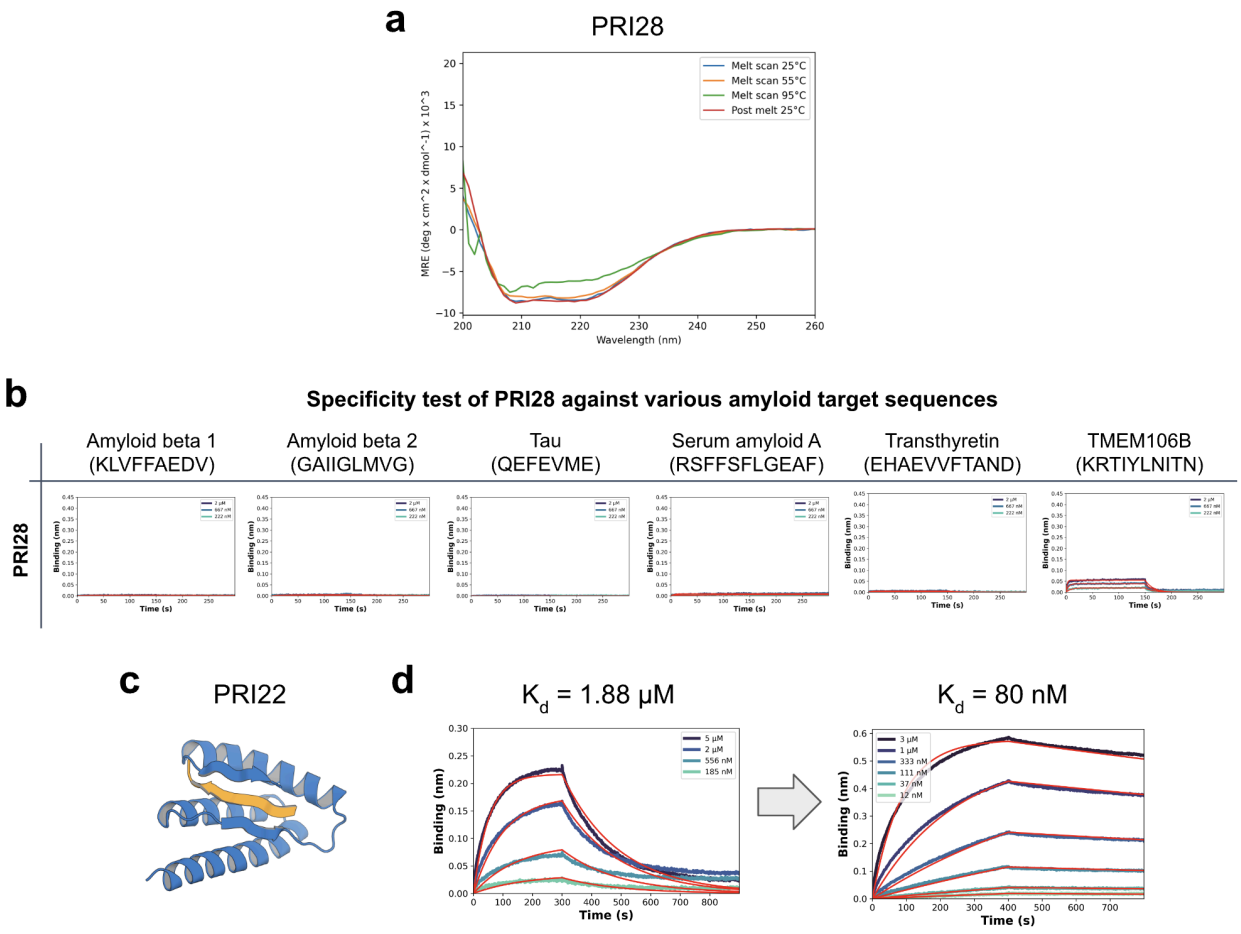
	Trajectories	Helix:Strand:loop	Pae_interaction <10, plddt_binder > 90
Sequence input	10k	5.7:3.8 :0.5	23
Strand specification	10k	0.54:8.9:0.6	1,192





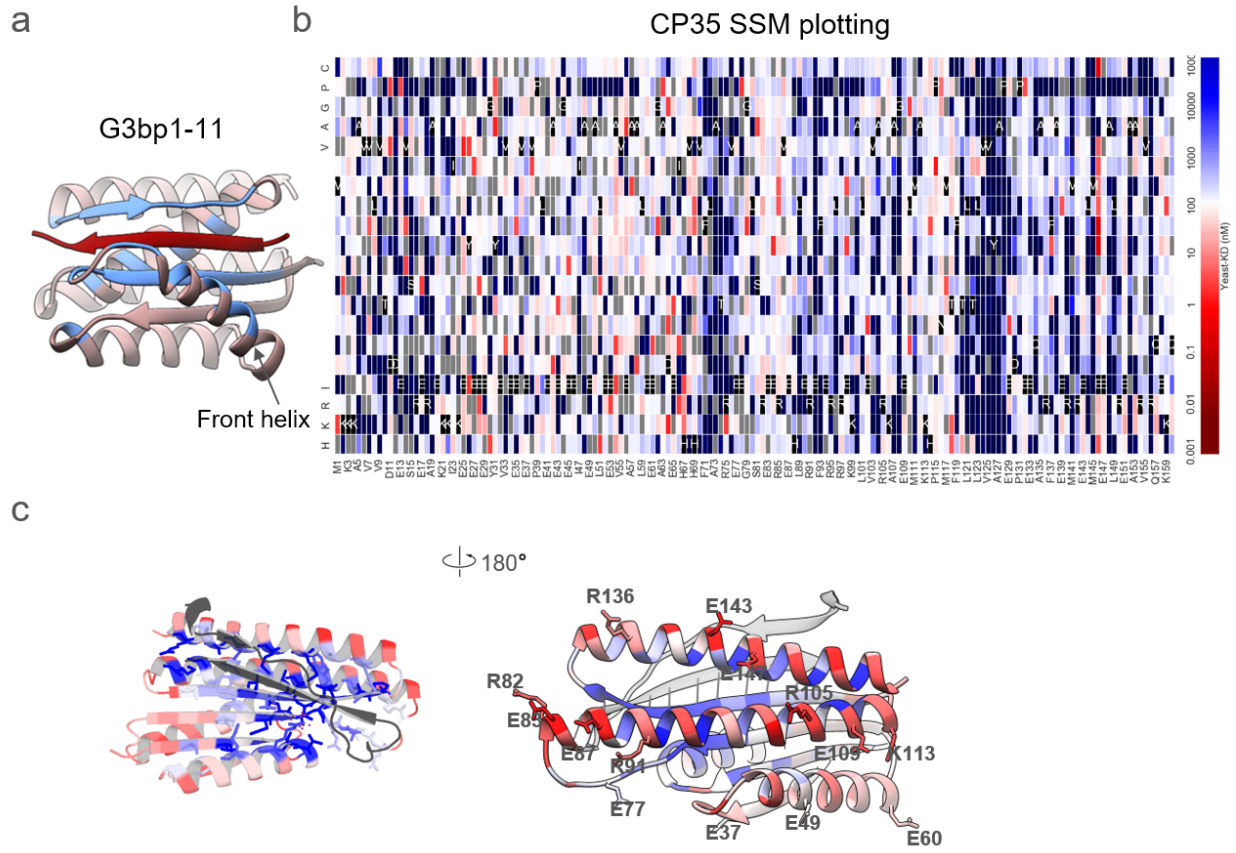
Supplementary figure 4. Diffusing de novo peptide binder design to G3BP1^{RBD}.

a, Comparative analysis of structural outcomes between sequence input and strand specification approaches in protein design. The table presents the number of trajectories (10k) and the distribution of secondary structures (Helix:Strand: Loop) for both methods. This table counts the successful cases where the Pae_interaction is under 10 and the plddt_binder score is over 90, noting 23 successes with sequence input and 1,192 with strand specification. This reflects an approximately 52-fold increase in efficacy with the strand specification method, highlighting its superior performance in achieving desired structural configurations. **b**,The 23 successful cases designed using sequence input RFdiffusion all feature targets in strand conformation. **c**, Design models and BLI data of the 4 initial hits of $G3BP1^{RBD}$ which was also the starting point of two-sided partial diffusion. **d and e,** Circular dichroism data show that the G3bp1-11 binder (d) and IL2RG binder (e) have helical secondary structure and are stable up to 95 °C (inset). **f**, The per residue pLDDT (predicted Local Distance Difference Test) plotting of CP35 binder in design.



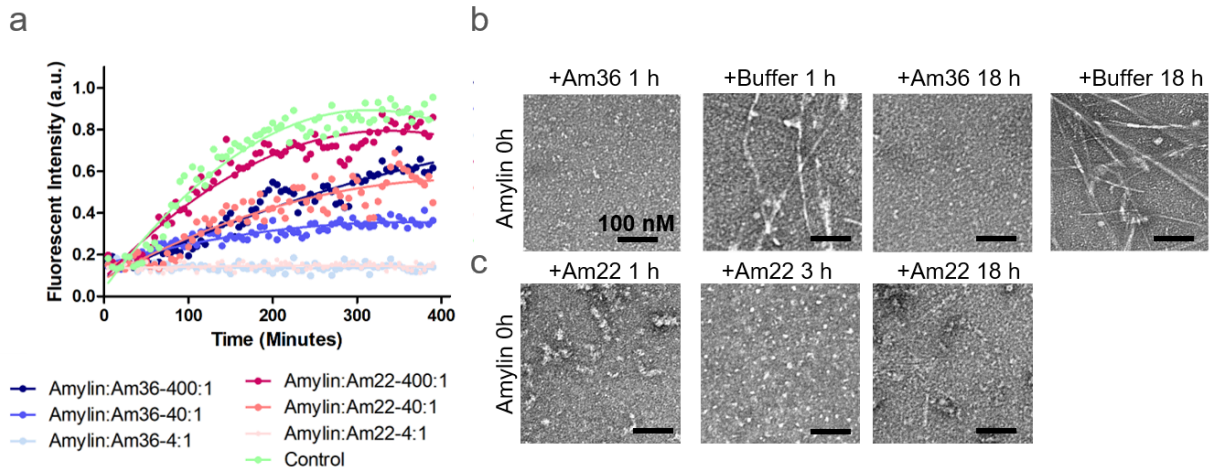
Supplementary figure 5. Diffusing de novo peptide binder design to prion protein.

a, Circular dichroism data show that the PRI28 binder has helical secondary structure and is stable up to 95 °C. **b**,The specificity test for prion binder PRI28 against various amyloid target sequences showed that PRI28 is highly specific, with some cross-reactivity observed only with TEME106B, related to Fig. 2h. **c,** The design model of PRI22, designed using target sequence information alone, is shown. **d**, The BLI data revealed that the binding affinity of PRI22 is 1.88 µM (left), which improved to 80 nM after two-sided partial diffusion (right).



Supplementary figure 6. SSM analysis of CP35.

a, The crystal structure of G3bp1-11, positioned 4 Å away from the target on the binder, is marked in blue. **b**, Full SSM maps for the design of CP35. **c**, Zoomed-in views of the residues presented in the surface region, as shown in Figure 3e.



Supplementary figure 7. Designs inhibit Amylin fibril formation and dissociate existing fibrils

a,Amylin binders Amylin-22 and Amylin-36 inhibit fibril formation in a concentration-dependent manner. The initial concentration of Amylin monomer was 10 μM, with subsequent additions of binders at 2.5 μM, 0.25 μM, and 0.025 μM, establishing molar ratios of binder to Amylin of 1:4, 1:40, and 1:400, respectively. **b-c** ,, Negative stain electron microscopy images were taken of 40 μM Amylin monomer samples following the addition of 10 μM Amylin-36 (b) and Amylin-22 (c) at 1 hour, 3 hours, and 18 hours, respectively. Scale bars, 100 nM.

IDP name	Targeted Sequence	Position
Amylin	KCNTATCATQRLANFLVHSSNNFG AILSSTNVGSNTY (37)	Full length
C peptide	EAEDLQVGQVELGGGPGAGSLQP LALEGSLQ (31)	Full length
VP48	DALDDFDLDMLGSDALDDFDLDML PADALDDFDLDMLGS (39)	Full length
G3BP1	KPGFGVGRGLAPR (13)	453-465
Common gamma chain	ERLCLVSEIP (10)	327-336
Prion	VNITIKQH (8)	180-187

Supplementary table 1. Summary of intrinsically disordered proteins (IDPs) in the study, detailing each protein's sequence and positional data within their respective structures.

Method

De novo peptide binder design given only sequence input using RFdiffusion and ProteinMPNN

For each target, approximately ten to fifty thousand diffused designs were generated given only sequence input of the target. The resulting library of backbones were sequence designed using ProteinMPNN, followed by AF2+initial guess (Bennett *et al.*, 2023). The resulting designs were filtered based on interface pAE, pLDDT. In addition, AF2 monomer was performed using only the binder sequence without the peptide to filter based on the monomer pLDDT of the binder and RMSD to the binder design model.Subsequently, FastRelax was executed to obtain Rosetta metrics. The resulting binders were then further filtered based on criteria including contact_molecular_surface, ddG, SAP score and the numbers of hydrogen bonds. Specific filtering criteria were carefully selected to narrow down the set to 48 to 96 designs for each target.

Two sided partial diffusion to optimize binders

Partial diffusion enables the input structure to be noised only up to a user-specified timestep instead of completing the full noising schedule. The starting point of the denoising trajectory is therefore not a random distribution. Rather, it contains information about the input distribution resulting in denoised structures that are structurally similar to the input. Unlike one sided directional partial diffusion, which solely diversifies the conformation of the binder while keeping the target fixed, two sided partial diffusion allows simultaneous conformational changes in both the target and the binder.The input designs were subjected to 15 noising timesteps out of a total of 50 timesteps in the noising schedule, and subsequently denoised. Approximately ten to fifty thousand partially diffused designs were generated for each target. The resulting library of backbones were sequence designed using ProteinMPNN, followed by AF2+initial guess (Bennett *et al.*, 2023). The resulting designs were filtered in the same way as the designs from the aforementioned sequence input diffusion process.

Integrating secondary structure specifications into RFdiffusion

To permit specification of the secondary structure (but not three-dimensional coordinates) of the peptide target, a modified version of RFdiffusion was trained that permits specification of the secondary structure of a region, along with its sequence. The training strategy largely followed that used to train previous RFdiffusion models [cite Watson & Vazquez-Torres], with some modifications. A summary is provided below.

Overview of "base" RFdiffusion Training: RFdiffusion [cite Watson] is a denoising diffusion probabilistic model (DDPM), which is fine-tuned from the RoseTTAFold structure prediction model [cite RF1 and RF2 papers]. In RFdiffusion, the N-Ca-C frame representation (translation and orientation) of protein backbones [cite AF2, RF1] is used, and, over 200 discrete timesteps, these backbone frames are corrupted following a defined forward noising process that noises these frames to distributions indistinguishable from random distributions (three-dimensional Gaussian distribution for translations, and uniform SO(3) distribution for rotations). RFdiffusion is trained to reverse this noising process, predicting the true (X0) protein structure at each timestep of prediction (starting from randomly sampled translations and rotations). Successive predictions are used to "self-condition" predictions through an inference trajectory, and mean squared error (MSE) losses minimize the error between forward and reverse processes. Full details of training are described in Watson et al [ref].

Modifications to permit secondary structure specification of the target: As in the original RFdiffusion fine-tuned for protein binder design, RFdiffusion was trained 50% of the time on single chains from the Protein Data Bank (PDB) < 384 amino acids in length, and 50% on hetero-complexes. In the latter case, one chain (< 250 amino acids in length) was designated the "binder", and when necessary the other "target" chain was radially cropped around the interface (to 384 – the length of the "binder" residues). For single chain examples, 20% of the time, the whole backbone was noised, and in the other 80% of cases 20-100% of the protein backbone was noised. For hetero-complex examples, the whole "binder" chain was noised. Additionally, and in contrast to the original RFdiffusion model trained for protein binder design, up to 50% of the noised monomer structure had sequence provided in the noised region. For hetero-complexes, up to 50% of the target chain backbone was also noised, while its sequence was provided to RFdiffusion. This permits RFdiffusion to condition on the sequence of the target chain in the absence of three-dimensional structure.

To permit specification of the secondary structure of the target (when three-dimensional coordinates are not provided), secondary structure and "block adjacency" [ref Watson] information were provided to RFdiffusion in exactly the manner described in Watson et al [ref]. Briefly, 50% of the time, RFdiffusion was provided with a (partially masked; 0-75%) secondary structure of the example protein chain/hetero-complex, and (an independently-sampled) 50% of the time a (partially masked; 0-75%) "block adjacency" of the protein chain/hetero-complex. Additionally, 50% of the time, the whole inter-chain "block adjacency" was masked in hetero-complex examples. This permits RFdiffusion to condition on a (partially) pre-specified secondary structure (and/or adjacency information) of the target peptide. This version of RFdiffusion was trained for seven epochs.

To design binders using RFdiffusion through secondary structure specification, foreach target, approximately ten thousand diffused designs were generated through sequence input of the target with the additional secondary structure specification. The resulting library of backbones were sequence designed using ProteinMPNN(Dauparas *et al.*, 2022), followed by AF2+initial guess (Bennett *et al.*, 2023). The resulting designs were filtered in the same way as the designs from the aforementioned sequence input diffusion process.

Backbone extension for VP48 binder design

During the design campaign, it was noticed not all designs provided sufficient interactions to the whole sequence of the target, especially the loopy regions. To explore and guide RFdiffusion to make more interactions around certain regions, we selected 20 AF2 passing designed complexes from the round one design campaign, based on the above criteria and manual selection. For each base design, we requested RFdiffusion to extend the binder backbone with 10-20 amino acids from either N terminal, or C terminal, or both (depending on where the loopy region was located. This was done with the inpaint flavor published in the original RFdiffusion work [ref]. 2,000 trajectories were performed each run, followed by the same MPNN and AF2 predictions as above.

Computational filtering

Precise metrics cutoffs changed for each design campaign to get to an orderable set, but largely focused on interface pAE <10, pLDDT >90, number of hydrogen bonds >11, RMSD < 0.5, sap score <45 and Rosetta ddG < -40 (Bennett *et al.*, 2023) .

Gene construction of peptide binders

The designed protein sequences were optimized for expression in *E. coli*. Linear DNA fragments (eBlocks, Integrated DNA Technologies) encoding design sequences included overhangs suitable for Golden Gate cloning into LM670 vector (Addgene #191552) for protein expression in *E.coli*. LM670 is a modified expression vector containing a Kanamycin resistance gene, a ccdB lethal gene between BsaI cut sites, and a C-terminal hexahistidine, commonly referred to as His tag.

Binding screening by Bio-layer interferometry (BLI) or co-lysis of binder and target peptide

For screening for all designs except the ones of partial diffusion design for Amylin-68n(Fig.2a), the designs were screened by BLI (method details described in below relative description). Linear gene fragments encoding binder design sequences were cloned into LM670 using Golden Gate assembly.Golden Gate subcloning reactions of peptide binders were constructed in 96-well PCR plates in 4µL volume. 1µL reaction mixtures were then transformed into a chemically competent expression strain (BL21 (DE3)). After 1 hour recovery in 100 µL SOC medium, the transformed cell suspensions were directly transferred into a 96-deep well plate containing 900 µL of LB media with Kanamycin. After overnight incubation in 37 ºC, 100 µL of growth culture were inoculated into 96-deep well plates containing 900 µL of auto-induction media (autoclaved TBII media supplemented with Kanamycin, 2mM MgSO4, 1X 5052). After overnight incubation(6 hours at 37 ºC followed by additional 18 hours at 30ºC), cells were harvested by centrifugation (15 min at 4000 x g). Bacteria were lysed for 15 minutes in 200 µL lysis buffer (1x BugBuster (Millipore#70921-4), 0.01 mg/mL DNAse, 1 tablet of pierce protease inhibitor tablet/50 mL culture). Lysates were clarified by centrifugation at 4000 g for 10 minutes, before purification on Ni-charged MagBeads (genscript #L00295; wash buffer: 25 mM Tris pH 8.0, 300 mM NaCl, 30 mM Imidazole; elution buffer: 25 mM Tris pH 8.0, 300 mM NaCl, 400 mM Imidazole). Subsequently, the elutions were directly subjected to a BLI test and the final

concentration is approximately 1 μM. The designs exhibiting binding signals were subsequently analyzed by BLI through titration.

For Amylin-68n, the designs from partial diffusion were expressed and purified using the same way as mentioned above. In addition to the designs, plasmids expressing target peptide fused with sfGFP (no His tag) were transformed into BL21 (DE3) cells, and overnight outgrowths were cultured in 5 mL of LB media with Kanamycin. After overnight incubation in 37 ºC and 250 rpm, growth cultures were inoculated into 50 mL auto-induction media. After overnight incubation in 37 ºC and 250 rpm, cells were harvested by centrifugation (15 min at 4000 x g), then resuspended in 20 mL lysis buffer (25 mM Tris-HCl, 150 mM NaCl, 0.1 mg/mL lysozyme, 10 μg/mL DNAse I, 1 mM PMSF). 100 μL of lysate of each binder were mixed with 100 μL of lysate of target peptide fused with sfGFP and incubated at room temperature for 15 min for co-lysis and target binding to the binders. Mixed lysates were applied directly to a 100 μL bed of Ni-NTA agarose resin in a 96-well fritted plate equilibrated with a Tris wash buffer. After sample application and flow through, the resin was thoroughly washed, and samples were eluted in 200 μL of a Tris elution buffer containing 300 mM imidazole. All eluates were sterile filtered with a 96-well 0.22μm filter plate (Agilent 203940-100) prior to size exclusion chromatography. Protein binders were then analyzed for target binding via sfGFP co-elution with the His-tagged binder. High-performance liquid chromatography (HPLC) analyses were conducted using an Agilent HPLC system (<product name>). Co-lysates were run on a Superdex200 Increase 5/150 GL column (Cytiva 28990945) with buffer of 25 mM Tris-HCl, 150 mM NaCl. To assess the binding interaction between the target and the binder, we monitored the elution profile of sfGFP using an absorbance wavelength of 395 nm, alongside a simultaneous measurement at 280 nm for total protein content to determine the extent of overlap between 395 nm and 280 nm, which indicates the binding interaction.

Medium scale protein expression and purification *E.coli* for hits from screening

For further validation, the initial hits were expressed at 50 mL scale via autoinduction for approximately 24 hours, in which the first 6 hours cultures were grown at 37 ºC and the remaining time at 22 ºC. Cultures were harvested at 4000 g for 10 minutes and resuspended in approximately 20 mL lysis buffer (25 mM Tris-HCl, 150 mM NaCl, 0.1 mg/mL lysozyme, 0.01

mg/mL DNAse, 1 mM PMSF,1 tablet of pierce protease inhibitor tablet/50 mL culture). Sonication was performed with a 4-prong head for 5 minutes total, 10 s pulse on-off at 80% amplitude. The resulting lysate was clarified by centrifugation at 14000 g for 30 minutes. Lysate supernatants were applied directly to a 1 mL bed of Ni-NTA agarose resin equilibrated. After sample application and flow through, the resin was thoroughly washed, and samples were eluted by an elution buffer containing 400 mM imidazole. After elution, protein samples were filtered and injected into an autosampler-equipped Akta pure system on a Superdex S75 Increase 10/300 GL column at room temperature. The SEC running buffer was 25mM Tris-HCl, 150mM NaCl pH 8. Protein concentrations were determined by absorbance at 280 nm using a NanoDrop spectrophotometer (Thermo Scientific) using their extinction coefficients and molecular weights obtained from their amino acid sequences.

Bio-layer interferometry (BLI) binding experiments

BLI experiments were performed on an Octet Red96 (ForteBio) instrument, with streptavidin coated tips (Sartorius Item no. 18-5019). Buffer comprised 1X HBS-EP+ buffer (Cytiva BR100669) supplemented with 0.1% w/v bovine serum albumin. Prior to target loading, each design was tested for binding against unloaded tips. 50 nM of biotinylated target protein was loaded on the tips for 50 s followed by a 60 s baseline measurement. After loading, all designs underwent a 60 s baseline, 300 s association and 200 s dissociation. Baseline measurements of unloaded tips were subtracted from their matched measurement of the loaded tip. The hits were taken forward for further titration experiments where concentration, association and dissociation times were chosen based on apparent affinity from the single point screen. Global kinetic fitting was used to determine KDs across the dilution series.

Circular dichroism (CD) experiments

For CD experiments, designs were diluted to 0.4mg/ml in 25 mM Tris-HCl and 150 mM NaCl. Spectra were acquired on a JASCO J-1500 CD Spectrophotometer. Thermal melt analyses were performed between 25 ^{o}C and 95 ^{o}C, measuring CD at 222 nm. All reported measurements were acquired within the linear range of the instrument.

Affinity enrichment of Amylin analyzed by LC-MS/MS

Bead preparation

Anti-amylin binder-coated beads were prepared by conjugating each amylin-targeted binder (Am16n-7, Am16n-12) to paramagnetic M280 Tosylactivated beads (Invitrogen, MA, USA). Each sample reaction conjugated 1 µg of binder to 225 µg of beads. Beads were blocked with a solution of 0.01% bovine serum albumin (BSA) in 0.2 M Tris to minimize non-specific interactions. An off-target binder-conjugated bead was included for quantification of non-specific binding. A BSA-blocked bead without a bound binder was used as a negative control and an anti-GPVGPSGPPGK (GPVG) peptide monoclonal antibody-conjugated bead was used as a positive control for the affinity binding step.

Sample preparation

Human amylin peptide (non-amidated) was purchased from Anaspec (Fremont, CA, USA) and reconstituted to 2 mg/mL in dimethylsulfoxide (DMSO). A secondary peptide stock (diluted into 50 µM in 5% acetonitrile, 0.1% formic acid, 0.01% BSA in water) was reduced with dithiothreitol (10 mM final concentration) and alkylated with iodoacetamide (30 mM final concentration). Excess iodoacetamide was quenched with additional dithiothreitol (5 mM final added concentration). This solution was diluted to a working stock of 10 µM with dilution solvent. Aliquots of the working stock were made in 1.5 mL LoBind tubes and stored at -20°C to avoid repeated freeze/thaw cycles.

Human specimens

Human plasma samples were composed of pooled de-identified leftover clinical samples obtained from the clinical laboratories at the University of Washington Medical Center. The use of de-identified leftover clinical samples was reviewed by the University of Washington Human Subjects Division (STUDY00013706).

Affinity enrichment

Amylin capture experiments were performed using four types of coupled beads (Am16n-7, Am16n-12, off-target binder, BSA-blocked) in phosphate-buffered saline (PBS) containing 0.1%

3-((3-cholamidopropyl) dimethylammonio)-1-propanesulfonate (CHAPS) as well as pooled normal human EDTA-anticoagulated plasma.

Samples were prepared by spiking the working stock of alkylated amylin to a final concentration of 20 nM in 100 µL of either PBS-CHAPS or pooled plasma. Additional PBS-CHAPS was added to each sample, followed by coupled beads. GPVG peptide and anti-GPVG monoclonal antibody-conjugated beads were added to each sample as a positive control. The mixtures were shaken for 1 hr at 900 rpm and room temperature (Thermomixer, Eppendorf, Framingham MA). The supernatant was removed and the beads were washed twice with 200 µL of PBS-CHAPS. Bound peptides were eluted in 50 µL of elution solvent (20% acetic acid, 10% acetonitrile, 10% DMSO, 0.001% BSA in water) with shaking for 8 min (900 rpm, room temperature). Each bead type (two anti-amylin binders, one off-target binder, one BSA-blocked) was assessed in separate samples and each was prepared in triplicate.

Sample analysis was performed by liquid chromatography-tandem mass spectrometry using a Shimadzu Nexera LC-XR HPLC (Columbia, MD, USA) coupled to a Sciex 6500+ triple quadrupole tandem mass spectrometer (Framingham, MA, USA) in multiple reaction monitoring (MRM) mode. Specifications for the liquid chromatography, mass spectrometer, and MRM methods are included in Supplementary Tables x, x, and x.

Data analysis

Data processing was performed with Skyline Daily (version 23.1.1.459). Chromatographic peak area was calculated by summing the peak area of all transitions for each peptide. The chromatographic peak areas observed during blank (elution solvent) injections were subtracted as background from sample peak areas before performing further data reduction. Signal from BSA and GPVG beads were for quality control of the assay and evaluated prior to processing of the experimental data.

Seven types of samples were analyzed:

1. Group A: Alkylated amylin peptide spiked directly into elution solvent served as the reference peak area for 100% recovery of amylin peptide.

2. Group B: Paramagnetic tosyl-activated beads conjugated to an off-target binder were incubated in PBS-CHAPS spiked with alkylated amylin. The peak area of this negative control was used to quantify nonspecific binding.

3. Group C: Amylin-targeted binders conjugated to paramagnetic tosyl-activated beads were incubated in PBS-CHAPS spiked with alkylated amylin. The peak areas of these samples were used to quantify the percent recovery of amylin by affinity enrichment.

4. Group D: An off-target binder conjugated to paramagnetic tosyl-activated beads was incubated with unspiked plasma. The peak area of this negative control was used to quantify the nonspecific signal from beads binding to plasma components.

5. Group E: Amylin-targeted binders conjugated to paramagnetic tosyl-activated beads were incubated with unspiked plasma. The peak areas observed in these samples were used to quantify the nonspecific signal from the binders binding to plasma components (i.e., assuming no non-amidated amylin in normal plasma).

6. Group F: An off-target binder conjugated to paramagnetic tosyl-activated beads was incubated with spiked plasma. The peak area of this negative control was used to quantify nonspecific binding.

7. Group G: Amylin-targeted binders conjugated to paramagnetic tosyl-activated beads were incubated with spiked plasma. The peak areas of these samples were used to quantify percent recovery of amylin by affinity enrichment.

The percent recovery of each binder-coated bead type was calculated using the following equations:

$$Percent\ recovery_{buffer} = \frac{Group\ C - Group\ B}{Group\ A}$$

$$Percent\ recovery_{Plasma} = \frac{(Group\ G - Group\ F) - (Group\ E - Group\ D)}{Group\ A}$$

Supplementary Table 2. Amylin transitions monitored

Peptide Sequence	Q1 (m/z) (charge state)	Q3 (m/z)	Ion type
KCNTATCATQRLANFLVHSSNNFGAILSSTNVGSNTY	976.90 (4+)	921.59	b_{26}^{3+}
		988.48	b_{28}^{3+}
		931.90	b_{36}^{3+}
		541.18	y_5^+
GPVGPSGPPGK	475.26 (2+)	795.44	y_9^+
		696.37	y_8^+
		639.35	y_7^+
		398.24	y_4^+
		301.19	y_3^+
		155.08	b_2^+
GPVGPSGPPGK[$^{13}C_6$, $^{15}N_2$]^	479.27 (2+)	704.38	y_8^+
		647.36	y_7^+
K^ = $^{13}C_6H14^{15}N_2O2$ (+8 Da)		406.25	y_4^+

Supplementary Table3. Liquid chromatography parameters		
Mobile phase	Phase A: 0.2% formic acid in water	
	Phase B: 0.2% formic acid in acetonitrile	
Column	Acquity UPLC HSS T3 1.8µm (C18, 2.1x50 mm, pore size 100 Å) (Waters, Milford, MA, P/N 186003539)	
Temperature	45°C	
Flow rate	0.3 mL/min	
Injection volume	10µL	
Gradient	0-0.5 min	20% B at 0.3 mL/min
	7.5 min	60% B at 0.3 mL/min

	9.5 min	98% B at 0.3 mL/min
	11.0 min	98% B at 0.3 mL/min
	11.1 min	20% B at 0.3 mL/min
	12.5 min	20% B at 0.3 mL/min

Supplementary Table 3. Mass spectrometry parameters	
Source Polarity	ESI+
Curtain Gas	35
Collision Gas	9
Ionspray Voltage	5500 V
Source Temperature	400°C
Ion Source Gas 1	40
Ion Source Gas 2	40

Preparation of SSM libraries

We performed SSM studies for some of the designed peptide–protein binding pairs to gain a better understanding of the peptide-binding modes, and to search for improved peptide binders. For CP35, we ordered a SSM library covering all the 159 amino acids. The chip synthesized DNA oligos for the SSM library were then amplified and transformed to EBY100 yeast together with a linearized pETCON3 vector. Each SSM library was subjected to an expression sort first, in which the low-quality sequences due to chip synthesizing defects or recombination errors were filtered out. The collected yeast population, which successfully expresses the designed mutants, will be regrown, and subjected to the next round of peptide-binding sorts. Two rounds of with-avidity sorts were applied at 1 μM concentration of C-peptides followed by 1 rounds of without-avidity sorts with C-peptide concentrations at 200nM, 40 nM, 8nM, 1.6nM and 0.32nM. The peptide-bound yeast populations were collected and sequenced using the Illumina NextSeq kit. The mutants were identified and compared to the mutants in the expression libraries.

Enrichment analysis was used to identify beneficial mutants and provide information for interpreting the peptide-binding modes. For each mutant, the fraction of cells collected in each of 5 titration sorts of decreasing concentration is measured. The SortingConcentration50 (SC50), the concentration where 50% of the expressing cells are collected, is calculated and plotted in heat maps for the SSM analysis.

X-ray crystallography

Crystallization experiments were conducted using the sitting drop vapor diffusion method. Initial crystallization trials were set up in 200 nL drops using the 96-well plate format at 20 ˚C. Crystallization plates were set up using a Mosquito LCP from SPT Labtech, then imaged using UVEX microscopes and UVEX PS-256 from JAN Scientific. Diffraction quality crystals formed in 0.1M succinic acid, sodium phosphate monobasic monohydrate, glycine mixture at pH 6 and 30% w/v PEG 1000 for Amylin-22. For g3bp1-11 diffraction quality crystals appeared in 0.05 M Calcium chloride dihydrate, 0.1 M BIS-TRIS pH 6.5, and 30% v/v Polyethylene glycol monomethyl ether 550.

Diffraction data was collected at the National Synchrotron Light Source II on beamline 17-ID-1 (AMF) for Amylin-22 and Advanced Light Source beamline 821 for g3bp1-11. X-ray intensities and data reduction were evaluated and integrated using XDS (Kabsch, 2010) and merged/scaled using Pointless/Aimless in the CCP4 program suite (Winn *et al.*, 2011). Structure determination and refinement starting phases were obtained by molecular replacement using Phaser (McCoy *et al.*, 2007) using the designed model for the structures. Following molecular replacement, the models were improved using phenix.autobuild; with rebuild-in-place to false, and using simulated annealing. Structures were refined in Phenix (Adams *et al.*, 2010)(4). Model building was performed using COOT (Emsley and Cowtan, 2004). The final model was evaluated using MolProbity (Williams *et al.*, 2018). Data collection and refinement statistics are recorded in Table 3. Data deposition, atomic coordinates, and structure factors reported in this paper have been deposited in the Protein Data Bank (PDB), http://www.rcsb.org/ with accession code 9CC5 and 9CC6.

Supplementary table 3. Data collection and refinement statistics.

	Amylin-22 (PDB Code: 9CC5)	G3bp1-11 (PDB Code: 9CC6)
Resolution range	33.32 - 1.87 (1.94 - 1.87)	31.43 - 2.4 (2.48 - 2.4)
Space group	$P\ 2_1\ 2_1\ 2_1$	$P\ 2_1$
Unit cell	33.33, 34.51, 127.68; 90, 90, 90	38.43, 42.39, 40.63; 90, 105.53, 90
Unique reflections	12855 (1372)	4698 (410)
Multiplicity	6.6 (6.8)	5.3 (5.0)
Completeness (%)	99.38 (99.28)	93.38 (92.0)
Mean I/sigma(I)	7.94 (1.22)	11.0 (2.2)
Wilson B-factor	28.60	38.45
R-merge	0.1598 (1.79)	0.033 (0.623)
R-pim	0.06738 (0.7403)	0.032 (0.152)
$CC_{1/2}$	0.999 (0.769)	0.993 (0.956)
Reflections used in refinement	12756 (1372)	4697 (410)
R-work	0.2256 (0.2949)	0.2199 (0.2709)
R-free	0.2721 (0.3400)	0.2567 (0.2856)
Number of non-hydrogen atoms	1335	1197
macromolecules	1289	1190
solvent	46	7
Protein residues	163	150
RMS(bonds)	0.015	0.003
RMS(angles)	1.31	0.60

Ramachandran favored (%)	95.60	98.63
Ramachandran allowed (%)	4.40	1.37
Ramachandran outliers (%)	0.00	0.00
Average B-factor	35	46
macromolecules	35	46
solvent	40	43

The highest-resolution shell are shown in parentheses.

Cell culture

HeLa cells were cultured in DMEM (Gibco, 11965-092) at 37 °C in a humidified atmosphere containing 5% CO2, supplemented with 10% (v/v) FetalClone II serum (Cytiva, SH3006603) and 1% penicillin–streptomycin (ThermoFisher, 15140122) .

Generation of IL2RG-knockout HeLa cells by CRISPR–Cas9 gene targeting

Pooled IL2RG-knockout HeLa cells was generated using the Gene Knockout kit V2 from Synthego, using multi-guide sgRNA targeting IL2RG (guide 1: CAUACCAAUAAUGCAGAGUG, guide 2: UCGAGUACAUGAAUUGCACU and guide 3: GAAACACUGAGGGAGUCAGU). The ribonucleoprotein complex with a ratio of 4.5:1 of sgRNA and Cas9 was delivered following the protocol of the SE Cell Line 4D-Nucleofector™ X Kit S (Lonza, V4XC-1032), using the nucleofection program CN-114 on the Lonza 4D X unit.

Transient transfection

Plasmids for IL2RG-30-mScarlet, IL2RG-EGFP were synthesized and cloned by Genscript USA, Inc. HeLa cells were seeded at 70–80% confluency in a chambered coverslip with 18 wells (ibidi, 81816). At the same time, HeLa cells were reverse-transfected using Lipofectamine 3000 transfection reagent (ThermoFisher, L3000008) according to the manufacturer's protocol.

Fluorescence imaging using 3D structured illumination microscopy

4-color, 3D images were acquired with a commercial OMX-SR system (GE Healthcare). Toptica diode lasers with excitation at 488 nm, and 568 nm were used. Emission was collected on three separate PCO.edge sCMOS cameras using an Olympus 60× 1.42NA PlanApochromat oil immersion lens. 512×512 images (pixel size 6.5 μm) were captured with no binning. Acquisition was controlled with AcquireSR Acquisition control software. Z-stacks were collected with a step size of 250 nm. Images were deconvolved in SoftWoRx 7.0.0 (GE Healthcare) using the ratio method and 200 nm noise filtering. Images from different color channels were registered in SoftWoRx using parameters generated from a gold grid registration slide (GE Healthcare).

Thioflavin-T (ThT) fluorescence assay

Amylin fibrils at various growth stages (0 h, 3 h and 24 h) with a concentration of 10 μM were adequately mixed with 10 μM ThT and added into 96-well-plates containing different types and concentrations of binders (Am75, Am36, Am22, Am68n). The samples were then incubated at 37 °C for 6-12 hours with 600 rpm orbital shaking. ThT fluorescence signals were measured using a Thermo Varioskan Flash Multi Detection Microplate Reader (0 h and 3 h) or a Perkin elmer EnSight Multifunctional Microplate Reader (24 h) with excitation wavelength at 440 nm and an emission wavelength at 482 nm.

Negative-stain electron microscopy(NS-EM) experiment

Samples for negative-stain electron microscopy were dropped onto freshly glow-discharged carbon-coated copper grids and incubated for 1 minute, and excess sample was removed by blotting on filter paper. The grids were then stained with 2 % (w/v) uranyl acetate for 1 minute, and excess uranyl acetate was blotted off. Finally, the grids were examined using a Tecnai Spirit transmission electron microscope (FEI) at an acceleration voltage of 120 kV.

Acknowledgements:

This research used resources (FMX/AMX) of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. The Center for BioMolecular Structure (CBMS) is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011). Amylin fibril inhibition and dissociation work is supported by the Chinese Academy of Sciences (CAS) (XDB37010100), and the basic Research Program Based on Major Scientific Infrastructures, CAS (JZHKYPT-2021-05). We appreciate the help provided by David Juergens in the training of the RFdiffusion model with Joseph L. Watson.

Author contributions

D.B. directed the work. C.L., K.W. and D.B. designed the research. C.L., H.C., K.W., and H.H. designed, screened and experimentally characterized the binders. J.L.W. developed the sequence input RFdiffusion and secondary structure specification algorithm used for IDP/IDR binder design. C.L. prepared samples for crystallography, A.K.B., A.K. and E.B. obtained all the crystal structures shown in this manuscript. H.C., C.L., and K.W. performed all-by-all specificity BLI. W.Y. constructed the SSM library. C.L. screened SSM library and analyzed the SSM result with the help from B.C., D.R.H. and X.W.. H.C. and J.D. experimentally validated the colocalization of IL2RG binders to the target in mammalian cells. S.S. and A.N.H carried out the LC–MS/MS peptide detection. X.Z. and P.Z. performed the Amylin fibrils formation inhibition and Amylin fibrils dissociation experiments. S.R.G., A.M. and M.L. carried out additional scaled-up protein

purification. M.B. helped with prion binder design. C.L., K.W., H.C. and D.B. wrote the manuscript with input from the other authors. All authors revised the manuscript.

Method references

[1] Bennett, N. R. et al. Improving de novo protein binder design with deep learning. Nat. Commun. 14, 2625 (2023).

[2] B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, and D. Baker. Hallucinating symmetric protein assemblies. Science, 378(6615):56–61, 2022.

Bennett, N.R., Coventry, B., Goreshnik, I., Huang, B.W., Allen, A., Vafeados, D., Peng, Y.P., Dauparas, J., Baek, M., Stewart, L.*, et al.* (2023). Improving de novo protein binder design with deep learning. Nat Commun *14*.

Danny, D.S., Ewa, A.A., Hannah, L.H., Enrico, R., Matthias, M.S., Georg, M., Maggie, A., Justin, D., Hannah, N., Alex, K.*, et al.* (2023). Design of amyloidogenic peptide traps. bioRxiv, 2023.2001.2013.523785.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I.M., Courbet, A., de Haas, R.J., Bethel, N.*, et al.* (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science *378*, 49-55.

He, J.F., Dai, J., Li, J., Peng, X.B., and Niemi, A.J. (2015). Aspects of structural landscape of human islet amyloid polypeptide. J Chem Phys *142*.

Iqbal, S., Jayyab, A.A., Alrashdi, A.M., and Reverté-Villarroya, S. (2023). The Predictive Ability of C-Peptide in Distinguishing Type 1 Diabetes From Type 2 Diabetes: A Systematic Review and Meta-Analysis.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zídek, A., Potapenko, A.*, et al.* (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-+.

Roberts, A.N., Leighton, B., Todd, J.A., Cockburn, D., Schofield, P.N., Sutton, R., Holt, S., Boyd, Y., Day, A.J., Foot, E.A.*, et al.* (1989). Molecular and

Functional-Characterization of Amylin, a Peptide Associated with Type-2 Diabetes-Mellitus. P Natl Acad Sci USA *86*, 9662-9666.

Sahtoe, D.D., Coscia, A., Mustafaoglu, N., Miller, L.M., Olal, D., Vulovic, I., Yu, T.Y., Goreshnik, I., Lin, Y.R., Clark, L.*, et al.* (2021). Transferrin receptor targeting by de novo sheet extension. Proc Natl Acad Sci U S A *118*.

Wei, Y., Quan, L., Zhou, T., Du, G., and Jiang, S. (2021). The relationship between different C-peptide level and insulin dose of insulin pump. Nutrition & Diabetes *11*, 7.

Westermark, P. (2011). Amyloid in the islets of Langerhans: Thoughts and some historical aspects. Upsala J Med Sci *116*, 81-89.

Wu, K., Bai, H., Chang, Y.T., Redler, R., McNally, K.E., Sheffler, W., Brunette, T.J., Hicks, D.R., Morgan, T.E., Stevens, T.J.*, et al.* (2023). De novo design of modular peptide-binding proteins by superhelical matching. Nature *616*, 581-589.

Zhang, G., Meng, L., Wang, Z., Peng, Q., Chen, G., Xiong, J., and Zhang, Z.A.-O. (2022). Islet amyloid polypeptide cross-seeds tau and drives the neurofibrillary pathology in Alzheimer's disease.

Minezaki Y, Homma K, Nishikawa K. Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment. J Mol Biol 2007; 368:902-13.

De Biasio A, Guarnaccia C, Popovic M, Uversky VN, Pintar A, Pongor S. Prevalence of intrinsic disorder in the intracellular region of human single-pass type I proteins: The case of the notch ligand Delta-4. J Proteome Res 2008; 7:2496-506.

Fernández-Quintero et al., Conformational selection of allergen-antibody complexes—surface plasticity of paratopes and epitopes, Protein Engineering, Design and Selection, 2020

Olejniczak, E.T., Ruan, Q., Ziemann, R.N. et al. (2010) Biopolymers, 93, 657–667.

Warren J. Leonard et al., The γc Family of Cytokines: Basic Biology to Therapeutic Ramifications, Immunity, 2019
Scheckel, C., Aguzzi, A. Prions, prionoids and protein misfolding disorders. Nat Rev Genet 19, 405–418 (2018). https://doi.org/10.1038/s41576-018-0011-4
Aguzzi et al. Molecular Mechanisms of Prion Pathogenesis, Annual Review of Pathology: Mechanisms of Disease , 2008
Prusiner, S. B. Novel proteinaceous infectious particles cause scrapie. Science 216, 136–144 (1982).
Bolton, D. C., McKinley, M. P. & Prusiner, S. B. Identification of a protein that purifies with the scrapie prion. Science 218, 1309–1311 (1982).
S.B. Prusiner, D.F. Groth, D.C. Bolton, S.B. Kent, L.E. Hood. Purification and structural studies of a major scrapie prion protein. Cell, 38 (1984)
Thody et al. Mechanism of aggregation and membrane interactions of mammalian prion protein. BBA-Biomembranes, 2018
Kabsch, W. XDS. Acta Crystallogr. D Biol. Crystallogr. 66, 125–132 (2010).
Winn, M. D. et al. Overview of the CCP 4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. 67, 235–242 (2011).
McCoy, A. J. et al. Phaser crystallographic software. J. Appl. Crystallogr. 40, 658–674 (2007).
Adams, P. D. et al. PHENIX : a comprehensive Python-based system for macromolecular
structure solution. Acta Crystallogr. D Biol. Crystallogr. 66, 213–221 (2010).
Emsley, P. & Cowtan, K. Coot : model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132 (2004).
Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation: PROTEIN SCIENCE.ORG. Protein Sci. 27, 293–315 (2018).

Chapter 4 – De novo design of modular peptide-binding proteins by superhelical matching

4.0 – preface, authors, and abstract

Preface:

Note: this chapter is borrowed directly from the published article in Nature, which is the original copy. I, Kejia Wu, am co-first author for this work.

Authors:

Kejia Wu[1,2,3,#], Hua Bai[1,2,10#], Ya-Ting Chang[4], Rachel Redler[4], Kerrie E. McNally[7], William Sheffler[1,2], TJ Brunette[1,2], Derrick R. Hicks[1,2], Tomos E Morgan[7], Tim J Stevens[7], Adam Broerman[1,2,6], Inna Goreshnik[1,2], Michelle DeWitt[1,2], Cameron M. Chow[1,2], Yihang Shen[11], Lance Stewart[1,2], Emmanuel Derivery[7*], Daniel Adriano Silva[1,2,8,9*], Gira Bhabha[4], Damian Ekiert[4,5], David Baker[1,2,10*]

1. Department of Biochemistry, University of Washington, Seattle, WA, USA.

2. Institute for Protein Design, University of Washington, Seattle, WA, USA.

3. Biological Physics, Structure and Design Graduate Program, University of Washington, Seattle, WA, USA

4. Department of Cell Biology, New York University School of Medicine, New York, NY, USA

5. Department of Microbiology, New York University School of Medicine, New York, NY, USA

6. Department of Chemical Engineering, University of Washington, Seattle, WA, USA.

7. MRC Laboratory of Molecular Biology, Cambridge CB2 0QH UK.

8. Division of Life Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

9. Monod Bio, Inc, Seattle, WA, USA.

10. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

11. Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA

[#]These authors contributed equally

*Corresponding author. E-mail: dabaker@uw.edu

dadriano@gmail.com

derivery@mrc-lmb.cam.ac.uk

Abstract

General approaches for designing sequence-specific peptide binding proteins would have wide utility in proteomics and synthetic biology. However, the design of peptide binding proteins is challenging as most peptides do not have defined structures in isolation, and hydrogen bonds must be made to the buried peptide backbone polar groups[1–3]. Inspired by natural and re-engineered protein-peptide systems[4–11], we set out to design proteins made out of repeating units that bind peptides with repeating sequences with a one to one correspondence between repeat units on the protein and peptide. We use geometric hashing to identify protein backbones and peptide docking arrangements compatible with bidentate hydrogen bonds between side chains on the protein and the peptide backbone[12]; the remainder of the protein sequence is then optimized for folding and peptide binding. We design repeat proteins to bind to six different tripeptide repeat sequences in polyproline II conformations; the proteins are hyperstable and bind 4-6 tandem repeats of their tripeptide targets with nanomolar to picomolar affinities in vitro and in living cells. Crystal structures reveal repeating interactions between protein and peptide interactions as designed, including ladders of protein sidechain to peptide backbone hydrogen bonds. By redesigning the binding interfaces of individual repeat units, specificity can be achieved for both synthetic and naturally occurring proteins especially containing proline rich disordered regions.

4.1 - Main

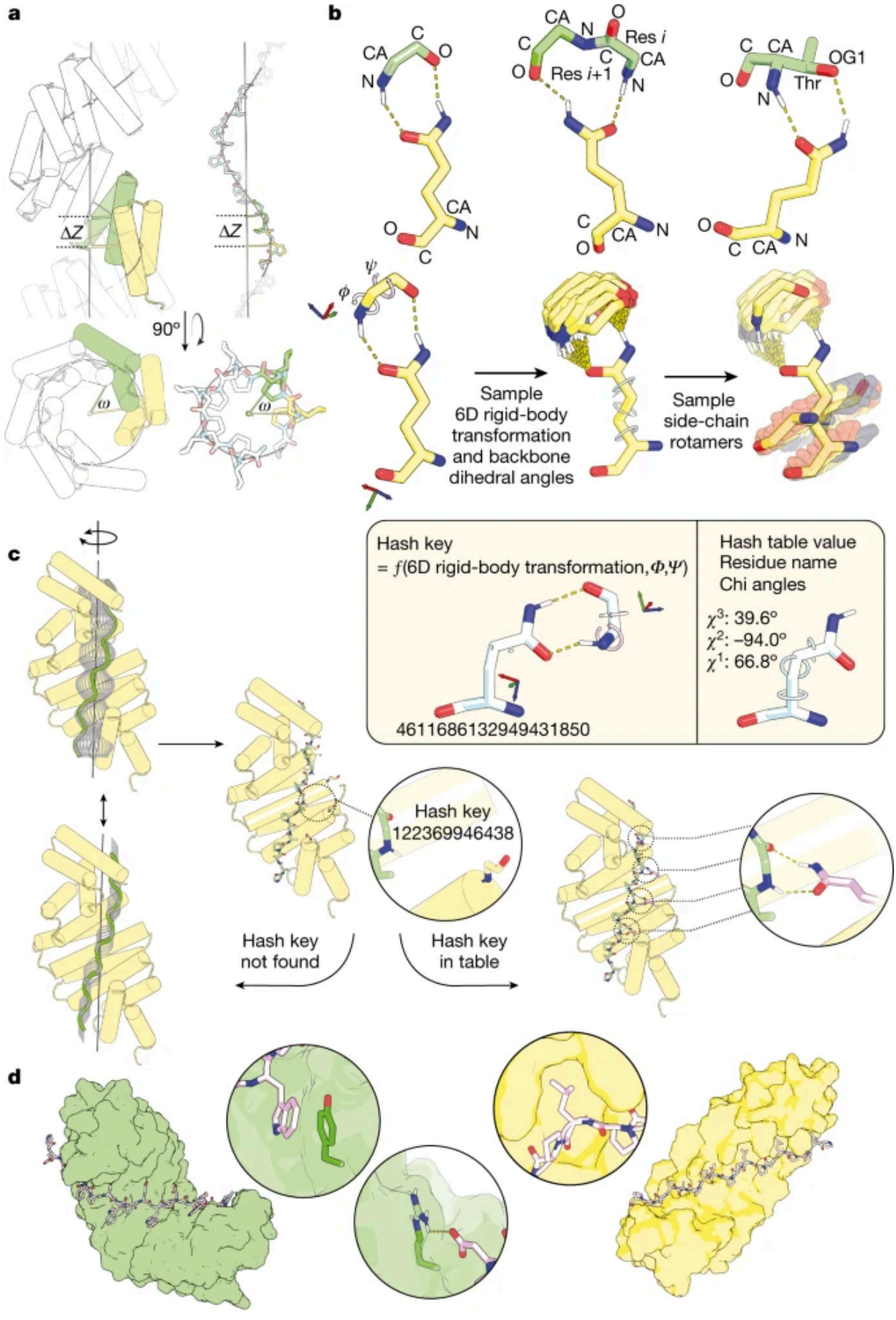
A number of naturally occurring protein families bind to peptides with repeating internal sequences[7, 9]. The Armadillo repeat proteins (ARM), which include the nuclear import receptors, bind to extended peptides with lysine and arginine rich sequences such that each repeat unit in the peptide fits into a repeat unit/module in the protein[5, 8]. The Plückthun group has demonstrated that the specificity of individual protein repeat units can be re-engineered, enabling broader peptide sequence recognition[6, 11, 13, 14]. While powerful, this approach is limited to binding peptides in backbone conformations compatible with the geometry of the armadillo repeat. Tetratricopeptide repeat proteins (TPRs) bind peptides with a variety of sequences and conformations, generally with relatively low affinity (~μM Kd; for exception see[15, 33]) and with deviations in peptide - protein register which complicates engineering for more general peptide recognition[4, 9, 10].

4.2- Design approach

We set out to generalize peptide recognition by modular repeat-protein scaffolds to arbitrary repeating peptide backbone geometries. This requires solving two main challenges: building protein structures with a repeat spacing and orientation matching that of the target peptide conformation, and ensuring the replacement of peptide-water hydrogen bonds in the unbound state with peptide-protein hydrogen bonds in the bound state. The first challenge is critical for modular and extensible sequence recognition: if individual repeat units in the protein are to bind

individual repeat units on the peptide in the same orientation, the geometric phasing of the repeat units on protein and peptide must be compatible. The second challenge is critical for achieving high binding affinity: in conformations other than the alpha and 3-10 helix, the NH and C=O groups make hydrogen bonds with water in the unbound state that need to be replaced with hydrogen bonds to the protein upon binding to avoid incurring a substantial free energy penalty[16].

To address the first challenge, we reasoned that a necessary (but not sufficient) criterion for in-phase geometric matching between repeating units on designed protein and peptide was a correspondence between the superhelices that the two trace out. All repeating polymeric structures trace out superhelices which can be described by three parameters: the translation (rise) along the helical axis per repeat unit, the rotation (twist) around this axis, and the distance (radius) of the repeat unit centroid from the axis (Fig. 1A)[17, 36]. As described in the methods, we generated large sets of repeating protein backbones sampling a wide range of superhelical geometries. We generated corresponding sets of repeating peptide backbones by randomly sampling di-peptide and tri-peptide conformations (avoiding intra-peptide steric clashes), and then repeating these four to six times to generate 12-24 residue peptides. We then searched for matching pairs of repeat protein and repeat peptide backbones, requiring that the rise be within 0.2A, the twist within 5 degrees, and the radius differ by at least 4Å (the difference in radius is necessary to avoid clashing between peptide and protein; the peptide can wrap either outside or inside the protein).


a
ΔZ
ΔZ
90°
ω
ω
b
C
CA
O
N
CA
O
N
C
Res i
Res i+1
C
CA
O
N
C
CA
N
O
C
CA
OG1
Thr
O
N
O
C
CA
N
ψ
ϕ
Sample
6D rigid-body
transformation
and backbone
dihedral angles
Sample
side-chain
rotamers
c
Hash key
= f(6D rigid-body transformation, Φ,Ψ)
4611686132949431850
Hash table value
Residue name
Chi angles
χ3: 39.6°
χ2: –94.0°
χ1: 66.8°
Hash key
122369946438
Hash key
not found
Hash key
in table
d

a, Like all repeating structures, repeat proteins and peptides form superhelices with constant axial displacement (Δ*Z*) and angular twist (*ω*) between adjacent repeat units (shown in green and yellow). For in-register binding, the protein and peptide parameters must match (for some integral multiple of repeat units). **b,** Construction of hash tables for privileged residue–residue interactions. Top row: classes of side-chain–backbone interactions for which hash tables were built. The side-chain amide group of asparagine or glutamine forms bidentate interactions with the N–H and C=O groups on the backbone of a single residue (left) or consecutive residues (middle), or with the backbone N–H group and side-chain oxygen atom of a serine or threonine residue (right). Second row: as illustrated for the case of the glutamine–backbone bidentate interaction, to build the hash table we perform Monte Carlo sampling over the rigid-body orientation between the terminal amide group and the backbone, and the backbone torsions *φ* and *ψ*, saving configurations with low-energy bidentate hydrogen bonds. For each configuration, the possible placements for the backbone of the glutamine are enumerated by growing side-chain rotamers back from the terminal amide. Third row: from the six rigid-body degrees of freedom relating the backbones of the two residues, together with the two *φ* and *ψ* torsion angle degrees of freedom, a hash key is calculated using an eight-dimensional hashing scheme. The hash key is then added to the hash table with the side-chain name and torsions as the value. CA, α-carbon; OG, γ-oxygen. **c,** To dock repeat proteins and repeat peptides with compatible superhelical parameters, their superhelical axes are first aligned, and the repeat peptide is then rotated around and slid along this axis. For each of these docks, for each pair of repeat protein–repeat peptide residues within a threshold distance, the hash key is calculated from the rigid-body transform between backbones and the backbone torsions of the peptide residue, and the hash table is interrogated. If the key is found in the hash table, side chains with the stored identities and torsion angles are installed in the docking interface. **d,** The sequence of the remainder of the interface is optimized using Rosetta for high-affinity binding. Two representative designed binding complexes are shown to highlight the peptide-binding groove

and the shape complementarity. The magnified views illustrate hydrophobic interactions (right), salt bridges (middle) and π–π stacks (left) incorporated during design.

To address the second challenge, we reasoned that bidentate hydrogen bonds between side chains on the protein and pairs of backbone groups or backbone and sidechain groups on the peptide could allow the burying of sufficient peptide surface area on the protein to achieve high affinity binding without incurring a large desolvation penalty[18]. As the geometric requirements for such bidentate hydrogen bonds are quite strict, we developed a geometric hashing approach to enable rapid identification of rigid body docks of the peptide on the protein compatible with ladders of bidentate interactions. To generate the hash tables for bidentate sidechain-backbone interactions, Monte Carlo simulations of individual sidechain functional groups making bidentate hydrogen bonding interactions with peptide backbone and/or sidechain groups were carried out using the Rosetta energy function[12], and a move set consisting of both rigid body perturbations and changes to the peptide backbone torsions (Fig. 1B; see Methods for details). For each accepted (low energy) arrangement, sidechain rotamer conformations were built backwards from the functional group to identify placements of the protein backbone from which the bidentate interaction could be realized. The results were stored in hash tables: for each placement, a hash key was computed from the rigid body transformation and peptide backbone and side chain torsion angles determining the position of the hydrogen bonding groups (for example the phi and psi torsion angles for a bidentate hydrogen bond to the NH and CO groups of the same amino acid), and the chi angles of the corresponding rotamer were stored in the hash for this key[18]. Hash tables were generated for ASN and GLN making bidentate interactions with the N-H and C=O groups on the backbone of a single residue or adjacent residues, ASP or GLU making bidentate interactions with the N-H groups of two successive amino acids, and for sidechain-sidechain pi-pi and cation-pi interactions (see Methods).

To identify rigid body docks that enable multiple bidentate hydrogen bonds between repeat protein and peptide, we took advantage of the fact that for matching two superhelical structures along their common axis, there are only two degrees of freedom: the relative translation and

rotation along this axis. For each repeat protein-repeat peptide pair, we carried out a grid search in these two degrees of freedom, sampling relative translations and rotations in ~1 Å and 10 degree increments (Fig. 1E). For each generated dock, we computed the rigid body orientation for each peptide-protein residue pair, and queried the hash tables to rapidly determine if bidentate interactions could be made; docks for which there were lower than a threshold number of matches were discarded. For the remaining docks, following building of the interacting side chains using the chi angle information stored in the hash, and rigid body minimization to optimize hydrogen bond geometry, we used Rosetta combinatorial optimization to design the protein and peptide sequences[20], keeping the residues identified in the hash matching fixed, and enforcing sequence identity between repeats in both peptide and protein (see Methods).

In initial calculations with unrestricted sampling of peptide conformations, designs were generated with a wide range of peptide conformations. Examples of repeat proteins designed to bind to extended beta strand, polypeptide II, and helical peptide backbones, as well as a range of less canonical structures are shown in Extended Data Fig. 1A-C. Reasoning that proline containing peptides would incur a lower entropic cost upon binding, we decided to start experimental characterization with designs containing at least one proline residue; in most such designs the peptide backbone is in or near the polyproline II portion of the Ramachandran map. Our design strategy requires matching the twist of the repeat unit of the peptide with that of the protein, and hence choosing a repeat length of the peptide that generates close to a full 360 degree turn requires less of a twist in the repeat protein; for the polyproline helix there are roughly 3 residues per turn and likely because of this we obtained more designs which target 3 residue than 2 residue proline containing repeat units. We selected for experimental characterization 43 designed complexes with near ideal bidentate hydrogen bonds between protein and peptide, favorable protein-peptide interaction energies[12], interface shape

complementary[21], and few interface unsatisfied hydrogen bonds[22], and which consistently retained more than 80% of the interchain hydrogen bonds in 20 ns molecular dynamics trajectories.

4.3 - Experimental characterization

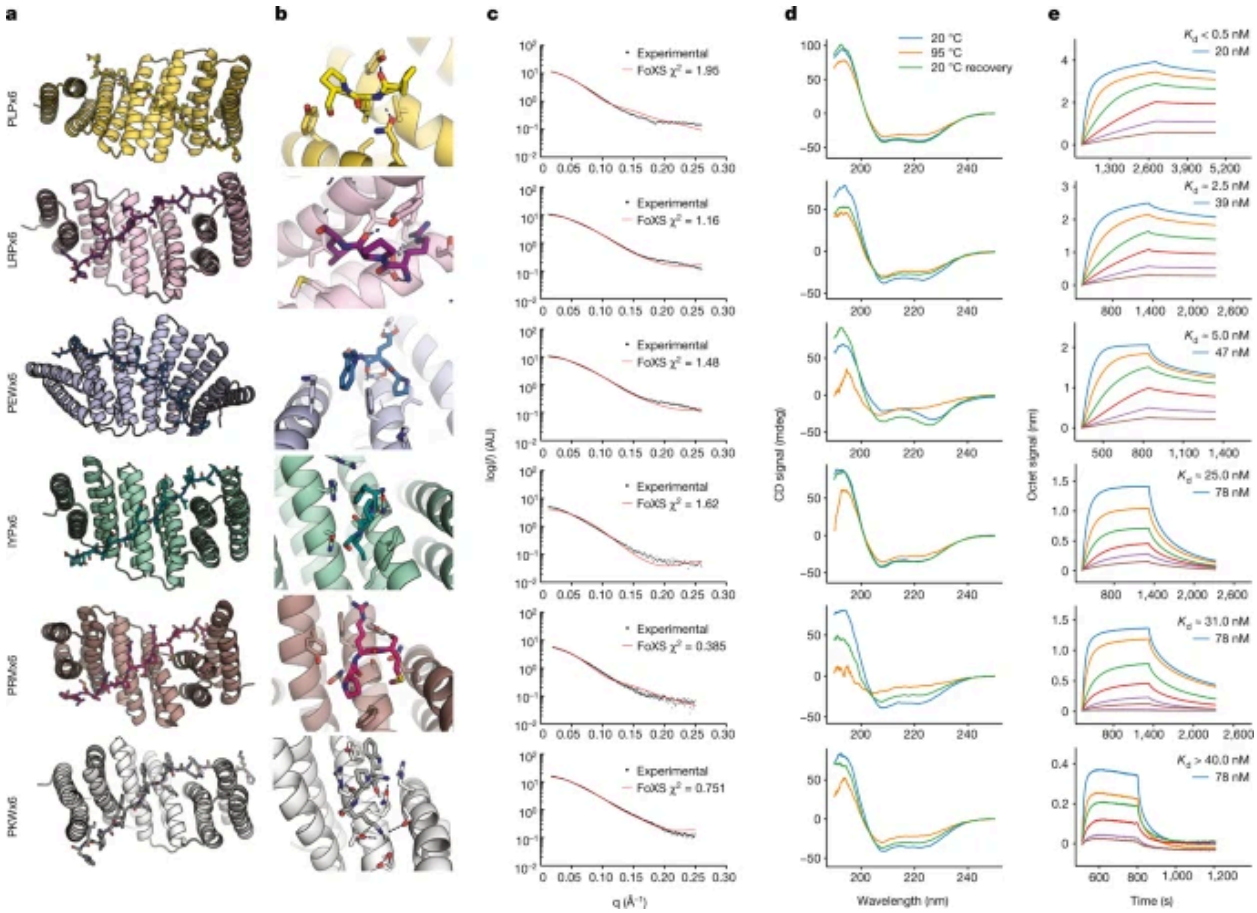
We obtained synthetic genes encoding the designed proteins with 6-His tags for purification and terminal biotinylation tags for fluorescent labeling, expressed the proteins in E. coli, and purified them by Ni-NTA chromatography. 30 of 49 were monomeric and soluble. To assess binding, the target peptides were displayed on the yeast cell surface [23], and binding to the repeat proteins was monitored by flow cytometry. To obtain readout of the peptide binding specificity of individual designs, we in parallel used large scale array based oligonucleotide synthesis to generate yeast display libraries encoding all 2 and 3 residue repeat peptides with 8 repeat units each, and used fluorescence activated cell sorting (FACS) followed by Sanger sequencing to identify the peptides recognized by each designed protein. Many of the designs bound peptides with sequences similar to those targeted but the affinity and specificity were both relatively low, with most of the successes for 3 residue repeat units (Extended Data Table. 1A).

Based on these results, we sought to increase the peptide sequence specificity of the computational design protocol, focusing on design of binders for peptides with 3 residue repeat units. First, we required that each non-proline residue in the peptide make specific contacts with the protein, and that the pockets and grooves engaging sidechains emanating on the two sides of the peptide were quite distinct. Second, following design, we evaluated the change in binding energy (Rosetta ddG)[24] for all single residue changes to the peptide repeating unit, and selected only designs for which the design target sequence made the most favorable interactions with

the designed protein. Third, we used computational Alanine scanning to remove hydrophobic residues on the protein surface not contributing to binding specificity to decrease non-specific binding[25]. Fourth, to assess the structural specificity of the designed peptide binding interface, we carried out Monte Carlo flexible backbone docking calculations, starting from large numbers of peptide conformations with superhelical parameters in the range of those of the proteins, and selected those designs with converged peptide backbones (RMSD<2.0 between the 20 lowest ddG designs) close to the design model (RMSD<1.5) (Extended Data Fig. 1D).

We tested 54 second-round designed protein-peptide pairs using the yeast flow cytometry assay described above. 42 of the designed proteins were solubly expressed in *E. coli*, and 16 designed bound their targets with considerably higher affinity and specificity than in the first round (Extended Data Table 1B, see Supplementary Table 3-4 for amino acid sequences of validated designed binders in this work). We selected six designs with diverse superhelical parameters and shapes, and a range of target peptides for more detailed characterization (Fig. 2). As evident in the design models (Fig. 2A), there is a one to one match between the six repeat units in the protein and in the target peptide (Fig. 2B illustrates a single unit interaction). Small Angle X-ray Scattering (SAXS) profiles[26, 27] were close to those computed from the design models, suggesting that the proteins fold into the designed shapes in solution (Fig. 2C, Extended Data Table 2B). Circular dichroism (CD) studies showed that all six are largely helical and thermostable up to 95°C (Fig. 2D). Bio-Layer Interferometry (BLI) characterization of binding to biotinylated target peptides immobilized on Octet sensor chips revealed dissociation constant (Kd) values ranging from <500 pM (below the instrument level of detection) to ~40nM; five out of six had dissociation half-life >= 500s, and for three of the six there was little dissociation after 2000s (Fig. 2E; little decrease in binding was observed after storage of the proteins for 30 days at 4°C, Extended Data Fig. 2A). The binding surfaces of several related designs were subjected to Site Saturation Mutagenesis (SSM)[28] on yeast; and following

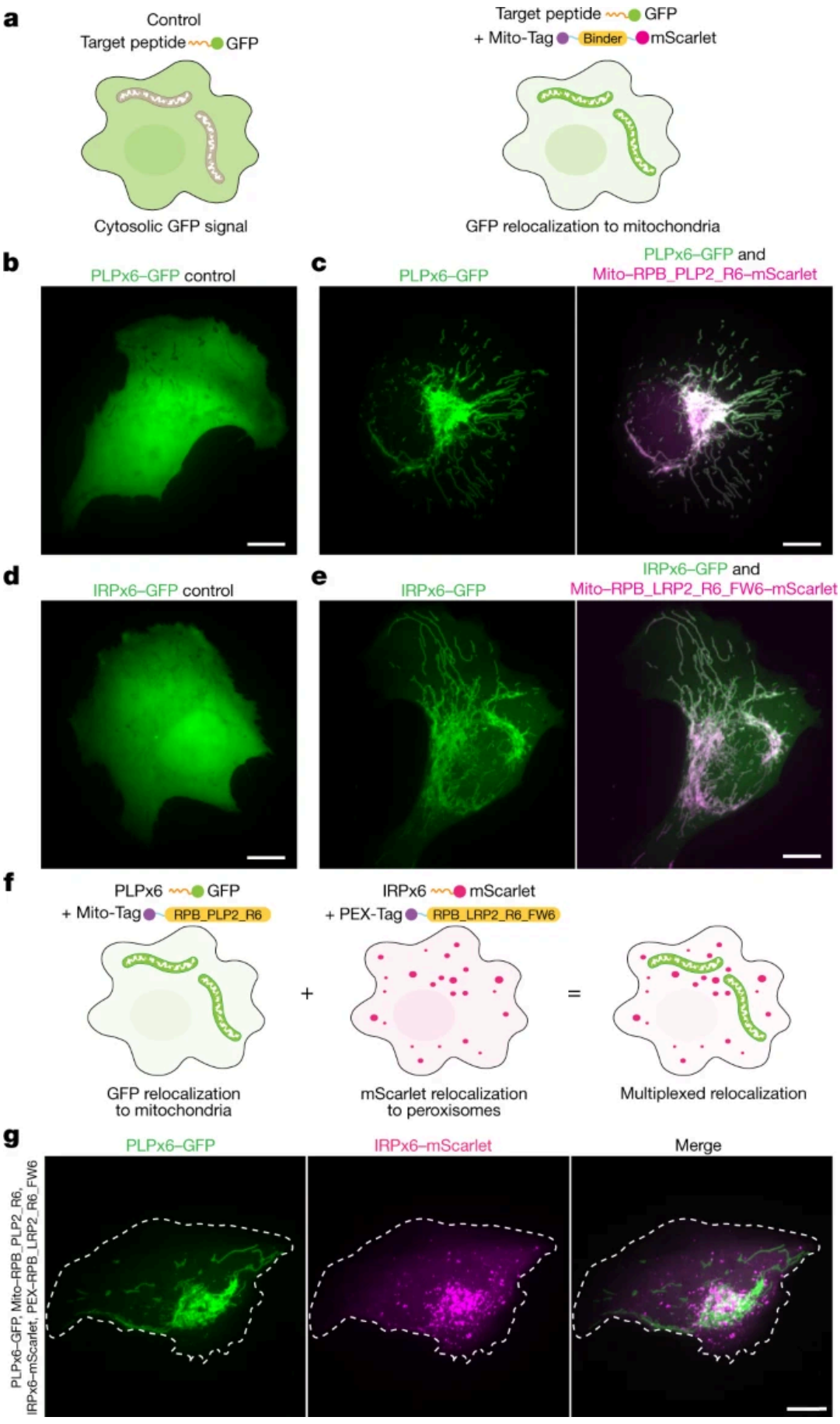
incorporation of 1-3 enriched substitutions, binding was observed by flow cytometry using only 10pM biotinylated cognate peptides (Extended Data Fig. 3).



a, Computational models of the designed six-repeat version of protein–peptide complexes. Designed proteins are shown in cartoons and peptides in sticks. **b,** Magnified views for single designed protein–peptide interaction units. Residues interacting across the interface are shown in sticks. **c,** Predicted SAXS profiles overlaid on experimental SAXS data points. The scattering vector q is on the *x* axis (from 0 to 0.25) and the intensity (*I*) is on the *y* axis on a logarithmic scale. AU, arbitrary units. **d,** Circular dichroism (CD) spectra at different temperatures (blue, 20 °C; orange, 95 °C; green, 95 °C followed by 20 °C). **e,** Bio-layer interferometry characterization of the binding of designed proteins to the corresponding peptide targets. Twofold serial dilutions were tested for each binder and the highest concentration is labelled.

The biotinylated target peptides were loaded onto streptavidin biosensors, and incubated with designed binders in solution to measure association and dissociation.

Many current cell biology approaches[29] involve tagging cellular target proteins with a protein or peptide, and then introducing into the same cell a protein which binds the tag with high affinity and specificity, but does not bind endogenous targets. A bottleneck in such studies is that binders obtained from antibody-scaffold (scFV or VHH) based library screens often do not fold properly in the reducing environment of the cytosol, resulting in loss of binding[30]. We reasoned that our binders would not have this limitation as they are designed for stability and lack disulfide bonds. As a proof of concept, we coexpressed the peptide PLPx6 fused to GFP and its cognate binder, RPB_PLP2_R6, a variant of RPB_PLP1_R6, fused to both mScarlet and a targeting sequence for the mitochondria outer membrane (Fig. 3A). While the PLPx6 peptide on its own was diffuse in the cytosol (Fig. 3B), upon coexpression with the binder, it was relocalized to mitochondria (Fig. 3C; Extended Data Fig. 2B). Thus the PLPx6/ RPB_PLP2_R6 pair retains binding activity in cells. Similar results were obtained for IRPx6-GFP and its cognate binder PXX13_FW6 (Fig. 3D, E).


a
Control
Target peptide GFP
Cytosolic GFP signal
Target peptide GFP
+ Mito-Tag Binder mScarlet
GFP relocalization to mitochondria
b
PLPx6–GFP control
c
PLPx6–GFP
PLPx6–GFP and
Mito–RPB_PLP2_R6–mScarlet
d
IRPx6–GFP control
e
IRPx6–GFP
IRPx6–GFP and
Mito–RPB_LRP2_R6_FW6–mScarlet
f
PLPx6 GFP
+ Mito-Tag RPB_PLP2_R6
GFP relocalization
to mitochondria
+
IRPx6 mScarlet
+ PEX-Tag RPB_LRP2_R6_FW6
mScarlet relocalization
to peroxisomes
=
Multiplexed relocalization
g
PLPx6–GFP
IRPx6–mScarlet
Merge
PLPx6–GFP, Mito–RPB_PLP2_R6,
IRPx6–mScarlet, PEX–RPB_LRP2_R6_FW6

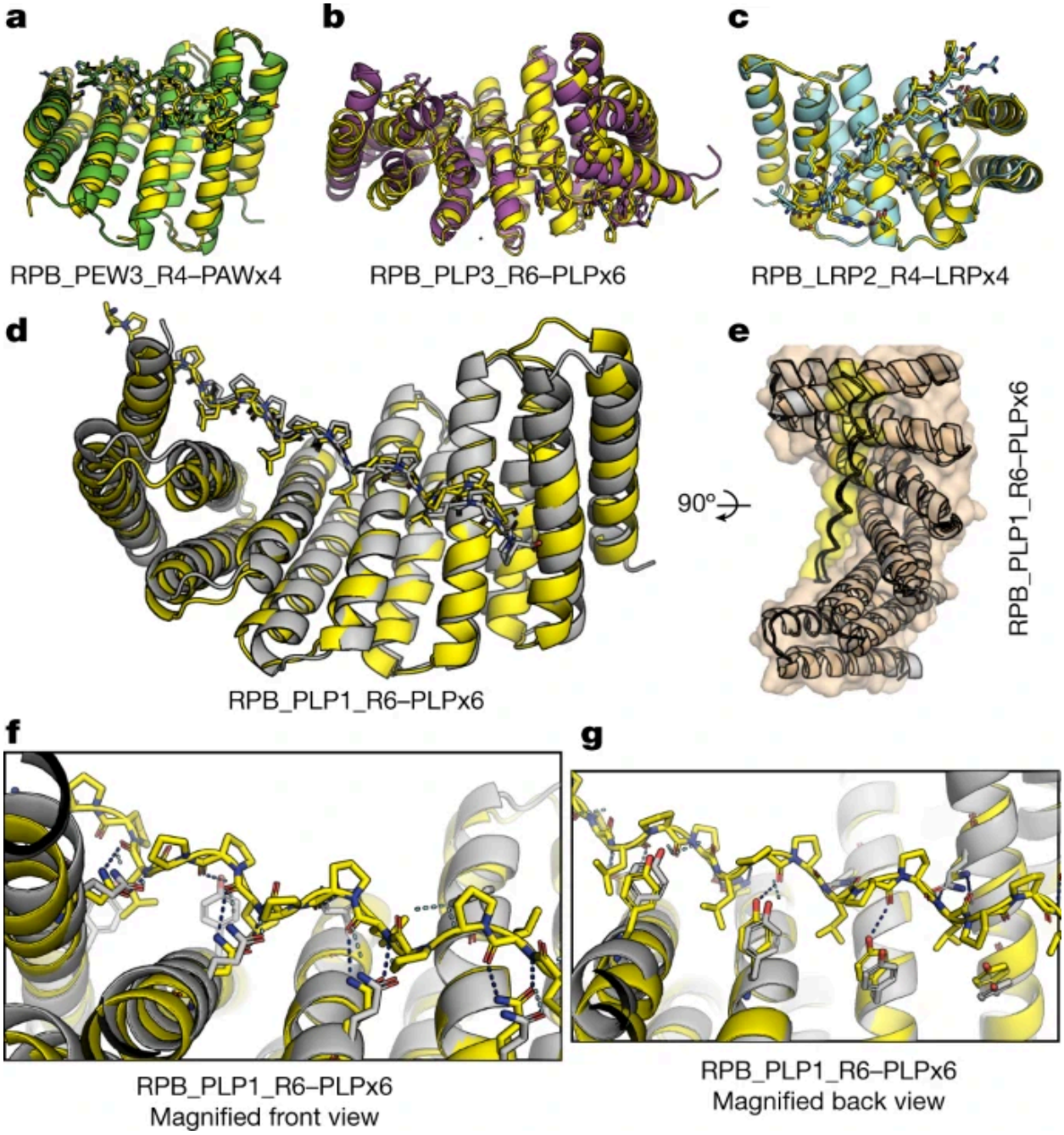
a, Experimental design. U2OS cells co-express the target peptide fused to GFP and a fusion between the specific binder fused to mScarlet and a mitochondria-targeting sequence (Mito-Tag). If binding occurs in cells, the GFP signal is relocalized to the mitochondria, whereas control cells that do not express the binder show a cytosolic GFP signal. **b–e,** In vivo binding. Live, spreading U2OS cells expressing PLPx6–GFP alone (b), IRPx6–GFP alone (d), PLPx6–GFP and Mito–RPB_PLP2_R6–mScarlet (c) or IRPx6–GFP and Mito–RPB_LRP2_R6_FW6–mScarlet (e) were imaged by spinning disk confocal microscopy (SDCM). Note that the GFP signal is cytosolic in the control but relocalized to the mitochondria after co-expression with the respective binder. **f,g,** In vivo multiplexing. f, Experimental design. U2OS cells co-express two target peptides, one fused to GFP and the other to mScarlet, and their corresponding specific binder fused to mitochondria- or peroxisome-targeting sequences. If orthogonal binding occurs, GFP and mScarlet signals should not overlap. g, Live, spreading U2OS cells co-expressing PLPx6–GFP, IRPx6–mScarlet, Mito–RPB_PLP2_R6 and PEX–RPB_LRP2_R6_FW6 imaged by SDCM. Note the absence of overlap between channels. Images correspond to maximum intensity z-projections ($\Delta z = 6$ µm). Dashed line indicates the cell outline. Scale bars, 10 µm.

If individual repeat units on the designed protein engage individual repeat units on the target peptide, binding affinity should increase with increasing the number of repeats. We investigated this with four of our designed systems, in two cases varying the number of protein repeats while keeping the peptide constant, and in the other two, varying the number of peptide repeats while keeping the protein constant. Six-repeat versions of RPB_LRP2_R6 and RPB_PEW2_R6 had higher affinity for eight-repeat LRP and PEW peptides than four-repeat versions without any decrease in specificity (Extended Data Fig. 4A). Similarly, six-repeat IYP and PLP peptides had higher affinity for six-repeat versions of the cognate designed repeat proteins (RPB_IYP1_R6, RPB_PLP1_R6) than four-repeat versions (Extended Data Fig. 4B). These results are consistent with one to one modular interaction between repeat units on the protein and peptide, and suggest a route to very high binding affinity by simply increasing the number of interacting repeat units. The ability to vary the affinity simply by varying the number of repeats could be useful in many contexts where competitive binding would be advantageous; for example for protein purification by affinity purification, a peptide with a larger number of repeats than that fused to the protein being expressed could be used for elution.

4.4 - High-resolution structural validation

To assess the structural accuracy of our design method, we used X-ray crystallography. We obtained high-resolution co-crystal structures of three first-round designs (RPB_PEW3_R4 - PAWx4, RPB_LRP2_R4 - LRPx4, RPB_PLP3_R6 - PLPx6) and one second-round design (RPB_PLP1_R6 - PLPx6) (Fig. 4); and a crystal structure of the unbound first-round design RPB_LRP2_R4 (Extended Data Fig. 5A; interface sidechain RMSDs for all crystal structures are in Extended Data Table 2A)). In the crystal structure of RPB_PLP3_R6 - PLPx6 design, the PLP units fit exactly into the designed curved groove formed by repeating tyrosine, alanine, and

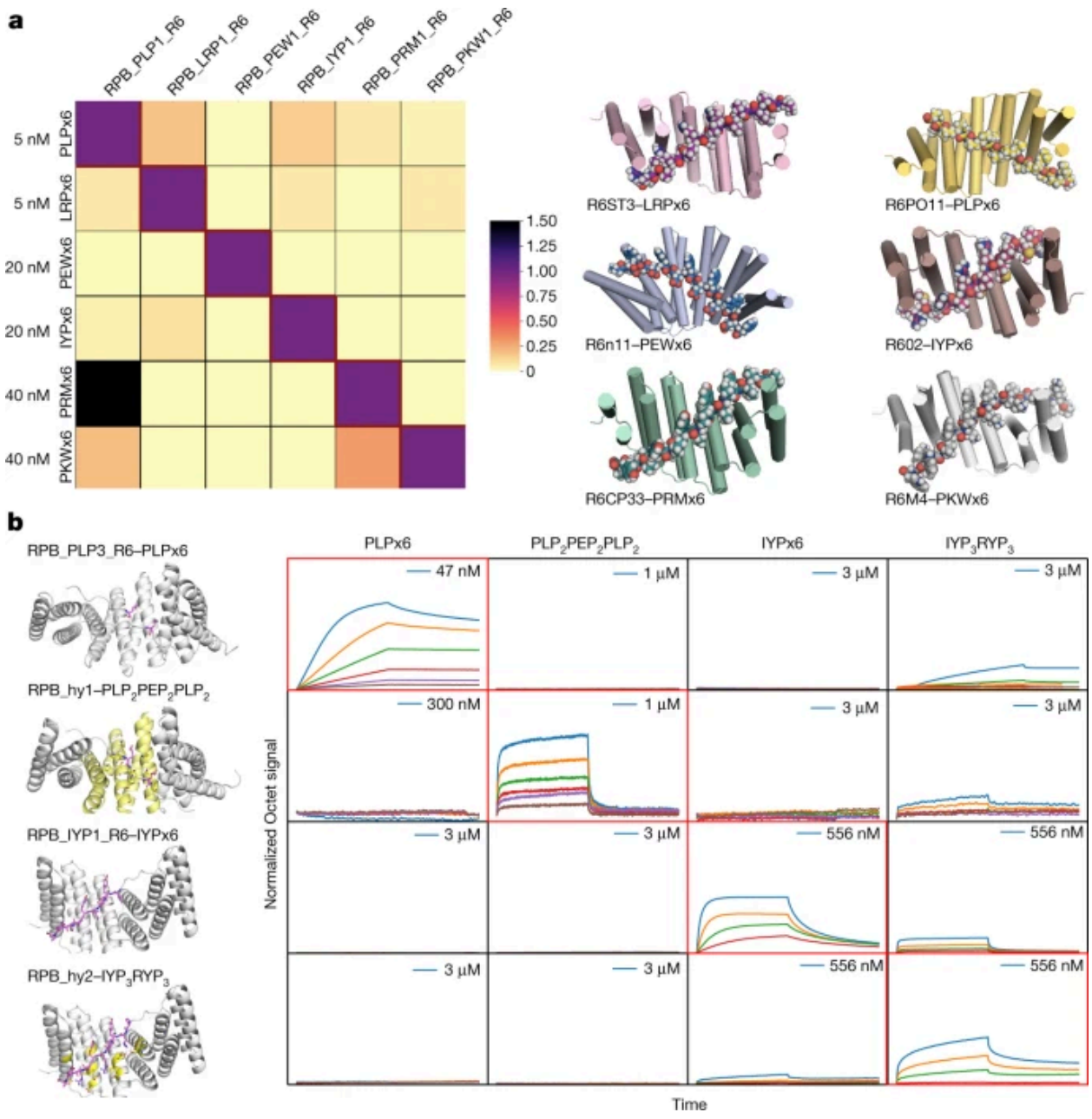
tryptophan residues matching the design model with near atomic accuracy (Cα RMSD for protein, protein-peptide interface, and full complex of 1.70, 2.00, and 1.64 Å; Fig. 4B, Extended Data Fig. 5B). In the co-crystal structure of RPB_PEW3_R4 - PAWx4, as in the design model, the PAW units bind to a relatively flat groove formed by repeating histidine residues and glutamine residues as designed (Fig. 4A, Extended Data Fig. 5A, 2.08 Å RMSD between design and crystal structure over the protein, median RMSD 2.12Å over the peptide and interface between crystal and docked peptide ensemble, Extended Data Fig. 5C). For RPB_LRP2_R4 - LRPx4, flexible backbone docking converged with the LRP units fitting in between repeating glutamine residues and phenylalanine residues as designed, and the peptide arginine sidechain sampling two distinct states associated with parallel and antiparallel protein binding modes (Extended Data Fig. 4C). The lowest energy docked structure was close to the crystal structure with Cα RMSDs of 1.15 Å, 0.98 Å and 1.16 Å for the protein alone, the peptide plus interface, and the entire complex (Fig. 3C, Extended Data Fig. 4C). SSM interface footprinting results were consistent with the design model and crystal structure (Extended Data Fig. 6), and a FtoW substitution that increases interactions across the interface substantially increases affinity (Extended Data Fig. 4D).



a–c, Superposition of computational design models (coloured) on experimentally determined crystal structures (yellow). a, RPB_PEW3_R4–PAWx4. b, RPB_PLP3_R6–PLPx6. c, RPB_LRP2_R4–LRPx4. d–g, RPB_PLP1_R6–PLPx6, **d,** Overview of the superimposition of the computational design model and the crystal structure. **e,** A 90° rotation of d. The complex is shown in surface mode (protein in orange and peptide in yellow) to highlight the shape

complementarity. **f,** Zoom in on the internal three units from d (front view). Glutamine residues from the protein in both the design and the crystal structure are shown as sticks to highlight the accuracy of the designed side-chain-to-backbone bidentate ladder. **g.** View from the side opposite to f. Tyrosine residues from the protein in both the design and the crystal structure are shown as sticks to highlight the accuracy of the designed polar interactions.

The 2.15 Å crystal structure of the 2nd round design RPB_PLP1_R6 - PLPx6 highlights key features of the computational design protocol. The PLPx6 peptide binds to the slightly curved groove primarily through polar interactions from tyrosine, hydrophobic interactions from valine, and sidechain-backbone bidentate hydrogen bonds from Glutamine exactly as designed (Fig. 4D-4G; RMSD of 1.11 Å for the protein peptide interface and 1.91 Å for the complex). All interacting side-chains from both the protein side and the peptide side in the computational design model are nearly perfectly recapitulated in the crystal structure. This design has near picomolar binding affinity (Fig. 2D) and high specificity for the PLP target sequence (Fig. 5A).



a, Left, to assess the cross-reactivity of each designed peptide binder in Fig. with each target peptide, biotinylated target peptides were loaded onto bio-layer interferometry streptavidin sensors and allowed to equilibrate, and the baseline signal was set to zero. The bio-layer interferometry tips were then placed into a solution containing proteins at the indicated concentrations for 500 s and washed with buffer, and dissociation was monitored for another 500 s. The heat map shows the maximum signal for each binder–target pair (cognate and

non-cognate) normalized by the maximum signal of the cognate designed binder–target pair. Right, surface shape complementarity of the cognate complexes. The peptides are in sphere representation. **b,** Modular pocket sequence redesign generates binders for peptide sequences that are not strictly repeating. Left, ribbon diagrams of base designs (rows 1 and 3) and versions with a matching subset of the protein and peptide modules redesigned. The ribbon diagrams show the cognate designed and redesigned assemblies; for example, the first row shows a six-repeat PLP binding design in complex with PLPx6, and the second row the same backbone with repeat units 3 and 4 redesigned to bind PEP instead of PLP, in complex with a PLP2PEP2PLP2 peptide. The redesigned peptide and protein residues are shown in purple sticks and yellow, respectively. Right, orthogonality matrix. Biotinylated target peptides were loaded onto biosensors, and incubated with designed binders in solution at the indicated concentrations. Red rectangle boxes indicate cognate complexes. Octet signal was normalized by the maximum signal of the cognate designed binder–target pair.

We next investigated the specificity of the six designs (Fig. 5A). The PLPx6, LRPx6, PEWx6, IYPx6, PKWx6 binders showed almost complete orthogonality in the 5~40 nM concentration range, with each design binding its cognate designed repeat peptide much more strongly than the other repeat peptides. For example, PLPx6 binds RPB_PLP1_R6 strongly at 5 nM, but shows no binding signal to RPB_IYP1_R6 at 40 nM, while PEWx6 binds RPB_PEW1_R6 but not RPB_PKW1_R6 at 20 nM. Some crosstalk was observed between the PRMx6 and LRPx6 binders perhaps involving the arginine residue which makes cation-pi interactions in both designs. We observe similar interaction orthogonality in cells: the IRPx6 and PLPx6 binders specifically direct localization of their cognate peptides to different compartments when coexpressed in the same cells (Fig. 3E, F).

As described thus far, our approach enables specific binding of peptides with perfectly repeating structures. To go beyond this limitation and enable targeting of a much wider range of non-repeating peptides, we investigated the redesign of a subset of the peptide repeat unit binding pockets to change their specificity. We broke the symmetry in the designed repetitive binding interface by redesigning both protein and peptide in one or more repeats of six-repeat complexes; the rest of the interface was kept untouched to maintain binding affinity. Following redesign, the peptide backbone conformation was optimized by Monte Carlo resampling and rigid body optimization (see Methods). Designs were selected for experimental characterization as described above, favoring those for which the new design had lower binding energy for the new peptide than the original peptide.

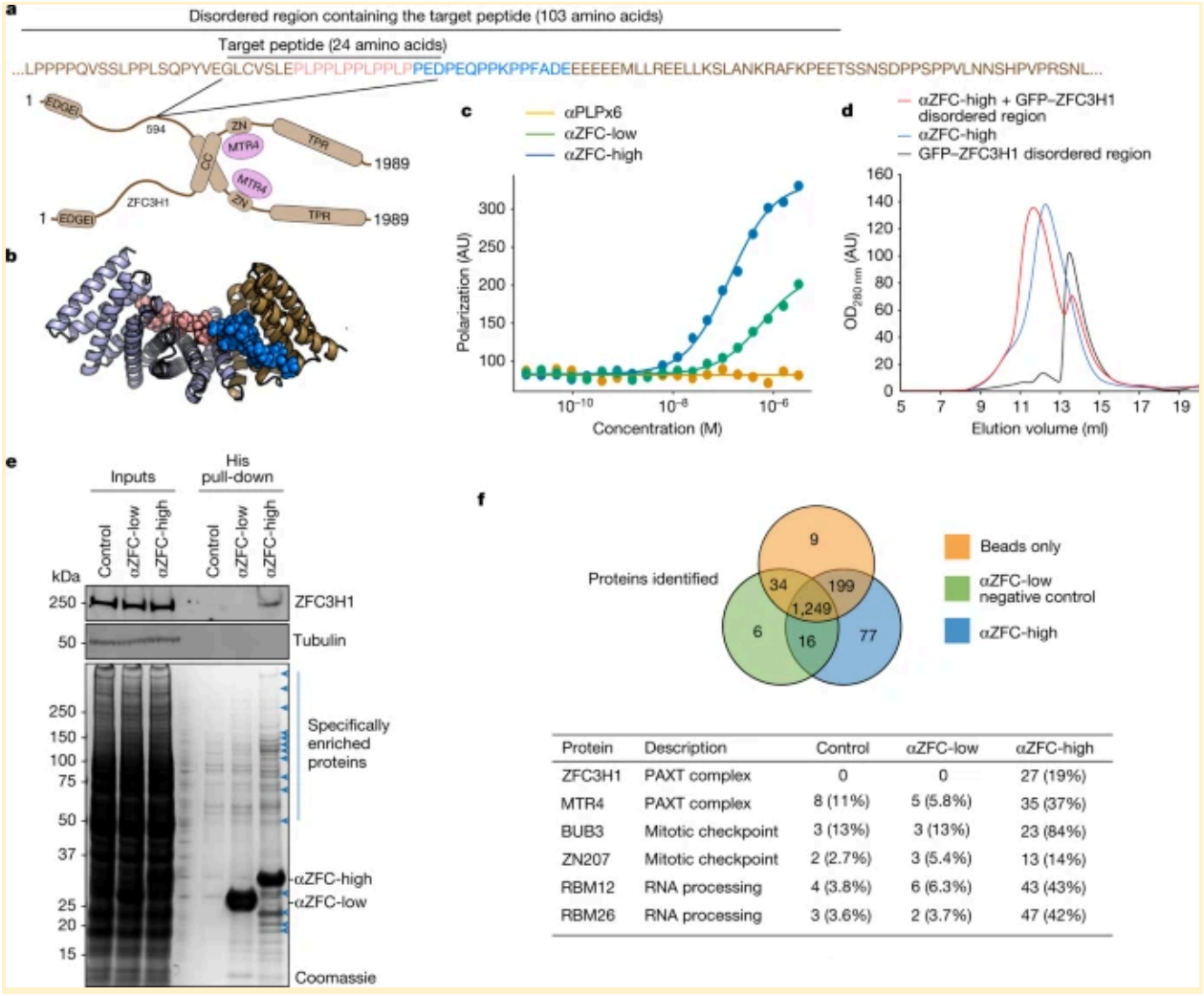
We redesigned the PLPx6 binder RPB_PLP3_R6 to bind two PEP units in the third and fourth positions (target binding sequence PLPPLPPEPPEPPLPPLP, or more concisely, $PLP_2PEP_2PLP_2$). The redesigned protein, called RPB_hyb1_R6, bound the redesigned peptide considerably more tightly in octet experiments, while the original design favored the original

perfectly repeating sequence, resulting in nearly complete orthogonality (Fig. 5B). We next designed another hybrid starting from the RPB_IYP1_R6 - IYPx6 complex, changing 3 of the IYP units to RYP to generate IYP_3RYP_3, and redesigning the corresponding binding pockets. The new design, RPB_hyb2_R6, selectively bound the intended cognate target as well (Fig. 5C). We measured binding of all four proteins against all four peptides, and observed quite high specificity of the designed repeat proteins for their intended peptide targets (Fig. 5B-5C).

4.5 - Generalization to native disordered regions

The ability to design hybrid binders against non-repetitive sequences opens the door to the *de novo* design of binders against endogenous proteins. Intrinsically disordered regions (IDR) have been very difficult to specifically target using other approaches, but are in principle good targets since binding is not complicated by folding. As a proof of concept, we focused on human ZFC3H1, a 226 kDa protein that together with MTR4 forms the heterotetrameric poly(A) tail exosome targeting (PAXT) complex, which directs a subset of long polyadenylated poly(A) RNAs for exosomal degradation (Fig. 6A and ref[31, 32]). We designed binders against ZFC3H1 residues 594-620 (PLP_4PEDPEQPPKPPF) which lie within a ~100 residue disordered region (Fig. 6A), by extending both the protein and peptide in the PLPx4 designed complex. On the peptide side, we kept the (PLP)x4 backbone fixed, and used Monte Carlo sampling with Ramachandran map biases to model the remaining sequence (PEDPEQPPKPPF); on the protein side, we extended the PLPx4 design with four additional repeats, designed binding interactions with each peptide conformer, and selected eight designs for experimental characterization. These eight designs were expressed, and seven found to bind the extended target peptide by biolayer interferometry (Extended Data Fig. 7A); the two highest affinity designs were found by fluorescence polarization to have Kd's of <200 nM and ~1.2 μM

(αZFC-high and αZFC-low; Fig. 6B,C), somewhat weaker than the synthetic constructs described above. Nevertheless, αZFC-high co-eluted with a 103 amino acid segment of the disordered region of ZFC3H1 containing the targeting sequence by Size-Exclusion Chromatography (SEC) (Fig. 6D), demonstrating that the binder can recognize the target peptide in a larger protein context. αZFC-high specifically pulled down the endogenous ZFC3H1 from human cell extracts when assessed by western blot with established antibodies (Fig. 6E, upper panel), while αZFC-low did not, which has similar size and surface composition and hence provides a control for non-specific association (see Extended Data Fig. 7B for replicates, Supplementary Figure 1 for gel source data and raw data for Western Blots, and Fig. 6F for independent identification of ZFC3H1 by mass spectrometry). Mass Spectrometry revealed that MTR4 was enriched in the αZFC-high pull down, demonstrating that the binder can recognize the native PAXT complex in a physiological context. We also detected in the αZFC-high pulldown, but not the αZFC-low pulldown, additional ZFC3H1 partners present in the Bioplex 3.0 interactome in multiple cell lines[33, 34], including BUB3 and ZN207, and multiple RNA binding proteins which likely associate with PAXT-RNA assemblies (Fig. 6F see source data for full proteomics dataset).



Protein	Description	Control	αZFC-low	αZFC-high
ZFC3H1	PAXT complex	0	0	27 (19%)
MTR4	PAXT complex	8 (11%)	5 (5.8%)	35 (37%)
BUB3	Mitotic checkpoint	3 (13%)	3 (13%)	23 (84%)
ZN207	Mitotic checkpoint	2 (2.7%)	3 (5.4%)	13 (14%)
RBM12	RNA processing	4 (3.8%)	6 (6.3%)	43 (43%)
RBM26	RNA processing	3 (3.6%)	2 (3.7%)	47 (42%)

a, Schematic model of the human PAXT complex composed of a heterotetramer of ZFC3H1 and MTR4. CC, coiled-coil domain; ZN, Zn-finger domain. Inset shows the sequence environment of the target sequence. **b,** Surface shape complementarity between the target peptide from ZFC3H1 (sphere) and the highest-affinity cognate binder, αZFC-high. **c,** Fluorescence polarization binding curves between the indicated ZFC3H1 binders and the target ZFC3H1 peptide (PLP)4PEDPEQPPKPP. As a negative control, we used the (PLP)x6 binder, RPB_PLP3_R6 (see Fig. 4). αZFC-high shows a higher binding affinity to the target peptide than αZFC-low, in contrast with RPB_PLP3_R6, which shows negligible binding. **d,** Superdex 200 10/300 GL SEC profiles of purified αZFC-high, a fusion between GFP and a 103-amino-acid fragment of the disordered region of ZFC3H1 containing the target sequence (see a), or a 1:1

mix of the two after two hours of incubation. OD280 nm, optical density at 280 nm. **e,** Top, HeLa cell extracts were subjected to pull-down using the indicated binders bound to Ni-NTA agarose beads, or naked beads as a control. Recovered proteins were processed for western blot against endogenous ZFC3H1 (or tubulin as a loading control). Bottom, Coomassie-stained SDS–PAGE gel of the samples analysed at the top. These panels are representative of $n = 3$ experiments. **f,** Proteomic analysis of the His-pull-down samples shown in e. Top, overlap between the proteins identified, setting a threshold of five peptides for correct identification. Bottom, examples of proteins identified (number indicates exclusive peptide count; protein coverage is indicated in parentheses). See Source Data for the full dataset. For gel source data, see Supplementary Fig. 1.

4.6 - Conclusion

Our results demonstrate that by matching superhelical parameters between repeating protein and peptide conformations, and incorporating specific hydrogen bonding and hydrophobic interactions between matched protein and peptide repeats, we can now design modular proteins that bind extended peptides with high affinity and specificity. The approach should be generalizable to a wide range of repeating peptide structures, and the ability to break symmetry by redesigning individual repeat units opens the door to more general peptide recognition. Our approach complements current efforts at achieving general peptide recognition by redesign of naturally occurring repeat proteins; an advantage of our approach is that a much broader range of protein conformations and binding site geometries can be generated by de novo protein design than by starting with a native protein backbone. Proteins binding repeating or nearly

repeating sequences could have applications as affinity reagents for diseases such as Huntington's which are associated with repeat expansions. Similarly, rigid fusion of protein modules designed to recognize different di, tri and tetra peptide sequences, using the approach described here, provides an avenue to achieving sequence specific recognition of entirely non-repeating sequences. The ability to design specific binders to proteins containing large disordered regions, demonstrated by the specific pull down of the PAXT complex (Fig. 6), should contribute to delineating the functions of this important but relatively poorly understood class of proteins and reduce reliance on animal immunization to generate antibodies, which can also suffer from reproducibility issues. Furthermore, the 100nM-range affinity we reached for this endogenous binder is compatible with other cellular applications, such as enzyme targeting for specific post-translational modifications in vivo[34], or for imaging probes, where a trade-off must always be found between high-affinity interaction for labeling specificity and low-affinity not to perturb protein function[35]. More generally, our results demonstrate the power of computational protein design for targeting peptides and intrinsically disordered regions not having rigid three dimensional structures, and as the designed proteins are expressed at quite high levels and very stable, we anticipate that these and further designs for a wider range of target sequences should find broad use in proteomics and other applications requiring specific peptide recognition.

REFERENCES

1. N. London, D. Movshovitz-Attias, O. Schueler-Furman, The Structural Basis of Peptide-Protein Binding Strategies. *Structure*. **18**, 188–199 (2010).
2. V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, R. B. Russell, Systematic Discovery of New Recognition Peptides Mediating

Protein Interaction Networks. *PLOS Biol.* **3**, e405 (2005).

3. V. Neduva, R. B. Russell, Peptides mediating interaction networks: new leads at last. *Curr. Opin. Biotechnol.* **17**, 465–471 (2006).

4. P. Ernst, A. Plückthun, Advances in the design and engineering of peptide-binding repeat proteins. *Biol. Chem.* **398**, 23–29 (2017).

5. M. A. Andrade, C. Petosa, S. I. O'Donoghue, C. W. Müller, P. Bork, Comparison of ARM and HEAT protein repeats11Edited by P. E. Wright. *J. Mol. Biol.* **309**, 1–18 (2001).

6. C. Reichen, S. Hansen, C. Forzani, A. Honegger, S. J. Fleishman, T. Zhou, F. Parmeggiani, P. Ernst, C. Madhurantakam, C. Ewald, P. R. E. Mittl, O. Zerbe, D. Baker, A. Caflisch, A. Plückthun, Computationally Designed Armadillo Repeat Proteins for Modular Peptide Recognition. *J. Mol. Biol.* **428**, 4467–4489 (2016).

7. E. Conti, J. Kuriyan, Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin α. *Structure*. **8**, 329–338 (2000).

8. E. Conti, M. Uy, L. Leighton, G. Blobel, J. Kuriyan, Crystallographic Analysis of the Recognition of a Nuclear Localization Signal by the Nuclear Import Factor Karyopherin α. *Cell*. **94**, 193–204 (1998).

9. N. Zeytuni, R. Zarivach, Structural and Functional Discussion of the Tetra-Trico-Peptide Repeat, a Protein Interaction Module. *Structure*. **20**, 397–405 (2012).

10. L. D. D'Andrea, L. Regan, TPR proteins: the versatile helix. *Trends Biochem. Sci.* **28**, 655–662 (2003).

11. P. Ernst, F. Zosel, C. Reichen, D. Nettels, B. Schuler, A. Plückthun, Structure-Guided Design of a Peptide Lock for Modular Peptide Binders. *ACS Chem. Biol.* **15**, 457–468 (2020).

12. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J. J.

Gray, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

13. S. Hansen, D. Tremmel, C. Madhurantakam, C. Reichen, P. R. E. Mittl, A. Plückthun, Structure and Energetic Contributions of a Designed Modular Peptide-Binding Protein with Picomolar Affinity. *J. Am. Chem. Soc.* **138**, 3526–3532 (2016).

14. C. Reichen, S. Hansen, A. Plückthun, Modular peptide binding: From a comparison of natural binders to designed armadillo repeat proteins. *J. Struct. Biol.* **185**, 147–162 (2014).

15. J. A. Cross, M. S. Chegkazi, R. A. Steiner, D. N. Woolfson, M. P. Dodding, Fragment-linking peptide design yields a high-affinity ligand for microtubule-based transport. *Cell Chem. Biol.* **28**, 1347-1355.e5 (2021).

16. P. J. Fleming, G. D. Rose, Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci.* **14**, 1911–1917 (2005).

17. T. J. Brunette, F. Parmeggiani, P.-S. Huang, G. Bhabha, D. C. Ekiert, S. E. Tsutakawa, G. L. Hura, J. A. Tainer, D. Baker, Exploring the repeat protein universe through computational protein design. *Nature*. **528**, 580–584 (2015).

18. L. Shimoni, J. P. Glusker, Hydrogen bonding motifs of protein side chains: descriptions of binding of arginine and amide groups. *Protein Sci. Publ. Protein Soc.* **4**, 65–74 (1995).

19. J. A. Fallas, G. Ueda, W. Sheffler, V. Nguyen, D. E. McNamara, B. Sankaran, J. H. Pereira, F. Parmeggiani, T. J. Brunette, D. Cascio, T. R. Yeates, P. Zwart, D. Baker, Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem.* **9**, 353–360 (2017).

20. J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B.

Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse, L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P.-S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J. Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koepnick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. Leaver-Fay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L. Malmström, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. Ó'Conchúir, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. R. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O. Schueler-Furman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D.-A. Silva, S. Smith, Y. Song, A. Stein, M. Szegedy, F. D. Teets, S. B. Thyme, R. Y.-R. Wang, A. Watkins, L. Zimmerman, R. Bonneau, Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods*. **17**, 665–680 (2020).

21. D. Kuroda, J. J. Gray, Shape complementarity and hydrogen bond preferences in protein–protein interfaces: implications for antibody modeling and protein–protein docking. *Bioinformatics*. **32**, 2451–2456 (2016).

22. B. Coventry, D. Baker, Protein sequence optimization with a pairwise decomposable penalty for buried unsatisfied hydrogen bonds. *PLOS Comput. Biol.* **17**, e1008061 (2021).

23. E. T. Boder, K. D. Wittrup, Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).

24. T. Kortemme, D. Baker, A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl. Acad. Sci.* **99**, 14116–14121 (2002).

25. T. Kortemme, D. E. Kim, D. Baker, Computational Alanine Scanning of Protein-Protein Interfaces. *Sci. STKE*. **2004**, pl2–pl2 (2004).

26. G. L. Hura, H. Budworth, K. N. Dyer, R. P. Rambo, M. Hammel, C. T. McMurray, J. A. Tainer, Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. *Nat. Methods*. **10**, 453–454 (2013).

27. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS) | Nature Methods, (available at https://www.nature.com/articles/nmeth.1353).

28. R. M. P. Siloto, R. J. Weselake, Site saturation mutagenesis: Methods and applications in protein engineering. *Biocatal. Agric. Biotechnol.* **1**, 181–189 (2012).

29. J. Helma, M. C. Cardoso, S. Muyldermans, H. Leonhardt, Nanobodies and recombinant binders in cell biology. *J. Cell Biol.* **209**, 633–644 (2015).

30. S. Moutel, N. Bery, V. Bernard, L. Keller, E. Lemesre, A. de Marco, L. Ligat, J.-C. Rain, G. Favre, A. Olichon, F. Perez, NaLi-H1: A universal synthetic library of humanized nanobodies providing highly functional antibodies and intrabodies. *eLife*. **5**, e16228 (2016).

31. A.-E. Foucher, L. Touat-Todeschini, A. B. Juarez-Martinez, A. Rakitch, H. Laroussi, C. Karczewski, S. Acajjaoui, M. Soler-López, S. Cusack, C. D. Mackereth, A. Verdel, J. Kadlec, Structural analysis of Red1 as a conserved scaffold of the RNA-targeting MTREC/PAXT complex. *Nat. Commun.* **13**, 4969 (2022).

32. N. Meola, M. Domanski, E. Karadoulama, Y. Chen, C. Gentil, D. Pultz, K. Vitting-Seerup, S. Lykke-Andersen, J. S. Andersen, A. Sandelin, T. H. Jensen, Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. *Mol. Cell*. **64**, 520–533 (2016).

33. D. K. Schweppe, E. L. Huttlin, J. W. Harper, S. P. Gygi, BioPlex Display: An Interactive Suite for Large-Scale AP–MS Protein–Protein Interaction Data. *J. Proteome Res.* **17**, 722–726 (2018).

34. E. L. Huttlin, R. J. Bruckner, J. Navarrete-Perea, J. R. Cannon, K. Baltier, F. Gebreab, M. P. Gygi, A. Thornock, G. Zarraga, S. Tam, J. Szpyt, B. M. Gassaway, A. Panov,

H. Parzen, S. Fu, A. Golbazi, E. Maenpaa, K. Stricker, S. Guha Thakurta, T. Zhang, R. Rad, J. Pan, D. P. Nusinow, J. A. Paulo, D. K. Schweppe, L. P. Vaites, J. W. Harper, S. P. Gygi, Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*. **184**, 3022-3040.e28 (2021).

32. N. Meola, M. Domanski, E. Karadoulama, Y. Chen, C. Gentil, D. Pultz, K. Vitting-Seerup, S. Lykke-Andersen, J.S. Andersen, A. Sandelin, T.H. Jensen, Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. *Mol. Cell*. **64**, 520-533 (2016).

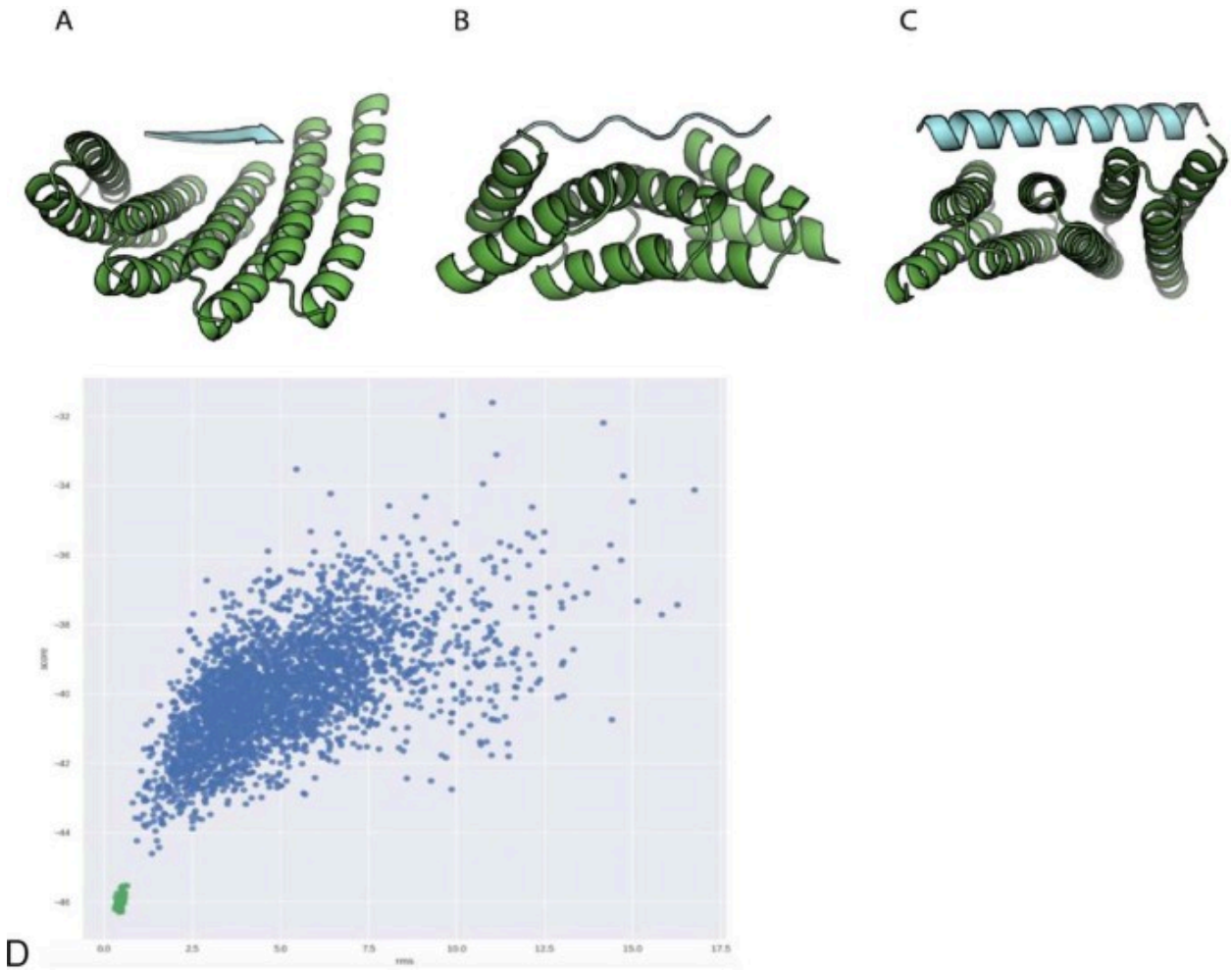
33. Rhys, G.G., Cross, J.A., Dawson, W.M. *et al.* De novo designed peptides for cellular delivery and subcellular localisation. *Nat Chem Biol* **18**, 999–1004 (2022).

34. Ramirez DH, Aonbangkhen C, Wu HY, Naftaly JA, Tang S, O'Meara TR, Woo CM. Engineering a Proximity-Directed O-GlcNAc Transferase for Selective Protein O-GlcNAcylation in Cells. *ACS Chem Biol.* **15(4)**, 1059-1066 (2020).

35. Kumari A, Kesarwani S, Javoor MG, Vinothkumar KR, Sirajuddin M. Structural insights into actin filament recognition by commonly used cellular actin markers. *EMBO J.* **39(14)**, e104006 (2020).

36. D. R. Hicks, M. A. Kennedy, K. A. Thompson, M. DeWitt, B. Coventry, A. Kang, A. K. Bera, T. J. Brunette, B. Sankaran, B. Stoddard, D. Baker, De novo design of protein homodimers containing tunable symmetric protein pockets. *Proc. Natl. Acad. Sci.* **119**, e2113400119 (2022).

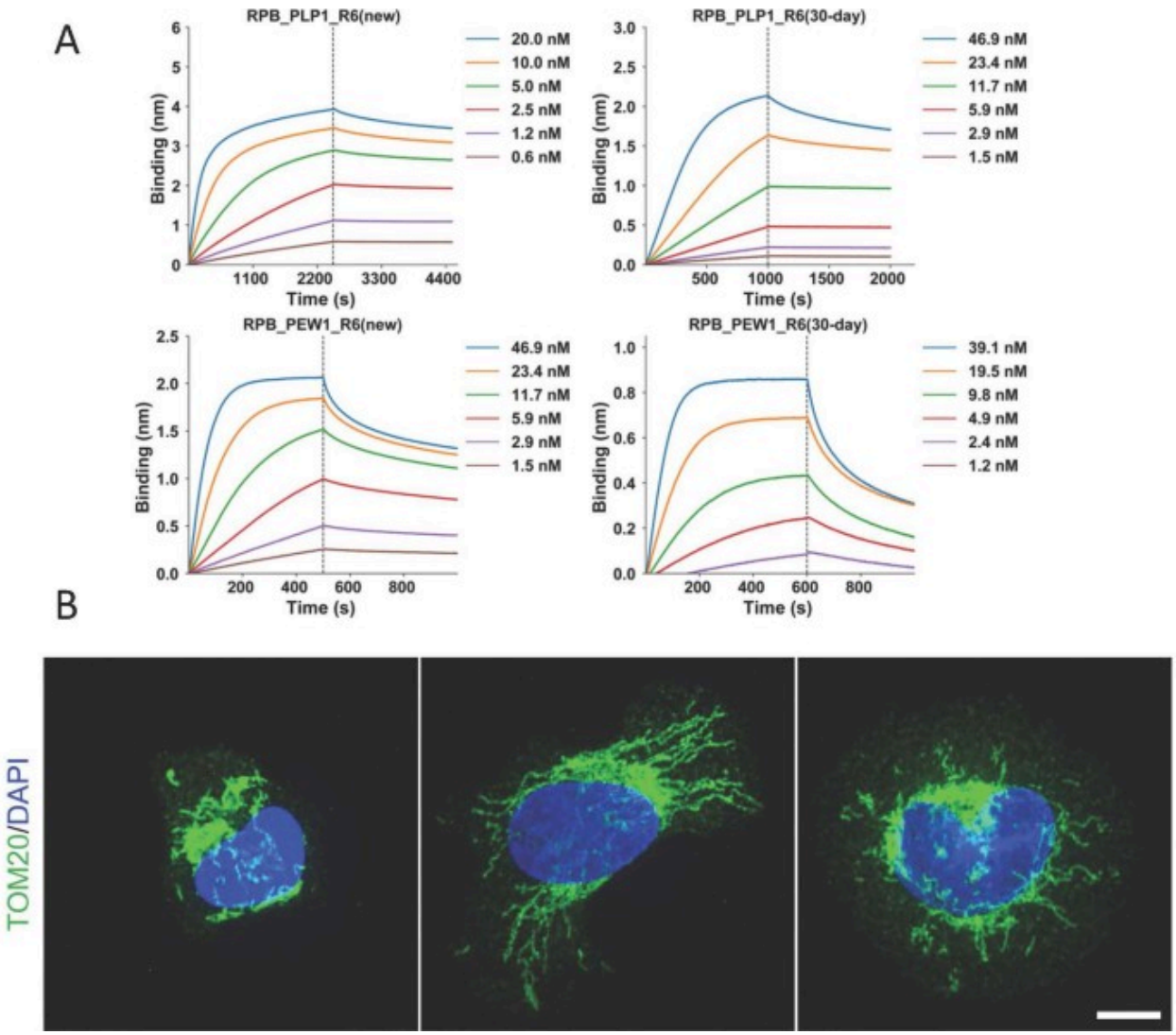
Supplementary data



Extended Data Fig. 1 Examples of computationally designed model geometry and convergence of backbone docking.

a–c, Examples of repeat proteins computationally designed to bind to extended beta strand (a), polypeptide II (b) and helical peptide backbones (c). d, Monte Carlo flexible backbone docking calculations after design to assess the structural specificity of the designed peptide-binding interface. It started from large numbers of peptide conformations randomly generated with superhelical parameters in the range of those of the proteins (usually 10,000–50,000 trajectories), and selected those designs with converged peptide backbones (RMSD < 2.0 among the top 20 designs with lowest DDG) close to the design model (RMSD < 1.5). Green

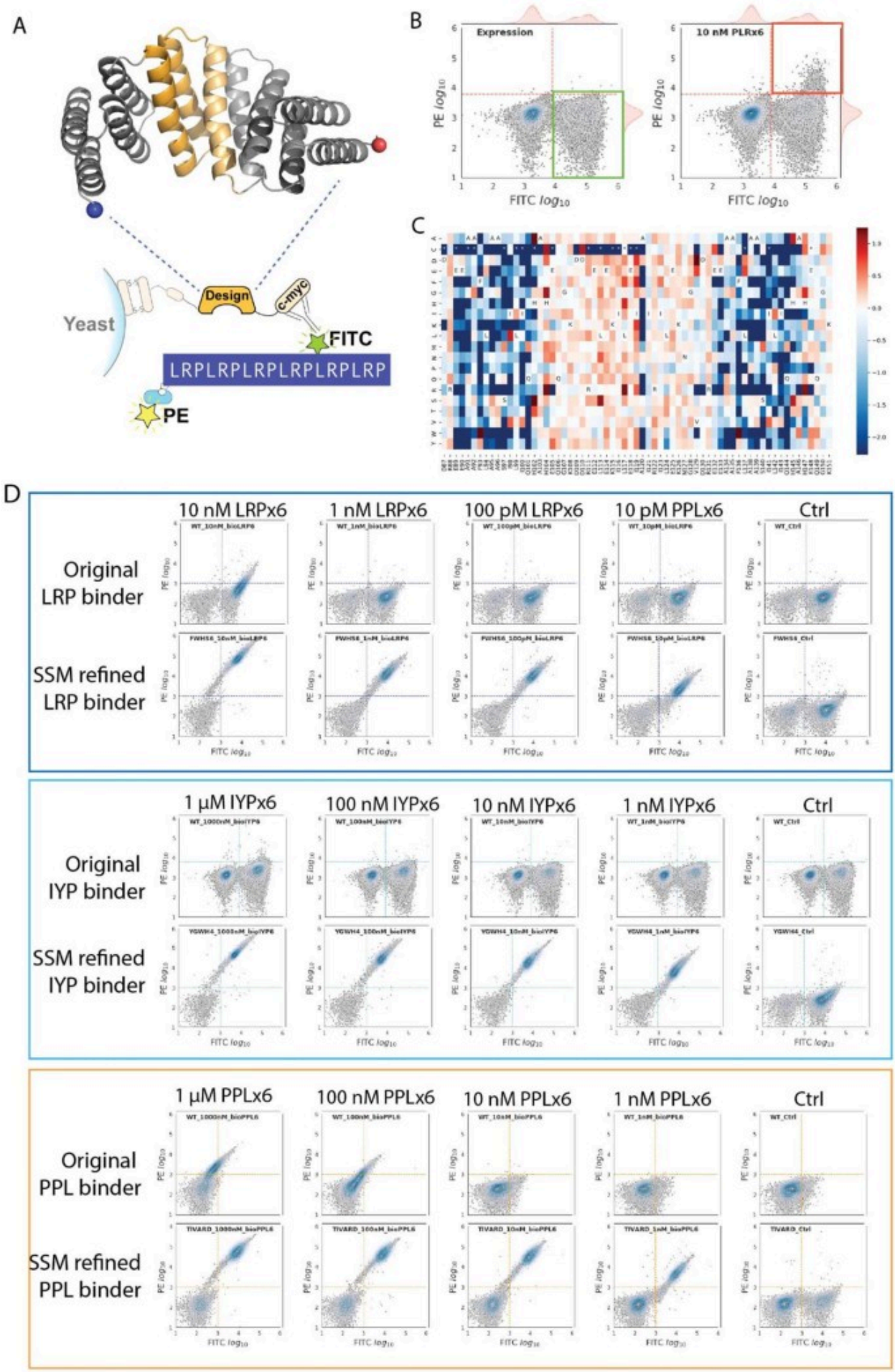
dots shown in the above example plot represent the converged designs picked by this threshold.



Extended Data Fig. 2 Comparison of binding affinities from freshly made and 30-day-old samples, and mitochondria immunostainings in control U2OS cells.

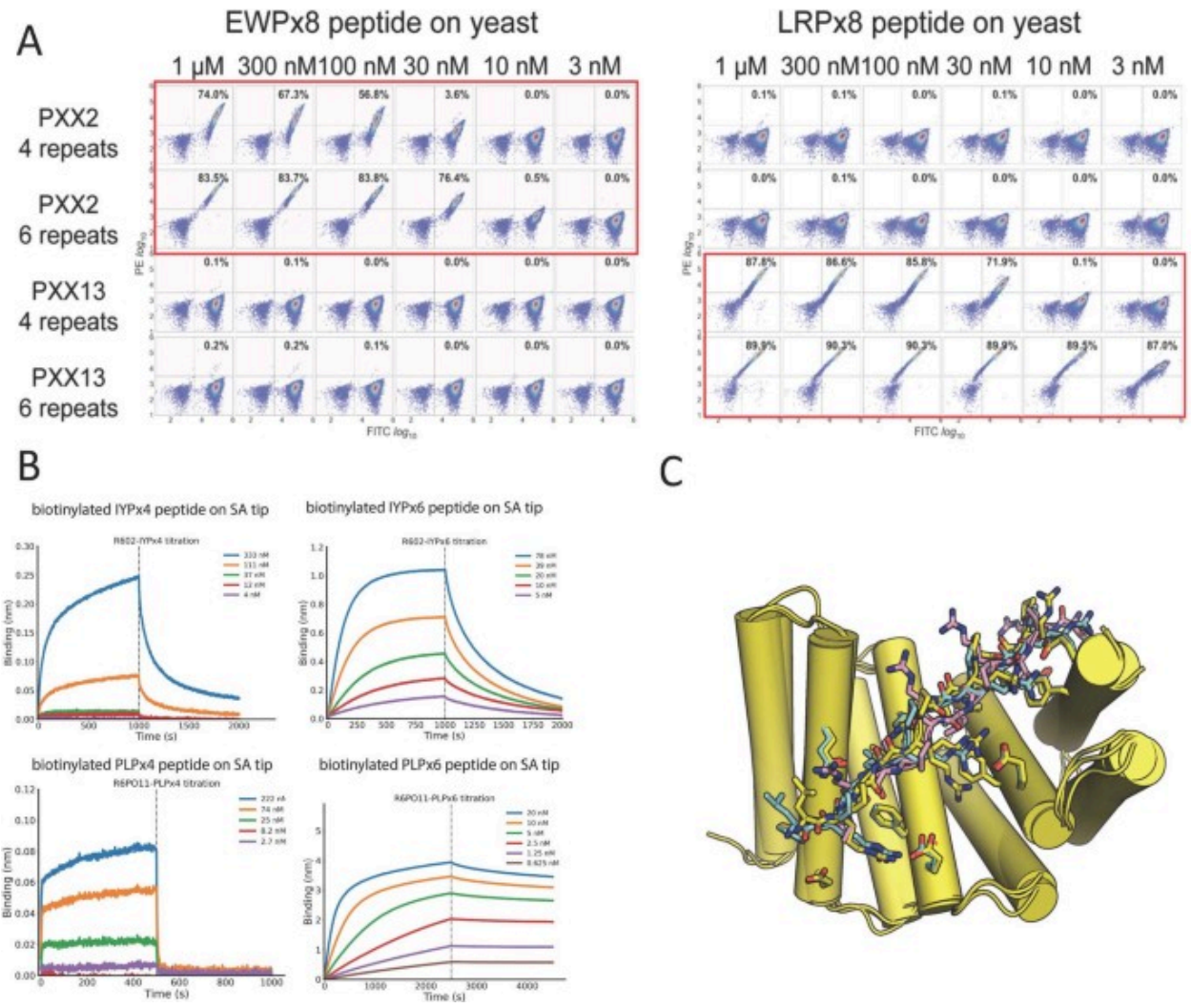
a, Little decrease in binding observed for designs RPB_PLP1_R6 and RPB_PEW1_R6 30-day-old in 4 °C. Bio-layer interferometry characterization of binding of designed proteins to the corresponding peptide targets. Twofold serial dilutions were tested for each binder, and the full tested concentration is labelled. The biotinylated target peptides were loaded onto the streptavidin (SA) biosensors, and incubated with designed binders in solution to measure association and dissociation. b, Mitochondria immunostainings in control U2OS cells. Wild-type U2OS cells were spread onto fibronectin coverslips as in Fig. 3, then fixed and processed for

immunofluorescence using TOM20 antibodies as a marker of mitochondria. Note that mitochondria appearance in these control cells is similar to that observed upon overexpression of designed binders fused to mitochondria-targeting sequences (Fig. 3). suggesting that these constructs do not affect mitochondria shape. Scale bar, 10 μm.


A
Yeast
Design
c-myc
FITC
LRPLRPLRPLRPLRPLRP
PE
B
Expression
10 nM PLRx6
PE log10
FITC log10
C
D
10 nM LRPx6
1 nM LRPx6
100 pM LRPx6
10 pM PPLx6
Ctrl
Original LRP binder
SSM refined LRP binder
1 µM IYPx6
100 nM IYPx6
10 nM IYPx6
1 nM IYPx6
Ctrl
Original IYP binder
SSM refined IYP binder
1 µM PPLx6
100 nM PPLx6
10 nM PPLx6
1 nM PPLx6
Ctrl
Original PPL binder
SSM refined PPL binder

Extended Data Fig. 3 SSMs libraries are constructed and screened for enhancing the peptide-binding abilities of designed repeat-peptide binders.

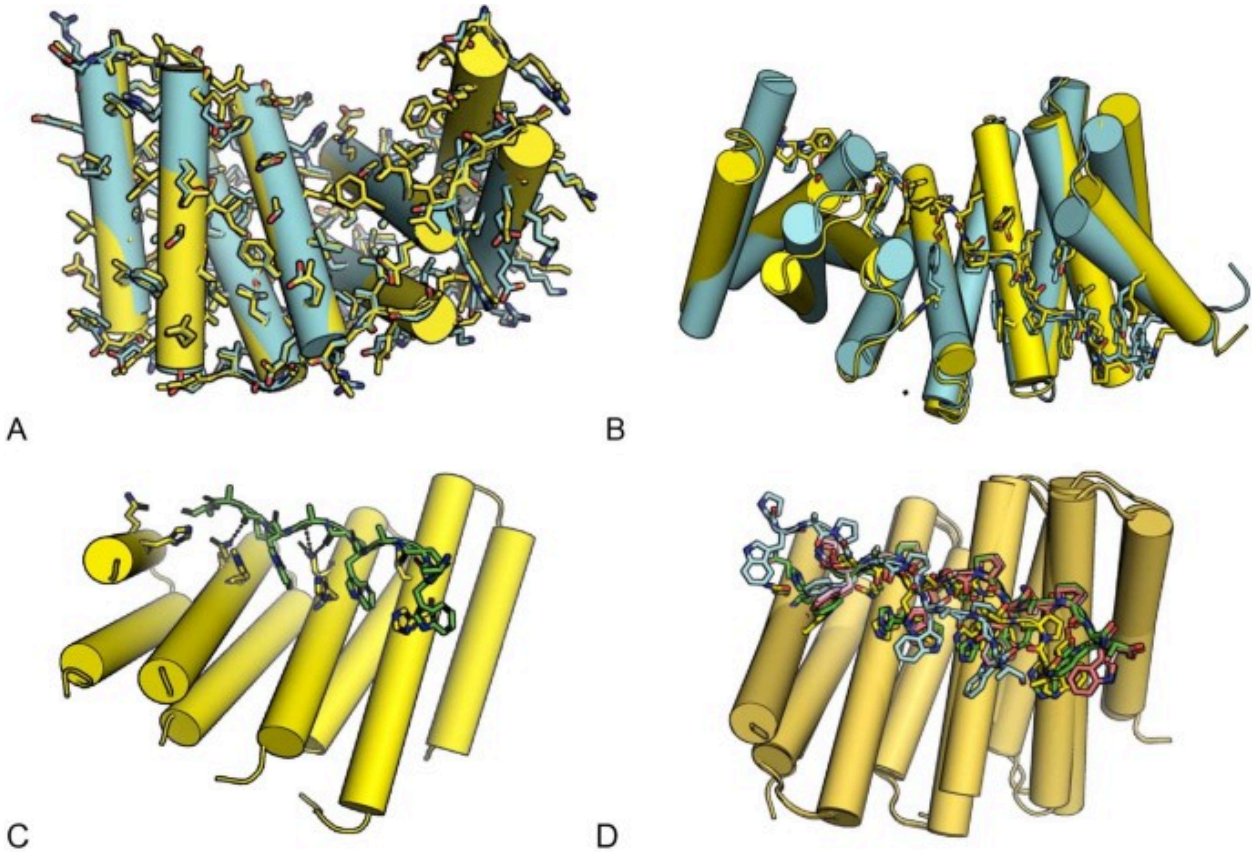
a, A schematic illustration of the mutagenesis region within the designed repeat protein, and the principles of the yeast surface display assay for peptide binding analysis. In short, the biotinylated repeat peptides (a six-repeat of LRP peptide is shown as an example) are synthesized and can be detected by SAPE, while the expression of designed protein on yeast surface are monitored by FITC-conjugated anti-Myc antibody. A double high signal of both PE and FITC, using flow cytometry, indicates the valid peptide-binding events. b, The SSM libraries are first subjected to expression sorting (left), in which there is no targeted peptide added. The yeast populations, which display well expressed SSM mutants, will show above threshold FITC signals, are collected (green box) for next-generation sequencing, and are regrown for the next rounds of sorting. In the next round sorting, the targeted peptide is incubated with the yeast library, and labelled by both FITC and SAPE (right). The FITC+PE+ population is collected for analysis (orange box). c, By using next-generation sequencing, enrichment analysis for each mutation is carried out, and a heat map for all mutations is generated. In this heat map, using a designed LRP binder SSM library as an example, the red shades indicate enrichment with incubating with the targeted peptide, and the blue shades indicate depletion. Several mutations show exceptional enhancement of the LRP repeat peptide-binding ability, such as F93W, H102S and others. d, Using the SSM library, we can markedly enhance the peptide-binding abilities of the designed peptide binder. Three example yeast display assays titrating the peptide concentrations are shown here. The top row of each example is using the originally designed peptide binder, and the bottom row is using the peptide binder containing the combinations of the best mutations discovered in the SSM library screenings. An approximately 1,000-fold increase of the peptide-binding ability can be achieved with the assistance of SSM libraries. Note, the ratio of yeast population in the upper right quadrant indicates the peptide-binding ability.



Extended Data Fig. 4 Comparison of binding affinities when changing repeat numbers from either binder or peptide side. and top five flexible backbone docks for the four-repeat LRP binder RPB_LRP2_R4–LRPx4.

a, Six-repeat versions of RPB_LRP2_R6 and RPB_PEW2_R6 had higher affinity for eight-repeat LRP and PEW peptides than four-repeat versions without any decrease in specificity in yeast surface display. Biotinylated repeat proteins (the six-repeat versions RPB_LRP2_R6 and RPB_PEW2_R6 and the four-repeat versions RPB_LRP2_R4 and RPB_PEW2_R4) were detected by SAPE, and the expression of the designed repeat peptide on yeast surface was monitored by FITC-conjugated anti-Myc antibody. Serial dilutions were tested for each binder, and the full tested concentration is labelled. b, Six-repeat IYP and PLP

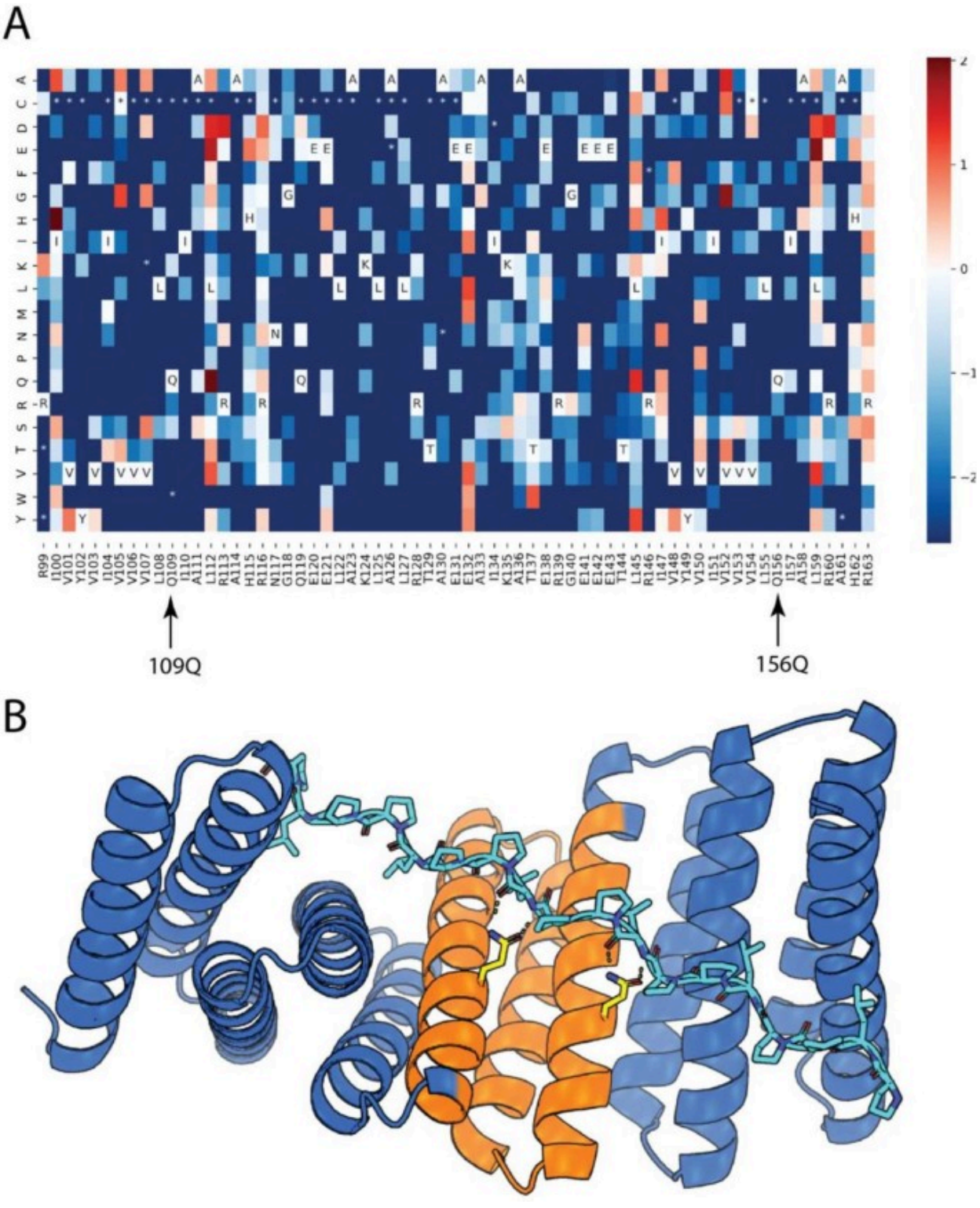
peptides had higher affinity for six-repeat versions of the cognate designed repeat proteins (RPB_IYP1_R6 and RPB_PLP1_R6) than four-repeat versions by bio-layer interferometry. The full tested concentration is labelled. The biotinylated target peptides were loaded onto the streptavidin (SA) biosensors, and incubated with designed binders in solution to measure association and dissociation. The dissociation rate was markedly increased when testing against the six-repeat peptides as compared to the four-repeat peptides, indicating a much tighter binding event. c, Top five complex PDBs for RPB_LRP2_R4–LRPx4 from the flexible docking generated ensemble. Green, pink and grey are the ones closest to the crystal structure (shown in yellow) with RMSD over the peptide and the binding residues ≈ 0.03 Å, whereas the cyan dock RMSD = 3.89 Å.



Extended Data Fig. 5 Crystal structures of the unbound RPB_LRP2_R4, bound RPB_PLP3_R6–PLPx6 and bound RPB_PEW3_R4 and its top five flexible backbone docks.

a, Crystal structure of the unbound first-round design RPB_LRP2_R4 (yellow) aligned with the design model (cyan). b, Crystal structure of the first-round complex RPB_PLP3_R6–PLPx6 (yellow) aligned with the design model (cyan). As is shown here, the peptide PLP units fit exactly into the designed curved groove formed by repeating tyrosine, alanine and tryptophan residues matching the design model with near atomic accuracy, with Cα RMSD of 1.70 Å for the binder apo, 2.00 Å for the peptide neighbour interface and 1.64 Å for the whole complex. c, Co-crystal structure of RPB_PEW3_R4–PAWx4. The PAW units bind to a relatively flat groove formed by repeating histidine residues and glutamine residues as designed (shown as sticks). d, Top five complex PDBs for RPB_PEW3_R4–PAWx4 from the flexible docking generated ensemble. Green, pink and grey are the ones closest to the crystal structure (shown in yellow)

with RMSD over the peptide and the binding residues ≈ 0.03 Å, whereas the cyan dock RMSD = 3.89 Å.



Extended Data Fig. 6 SSM binding interface footprinting results were consistent with the design model and crystal structure.

a, Using a PPL repeat-peptide binder as an example, a heat map presenting enrichment analysis for each mutation is generated. In each cell, the red colour indicates enrichment, and the blue colour indicates depletion. Wild-type sequences are indicated in the cells labelled with amino-acid one-letter codes. The mutants missing in the expression library are labelled with asterisks. Two positions (109Q and 156Q) are highlighted as examples showing conserved positions. Almost all mutations other than the wild type in these two positions are greatly depleted. b, Illustration shows the SSM region (orange), and the two conserved positions (109Q and 156Q in yellow).

Design Methods

DHR Scaffolds generation

Each Designed Helical Repeat (DHR) scaffold is formed by a helix-loop-helix-loop topology that is repeated four or more times[37]. The helices range from 18 to 30 residues and the loops from 3 to 4 residues. The DHR design process goes through backbone design, sequence design and computational validation by energy landscape exploration. To match the peptides the designs were required to have a twist (omega) between 0.6 and 1.0 radians, a radius of 0 to 13 Å and a rise between 0 and 10 Å. The geometry of a repeat protein can be described by the radius of the super-helix, the axial displacement and the twist[38].

The backbone is designed using Rosetta fragment assembly guided by motifs[39]. Backbone coordinates are built up through 3,200 Monte Carlo fragment assembly steps with fragments harvested from a non-redundant set of structures from the PDB. Following each fragment insertion, the rigid body transform is propagated to the downstream repeats. The score that guides fragment assembly is composed of Van der Waal interactions, packing, backbone dihedral angles, and Residue-Pair-Transform(RPX) motifs[39]. RPX motifs are a fast way to measure the full-atom hydrophobic packability of the backbone prior to assigning side chains. After design, backbones are screened for native-like features. The loops are required to be within 0.4Å of a naturally occurring loop or rebuilt. Structures with helices above 0.14Å appear bent and kinked and are discarded. And poorly packed structures where <4 helices are in contact with each-other are filtered.

Sequence is designed using Rosetta for each backbone that passes filtering. Design begins in a symmetric mode where each repeat is identical using the RepeatProteinRelax mover. Core residues are restricted to be hydrophobic and surface residues hydrophilic using the layer

design task operators. Sequence is biased toward natural proteins with similar local structure using the structure profile mover. After the symmetrical design is complete the N-terminal and C-terminal repeats are redesigned to eliminate exposed hydrophobics. Designs with poor core packing as measured by Rosetta holes < 0.5 are then filtered[40].

The designs are computationally validated using the Rosetta *ab initio* structure prediction on Rosetta@Home[41]. Rosetta *ab initio* verifies that the design is lower energy state than the thousands of alternatives conformations sampled. Simulating a protein using Rosetta@Home can take several days on hundreds of CPUs. To speed this up we used machine learning to filter designs that were most likely to fail[38].

Backbone generation of curved repeat protein monomers in Poly-Proline II conformation

A second round of designs was made to ensure the distance between helices match the 10.9Å. distance between prolines in the Poly-Proline II. To design these backbones we used AtomPair constraints between the first helix of each repeat. The atom pair constraints were set to 10.9Å with a tolerance of 0.5 Å. For these designs we found the topologies that most efficiently produced structures that matched the AtomPair constraints had a helix length of 20 or 21 residues and a loop range of three residues.

Peptide binders design

Modular peptide docking and Hashing

To construct hash tables storing the pre-computed privileged residue interactions, we first survey the non-redundant PDB database and extract the intended interacting residues as seeds. For each seeding interaction residue pairs, random perturbations were applied to search for alternative relative conformations of the interacting residues. In the case of the sidechain-backbone bidentate interactions, the backbone residues were applied a random rigid

body perturbations with a random set of euler angles drawing from a normal distribution with 0° as the mean and 60° as the standard deviation, as well as a random set of translation distances in 3D space drawing from a normal distribution with 0 Å as the mean and 1 Å as the standard deviation. At the same time, the backbone torsion angles Φ and Ψ of the backbone residue were randomly modified to values draw from a ramachandran density plot based on structures from PDB database. The transformed set of residues losing the intended interactions were discarded. The transformed residues keeping the interactions will be collected. Then the side chains of the sidechain residues were replaced with all reasonable rotamers, to further diversify the samples of the sets of interacting residues. Finally, the geometry relationship of each set of residues keeping the intended interactions were subjected to an 8D hash function (6D rigid body transformation plus two torsion angles), and represented with a 64 bit unsigned integer as the key of an entry in the hash table. The identity and the side chain torsion angles (Xs) of the side chain residues were treated as the value of the entry in the hash table. The similar processes were carried out to build different hash tables for various interactions, with minor alterations. For example, for pi-pi and cation-pi interactions, only a 6D hash function was used, because there is no need for the perturbation and consideration of the backbone torsions. For ASN, GLN, ASP or GLU interacting two residues on the backbone, a 10D hash table was applied for representing the geometry relationship, and in these cases, the geometries of the N-H and C=O groups on the backbone were treated as 5D rays.

To sample repeat peptides matching the superhelical parameters of the DHRs, we randomly generate a set of backbone torsion angles φ and ψ, for example, [φ_1, ψ_1, φ_2, ψ_2, φ_3, ψ_3] for repeats of tri-peptide. If there any pair of φ and ψ angles get a high Rosetta Ramachandran score above the threshold -0.5, it means that this pair of torsion angles are likely to introduce intra-peptide steric clashes, and we will randomly regenerate a new pair of φ and ψ angles until they are reasonable according to the Rosetta Ramachandran score. Next, we will set the

backbone torsion angles of the repeat peptide using this set of φ and ψ angles repetitively across the 8-repeats. And we will calculate the superhelical parameters using the 3D coordinates of adjacent repeat units of the repeat peptide. The repeat peptides matching the superhelical parameters of any one of the curated DHRs, will be saved for the docking step.

To dock cognate repeat proteins and repeat peptides, with matching superhelical parameters, they are first aligned to the z axis by their own superhelical axes. Next step, a 2D grid search (rotation around and translation along the z axis) were carried out to sample compatible positions of the repeat peptide in the binding groove of the repeat protein. Once a reasonable dock is generated without steric clash, the relevant hash function will be used to iterate through all potential peptide-protein interacting residue sets, to calculate the hash keys. If a hash key exists in the hash table, the interacting side chain identities and torsion angles would be pulled out immediately and installed on all equivalent positions of this repeat peptide-repeat protein docking conformation. The docked peptide-DHR pair will be saved for the interface design step if the peptide-DHR hydrogen-bond interactions are satisfied.

Peptide binding interface design

If a single dock was accepted with the designed repetitive peptide-DHR hydrogen-bond, the peptide was first trimmed to the exact same repeat number as the DHR (e.g., 4-repeat or 6-repeat). After that, for both peptide and DHR sides, each amino acid was set linked to its corresponding amino acids on the same position in each repeat unit. This was to make sure all the following design steps would be carried out with the exact same symmetry inside of both the DHR and peptide.

During our design cycles, the interface neighbor distance is set as 9Å as the whole designable range around the DHR-peptide binding interface, and 11Å as the whole minimization range. Three rounds of full hydrophobic fastdesign[39] followed by hydropathic fastdesign were carried out, with each hydrophobic or hydrophilic fastdesign repeating twice. The Rosetta score function beta_nov16 was chosen in all design cycles. In the produced complex, the peptide itself with an average score (three calculations were carried out) larger than 20.0 or the complex scored larger than -10.0 were rejected directly.

After the preliminary design was done, we carried out two types of sanity checks to further optimize the designed peptide sequence, as well as the designed DHR interface. Specifically, for the peptide side, in the tri-peptide repeat units, every two amino acids other than Proline were scanned for a possible mutation to all twenty amino acids except cysteine, unless a certain originally designed peptide amino acid is making the hashed sidechain-backbone hydrogen-bond, or sidechain-sidechain hydrogen-bond, or sidechain-sidechain-backbone hydrogen-bond with the DHR interface. DDG (binding energy for the peptide-DHR complex) was compared before and after this peptide side mutation; and the mutation was accepted if the delta DDG (DDG_after - DDG_before) was larger than 1.0. Similarly, we also checked the designed DHR interface by mutation. The whole DHR was scanned. For the designed hydrophobic amino acids which were originally hydrophilic, a delta DDG of -5.0 was set as the threshold to be accepted as a necessary design which made enough binding contribution. For the designed hydropathic amino acids, a delta DDG of -2.0 was used as threshold.

For experimental characterization, we selected designed complexes with near ideal bidentate hydrogen bonds between protein and peptide, favorable protein-peptide interaction energies (DDG <= -35.0), interface shape complementary (Iface_SCval >= 0.65), tolerable interface

unsatisfied hydrogen bonds (Iface_HbondsUnsatBB <= 2, Iface_HbondsUnsatSC <=4) and low peptide apo energies (ScoreRes_chainB <= 0.9).

Forward Docking

As for the selected designed complexes from our round-two experiments, forward docking was performed to ensure the specificity in silico. For each designed complex, 10,000 arbitrary peptide conformations were generated as above, using the designed sequence. The same docking protocol was conducted as described in the docking stage, against the untouched designed DHR. Fastrelax[42] was then performed for the 10,000 docks, and the DDG vs. peptide backbone RMSD was plotted to check the convergence of the complex. Only the “converged” complexes were selected for experimental characterization, e.g., i) peptide backbone RMSD<2.0Å among the top 20 designs with lowest DDG during Forward Docking and ii) the averaged peptide backbone of the top 20 designs was close to the original design model (RMSD<1.5Å).

SSM library preparation

We carried out the site saturation mutagenesis (SSM) studies for some of the designed peptide-protein binding pairs to gain better understanding of the peptide binding modes, and to search for improved peptide binders. For each designed repeat protein, we ordered a SSM library covering the central span of 65 amino acids within the whole repeat protein, due to the chip DNA size limitation. This span roughly equals one and half repeating units, across three helices. The chip synthesized DNA oligos for the SSM library then amplified and transformed to EBY100 yeast together with a linearized pETCON3 vector including the encoding regions of the rest of the designed repeat protein. Each SSM library was subject to an expression sort first, in which the low-quality sequences due to chip synthesize defects or recombination errors were

filtered out. The collected yeast population, which successfully expresses the designed repeat protein mutants, will be re-grown, and be subject to the further round of peptide binding sorts. The next-generation sequencing results of this yeast population will also serve as the reference data for SSM analysis. The next round of without-avidity peptide binding sorts used various concentrations of the target peptide, depending on the initial peptide binding abilities, ranging from 1 nM to 1000 nM. The peptide-bound yeast populations were collected and sequenced by using Illumina NextSeq kit. The mutants were identified and compared to the mutants in the expression libraries. The enrichment analysis was carried out to identify the beneficial mutants and provide information for interpreting the peptide binding modes. For each mutant, its enrichment value is calculated by dividing its ratio in peptide-bound population by its ratio in expression population. The enrichment value is then subject to a log10 transformation, and plotted in heat maps for the SSM analysis.

Design of binders against endogenous targets

To evaluate which endogenous proteins could be currently targeted with our method (Fig.6), we developed a python code to search databases for subsequences matching permutations of the set of amino acid triplets we designed binders for in this study (i.e LRP PEW PLP IYP PKW IRP LRT LRN LRQ RRN PSR PRQ). This code can be accessed freely (https://github.com/tjs23/prot_pep_scan). We then ranked all outputs to find the longest subsequence possible, and manually inspected the candidates to find subsequences landing in disordered regions. Doing this analysis on the human proteome suggested that ZFC3H1 could be a good target for two main reasons: 1) this protein possess the sequence (PLP)x4 within a large disordered domain, with downstream sequence (PEDPEQPPKPPF) within the reach of our binder design method and 2) the protein is well studied, and, in particular, commercial, highly specific and validated antibodies exist against it.

Materials and Methods

Synthetic gene constructs

All genes in this work were ordered from either Integrated DNA Technologies (IDT) or Genscript. For both the first and second round designs, a His-tag containing TEV protease cleavage site and short linkers were added to the N-terminus of protein sequences. For the protein lacking a Tryptophan residue, a single Tryptophan was added to the short N-terminal linker following the TEV protease cleavage site to help with protein concentration quantification by A280. The protein sequence along with linker (MGSSHHHHHHHHSSGGSGGLNDIFEAQKIEWHEGGSGGSENLYFQSG or LEHHHHHH) was reverse translated into DNA using a custom python script that attempts to maximize host-specific codon adaptation index[43] and IDT synthesize-ability, which includes optimizing whole gene and local GC content as well as removing repetitive sequences. Finally, a TAATCA stop codon was appended to the end of each gene. Genes were delivered cloned into pET-29b+ between NdeI/XhoI restriction sites. For the second round designs, the designed amino acid sequences were inserted into pET-29b+ between NdeI/XhoI restriction sites directly.

For the ZFC3H1-103 disordered region, the 103 amino acids harboring the key targeting sequence (LPPPPQVSSLPPLSQPYVEGLCVSLEPLPPLPPLPPLPPEDPEQPPKPPFADEEEEEEMLLREELLKSLANKRAFKPEETSSNSDPPSPPVLNNSHPVPRSNL) was cloned into a customized vector with sfGFP at N-terminal and His6 at C-terminal with linker (GGSGSG) in between.

Protein expression and purification

Proteins were transformed into Lemo21(DE3) E. coli from New England Biolabs (NEB) and then expressed as 50 ml cultures in 250 ml flasks using Studiers M2 autoinduction media with 50 ug/mL kanamycin. The cultures were either grown at 37°C for ~6-8 hours and then ~18°C overnight (~14 hours) or at 37°C the entire time ~14 hours. Cells were pelleted at 4,000g for 10 minutes, after which the supernatant was discarded. Pellets were resuspended in 30 ml lysis buffer (25 mM Tris HCl pH 8, 150 mM NaCl, 30 mM imidazole, 1mM PMSF, 0.75% CHAPS, 1 mM DNase, 10mM Lysozyme, with Thermo Scientific Pierce protease inhibitor tablet). Cell suspensions were lysed by microfluidizer or sonication, and the lysate was clarified at 20,000g for ~30 minutes. The His-tagged proteins were bound to Ni-NTA resin (Qiagen) during gravity flow and washed with a wash buffer (25 mM Tris HCl pH 8, 150 mM NaCl, 30 mM imidazole). Protein was eluted with an elution buffer (25 mM Tris HCl pH 8, 150 mM NaCl, 300 mM imidazole). For the first round designs, the His-tag was removed by TEV cleavage, followed by IMAC purification to remove TEV protease. The flowthrough was collected and concentrated prior to further purification by SEC/FPLC on a superdex 200 increase 10/300 GL column in TBS (25 mM Tris pH 8.0, 150 mM NaCl).

Circular dichroism

Circular dichroism spectra were measured with an AVIV Model 420 DC or Jasco J-1500 CD spectrometer. Samples were 0.25 mg/mL in TBS (25 mM Tris pH 8.0, 150 mM NaCl), and a 1-mm path length cuvette was used. The CD signal was converted to mean residue ellipticity by dividing the raw spectra by $N \times C \times L \times 10$, where N is the number of residues, C is the concentration of protein, and L is the path length (0.1 cm).

Size exclusion chromatography with multi-angle light scattering

Purified samples after the initial SEC run, samples were pooled then concentrated or diluted as needed to a final concentration of 2 mg/mL. 100 uL of each sample was then run through a high-performance liquid chromatography system (Agilent) using a Superdex 200 10/300 GL column. These fractionation runs were coupled to a multi-angle light scattering detector (Wyatt) in order to determine the absolute molecular weights for each designed protein as described previously[44].

Small angle X-ray scattering

Small-Angle X-ray Scattering (SAXS) was collected at the SIBYLS High Throughput SAXS Advanced Light Source in Berkeley, California[45]. Beam exposures of 0.3 s for 10.2 s resulted in 33 frames per sample. Data was collected at low (~1.5 mg/mL) and high (~2-3 mg/mL) protein concentrations in SAXS buffer (25mM Tris pH 8.0, 150mM NaCl, 2% glycerol). The sibyls website ("SAXS FrameSlice" n.d.) was used to analyze the data for high and low centration samples and average the best dataset. If there was obvious aggregation over the 33 frames, only the data points before aggregation arose were used in the Gunier region, otherwise, all data was included for the Gunier region. All data was used for Porod and Wide regions. The averaged file was used with scatter.jar to remove data points with outlier residuals in the Gunier region. Finally, the data was truncated at 0.25 q. This dataset was then compared to the predicted SAXS profile based on the design model using the FoxS SAXS server("FoXS Server: Fast X-Ray Scattering" n.d.), and volatility ratio (Vr) was calculated to quantify how well the predicted and data matched the experimental data. Proteins with Vr of less than 2.5 were considered to be folded to the designed quaternary shape.

Biolayer interferometry

Biolayer interferometry binding data were collected in an Octet RED96 (ForteBio) and processed using the instrument's integrated software. To measure the affinity of peptide binders, N-terminal biotinylated (biotin-Ahx) target peptides with a short linker (GGS) were loaded onto streptavidin-coated biosensors (SA ForteBio) at 50-100 nM in binding buffer (10 mM HEPES (pH 7.4), 150 mM NaCl, 3 mM EDTA, 0.05% surfactant P20, 0.5% non-fat dry milk) for 120 s. Analyte proteins were diluted from concentrated stocks into the binding buffer. After baseline measurement in the binding buffer alone, the binding kinetics were monitored by dipping the biosensors in wells containing the target protein at the indicated concentration (association step) and then dipping the sensors back into baseline/buffer (dissociation).

Yeast surface display

S. cerevisiae EBY100 strain cultures were grown in C-Trp-Ura media and induced in SGCAA media following the protocol in (reference). Cells were washed with PBSF (PBS with 1% BSA) and labeled with biotinylated designed proteins using two labeling methods, with-avidity and without-avidity labeling. For the with-avidity method, the cells were incubated with biotinylated RBD, together with anti-c-Myc fluorescein isothiocyanate (FITC, Miltenyi Biotech) and streptavidin– phycoerythrin (SAPE, ThermoFisher). The concentration of SAPE in the with-avidity method was used at ¼ concentration of the biotinylated RBD. The with-avidity method was used in the first few rounds of screening against the repeat peptide library to fish out weak binder candidates. For the without-avidity method, the cells were firstly incubated with biotinylated designed proteins, washed, secondarily labeled with SAPE and FITC.

Crystallography

Crystallization and structure determination for RPB_PEW3_R4-PAWx4

Purified RPB_PEW3_R4 protein + PAWx4 peptide at a concentration of 36 mg/mL was used to conduct sitting drop, vapor-diffusion crystallization trials using the JCSG Core I-IV screens (NeXtal Biotechnologies). Crystals of RPB_PEW3_R4-PAWx4 grew from drops consisting of 100 nL protein plus 100 nL of a reservoir solution consisting of 0.1 M MES pH 5.0, 30% (w/v) PEG 6K at 4°C, and were cryoprotected by supplementing the reservoir solution with 5% ethylene glycol. Native diffraction data was collected at APS beamline 23-ID-D, indexed to $P2_12_12_1$ and reduced using XDS[48] (Table S1). The structure was phased by molecular replacement using Phaser[48]. A set of ~50 lowest energy predicted models from Rosetta were used as search models. Several of these models gave clear solutions, which were adjusted in Coot[49] and refined using Phenix[50]. Model refinement in $P2_12_12_1$ initially resulted in unacceptably high values for R_{free} - R_{work}. Refinement was therefore first performed in lower symmetry space groups (P1 and $P2_1$). In the late stages of refinement, these P1 and $P2_1$ models were refined against the $P2_12_12_1$, which ultimately yielded acceptable, albeit somewhat higher R-factors. Data collection and refinement statistics are provided in Extended Data Table 1.

Crystallization and structure determination for RPB_PLP3_R6-PLPx6

Purified RPB_PLP3_R6 protein + PLPx4 peptide at a concentration of 70 mg/mL was used to conduct sitting drop, vapor-diffusion crystallization trials using the JCSG Core I-IV screens (NeXtal Biotechnologies). Crystals of RPB_PLP3_R6-PLPx6 grew from drops consisting of 100 nL protein plus 100 nL of a reservoir solution consisting of 2.4 M (NH4)2SO4, 0.1 M Na3Cit pH4 at 18 °C, and were cryoprotected by supplementing the reservoir solution with 2.2M sodium

malonate pH4. Native diffraction data was collected at APS beamline 23-ID-D, indexed to I422 and reduced using XDS[47] (Table S1). The structure was phased by molecular replacement using Phaser[48]. A set of ~28 lowest energy predicted models from Rosetta were used as search models. Several of these models gave clear solutions, which were adjusted in Coot[49] and refined using Phenix[50]. Data collection and refinement statistics are provided in Extended Data Table 1.

Crystallization and structure determination for RPB_LRP2_R4-LRPx4

Purified RPB_LRP2_R4 protein + LRPx4 peptide at a concentration of 21.4 mg/mL was used to conduct sitting drop, vapor-diffusion crystallization trials using the JCSG Core I-IV screens (NeXtal Biotechnologies). Crystals of RPB_LRP2_R4-LRPx4 grew from drops consisting of 100 nL protein plus 100 nL of a reservoir solution consisting of 0.1 M HEPES pH7, 10 %(w/v) PEG6000 at 18 °C, and were cryoprotected by supplementing the reservoir solution with 25 % Ethylene glycol. Native diffraction data was collected at APS beamline 23-ID-B, indexed to P32 2 1 and reduced using XDS[47] (Table S1). The structure was phased by molecular replacement using Phaser[48]. The coordinates of apo RPB_LRP2_R4 from the proteolyzed/filament structure were used as a search model. The resulting model was adjusted in Coot[49] and refined using Phenix[50]. Like the apo structure, this crystal structure of RPB_LRP2_R4 also contained “infinitely long filaments in the crystal, this time with peptide bound. Data collection and refinement statistics are provided in Extended Data Table 1.

Crystallization and structure determination for RPB_PLP1_R6-PLPx6

Purified RPB_PLP1_R6 protein + PLPx6 peptide at a concentration of 143 mg/mL was used to conduct sitting drop, vapor-diffusion crystallization trials using the JCSG Core I-IV screens (NeXtal Biotechnologies). Crystals of RPB_PLP1_R6-PLPx6 grew from drops consisting of 100 nL protein plus 100 nL of a reservoir solution consisting of 0.2 M NaCl, 20 %(w/v) PEG3350 at 4

°C, and were cryoprotected by supplementing the reservoir solution with 15 % Ethylene glycol. Native diffraction data was collected at APS beamline 23-ID-B, indexed to H32 and reduced using XDS[47] (Table S1). The structure was phased by molecular replacement using Phaser[48]. A set of ~230 lowest energy predicted models from Rosetta were used as search models. Several of these models gave clear solutions, which were adjusted in Coot[49] and refined using Phenix[50]. In the later stages of refinement, two copies of the 6xPLP peptide were built into clearly defined electron density in the asymmetric unit. The first copy adopts the expected location based on the design, and makes the designed interactions with RPB_PLP1_R6. The density for this peptide and the final atomic model (19 amino acid residues) are slightly longer than the peptide used in crystallization (18 residues); this is likely due to “slippage”/misregistration of the peptide relative to the RPB_PLP1_R6-PLPx6 in many unit cells, resulting in density longer than the peptide itself. A second copy of the peptide lies across a 2-fold symmetry axis at ~50% occupancy, resulting in the superposition of this peptide with a symmetry-derived copy of itself running in the opposite direction. Despite this, the locations of each Pro/Leu side chain unit was reasonably well defined. However, it seems unlikely that the binding of the peptide at this second site would occur readily in solution. Data collection and refinement statistics are provided in Extended Data Table 1.

Crystallization and structure determination for RPB_PLP1_R6, alternative conformation 1

Purified RPB_PLP1_R6 protein + PLPx6 peptide at a concentration of 166 mg/mL was used to conduct sitting drop, vapor-diffusion crystallization trials using the JCSG Core I-IV screens (NeXtal Biotechnologies). Crystals of RPB_PLP1_R6-PLPx6 grew from drops consisting of 100 nL protein plus 100 nL of a reservoir solution consisting of 0.02 M CaCl2, 30 % (v/v) MPD, 0.1 M NaAcet pH 4.6 at 18 °C, and were cryoprotected by supplementing the reservoir solution with 5 % MPD. Native diffraction data was collected at APS beamline 23-ID-B, indexed to P22121 and reduced using XDS[47] (Table S1). The structure was phased by molecular replacement using

Phaser[48], using the coordinates for RPB_PLP1_R6-PLPx6 (alternative conformation 1) as a search model. The model was adjusted in Coot[49] and refined using Phenix[50]. In the later stages of refinement, one copy of the 6xPLP peptide was model at a site of crystal contact, where it is sandwiched between adjacent subunits in a way that is likely only bound in the crystal lattice. Data collection and refinement statistics are provided in Extended Data Table 2.

Crystallization and structure determination for RPB_PLP1_R6, alternative conformation 2

Purified RPB_PLP1_R6 protein +PLPx6 peptide at a concentration of 166 mg/mL was used to conduct sitting drop, vapor-diffusion crystallization trials using the JCSG Core I-IV screens (NeXtal Biotechnologies). Crystals of RPB_PLP1_R6-PLPx6 grew from drops consisting of 100 nL protein plus 100 nL of a reservoir solution consisting of 40 % (v/v) MPD, 0.1 M Na Phos Cit pH 4.2 at 18 °C, and were cryoprotected by supplementing the reservoir solution. Native diffraction data was collected at APS beamline 23-ID-B, indexed to P22121 and reduced using XDS[47] (Table S1). Initial attempts to phase by molecular replacement using Phaser[48] and ~500 predicted models from Rosetta and RoseTTAfold failed to yield any clear solutions. Similarly, several thousand truncations of these models (containing all combinations of 1, 2, 3, 4, or 5 of the 6 repeat units), also failed to give clear solutions. To try to identify correct but low scoring solutions in the output of these trials, we ran SHELXE autobuilding and density modification on a large number of these potential solutions. Ultimately, we were able to identify an MR solution with 2 of 6 repeats correctly placed that allowed the autobuilding of a polyalanine model and an interpretable map, which could be further improved by iterative rounds of rebuilding in Coot[49] and refinement using Phenix[50]. Ultimately, the final model revealed that in this crystal form and a similar crystallization condition (RPB_PLP1_R6, alternative conformation 1, above), RPB_PLP1_R6 adopted an alternative fold. Data collection and refinement statistics are provided in Extended Data Table 2.

Crystallization and structure determination for RPB_LRP2_R4

Purified RPB_LRP2_R4-LRPx4 protein at a concentration of 33 mg/mL was used to conduct sitting drop, vapor-diffusion crystallization trials using the JCSG Core I-IV screens (NeXtal Biotechnologies). Crystals of RPB_LRP2_R4 grew from drops consisting of 100 nL protein plus 100 nL of a reservoir solution consisting of 0.2 M K2HPO4, 20% (w/v) PEG 3350 at 18°C, and were cryoprotected by supplementing the reservoir solution with 15% ethylene glycol. Native diffraction data was collected at APS beamline 23-ID-B, indexed to P 32 2 1 and reduced using XDS[47] (Table S1). The structure was phased by molecular replacement using Phaser[48]. A set of ~50 lowest energy predicted models from Rosetta, as well as a variety of truncated models, were used as search models. Several of these models gave clear solutions, which were adjusted in Coot[49] and refined using Phenix[50]. Four helical repeat modules were present in the asymmetric unit. However, unexpectedly, side chain density for all four repeats were very similar to one another and matched the sequence of the internal helical repeats, but not the N- and C-terminal capping repeats, which are slightly different from the internal ones. In addition, these four repeat units pack tightly against adjacent, symmetry-related molecules such that they form an “infinitely long” repeat protein running throughout the crystal. Careful examination of the junction between each repeat unit revealed no clear breaks in electron density; the density for the backbone is continuous through the asymmetric unit, and continuous with the symmetry related molecules near the N- and C-terminus of the molecule in the asymmetric unit. Rather than truly forming an infinitely long polymer, we suspect that proteolytic cleavage of the RPB_LRP2_R4 (either during purification or crystallization led to the removal of the N- and/or C-terminal caps in many molecules, which could allow the internal repeats from separate molecules to polymerize to form fibers in the crystal. Heterogeneity in these cleavage products and how they assemble into the crystal lattice (misregistration) could consequently explain the “continuous” filaments of this repeat protein that we observe in these crystals. Data collection and refinement statistics are provided in Extended Data Table 2.

Cell Studies

Plasmids

For expression in cells, constructs were synthesized by Genescript and cloned into a modified pUC57 plasmid (Genscript) allowing mammalian expression under a EF1a promoter. Target peptides were cloned as C-terminal fusions with a linker (GAGAGAGRP) followed by EGFP. Binders were expressed as fusions with the first 34 residues of the Mas70p protein (Mito-tag), shown to efficiently relocalise proteins to mitochondria in mammalian cells[51] in N-terminal and with mScarlet in C-terminal[52]. Plasmids encoding the GFP-tagged peptide and the mscarlet-tagged binder were then cotransfected into cells.

Alternatively, for in vivo demonstration of the multiplexed binding between different peptides and their cognate binders (Fig. 3F, G), bicistronic plasmids were generated expressing the binder flanked with a Mito-tag (respectively, PEX-tag, the first 66 residues of human PEX3, targeting to peroxisomes[53]) followed by a stop codon, then an Internal ribosome entry site (IRES) sequence and the target peptide tagged with EGFP (respectively mScarlet). Cells were then cotransfected with both bicistronic plasmids to express all four proteins.

Cells

U2OS FlipIn Trex cells (kind gift from Stephen C. Blacklow) and HeLa FlpIn Trex cells (kind gift from Simon Bullock), were cultured in DMEM (Corning) supplemented with 10% fetal bovine serum (Gibco) and 1% Pen/Strep (Gibco) at 37°C with 5% CO_2. Cells were transfected with

Lipofectamine 3000 (Invitrogen) according to the manufacturer's instructions and imaged after 1 day of expression.

Live-cell imaging

For live cell imaging (Fig.3), U2OS FlipIn Trex cells were plated on glass-bottom dishes (World Precision Instruments, FD35) coated with fibronectin (Sigma, F1141, 50 µg/ml in PBS), for 1 hours at 37°C DMEM-10% serum. Medium was then changed to Leibovitz's L-15 medium (Gibco) supplemented with 20 mM HEPES (Gibco) for live cell imaging. Imaging was performed onto a custom spinning disk confocal instrument composed of Nikon Ti stand equipped with perfect focus system, a fast Z piezo stage (ASI) and a PLAN Apo Lambda 1.45 NA 100X objective, and a spinning disk head (Yokogawa CSUX1). Images were recorded with a Photometrics Prime 95B back-illuminated sCMOS camera run in pseudo global shutter mode and synchronized with the spinning disk wheel. Excitation was provided by 488, 561 lasers (all Coherent OBIS mounted in a Cairn laser launch) and imaged using dedicated single bandpass filters for each channel mounted on a Cairn Optospin wheel (Chroma 525/50 for GFP and Chroma 595/50 for mScarlet). To enable fast 4D acquisitions, an FPGA module (National Instrument sbRIO-9637 running custom code) was used for hardware-based synchronization of the instrument, in particular to ensure that the piezo z stage moved only during the readout period of the sCMOS camera. Temperature was kept at 37ºC using a temperature control chamber (MicroscopeHeaters.Com, Brighton UK). System was operated by Metamorph.

Immunofluorescence

For immunofluorescence of mitochondria (Extended Data Fig. 2B), U2OS FlpIn Trex cells (kind gift from Stephen C. Blacklow) were spread on glass-bottom dishes coated with fibronectin as above. Cells were washed with PBS then fixed in 4% PFA for 20 minutes at room temperature. Following fixation, cells were washed in with PBS and then permeabilized with 0.1% Triton X-100 in PBS for 5 minutes at room temperature. Cells were washed again with PBS and blocked in 1% BSA in PBS for 15 minutes. Cells were then incubated with TOM20 antibody (Santa Cruz, sc-17764, used at 1/200 dilution), diluted in 1% BSA in PBS, for 1 hour at room temperature. Cells were washed 3 times with PBS and then incubated with DAPI (Roche #10236276001) and anti-mouse Alexa Fluor 488, diluted at 1/400 in 1% BSA in PBS, for one hour at room temperature. Cells were washed a final 3 times in PBS and then imaged using the spinning disk confocal described above.

Pull down of endogenous proteins from extracts using designed binders

For pull down of endogenous ZFC3H1 from human cell extracts, HeLa FlpIn Trex cells were lysed in lysis buffer (25 mM HEPES, 150 mM NaCl, 0.5% Tx100, 0.5% NP-40, 20 mM imidazole, pH 7.4 supplemented with Roche EDTA free protease inhibitor tablets). Lysate was incubated on ice for 10 minutes to continue lysis and then were spun at 4000 x *g* for 15 minutes at 4°C. The supernatant was incubated with pre-washed Ni-NTA agarose (Qiagen, 30210 318/AV/01) for 1 hour with rocking at 4°C to remove/reduce proteins in the lysate that bind to the resin non-specifically. For each condition, 50 µl of fresh Ni-NTA agarose resin was washed twice in lysis buffer. Equimolar amounts of purified his-tagged binder, or as a control an equal volume of buffer, was added to the Ni-NTA agarose. The pre-cleared HeLa lysate was split evenly between the 3 conditions. An input was taken of each condition, and the tubes were incubated for 2 hours at 4°C with rocking. Beads were then washed twice in lysis buffer and twice in wash

buffer (25 mM HEPES, 150 mM NaCl, 20 mM imidazole pH 7.4). Proteins were then eluted from the beads in elution buffer (25 mM HEPES, 150 mM NaCl, 500 mM imidazole, pH 7.4). Inputs and elutions were ran on a NuPage 3-8% Tris-Acetate gel (Invitrogen, EA0375) and transferred to a nitrocellulose membrane using the iBlot system (ThermoFischer). Membranes were blocked in 5% (w/v) milk in TBS-TWEEN (10 mM Tris-HCl, 120 mM NaCl, 1% (w/v) TWEEN20, pH 7.4) for 30 mins at room temperature with gentle shaking. Rabbit anti-ZFC3H1 (Sigma, HPA007151, used at 1:250) and Mouse anti-alpha tubulin 488 (Clone DMA1, Sigma T6199, directly labelled with Abberior® STAR 488, NHS ester leading to a 4.5 dye/antibody degree of labelling, and used at 0.1 µg/mL final concentration) were diluted 1% (w/v) milk in TBS-TWEEN and incubated with the membrane overnight at 4°C with gentle shaking. The membrane was washed 3x in TBS-TWEEN then incubated with goat anti-Rabbit Alexa 555 (Invitrogen, A32732, 1:2000) for one hour at room temperature with gentle shaking. The membrane was washed twice with TBS-TWEEN and a final wash with TBS-TWEEN with 0.001% SDS. Membranes were imaged using a ChemiDoc system (BioRad). Alternatively, the same samples were analyzed using 4-12% Bis-Tris gels (Invitrogen NP0323BOX) and stained with Instant blue coomassie stain (Sigma ISB1L). Note that αZFC-high is also able to pull down endogenous ZFC3H1 from human cell extracts when 50 mM rather than 150 mM NaCl was used in all buffers (Extended Data Fig. 7B).

Mass Spectrometry

Each line of the polyacrylamide gel presented in Fig.6C was cut into six pieces (1-2 mm) and prepared for mass spectrometric analysis by manual *in situ* enzymatic digestion (the gel area containing the binder was omitted from the analysis to avoid saturation of the detector by overabundance of binder peptides). Briefly, the excised protein gel pieces were placed in a well

of a 96-well microtiter plate and destained with 50% v/v acetonitrile and 50 mM ammonium bicarbonate, reduced with 10 mM DTT, and alkylated with 55 mM iodoacetamide. After alkylation, proteins were digested with 6 ng/μL Trypsin (Promega, UK), 0.1% Protease Max (Promega, UK) overnight at 37 °C. The resulting gel pieces were extracted with ammonium bicarbonate (100 μL, 100 mM) and ammonium bicarbonate/acetonitrile (50/50, 100 μL) before being dried down *via* vacuum. Clean-up of peptide digests was carried out with HyperSep SpinTip P-20 (ThermoScientific, USA) C18 columns, using 80% acetonitrile as the elution solvent before being dried down again. The resulting peptides were and were extracted in 0.1% v/v trifluoroacetic acid acid, 2% v/v acetonitrile. The digest was analyzed by nano-scale capillary LC-MS/MS using an Ultimate U3000 HPLC (ThermoScientific Dionex, San Jose, USA) to deliver a flow of 250 nL/min. Peptides were trapped on a C18 Acclaim PepMap100 5 μm, 100 μm x 20 mm nanoViper (ThermoScientific, USA) before separation on PepMap RSLC C18, 2 μm, 100 A, 75 μm x 75 cm EasySpray column (ThermoScientific, USA). Peptides were eluted on a 90 minute gradient with acetonitrile and interfaced *via* an EasySpray ionization source to a quadrupole Orbitrap mass spectrometer (Q-Exactive HFX, ThermoScientific, USA). MS data were acquired in data dependent mode with a Top-25 method, high resolution scans full mass scans were carried out (R = 120,000, *m/z* 350 – 1750) followed by higher energy collision dissociation (HCD) with collision energy 27 % normalized collision energy. The corresponding tandem mass spectra were recorded (R=30,000, isolation window *m/z* 1.6, dynamic exclusion 50 s). LC-MS/MS data were then searched against the Uniprot human proteome database, using the Mascot search engine programme (Matrix Science, UK)[54]. Database search parameters were set with a precursor tolerance of 10 ppm and a fragment ion mass tolerance of 0.1 Da. One missed enzyme cleavage was allowed and variable modifications for oxidation, carboxymethylation, and phosphorylation. MS/MS data were validated using the Scaffold programme (Proteome Software Inc., USA)[55]. All data were additionally interrogated manually. To generate the Venn diagram in Fig.6F, we considered a threshold of minimum 5 peptides to

consider that a protein had been identified. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[56] partner repository with the dataset identifier PXD038492 and 10.6019/PXD038492. See also source data for the annotated full dataset.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The atomic coordinates and experimental data of RPB_PEW3_R4-PAWx4, RPB_PLP3_R6-PLPx6, RPB_LRP2_R4-LRPx4, RPB_PLP1_R6-PLPx6, RPB_PLP1_R6-PLPx6 (alternative conformation 1), RPB_PLP1_R6-PLPx6 (alternative conformation 2) and RPB_LRP2_R4 (pseudopolymeric) have been deposited in the RCSB PDB with the accession numbers 7UDJ, 7UE2, 7UDK, 7UDL, 7UDM, 7UDN, and 7UDO respectively. The Rosetta macromolecular modeling suite (https://www.rosettacommons.org) is freely available to academic and non-commercial users. Commercial licenses for the suite are available through the University of Washington Technology Transfer Office. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD038492 and 10.6019/PXD038492**.**

Code availability

The Rosetta macromolecular modeling suite (https://www.rosettacommons.org) is freely available to academic and non-commercial users. Commercial licenses for the suite are available through the University of Washington Technology Transfer Office. The design scripts and main PDB models, computational protocol for data analysis, experimental data and analysis scripts, all the design models and NGS results used in this paper can be downloaded from file servers hosted by the Institute for Protein Design:
https://files.ipd.uw.edu/pub/2022_de_novo_design_of_modular_peptide_binding_proteins_by_superhelical_matching/

The code to identify proteins in databases containing any linear combination of amino acid triplets given as an input can be found on github (https://github.com/tjs23/prot_pep_scan) .

REFERENCES

37. Brunette, TJ et al. Exploring the repeat protein universe through computational protein design. Nature 528, 580–584 (2015).

38. Brunette, TJ et al. Modular repeat protein sculpting using rigid helical junctions. Proc. Natl. Acad. Sci. U. S. A. 117, 8870–8875 (2020).

39. Fallas, J. A. et al. Computational design of self-assembling cyclic protein homo-oligomers. Nat. Chem. 9, 353–360 (2017).

40. Sheffler, W. & Baker, D. RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Sci. 18, 229–239 (2009).

41. Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. Science 309, 1868–1871 (2005).

42. Tyka MD, Keedy DA, André I, et al. Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol. 405(2):607-618 (2011).

43. Sharp P. M., Li W., The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Research, 15(3), 281–1295 (1987).

44. Fallas, J., Ueda, G., Sheffler, W. et al. Computational design of self-assembling cyclic protein homo-oligomers. Nature Chem 9, 353–360 (2017).

45. Dyer, K. N., Hammel, M., Rambo, R. P., Tsutakawa, S. E., Rodic, I., Classen, S., Tainer, J. A., & Hura, G. L. High-throughput SAXS for the characterization of biomolecules in solution: a practical approach. Methods in molecular biology (Clifton, N.J.), 1091, 245–258 (2014).

46. Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 27, 293–315 (2018).

47. Kabsch, W. XDS. Acta Crystallogr. D Biol. Crystallogr. 66, 125–132 (2010).

48. McCoy, A. J. et al. Phaser crystallographic software. J. Appl. Crystallogr. 40, 658–674 (2007).

49. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. 66, 486–501 (2010).

50. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. 66, 213–221 (2010).

51. Kessels, M. M., & Qualmann, B. Syndapins integrate N-WASP in receptor-mediated endocytosis. The EMBO journal, 21(22), 6083–6094 (2002).

52. Bindels, D., Haarbosch, L., van Weeren, L. et al. mScarlet: a bright monomeric red fluorescent protein for cellular imaging. Nat Methods 14, 53–56 (2017).

53. Fakieh, M. H., Drake, P. J., Lacey, J., Munck, J. M., Motley, A. M., & Hettema, E. H. Intra-ER sorting of the peroxisomal membrane protein Pex3 relies on its luminal domain. Biology open, 2(8), 829–837 (2013).

54. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 20, 3551–3567 (1999).

55. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392 (2002).

56. Perez-Riverol Y, Bai J, Bandla C, Hewapathirana S, García-Seisdedos D, Kamatchinathan S, Kundu D, Prakash A, Frericks-Zipper A, Eisenacher M, Walzer M, Wang S, Brazma A, Vizcaíno JA (2022). The PRIDE database resources in 2022: A Hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res 50(D1):D543-D552 (PubMed ID: 34723319).

Acknowledgements

Funding: This work was supported by The Audacious Project at the Institute for Protein Design (D.B., K.W., M.D., D.A.S., A.B.), The Michelson Found Animals Foundation Grant Number GM15-S01 (L.S., K.W., D.B.), the National Institute on Aging grant 5U19AG065156-02 (D.H., K.W., D.B.), the Howard Hughes Medical Institute (D.B., W.S., H.B.), The Open Philanthropy Project Improving Protein Design Fund (A.C., R.R., C.M.C., G.B., D.E., D.B.), The Donald and Jo Anne Petersen Endowment for Accelerating Advancements in Alzheimer's Disease Research (T.J.B., D.B.), a donation from AMGEN to the Institute for Protein Design (I.G.), the Medical Research Council (MC_UP_1201/13 to E.D., T.E.M, T.J.S.), the Human Frontier Science Program (CDA00034/2017-C to E.D.) and a Sir Henry Wellcome Postdoctoral Fellowship (220480/Z/20/Z to K.E.M.).

Competing Interests

Each contributor attests that they have no competing interests relating to the subject contribution, except as disclosed. K.W., H.B., D.R.H., T.J.B., K.E.M., T.J.S, T.E.M, A.C., R.R., G.B., D.E., L.S., E.D, D.A.S., W.S., I.G. and D.B. are co-inventors on a patent application that incorporate discoveries described in this article.

We thank B. Wicky, A. Ljubetic, and I. Lutz for advice on the split luciferase assay for the second-round design screening, C. Xu for help trouble-shooting experiments, T. Schlichtharle for discussion, L. Cao and I. Goreshnik for advice on biolayer interferometry, H. Pyles for advice on circular dichroism (CD) and designed helical repeat proteins (DHR),Ramanujan Hegde for

the suggestion to target disordered regions of endogenous proteins, K. Van Wormer and A. Curtis Smith for laboratory support during COVID-19.

Author contributions

K.W. and H.B. contributed equally to this work; K.W., D.A.S. and D.B. designed the research; D.A.S. and D.B. developed the preliminary computational method and hash database; W.S. contributed to the hash database development; K.W. updated the computational method with the help from D.A.S and H.B.; H.B. updated the hash database to be more general; Y.S. helped and contributed to the first hash database development; K.W. and T.J.B. designed the polyproline 2 DHR scaffold library using the method developed by D.R.H.; K.W. designed the binders with the help from H.B; H.B. and K.W. performed the yeast screening, expression and binding experiments with the help from I.G. for the first-round design characterization; K.W. performed biolayer interferometry, Octet assays for the second-round design characterization; H.B. constructed and screened site saturation mutagenesis libraries (SSMs). A.C., R.R., G.B., D.E. solved the structures of RPB_PEW3_R4-PEWx4, RPB_PLP3_R6-PLPx6, RPB_LRP2_R4-LRPx4 and RPB_PLP1_R6-PLPx6; K.E.M. designed and performed all cell experiments in this work, in particular the multiplex binding assay and the demonstration of the endogenous binder for ZFC3H1. E.D. identified ZFC3H1 as a good target for the development of an endogenous binder with help from T.J.S. T.E.M. performed mass spectrometry analysis ; A.B. helped with the modular binding assay; M.D. and C.M.C. helped with preparing protein samples for crystallography. All authors analyzed data. L.S., D.A.S. and D.B. supervised research. K.W. and D.B. wrote the manuscript with the input from the other authors. All authors revised the manuscript.

Chapter 5 – De novo design of Ras selective binders

5.0 – preface, authors, and abstract

Preface:

Note: this chapter is borrowed directly from a submitted manuscript prepared at the same time, which is the original copy. I, Kejia Wu, am co-corresponding author for this work.

Authors:

Jason Z. Zhang[1-3,8], Alexa Rane Batingana[4], Xinting Li[1,2], Caixuan Liu[1,2], Hanlun Jiang[1,2,5], Kevin Shannon[4,6], Benjamin J. Huang[4,6], Kejia Wu[1,2,7,8] ,David Baker[1-3,8]

[1]Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States

[2]Institute for Protein Design, University of Washington, Seattle, Washington 98195, United States

[3]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, United States

[4]Department of Pediatrics, University of California, San Francisco, San Francisco, CA

[5]Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, California 94720, United States

[6]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA

[7]Biological Physics, Structure and Design Graduate Program, University of Washington, Seattle, Washington 98195, United States

[8]Co-corresponding authors, email: jzz0428@uw.edu, kejiawu@uw.edu, dbaker@uw.edu

Abstract:

The proto-oncogene Ras which governs diverse intracellular pathways has four major isoforms (KRAS4A, KRAS4B, HRAS, and NRAS) with substantial sequence homology and similar *in vitro* biochemistry. There is considerable interest in investigating the roles of these independently as their association with different cancers vary, but there are few Ras isoform-specific binding reagents as the only significant sequence differences are in their disordered and highly charged C-termini which have been difficult to elicit antibodies against. To overcome this limitation, we use deep learning-based methods to *de novo* design Ras isoform-specific binders (RIBs) for all major Ras isoforms that specifically target the Ras C-terminus. The RIBs bind to their target Ras isoforms both *in vitro* and in cells with remarkable specificity, disrupting their membrane localization and inhibiting Ras activity, and should contribute to dissecting the distinct roles of Ras isoforms in biology and disease.

5.1 - Introduction

The Ras family of GTPases modulates the mitogen activated protein kinase (MAPK) and other intracellular signaling pathways essential for cell growth and survival, and mutations in Ras are prevalent in many human cancers. The four major Ras isoforms–KRAS4A, KRAS4B, HRAS, and NRAS–have high sequence homology and similar biochemical properties. Despite their similarities, the isoforms are differentially mutated in different cancers[1], play different roles in drug resistance[2], and have distinct subcellular locations[3,4] due to their divergent disordered and highly charged C-termini. Whereas the structured portion of Ras GTPases has 90% sequence homology across the major Ras isoforms, the homology over the C-termini is only 8% (**Figure 1a-b**). While all isoforms can localize to the plasma membrane, NRAS, HRAS, and KRas4A are reversibly palmitoylated on their hypervariable C-termini, enabling endomembrane

localization[3,4]. It has been challenging to develop Ras isoform specific reagents[5]; those that are available often result in multiple bands in immunoblotting experiments and are insufficient for more sensitive assays such as immunostaining. Despite decades of research into the different Ras isoforms, their specific signaling activities and functional roles remain unclear due to the lack of isoform-selective molecular tools, such as selective affinity reagents, complicating precise functional studies.

Recent advances in protein design[6–8] now enable the design of binders to a wide range of protein targets. Designing Ras isoform specific binders requires targeting the disordered and highly charged C-terminus of Ras (**Figure 1a**) as it is the only region that differs between the Ras isoforms. Targeting highly polar, native protein regions such as these C-termini of Ras (e.g. KRAS4B C-terminus is 86% polar residues, 63% charged residues) has been difficult[9] due to the biochemical challenge of binder interaction competing with water interaction. Precise and extensive polar interactions are required for designing Ras binders, otherwise we need to pay substantial enthalpic penalty for striping away water-mediated hydrogen bonds and entropic penalty for stabilizing intrinsically disordered C-terminus of Ras. We reasoned that recently developed methods for designing binders to intrinsically disordered regions[10,11] could enable design of Ras isoform specific binders superior to currently available antibodies. These reagents could be valuable tools for dissecting the roles of the different isoforms in cellular function and disease.

5.2 - Computational design of Ras isoform selective binders (RIBs)

We explored two different peptide binder design strategies for designing Ras isoform specific binders (RIBs). The first is “side-chain centered”, i.e., threads the sequences of disordered regions (IDRs) through ~800 protein templates from logos[11] (logos library A) with specific amino acid recognition pockets arranged to bind diverse sequences in a range of extended conformations. The second is “backbone centered”, i.e., generates backbones via either “scaffolded RFDiffusion” by threading the sequences into a subset of ~200 β-sheet containing wrapping-up scaffolds also from logos[11] (logos library B) or from random noise using sequence input RFdiffusion[10] in which both the backbone conformation of the binder and the target were widely sampled. In both cases, the top scored designs would be optimized by the “refinement” protocol from the standard logos pipeline[11], i.e., partial RFDiffusion[12], motif RFDiffusion[11], and parametric perturbation[13]. Sequences are all designed using ProteinMPNN[7] (see Methods for details). Designs are selected for experimental characterization based on Alphfold 2 (AF2)[14] prediction confidence metrics (pae interaction), predicted binding affinity (Rosetta ΔΔG), and extent of buried polar atoms without hydrogen binding (buried unsaturated residues)[11] (**Figure S1**). To achieve specificity, we prioritized binders with extensive polar interactions with the Ras C-terminus, but disregarding the last 4 residues which either get cleaved or are farnesylated.

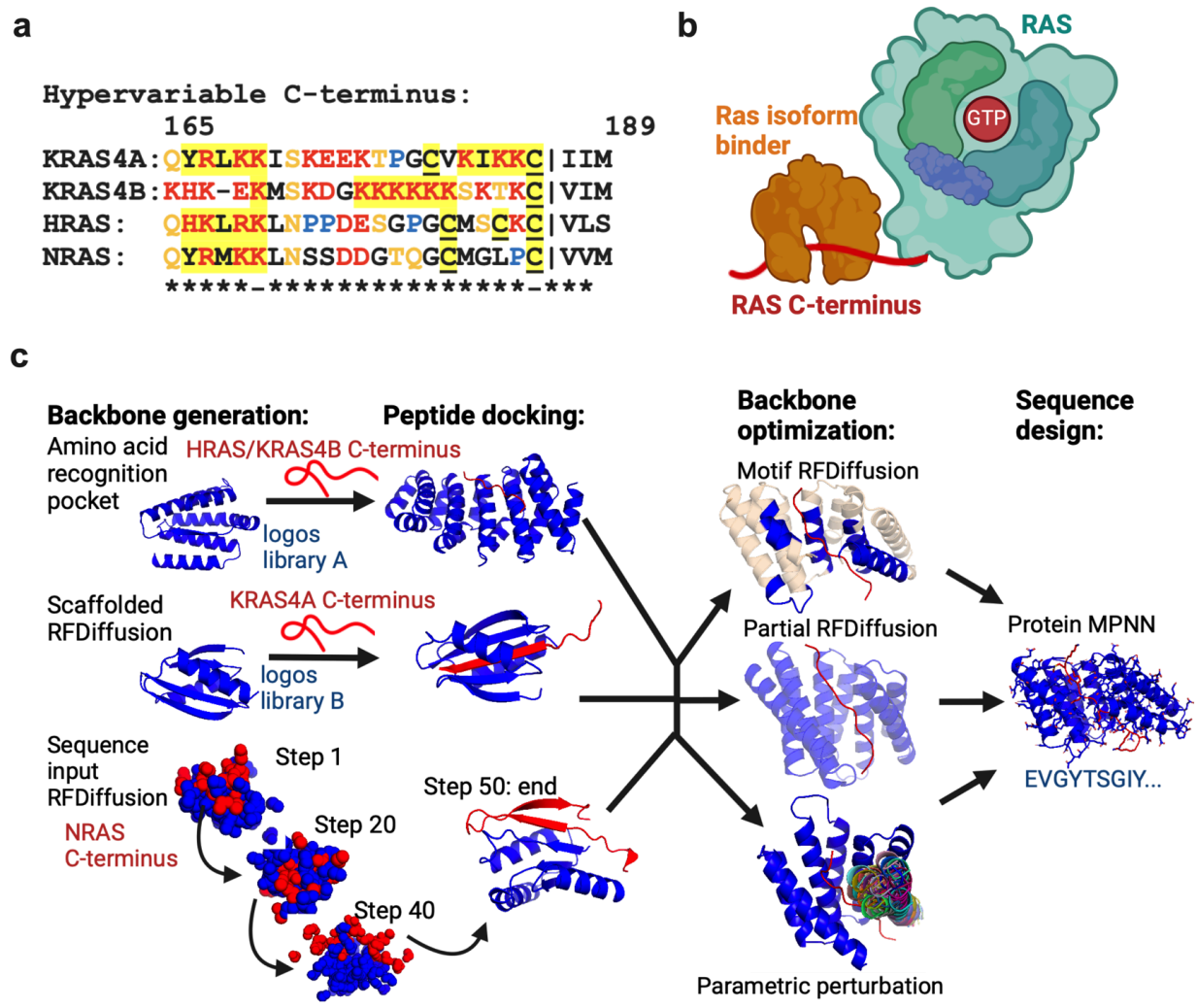


Figure 1: Computational design methods.

a) Sequence alignment of the C-termini of the major Ras isoforms. Residues in red are charged, in orange are polar, and in blue are prolines. Highlighted in yellow are residues involved in membrane interaction. Underlined residues undergo lipid modification.

b) Schematic representation of RIBs binding to the Ras C-terminus.

c) Computational design. Backbone generation either through an amino acid recognition pocket based approach, scaffolded RFDiffusion, or sequence input diffusion. These peptide-scaffold backbones are then optimized for either through motif RFDiffusion (subsets of the backbone kept fixed), partial RFDiffusion (slightly altering the overall scaffold backbone), or parametric

perturbation (rotating/moving a portion of the scaffold backbone in a specified manner). Next, ProteinMPNN was used to optimize the amino acid sequence for a given scaffold backbone. Scaffold-target complexes are evaluated using AlphaFold2 and Rosetta metrics to determine what to experimentally characterize.

Initially, we attempted to design binders to all the Ras isoforms using all methods. We found that both scaffolded and sequence input RFDiffusion methods primarily produced binder backbones that induced the target to conform to regular secondary structures. This worked well with the KRAS4A and NRAS C-termini which were induced to form β-strands that formed an extended β-sheet with the binder, but not for the KRAS4B and HRAS C-termini which are less compatible with regular secondary structure (KRAS4B has a highly charged 6x lysine region and HRAS contains several unevenly spaced proline residues). The amino acid recognition pocket based approach was more successful in generating good scoring designs for all the Ras isoforms as it does not require any secondary structure propensity in the target. We selected for experimental characterization 8,317 (5,254 from sequence input and 3,063 from scaffolded) and 3,078 (343 from sequence input and 2,735 from scaffolded) designs made using RFdiffusion for KRAS4A and NRAS, and 2,556 and 1,264 designs made using the amino acid recognition pocket based approach for KRAS4B and HRAS.

5.3 - In vitro testing of *de novo* designed RIBs

We obtained synthetic DNA encoding the selected RIB designs and identified binders using yeast display[15,16] (see Methods for details, **Figure S2a**-**b**). For KRAS4B and NRAS, the initial RIB library showed specific binding to the C-terminus but did not bind to full length Ras, perhaps due to steric interference with the full length proteins, which were not considered in our original

design calculations. We carried out a second round of binder scaffold optimization via motif and partial RFDiffusion requiring compatibility with both the C-terminus and full length protein (**Figure S2c**), which resulted in smaller scaffolds clashing less with the target. For example, the helix of the KRAS4B C-terminus binder is predicted to clash with a portion of the structured domain of KRAS4B and hence portions of that helix were re-designed to accommodate binding for full length KRAS4B (**Figure S2d)**. Ultimately, yeast display selection yielded ~250 binders for KRAS4A, ~90 binders and NRAS, ~50 binders for HRAS, and ~20 binders for KRAS4B. For KRAS4A, binders were obtained from scaffolded RFDiffusion where a portion of the KRAS4A C-terminus adopts a β-strand conformation which forms extended β-sheet hydrogen bonds with the binder's β-sheet (**Figure 2a and S3a**). In contrast, NRAS binders came from sequence input RFDiffusion and induced NRAS to adopt a conserved conformation where it formed multiple β-strands to create extended β-sheets with the binder (**Figure 2a and S3a**). The successful KRAS4B and HRAS binders came from the amino acid recognition pocket approach and induced KRAS4B or HRAS to adopt a non-regular secondary structure with an extended interface (**Figure 2a and S3a**).

We next expressed and purified the top 10 RIB hits from yeast display per target for *in vitro* characterization. Corroborating the yeast display results, RIBs specifically bound to their cognate full length target (**Figure 2b**) with affinities ranging from 35-1300nM (**Figure S3b-c, Table S1**) measured using bilayer interferometry. The RIBs bound to their intended target in a nucleotide-independent manner, which is expected as the C-terminus of Ras is distal from the nucleotide binding site[1] in the structured portion (**Figure S3d**). All 10 RIBs (numbered based on BLI-derived kd ranking) for each target that were tested experimentally bound both *in vitro* and in cell experiments (**Table S1**), and the RIBs with the most consistent and specific binding throughout the experiments are shown in **Figure 2** and are used for the rest of the paper (**Figures 3-4 and S5-7**). Of note, the binders with tight affinity (kd<100nM, **Table S1**) were not

necessarily the most specific when applied in cell experiments. In the design model of the scaffolded RFDiffusion generated KRAS4A_RIB_7:KRAS4A complex, the C-terminal section of the KRAS4A C-terminus forms a β-strand that pairs with β-strands deep within the binder and is packed by surrounding helices. In the design model of the amino acid recognition pocket-derived KRAS4B_RIB_2:KRAS4B and HRAS_RIB_9:HRAS complex, a portion of the KRAS4B or HRAS C-terminus forms an extended unstructured interface with the binder. In the design model of the sequence input RFDiffusion-generated NRAS_RIB_6:NRAS complex, the NRAS C-terminus forms two β-strands (innermost β-strand is NRAS C-terminus) which is incorporated into an extended β-sheet with the binder and is packed by surrounding helices.

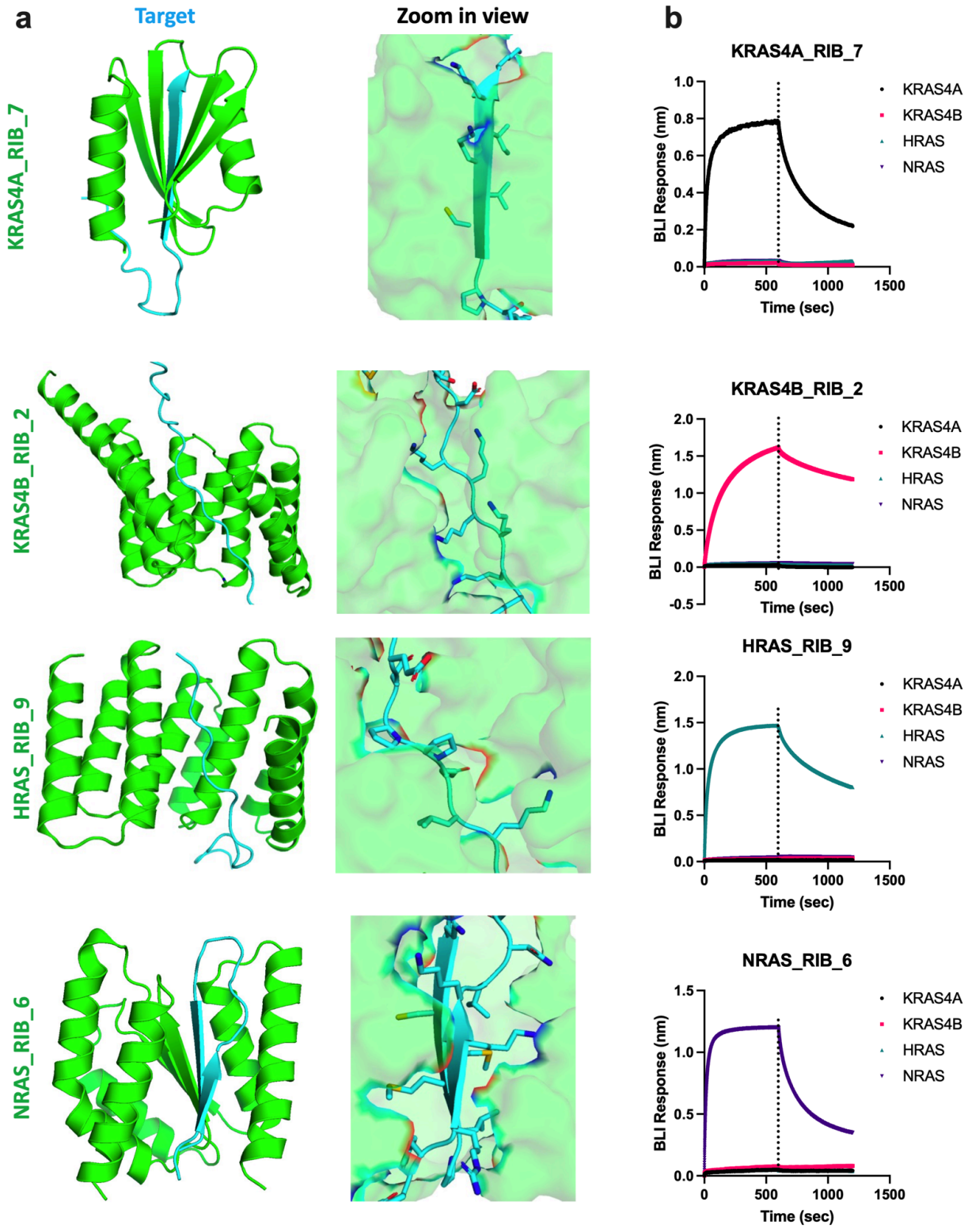


Figure 2: RIBs specifically bind to their target *in vitro*.

a) Design models of the RIBs (green) in complex with their Ras C-termini targets (cyan).

b) Representative biolayer interferometry (BLI) results of RIBs (1μM) interacting with the different Ras isoforms full length protein. In each case, binding is specific for the intended target. Each experiment was repeated 3 times.

5.4 - RIBs identify RAS isoforms in mammalian cells

To test the ability of RIBs to selectively bind Ras isoforms in mammalian cells, we used mouse embryonic fibroblast cell lines (MEFs) in which only one of the major wildtype (WT) Ras[17,18] isoforms is expressed at a time (e.g. KRAS4A Rasless MEFs express only WT KRAS4A). To test for RIB specificity, we carried out the analogue of an immunoblot in which Rasless MEF cell lysates separated on SDS gels were probed with biotinylated RIBs (see Methods for details) followed by labeling by AlexFluor 488-conjugated streptavidin. For each RIB, little binding was observed for the non-cognate cell lines, and a single band with the molecular weight of Ras was observed in the cognate cell line (**Figure 3a**). To determine whether RIBs can detect endogenous levels of the target, we carried out siRNA knockdown experiments (**Figure 3b**) in unmodified WT MEFs which contain all the major Ras isoforms. RIB-mediated band signal was lost only when its target was knocked down (e.g. only NRAS siRNA eliminates NRAS RIB-mediated band). To test whether RIBs can detect endogenous levels of Ras isoforms, we probed cell lines with different Ras genetic backgrounds and observed only one major band was seen around 20 kDa, the expected molecular weight of Ras (21 kDa) (**Figure S4**). In contrast, the most specific known Ras isoform antibodies[5] displayed many non-specific bands across the same cell lines (**Figure S4**), highlighting the specificity of RIBs compared to other reagents.

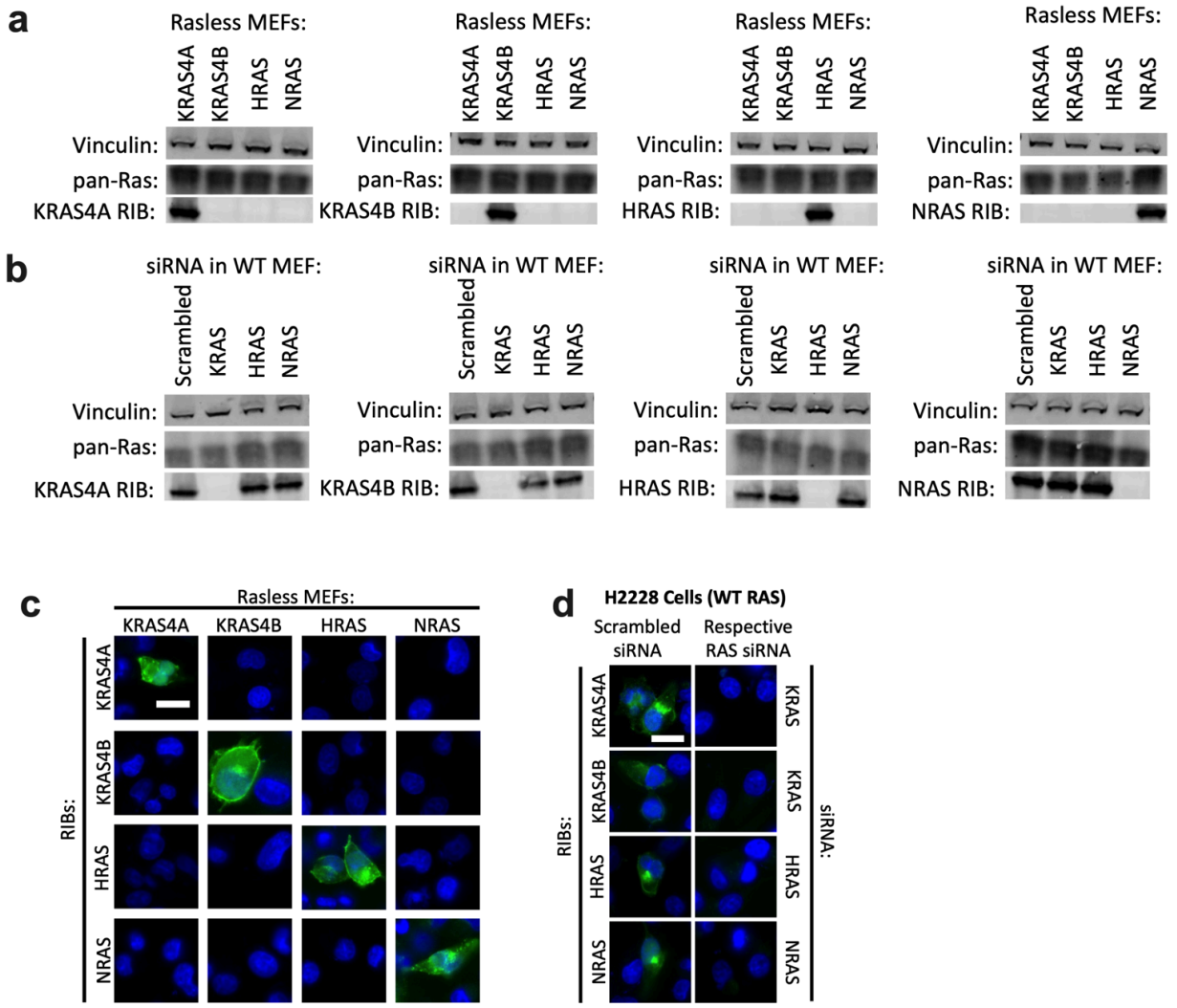


Figure 3: RIBs specifically bind to their targets in cells.

a) Lysates from Rasless MEFs, which express only one WT Ras isoform at a time, were run on SDS-PAGE gels and the Ras isoforms were probed with 10μM biotinylated RIBs or antibodies. Shown are representative blots from 3 independent experiments.

b) Wildtype (WT) MEF cells were transfected with either Scrambled or Ras isoform specific siRNA. Two days later, cells were lysed, ran on SDS-PAGE gels, and probed with biotinylated RIBs or other antibodies. Shown are representative blots from 3 independent experiments.

c) Rasless MEFs were probed via biotinylated RIBs (green) and DAPI (blue). Shown are representative epifluorescence images from 3 independent experiments. Scale bar=10μm.

d) Wildtype (WT) H2228 cells were transfected with either Scrambled or Ras isoform specific siRNA. Two days later, cells were fixed, permeabilized, and probed with biotinylated RIBs (green) and DAPI (blue). Shown are representative epifluorescence images from 3 independent experiments. Scale bar=10μm.

5.5 - Imaging Ras isoforms in cells

We investigated whether the RIBs are sufficiently specific and sensitive enough for fluorescence imaging assays. We applied biotinylated RIBs to fixed and permeabilized Rasless MEFs and used AlexFluor 488-conjugated streptavidin for imaging detection. The RIBs only stain their respective Rasless MEFs (e.g. KRAS4A RIB fluorescence signal is only seen in KRAS4A Rasless MEFs) (**Figure 3c**), with no significant signal in the other Rasless MEFs (**Figure S5a**). As found in other studies[5], only a subset of cells showed significant staining signal possibly due to heterogeneity in Ras isoform expression. The specific fluorescence labeling with the RIBs is greater than that of the previously described isoform-specific antibodies[5] which do not show significant immunostaining signal in the corresponding Rasless MEF cell line (**Figure S5b**). Only the NRAS antibody showed significant immunostaining signal, but knockdown of NRAS via siRNA did not diminish immunostaining signal, suggesting lack of specificity (**Figure S5b**).

The fluorescence staining results suggested differences in localization with KRAS4B having the most plasma membrane localization and NRAS the most endomembrane signal, consistent with previous work[2,3]. However, localization using the RIBs could be biased if RIB binding is reduced by Ras C-terminal palmitoylation since we did not include covalent modification by palmitoylation in our initial design computations and non-palmitoylated Ras is thought to be primarily in the golgi and ER while palmitoylated Ras is in the plasma membrane[3]. To test for

this potential bias, we probed Rasless MEFs both with our biotinylated RIBs and a pan-Ras antibody, which binds to the structured domain of Ras thus should not interfere with RIB binding and should only label the expressed Ras isoform in these Rasless MEFs (**Figure S6a**). The signal from KRAS4A, KRAS4B, and HRAS RIBs mostly co-localized with the signal from pan-Ras immunostaining, suggesting that the RIBs do not have a significant bias in subcellular binding location. The NRAS RIB led to more endomembrane signal and less plasma membrane signal than pan-Ras immunostaining. The predicted complex structures of NRAS RIB with NRAS C-terminus shows that the NRAS palmitoylation site C181[3,21,23] is buried within the interface, thus we hypothesized that the lower signal colocalization could be due to the NRAS RIBs binding primarily to the non-palmitoylated version of NRAS, which should be more endomembrane localized. In BRAFV600E-expressing Rasless MEFs, cells which do not express any of the major Ras isoforms, either ectopically expressing NRAS WT or palmitoylation-resistant C181S, NRAS RIBs pulled down NRAS C181S[3,21,23] to a greater extent than NRAS WT (**Figure S6b**), suggesting indeed that NRAS RIBs preferentially bind to the non-palmitoylated forms of NRAS.

To assess whether our RIBs can label in intact cells endogenous levels of Ras isoforms, H2228 cells which contain all major Ras isoforms were transfected with Ras isoform selective siRNAs. While each of the RIBs had detectable staining signal when scrambled siRNA was transfected, knockdown of the targeted Ras isoform in H2228 cells eliminated this signal (**Figure 3d**), demonstrating that our RIBs can measure endogenous Ras isoforms. This has not been achievable with previously described isoform-specific antibodies[5].

5.6 - Overexpression of RIBs disrupts Ras localization and signaling

We next explored the effects of ectopically expressing the RIBs inside cells. As the RIBs bind to the Ras C-terminus which is important for membrane localization, these RIBs could disrupt their membrane localization. Rasless MEFs expressing their respective RIBs shifts the Ras isoform from being in the membrane to the cytosolic fraction (**Figure 4a**). Furthermore, Rasless MEFs expressing GFP tagged RIB (RIB-GFP) and immunostained with pan-Ras antibodies, which report on the localization of the only Ras isoform expressed, showed a significant shift in Ras localization from membranes to the cytosol for all isoforms (**Figure S7a**). Quantitative membrane co-localization analysis confirmed that membrane localization of Ras was statistically significantly decreased when RIBs were expressed (**Figure S7b**). These results suggest that RIB overexpression disrupts Ras membrane localization.

The canonical view is that Ras requires a membrane to be in its active GTP loaded state[1,19,20], but a membrane-less cytosolic pool of Ras[22,24] could also be active. As RIB expression enhances the cytosolic localization of Ras, we explored whether this perturbation in Ras localization affects signaling. We used a previously constructed Ras activity biosensor (Ras-LOCKR-S[24]) to measure Ras-GTP loading in live cells with single cell resolution. Rasless MEFs expressing either mCherry tagged RIB or empty vector and co-expressing untargeted Ras-LOCKR-S to measure whole cell Ras activity were imaged (**Figure 4b**). For all the Ras isoforms tested, the Ras-LOCKR-S FRET ratios (yellow over cyan (Y/C)), which reports on Ras-GTP loading, were lower in cells overexpressing RIB (“0 min EGF” in **Figure 4b**). However, upstream Ras activation by 100ng/mL epidermal growth factor (EGF) still induced some transient Ras activation (“3 min EGF” in **Figure 4b**) when RIBs were expressed, and this is true

for all Ras isoforms tested. While KRAS4A and KRAS4B Rasless MEFs displayed transient activation of Ras with Ras signaling going to or even below baseline Ras activity, NRAS and HRAS Rasless MEFs showed transient Ras activation where Ras activity still lingers 10 minutes after EGF stimulation[24,25] (**Figure 4b**). These differences in Ras activity dynamics may be due to the differential localization of these Ras isoforms. A negative control (NC) version of Ras-LOCKR-S[24] that is a point mutant away from Ras-LOCKR-S but renders the biosensor insensitive to Ras-GTP displayed no difference in FRET ratios between RIB versus empty vector expression (**Figure S7a**).

Looking downstream of Ras activity, we probed for Erk activity changes during RIB expression by probing for phospho-Erk (pErk) levels. Rasless MEFs expressing their respective RIB displayed decreased pErk/Vinculin levels compared to WT MEFs transfected with empty vector (**Figure 4c**). However, EGF stimulation in these same conditions led to increases in pErk/Vinculin levels that were indistinguishable from WT MEFs transfected with empty vector (**Figure S7d**), corroborating our Ras-LOCKR-S data showing that Ras activity can still be stimulated when RIBs are expressed. This Erk activation observed when expressing RIBs is dependent on Ras farnesylation as dual farnesyl transferase and geranylgeranyl transferase-1 inhibition greatly diminishes pErk levels (**Figure S7e**), aligning with previous reports indicating Ras activation and downstream signaling requires Ras farnesylation (the last 3 or 4 C-terminal residues which get farnesylated and cleaved were not included in RIB design, thus we expect that RIB binding to Ras isoforms is not influenced by the farnesylation). Probing with a pan-Ras antibody showed no significant differences in Ras levels (**Figure 4c and S7d**), suggesting that these decreases in basal pErk/Vinculin levels is not due to Ras protein expression differences. To verify that RIB overexpression only affects their respective Ras isoform, RIBs were overexpressed in all the Rasless MEFs and probed for Ras activity levels via Ras-LOCKR-S or Erk activation via pErk/Erk levels (**Figure S7f**). Only RIB expression in their respective Rasless

MEF decreased Ras activity and Erk activation, demonstrating the specificity of the designs. Overall, these results corroborate the dogma that Ras-GTP loading is enhanced when in proximity to membranes, but also suggest that cytosolic Ras can be activated and lead to downstream signaling.

Membrane interaction of Ras is important for its activation and downstream signaling, which ultimately alters cell growth and proliferation. As RIBs can disrupt Ras localization and signaling, we wondered if RIB expression can alter cell proliferation. We focused on NRAS as our NRAS RIBs have a clear preference for non-palmitoylated NRAS (**Figure S6B**) and NRAS depalmitoylation has emerged as a promising therapeutic strategy for melanoma, hematologic cancers, and other *NRAS*-mutant cancers. We expressed NRAS RIB in engineered isogenic MOLM-13 cell lines that are dependent on doxycycline-induced expression of mutant NRAS or KRAS4B. We found that NRAS RIB selectively inhibited the growth of mutant NRAS expressing MOLM-13 cell lines (**Figure 4d**) but not KRAS4B expressing MOLM-13 cell lines nor MOLM-13 cells not given doxycycline (**Figure 4d and S8**). These results indicate that perturbing the localization of Ras can influence Ras activation, downstream signaling, and cell growth.

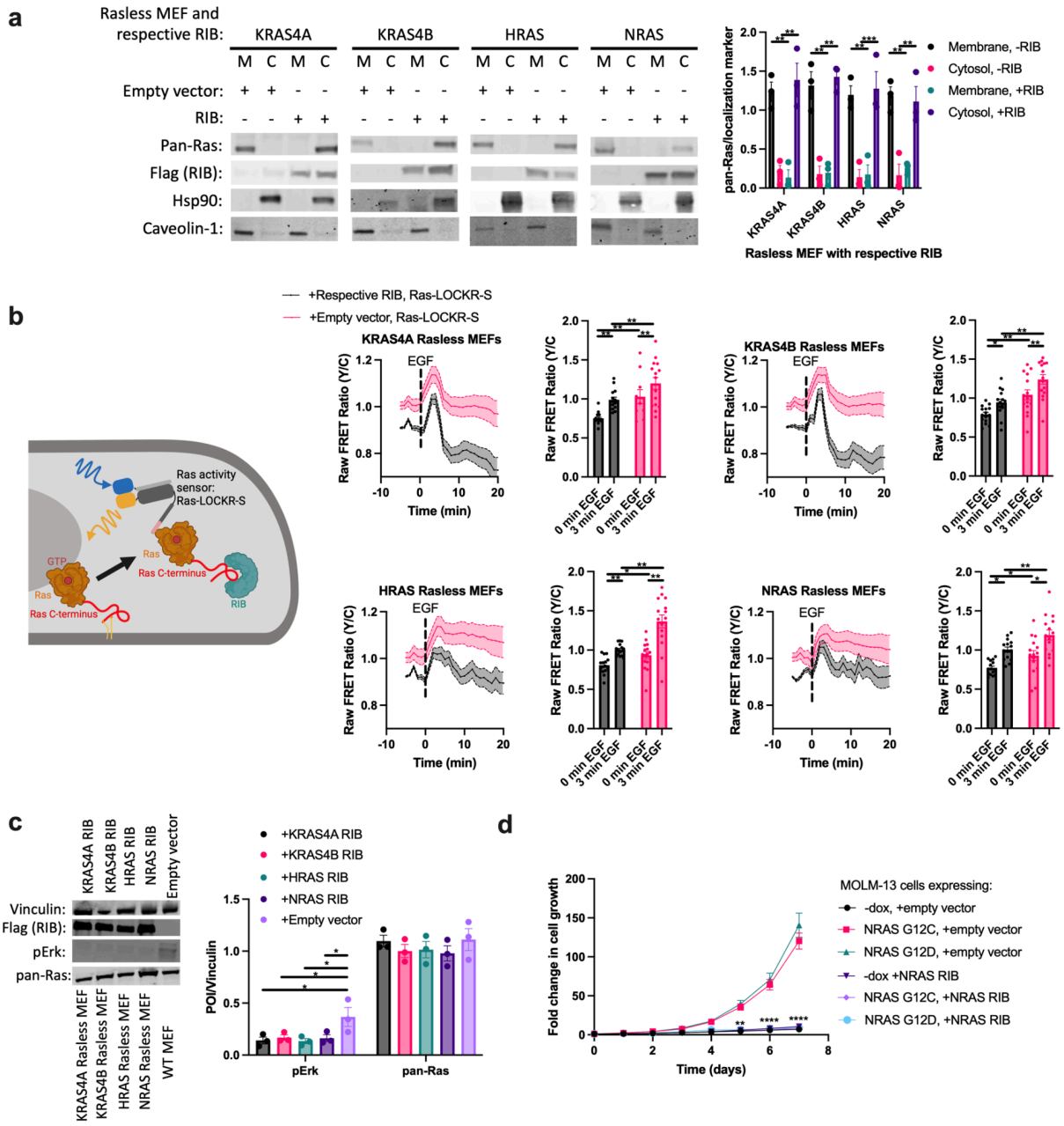


Figure 4: RIB expression alters Ras localization and signaling.

a) Rasless MEFs were transfected with either empty vector (-RIB) or their respective RIB (+RIB) for 1 day, lysed, and underwent membrane fractionation and immunoblotting analysis. Left: representative immunoblot. Right: densitometry quantification of immunoblots comparing pan-Ras signal with the signal from the localization marker (Hsp90 for cytosol, Caveolin-1 for membrane) (n=3 experiments). Statistics: two-way ordinary ANOVA.

b) Rasless MEFs were transfected with Ras-LOCKR-S and either RIB fused to mCherry or empty vector. Then cells were imaged and stimulated with 100ng/mL Epidermal Growth Factor (EGF). Top left: Schematic of experiment. Left: Time-course imaging of raw FRET ratios (yellow over cyan (Y/C)) representative of three independent experiments. Right: Bar graph quantification of the Ras-LOCKR-S raw FRET ratios across all experiments (n=15 cells per experiment). Statistics: two-way ordinary ANOVA.

c) Rasless MEFs were transfected with their respective RIB for 1 day. These cells were compared to WT MEFs. These cells then underwent immunoblotting analysis. Right: densitometry quantification of immunoblots (n=3 experiments). Statistics: two-way ordinary ANOVA.

d) MOLM-13 cell lines expressing either NRAS G12C or G12D were treated with or without doxycycline to induce mutant NRAS expression. Cells were also transfected with NRAS RIB expressing plasmid or empty vector and co-treated with AC220 (quizartinib, FLT3 inhibitor) (see Methods for details). Cells were then counted over a 7 day period (n=3 experiments). Statistics: one-way ordinary ANOVA.

5.7 - Discussion

The Ras isoforms exhibit distinct subcellular distributions and mutational preferences in cancer yet tools to differentiate these Ras isoforms have been lacking due to their high sequence similarity except at their C-terminus. The *de novo* designed RIBs bind to these highly charged and disordered C-termini with higher specificity than any reagents described to date[5] both *in vitro* and in cells. Expression of the RIBs enhanced the cytosolic localization of these Ras isoforms which led to decreased Ras activity, corroborating the notion that membrane association of Ras is important for its function. Given the central importance of Ras in signaling

and disease, our Ras isoform selective binders should be useful for a wide range of applications including isoform specific inhibitors (**Figure 4b-c**), affinity handles for targeted degradation, diagnostic markers for cancer patient samples, and Ras isoform specific biosensors.

Acknowledgements:

We acknowledge funding from HHMI (J.Z.Z. and D.B.), Helen Hay Whitney Foundation (J.Z.Z.), NCI K99-CA293001 (J.Z.Z.), the Croucher Fellowship (H.J.), the Audacious Project at the Institute for Protein Design (J.Z.Z, D.B.), NIH K08-CA256489 (B.J.H.), and NIH R01 CA193994 (K.S., B.J.H.). We thank D.J. Maly, F. McCormick, and D. Esposito for fruitful discussion of the Ras results, C.M. Dobbins and S. Cheng for help with mammalian cells and cell culture, and the RAS Initiative at the Frederick National Lab for providing the MEF cells.

Author contributions:

J.Z.Z. conceived of the project. K.W. supervised design strategies for making Ras isoform specific binders in this work, designed and assembled templates for KRAS4B, designed the scaffolds used in the logos pipeline. J.Z.Z. computationally designed these proteins with help from K.W., H.J. and C.L.. J.Z.Z. and D.B. supervised, designed, and interpreted the experiments. J.Z.Z. performed all experiments. X.L. made the Ras isoform peptides. J.Z.Z., and D.B. wrote the original draft. All authors reviewed and commented on the manuscript.

Competing interest:

The authors claim no competing interests.

References:

1. Nissley, D. V, and McCormick, F. (2022). RAS at 40: update from the RAS initiative. Cancer Discov *12*, 895–898.

2. Zhang, J.Z., Ong, S.-E., Baker, D., and Maly, D.J. (2024). Single-cell sensor analyses reveal signaling programs enabling Ras-G12C drug resistance. Nat Chem Biol. https://doi.org/10.1038/s41589-024-01684-4.

3. Choy, E., Chiu, V.K., Silletti, J., Feoktistov, M., Morimoto, T., Michaelson, D., Ivanov, I.E., and Philips, M.R. (1999). Endomembrane trafficking of ras: the CAAX motif targets proteins to the ER and Golgi. Cell *98*, 69–80.

4. Amendola, C.R., Mahaffey, J.P., Parker, S.J., Ahearn, I.M., Chen, W.-C., Zhou, M., Court, H., Shi, J., Mendoza, S.L., Morten, M.J., et al. (2019). KRAS4A directly regulates hexokinase 1. Nature *576*, 482–486. https://doi.org/10.1038/s41586-019-1832-9.

5. Waters, A.M., Ozkan-Dagliyan, I., Vaseva, A. V, Fer, N., Strathern, L.A., Hobbs, G.A., Tessier-Cloutier, B., Gillette, W.K., Bagni, R., Whiteley, G.R., et al. (2017). Evaluation of the selectivity and sensitivity of isoform- and mutation-specific RAS antibodies. Sci Signal *10*, eaao3332. https://doi.org/10.1126/scisignal.aao3332.

6. Watson, J.L., Juergens, D., Bennett, N.R., Trippe, B.L., Yim, J., Eisenach, H.E., Ahern, W., Borst, A.J., Ragotte, R.J., Milles, L.F., et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature *620*, 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.

7. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I.M., Courbet, A., de Haas, R.J., Bethel, N., et al. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science (1979) *378*, 49–56. https://doi.org/10.1126/science.add2187.

8. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science (1979) *373*, 871–876. https://doi.org/10.1126/science.abj8754.

9. Wu, K., Bai, H., Chang, Y.-T., Redler, R., McNally, K.E., Sheffler, W., Brunette, T.J., Hicks, D.R., Morgan, T.E., Stevens, T.J., et al. (2023). De novo design of modular peptide-binding proteins by superhelical matching. Nature *616*, 581–589. https://doi.org/10.1038/s41586-023-05909-9.

10. Liu, C., Wu, K., Choi, H., Han, H., Zhang, X., Watson, J.L., Shijo, S., Bera, A.K., Kang, A., Brackenbrough, E., et al. (2024). Diffusing protein binders to intrinsically disordered proteins. bioRxiv, 2024.07.16.603789. https://doi.org/10.1101/2024.07.16.603789.

11. Wu, K., Jiang, H., Hicks, D.R., Liu, C., Muratspahić, E., Ramelot, T.A., Liu, Y., McNally, K., Gaur, A., Coventry, B., et al. (2024). Sequence-specific targeting of intrinsically disordered protein regions. bioRxiv, 2024.07.15.603480. https://doi.org/10.1101/2024.07.15.603480.

12. Vázquez Torres, S., Leung, P.J.Y., Venkatesh, P., Lutz, I.D., Hink, F., Huynh, H.-H., Becker, J., Yeh, A.H.-W., Juergens, D., Bennett, N.R., et al. (2024). De novo design of high-affinity binders of bioactive helical peptides. Nature *626*, 435–442. https://doi.org/10.1038/s41586-023-06953-1.

13. Jiang, H., Jude, K.M., Wu, K., Fallas, J., Ueda, G., Brunette, T.J., Hicks, D.R., Pyles, H., Yang, A., Carter, L., et al. (2024). De novo design of buttressed loops for sculpting protein functions. Nat Chem Biol *20*, 974–980. https://doi.org/10.1038/s41589-024-01632-2.

14. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

15. Cao, L., Goreshnik, I., Coventry, B., Case, J.B., Miller, L., Kozodoy, L., Chen, R.E., Carter, L., Walls, A.C., Park, Y.-J., et al. (2020). De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science (1979) *370*, 426 LP – 431. https://doi.org/10.1126/science.abd9909.

16. Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J.S., Jude, K.M., Marković, I., Kadam, R.U., Verschueren, K.H.G., et al. (2022). Design of protein-binding proteins from the target structure alone. Nature *605*, 551–560. https://doi.org/10.1038/s41586-022-04654-9.

17. Drosten, M., Dhawahir, A., Sum, E.Y.M., Urosevic, J., Lechuga, C.G., Esteban, L.M., Castellano, E., Guerra, C., Santos, E., and Barbacid, M. (2010). Genetic analysis of Ras signalling pathways in cell proliferation, migration and survival. EMBO J *29*, 1091–1104.

18. Burgan, W., Fer, N. (2024). Creation of an Isogenic H/N/KRAS-Less Mouse Embryonic Fibroblast Cell Line Panel Derived from a Size-Sorted Diploid Clone. Methods Mol Biol. *2797*, 323-336.

19. Lynch, S.J., Snitkin, H., Gumper, I., Philips, M.R., Sabatini, D., and Pellicer, A. (2015). The differential palmitoylation states of N‐Ras and H‐Ras determine their distinct Golgi subcompartment localizations. J Cell Physiol *230*, 610–619.

20. Augsten, M., Pusch, R., Biskup, C., Rennert, K., Wittig, U., Beyer, K., Blume, A., Wetzker, R., Friedrich, K., and Rubio, I. (2006). Live‐cell imaging of endogenous Ras‐GTP illustrates predominant Ras activation at the plasma membrane. EMBO Rep *7*, 46–51.

21. Zambetti, N.A., Firestone, A.J., Wong, J.C., Long, A., Inguva, A., Suciu, R.M., Cravatt, B.F., Haigis, K.M., and Shannon, K.M. (2017). Genetic Validation of the Palmitoylation/Depalmitoylation Cycle As a Drug Target in NRAS Mutant Hematologic Malignancies In Vivo. Blood *130*, 1221. https://doi.org/https://doi.org/10.1182/blood.V130.Suppl_1.1221.1221.

22. Tulpule, A., Guan, J., Neel, D.S., Allegakoen, H.R., Lin, Y.P., Brown, D., Chou, Y.-T., Heslin, A., Chatterjee, N., Perati, S., et al. (2021). Kinase-mediated RAS signaling via membraneless cytoplasmic protein granules. Cell *184*, 2649-2664.e18. https://doi.org/10.1016/j.cell.2021.03.031.

23. Remsberg, J.R., Suciu, R.M., Zambetti, N.A., Hanigan, T.W., Firestone, A.J., Inguva, A., Long, A., Ngo, N., Lum, K.M., and Henry, C.L. (2021). ABHD17 regulation of plasma membrane palmitoylation and N-Ras-dependent cancer growth. Nat Chem Biol *17*, 856–864.

24. Zhang, J.Z., Nguyen, W.H., Greenwood, N., Rose, J.C., Ong, S.-E., Maly, D.J., and Baker, D. (2024). Computationally designed sensors detect endogenous Ras activity and signaling effectors at subcellular resolution. Nat Biotechnol. https://doi.org/10.1038/s41587-023-02107-w.

25. Chiu, V.K., Bivona, T., Hach, A., Sajous, J.B., Silletti, J., Wiener, H., Johnson, R.L., Cox, A.D., and Philips, M.R. (2002). Ras signalling on the endoplasmic reticulum and the Golgi. Nat Cell Biol *4*, 343–350.

Supplementary materials

Amino acid recognition pocket based design and scaffolded RFDiffusion

As described in the previous work featuring the logos pipeline[9,11], the recognition pocket approach took advantage of the overall flexible nature of disordered regions as well as the base of amino acid recognition in a string of polymers. In short, this approach would thread the query target into a diverse and robust family of ~1,000 de novo protein-peptide complex templates, made up of arbitrary combos of 20 amino acid recognition pockets assembled into computationally designed universal binding modes. Initial docks were judged by how well fit each individual amino acid from the query target was, and how many of those fits were out of non-fits. Top docks were then identified by the pipeline, which were considered as the best design starting points. Sequence design, structural prediction, RFdiffusion refinement were then applied sequentially, until new customizable binding proteins were made and predicted well to the query target of interest.

Sequence Design

All accepted docks are sent to ProteinMPNN with a customized script that applies empirical weights to certain amino acids. Two iterations of ProteinMPNN and FastRelax are performed to adjust chain B backbones further and optimize the binder SAP score. Sequences with the lowest ProteinMPNN scores are filtered and sent to our customized AF2 package. This filtering process varies depending on the target's difficulty level. Criteria can be adjusted based on individual targets and design campaigns.

Predictions

After two cycles of sequence design and AF2 (prediction-guided sequence design), passing designs undergo AlphaFold-multimer for final predictions. For highly polar or proline-rich targets, AlphaFold-multimer proved more predictive than AlphaFold-initial guess. AlphaFold-monomer is then applied to assess designed monomer predictions and select the final designs for experimental testing.

Scaffolded RFDiffusion

A subset of ~200 β-sheet containing templates derived from the amino acid recognition pocket based design of template backbones (logos library B) were used as the starting point. These templates then went through the same pipeline as described above of peptide threading, sequence design, predictions, and then diffusion refinement.

KRAS4B binder design

Following the same computation methodology as the original logos pipeline, manual refinement was performed on designs against this target. Due to the highly charged nature of this target (fill in the AA), we reasoned that a set of customized assembly of positive charged pockets (which was unenriched in the original logos library) would boost design success rate. Therefore, all the

mono-, di- peptide pockets containing target of positively charged amino acids (i.e., K, R, H) were identified, same pocket assembly strategy as in previous work [cite] was applied, along with another set of logos templates toward targets containing >60% of positively charged amino acids. For these two sets, chain B (target) was replaced with the KRAS4B target. The same diffusion refinement (i.e., motif diffusion for 8,000 trajectories each template, one-sided partial diffusion for 1,000 trajectories each template) was performed followed by ProteinMPNN and Alphafold2.

Sequence input RFDiffusion based design

For each target, approximately 10,000-50,000 diffused designs were generated based solely on the sequence input of the target. The resulting backbone library was subjected to sequence design using ProteinMPNN, followed by AF2 initial guess. These designs were then filtered based on interface pAE and pLDDT. Additionally, AF2 monomer analysis was conducted using only the binder sequence without the peptide, allowing for filtering based on the binder's monomer pLDDT and RMSD relative to the binder design model. Subsequently, FastRelax was employed to obtain Rosetta metrics. The resulting binders were further filtered according to specific criteria, including contact molecular surface, ddG, SAP score, and the number of hydrogen bonds. These filtering criteria were meticulously chosen to narrow the selection down.

Backbone optimization with RFDiffusion and parametric perturbation

The process of pocket recognition and assembly played a crucial role in refining precise interactions with individual amino acids in a sequence-specific manner. This was particularly important when dealing with challenging polar or charged targets or targets that would not form regular secondary structures upon binding.

Motif RFDiffusion

As described in the previous work of *logos* pipeline[11], if the design did not meet the complex prediction criteria mentioned above, or if a larger number of tested designs was preferred, motif RFDiffusion was employed to optimize the fit between less-than-ideal interacting pockets and binder backbones. In this work, especially for the need of yeast display screening thousands of binders a time, we used motif RFdiffusion more dramatically than originally developed to meanwhile shrink the binder size. Here, we retained a significant portion of the original binding interactions and motifs mostly around the central part of the RIBs to preserve the advantages of our general platform. Meanwhile, we deleted the terminal region and let RFdiffusion regrow with a smaller length. As part of this strategy, we developed partial RFdiffusion with small steps and multi-motif-constrained RFdiffusion (motif RFdiffusion).

Partial RFDiffusion (One-Sided)

As in other cases, RFdiffusion was modified to allow the input structure to be noised only up to a user-specified time step, rather than completing the full noising schedule. Consequently, the denoising trajectory's starting point retained information about the input distribution, leading to denoised structures that remained structurally similar to the original input. In our work, the time step was set lower to maintain structural similarity, given that the high design resolution of hydrogen bonding interactions (especially with polar amino acids) was deemed critical. Specifically, 10-18 noising timesteps (10, 12, 15, 18) out of a total of 50 were selected for this paper's work. For each target parameter, 200 to 3,000 partially diffused designs were generated.

The new backbones were then subjected to ProteinMPNN sequence design and AF2 predictions, as previously described. These designs were filtered using the same criteria, with the observation that partial RFdiffusion tended to increase the overall number of passing designs. However, it was also noted that the original hydrogen bonding networks (particularly

bidentate hydrogen bonds to the target backbone) were sometimes disrupted without being detected by AF2. To address this, we implemented an additional filter, ‘buns’ (buried unsatisfied hydrogen bonds) <=1, to explicitly count unsatisfied heavy atoms. Depending on the target’s polarity, this step could filter out 50-80% of designs.

Parametric perturbation

Helices that were suboptimal in terms of interaction with the target could be isolated from the binder template and underwent defined x/y/z translations and rotations. Rosetta-based scoring of these translations and rotations filtered perturbations that were either non-productive (moved the helix too far away from the target) or clashed with the target. Then motif RFDiffusion was used to connect the perturbed helix to the rest of the binder template.

Peptide generation

All Fmoc-protected amino acids were obtained from P3 Biosystems; the coupling reagent HATU, DIPEA, and other chemicals unless otherwise stated were obtained from Sigma-Aldrich; DIC was obtained from Oakwood Chemicals; acetonitrile (ACN), methylene chloride (DCM), diethyl ether, and DMF were obtained from Fisher Scientific. Preloaded Wang resin and OxymaPure were obtained from CEM.

The linear peptide sequences were synthesized via solid-phase peptide synthesis (SPPS) on a LibertyBlue microwave synthesizer (CEM) at 0.1mmol scale on preloaded Wang resin. The complete linear peptide sequence was then coupled by hand to Fmoc-6-aminohexanoic acid (AHX) (3eq), HATU (3 eq), and DIPEA (5eq) for 3h, washed with DMF, then treated with 20% piperidine in DMF for 2x15m to remove Fmoc. The resin was then washed with DMF and swelled in 50/50 DMSO/DMF prior to coupling with biotin (3eq), HATU (eq), and DIPEA (5eq) for 3h, then washed with DMF and DCM prior to total deprotection and cleavage from resin using a

cleavage cocktail (92.5:2.5:2.5:2.5 TFA:triisopropylsilane:H_2O:2,2'-(ethylenedioxy)diethanethiol, v/v/v/v) for 3h. The peptide cleavage solution was concentrated *in vacuo* and precipitated in ice-cold diethyl ether, centrifuged to pellet the peptide crude (7000g, 4°C, 10m), and dried under N_2. The peptide crude was resuspended in minimal water and acetonitrile and purified by RP-HPLC on a semi-prep Agilent 1260 Infinity with a linear gradient from 10-45% A->B (A: water with 0.1% TFA, B: ACN with 0.1% TFA) on a Zorbax C18-300SB 5 µm 9.4x250mm column (Agilent), and analyzed via LC/MS-TOF on an Agilent G6230B. Pure peptide fractions were combined and lyophilized.

DNA library preparation

All protein sequences were padded to a uniform length by adding a (GGGS)n linker at the C terminal of the designs, to avoid the biased amplification of short DNA fragments during PCR reactions. The protein sequences were reversed translated and optimized using DNAworks2.0 with the S. cerevisiae codon frequency table. Homologous to the pETCON plasmid Oligo libraries encoding the designs were ordered from Twist Bioscience. Combinatorial libraries were ordered as IDT (Integrated DNA Technologies) ultramers with the final DNA diversity ranging from 1×10^6 to 1×10^7.

All libraries were amplified using Kapa HiFi Polymerase (Kapa Biosystems) with a qPCR machine (BioRAD CFX96). In detail, the libraries were firstly amplified in a 25 µL reaction, and PCR reaction was terminated when the reaction reached half the maximum yield to avoid over-amplification. The PCR product was loaded to a DNA agarose gel. The band with the expected size was cut out and DNA fragments were extracted using QIAquick kits (Qiagen, Inc.). Then, the DNA product was re-amplified as before to generate enough DNA for yeast transformation. The final PCR product was cleaned up with a QIAquick Clean up kit (Qiagen,

Inc.). For the yeast transformation, 2-3 μg of digested modified pETcon vector (pETcon3) and 6 μg of insert were transformed into EBY100 yeast strain using the protocol as described before.

DNA libraries for deep sequencing were prepared using the same PCR protocol, except the first step started from yeast plasmid prepared from 5×10^7 to 1×10^8 cells by Zymoprep (Zymo Research). Illumina adapters and 6-bp pool-specific barcodes were added in the second qPCR step. Gel extraction was used to get the final DNA product for sequencing. All libraries include the native library and different sorting pools were sequenced using Illumina NextSeq/MiSeq sequencing.

For mammalian cell expression, all plasmids use the pcDNA3.1 backbone and were produced by GenScript.

Yeast surface display

S. cerevisiae EBY100 strain cultures were grown in C-Trp-Ura media and induced in SGCAA media following the protocol as described before. Cells were washed with PBSF (PBS with 1% BSA) and labeled with biotinylated target using two labeling methods, with-avidity and without-avidity labeling. For the with-avidity method, the cells were incubated with biotinylated target, together with anti-c-Myc fluorescein isothiocyanate (FITC, Miltenyi Biotech) and streptavidin–phycoerythrin (SAPE, ThermoFisher). The concentration of SAPE in the with-avidity method was used at 1/4 concentration of the biotinylated target. The with-avidity method was used in the first few rounds of screening of the original design to fish out weak binder candidates. For the without-avidity method, the cells were firstly incubated with biotinylated target, washed, secondarily labeled with SAPE and FITC. For these designs, the first two to four rounds of sorts were applied with 1 μM concentration of the Ras C-terminus with biotin at the C-terminal end of the peptide. The remaining subsequent sorts were done with

varying concentrations (1 nM - 1 μM) of full length RAS. The final sorting pools of the combinatorial libraries were sequenced using Illumina NextSeq/MiSeq sequencing. All FACS data was analyzed in FlowJo.

Cell culture and transfection

HEK293T, HeLa, MEF, Rasless MEF (isogenic cell lines), MIA PaCa-2, H358, H441, Rh36, RD, H1299 cell lines were cultured in Dulbecco's modified Eagle medium (DMEM) containing 1 g L^{-1} glucose and supplemented with 10% (v/v) fetal bovine serum (FBS) and 1% (v/v) penicillin–streptomycin (Pen-Strep). H441, H2228, and MOLM-13 cells were grown in RPMI-1640 cell media containing 1 g L^{-1} glucose and supplemented with 10% (v/v) fetal bovine serum (FBS) and 1% (v/v) penicillin–streptomycin (Pen-Strep). For MOLM-13 cells, 2mM glutamine was also added to the media. All cells were grown in a humidified incubator at 5% CO_2 and at 37°C.

Before transfection, all cells were plated onto sterile poly-D-lysine coated plates or dishes and grown to 50%–70% confluence. All cells were transfected with Lipofectamine LTX. Subsequently, all cells were grown for an additional 1-2 days before experimental testing. All cells underwent serum starvation for 16 hours before downstream assay analysis. See **Table S2** for details of reagents.

General procedures for bacterial protein production and purification

The *E. coli* Lemo21(DE3) strain was transformed with a pET29b$^+$ plasmid encoding the synthesized gene of interest. Cells were grown for 24 hours in liquid broth medium supplemented with kanamycin. Cells were inoculated at a 1:50 ml ratio in the Studier TBM-5052 autoinduction medium supplemented with kanamycin, grown at 37°C for 2–4 hours and then grown at 18°C for an additional 18 hours. Cells were collected by centrifugation at 4,000 *g* at 4

°C for 15 min and resuspended in 30 ml lysis buffer (20 mM Tris-HCl, pH 8.0, 300 mM NaCl, 30 mM imidazole, 1 mM PMSF and 0.02 mg ml^{-1} DNase). Cell resuspensions were lysed by sonication for 2.5 min (5 s cycles). Lysates were clarified by centrifugation at 24,000 *g* at 4°C for 20 min and passed through 2-ml Ni-NTA nickel resin pre-equilibrated with wash buffer (20 mM Tris-HCl, pH 8.0, 300 mM NaCl and 30 mM imidazole). The resin was washed twice with 10 column volumes (Cversus) of wash buffer, and then eluted with 3 Cversus elution buffer (20 mM Tris-HCl, pH 8.0, 300 mM NaCl and 300 mM imidazole). The eluted proteins were concentrated using Ultra-15 Centrifugal Filter Units and further purified by using a Superdex 75 Increase 10/300 GL size exclusion column in TBS (25 mM Tris-HCl, pH 8.0, and 150 mM NaCl). Fractions containing monomeric protein were pooled, concentrated and snap-frozen in liquid nitrogen and stored at −80°C. See **Table S2** for details of reagents.

To biotinylate RIBs, an AviTag was added at their C-terminus. Biotinylation of purified RIBs fused with AviTags was performed using the BirA bulk kit according to manufacturer's protocol (Avidity LLC). Briefly, biotinylation reactions (pH 8.0; 1:1 ratio) were performed for 1 hour at 4°C on an orbital shaker and then excess biotinylation reagent was removed using Superdex 200 Increase 10/300 GL (depending on kDa of protein) size exclusion column in TBS (25 mM Tris-HCl, pH 8.0, and 150 mM NaCl).

Biolayer interferometry

Protein–protein interactions were measured by using an Octet RED96 System (ForteBio) using streptavidin-coated biosensors (ForteBio). Each well contained 200 µL of solution, and the assay buffer was HBS-EP+ buffer (GE Healthcare Life Sciences, 10 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.05% (v/v) surfactant P20) plus 0.5% non-fat dry milk blotting grade blocker (BioRad). The biosensor tips were loaded with analyte peptide or protein at 20 µg mL^{-1} for 300 s (threshold of 0.8 nm response), incubated in HBS-EP+ buffer for 60 s to acquire the

baseline measurement, dipped into the solution containing cage and/or key for 1800 s (association step) and dipped into the HBS-EP+ buffer for 1800 s (dissociation steps). The binding data were analyzed with the ForteBio Data Analysis Software version 9.0.0.10.

Immunostaining

All cell lines were seeded onto 24-well glass-bottom plates. After transfection and drug addition, cells were fixed with 4% PFA in 2x PHEM buffer (60 mM PIPES, 50 mM HEPES, 20 mM EGTA, 4 mM $MgCl_2$, 0.25 M sucrose, pH 7.3) for 10 min, permeabilized with 100% methanol for 10 min, washed with PBS 3x, blocked in 1% BSA in PBS for 30 min, incubated with biotinylated RIBs and antibodies for 2 hours at room temperature, washed with PBS 3x, incubated with DAPI and fluorescently labeled reagents (streptavidin and BioTracker) for 1 hour at room temperature and aluminum foil cover. Cells were then washed with PBS 3x and mounted for epifluorescence imaging. All images were analyzed in ImageJ. See **Table S2** for details of reagents.

Immunoblotting and Immunoprecipitation

Cells expressing indicated constructs and incubated with indicated drugs were plated, transfected, and labeled as described in figure legends. For cells that required membrane fractionation were treated with the Cell Fractionation Kit from Cell Signaling Technologies according to the manufacturer's protocol. Cells were then transferred to ice and washed 2x with ice cold DPBS. Cells were then detached from the well by addition of 1x RIPA lysis buffer (50 mM Tris pH 8, 150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, 1% Triton X-100, 1x protease inhibitor cocktail, 1 mM PMSF, 1mM Na_3VO_4, 1% NP-40) and either scraping of cells or rotation on shaker for 30 min at 4ºC. Cells were then collected and vortexed for at least 5 s every 10 min for 20 min at 4ºC. Cells were then collected and clarified by centrifugation at 13,000g for 10 minutes at 4°C. The supernatant was collected and underwent Pierce BCA assay to quantify total protein amounts.

For immunoblotting, whole cell lysate protein amounts were normalized across samples in the same gel, mixed with 4x loading buffer prior to loading, incubated at 95°C for 5 min and then 4°C for 5 min, and separated on Any kDa SDS-PAGE gels. Proteins separated on SDS-page gels were transferred to nitrocellulose membranes via the TransBlot system (BioRad). The blots were then blocked in 5% milk (w/v) in TBST (Tris-buffered saline, 0.1% Tween 20) for 1 hour at room temperature. Blots were washed with TBST 3x then incubated with indicated primary antibodies or biotinylated RIBs in 1% BSA (w/v) in TBST overnight at 4°C. Blots were then washed with TBST 3x and incubated with LICOR dye-conjugated secondary antibodies (LICOR 680/800 or streptavidin-LICOR 800) and fluorescently labeled streptavidin in 1% BSA (w/v) in TBST for 1 hour at room temperature. The blots were washed with TBST 3x and imaged on an Odyssey IR imager (LICOR). Quantitation of Western blots was performed using ImageJ on raw images. See **Table S2** for details of reagents including antibody dilutions.

For immunoprecipitation, streptavidin beads were loaded by three lysis buffer washes before the addition of 1 mg ml^{-1} of biotinylated RIBs at 4 °C on an orbital shaker for 3 h. Beads were then washed two times in lysis buffer. Whole-cell lysate protein amounts were normalized across samples and protein samples were added to beads (at least 100 µg per sample) either at room temperature for 1 h for streptavidin beads or at 4 °C on an orbital shaker overnight. Beads were then washed two times in lysis buffer and one time in TBS and then mixed with 4× loading buffer. The remaining portion of the protocol was the same as for immunoblotting.

Epifluorescence imaging

Cells were washed twice with FluoroBrite DMEM imaging media and subsequently imaged in the same media in the dark at room temperature. Epifluorescence imaging was performed on a Yokogawa CSU-X1 spinning dish confocal microscope with either a Lumencor Celesta light

engine with 7 laser lines (408, 445, 473, 518, 545, 635, 750 nm) or a Nikon LUN-F XL laser launch with 4 solid state lasers (405, 488, 561, 640 nm), 40x/0.95 NA objective or 60x/1.4 NA oil immersion objective and a Hamamatsu ORCA-Fusion scientific CMOS camera, both controlled by NIS Elements 5.30 software (Nikon). The following laser and filter combinations (center/bandwidth in nm) were used: GFP: EX473 EM525/36, RFP: EX545 EM605/52, BFP: EX445 EM525/36. Exposure times were 500ms for all channels, with no EM gain set and no ND filter added. All epifluorescence experiments were subsequently analyzed using Image J. Brightfield images were acquired on the ZOE Fluorescent Cell Imager (BioRad). See **Table S2** for details of reagents.

Co-localization analysis

For co-localization analysis, cell images were individually thresholded and underwent Coloc 2 analysis on ImageJ. Mander's coefficient, which ranges from 0 to 1 with 1 being 100% colocalized, is measuring the spatial overlap of one imaging channel with another imaging channel. Pearson's coefficient compares the pixel intensity of one channel with another channel. Pearson's coefficient values can range from -1 to 1 with -1 meaning inversely proportional and 1 meaning same pixel intensities.

MOLM-13 cell line generation and cell growth assays

MOLM-13 KRAS and NRAS mutant cell lines were generated as previously described (PMID 37681415 and PMID 39255801). Briefly, the plasmids pCW57.1 (Addgene 41393), pDONR223 KRAS WT (Addgene 81751), and pDONR223 NRAS WT (Addgene 82151) were used to generate doxycycline-inducible KRAS and NRAS constructs. mCherry was cloned (Addgene 60954) to the N-terminus of KRAS or NRAS on the pCW57.1 backbone and site-directed mutagenesis was performed to generate KRAS and NRAS mutations. MOLM-13 cells (DSMZ) were transduced with lentivirus generated from these constructs and sorted for mCherry

positivity after treatment with doxycycline 2 μg/mL. MOLM-13 cells were transfected with plasmids and treated with doxycycline 2 μg/mL (if indicated). 6 hours later, 10 nM AC220 (quizartinib) was added and this addition marks the start of the cell counting (t=0 days). Cell counts were obtained using a Countess 3 automated cell counter (ThermoFisher).

Graphics

All schematics were generated using BioRender.

Statistics and reproducibility

No statistical methods were used to predetermine the sample size. No sample was excluded from data analysis, and no blinding was used. All data were assessed for normality. For normally distributed data, pairwise comparisons were performed using unpaired two-tailed Student's t tests, with Welch's correction for unequal variances used as indicated. Comparisons between three or more groups were performed using ordinary one-way or two-way analysis of variance (ANOVA) as indicated. For data that were not normally distributed, pairwise comparisons were performed using the Mann-Whitney U test, and comparisons between multiple groups were performed using the Kruskal-Wallis test. All data shown are reported as mean ± SEM and error bars in figures represent SEM of biological triplicates. All data were analyzed and plotted using GraphPad Prism 8 including non-linear regression fitting.

Data availability

The data that support the findings of this study are available from Figshare.

Code availability

The code and design models used in this study are available from Zenodo.

Supplemental Figures:

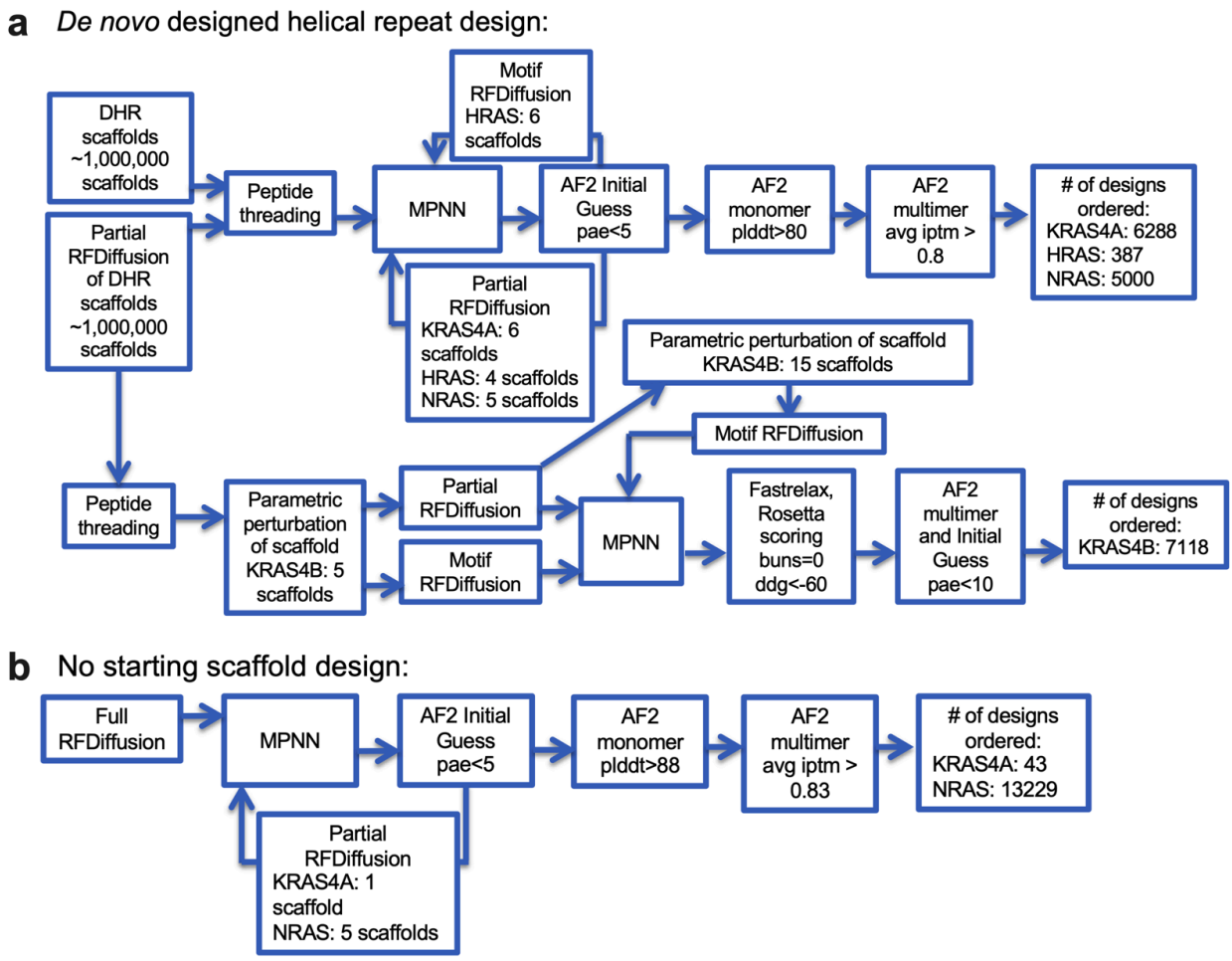
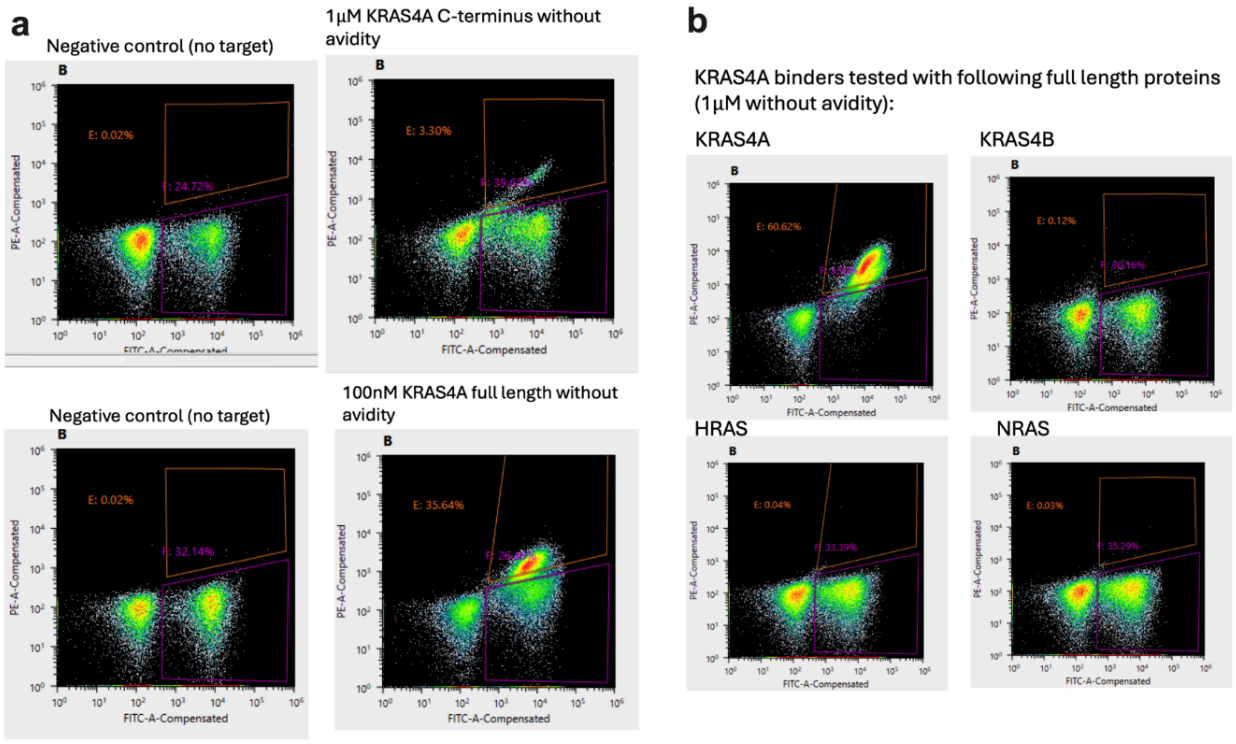
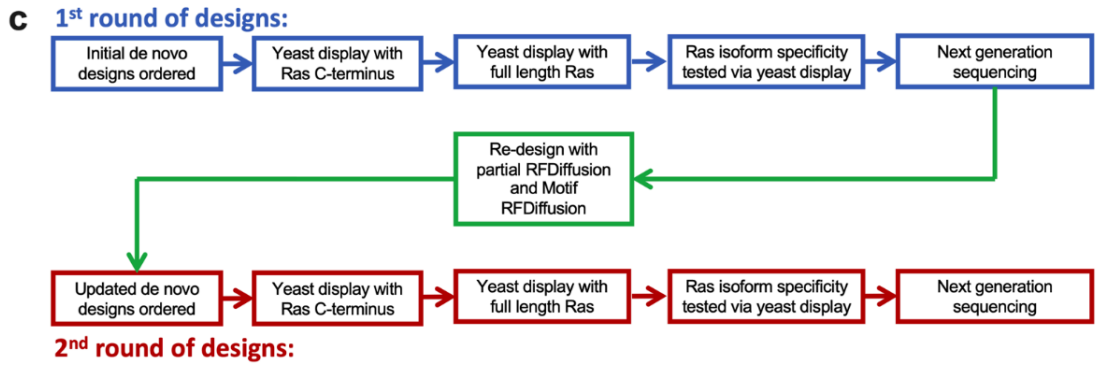


Figure S1: Details of computational design pipeline

a-b) Details of the computational design workflow for designing RIBs either using (**a**) *de novo* designed helical repeat proteins for the logos pipeline as starting scaffolds. (**b**) RIBs were also designed without a particular starting scaffold utilizing full two-sided RFDiffusion with both strand and helix specification.


a
Negative control (no target)
E: 0.02%
F: 24.72%
1μM KRAS4A C-terminus without avidity
E: 3.30%
Negative control (no target)
E: 0.02%
F: 32.14%
100nM KRAS4A full length without avidity
E: 35.64%
b
KRAS4A binders tested with following full length proteins (1μM without avidity):
KRAS4A
E: 60.62%
KRAS4B
E: 0.12%
HRAS
E: 0.04%
NRAS
E: 0.03%
PE-A-Compensated
FITC-A-Compensated


c
1st round of designs:
Initial de novo designs ordered
Yeast display with Ras C-terminus
Yeast display with full length Ras
Ras isoform specificity tested via yeast display
Next generation sequencing
Re-design with partial RFDiffusion and Motif RFDiffusion
Updated de novo designs ordered
Yeast display with Ras C-terminus
Yeast display with full length Ras
Ras isoform specificity tested via yeast display
Next generation sequencing
2nd round of designs:

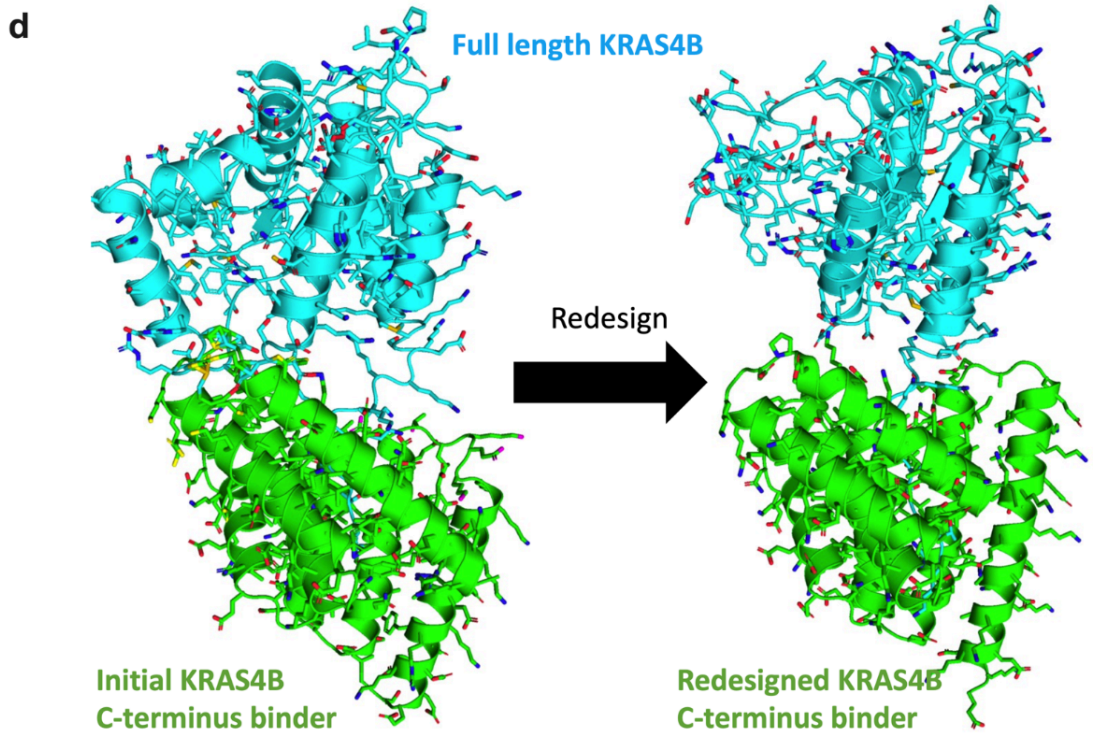
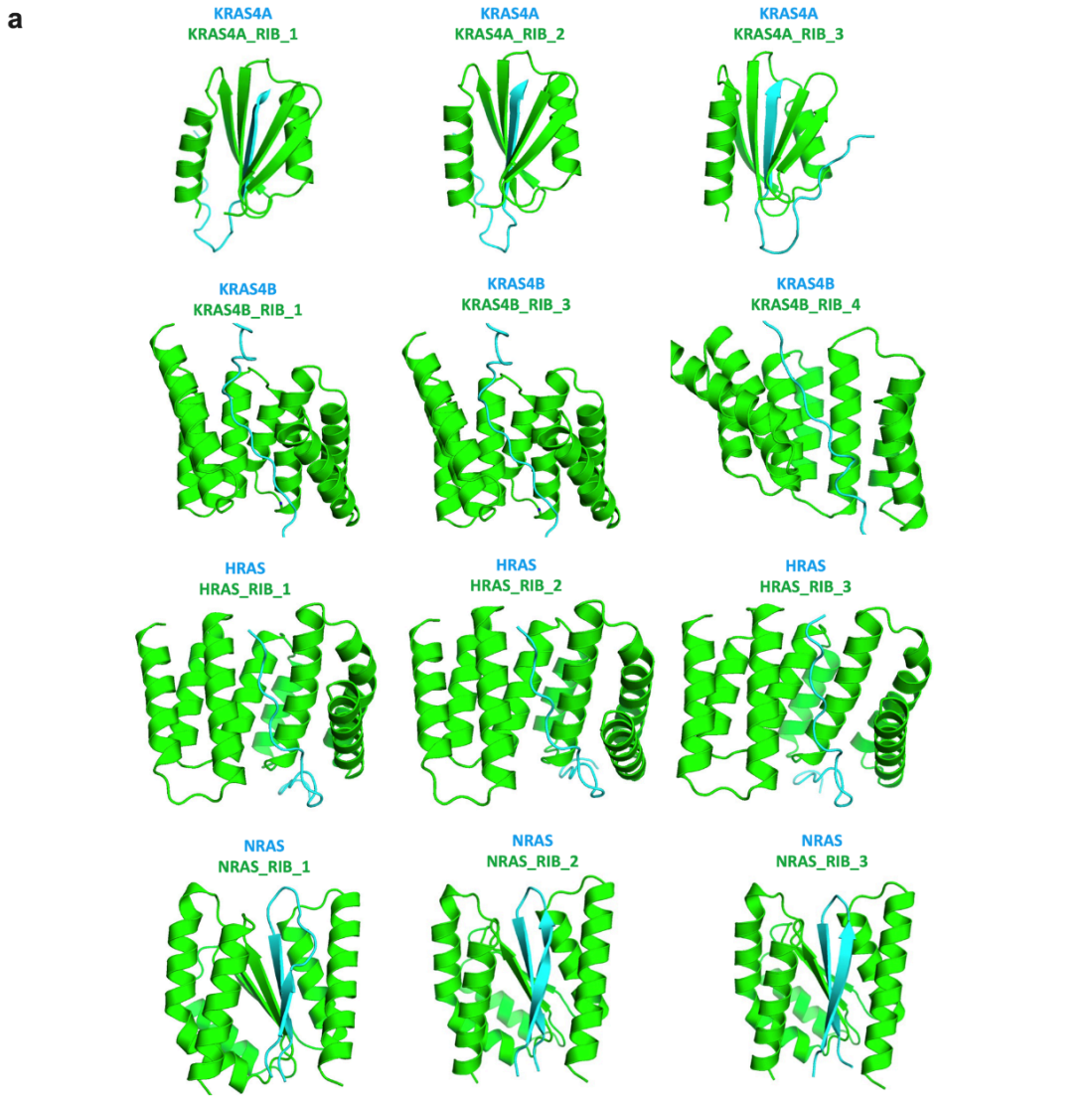

d
Full length KRAS4B
Redesign
Initial KRAS4B C-terminus binder
Redesigned KRAS4B C-terminus binder

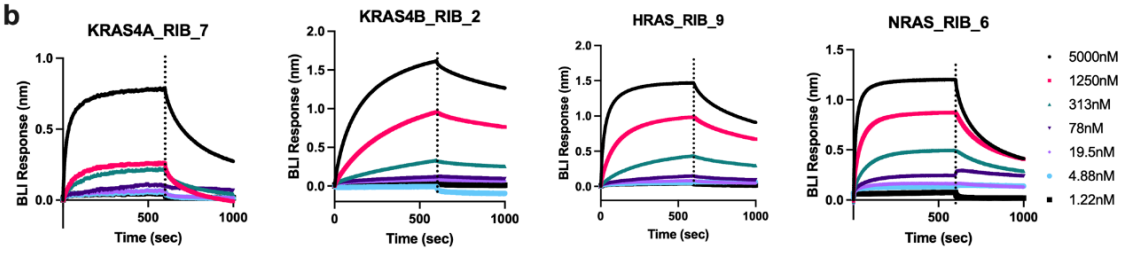
Figure S2: High throughput yeast display testing of designs.

a-b) Representative flow cytometry plots of designs tested via yeast display (see Methods for details). X-axis represents binder display and Y-axis represents target binding. Negative controls are when no target was added. (**a**) Either biotinylated C-terminus or full length protein was incubated with yeast cells expressing on their surface the designs. (**b**) Testing of Ras isoform specificity was initially done by incubating biotinylated full length Ras isoforms with yeast display by incubating yeast cells expressing designs. For all cases, none of the design libraries had significant binding signal to off-target Ras isoforms.

c) Experimental and computational workflow for designing RIBs. The first round of computational design focused only on the Ras C-terminus and were tested for binding to Ras C-terminus and full length protein via yeast display. For the RIBs that only bound to the C-terminus (KRAS4B and NRAS), the binder backbones were optimized via partial RFDiffusion and motif RFDiffusion for both the C-terminus (to tighten target binding) and the full length protein (to account for steric clashes). After this second round of computational design, these designs were again tested for binding to Ras C-terminus and full length protein via yeast display.

d) AlphaFold2 structure predictions of KRAS4B RIBs. Left: KRAS4B C-terminus binder which is predicted to sterically clash with the folded domain of KRAS4B. Right: After redesign of these KRAS4B C-terminus binder, a new KRAS4B RIB design is predicted to no longer clash with full length KRAS4B.





c

	KD (M)	KD Error	kon(1/Ms)	kon Error	kdis(1/s)	kdis Error
KRAS4A RIB	3.80E-07	3.54E-09	1.42E+03	8.56E+00	5.39E-04	3.81E-06
KRAS4B RIB	4.34E-07	8.38E-09	7.02E+03	1.30E+02	3.05E-03	1.72E-05
HRAS RIB	1.77E-07	1.10E-09	5.89E+03	3.12E+01	1.04E-03	3.36E-06
NRAS RIB	5.95E-07	2.37E-09	1.71E+03	3.06E+02	1.01E-03	1.21E-05

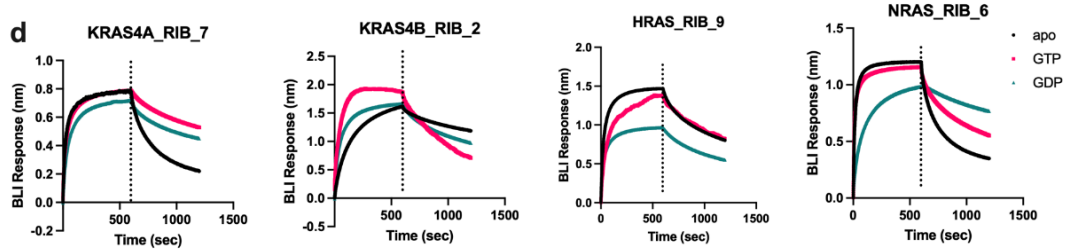


Figure S3: *In vitro* binding of RIBs to their target.

a) Design models of representative RIBs tested for each target.

b) BLI results of RIBs binding to their full length target. Results are representative of 3 independent experiments. The BLI data is for the RIB that was used for the rest of the paper.

c) Estimated binding kinetics and affinity of RIBs with their full length target.

d) BLI results of RIBs (5μM) binding to their full length target either without nucleotide (apo), 5μM GTP (GTP), or 5μM GDP (GDP). Results are representative of 3 independent experiments.

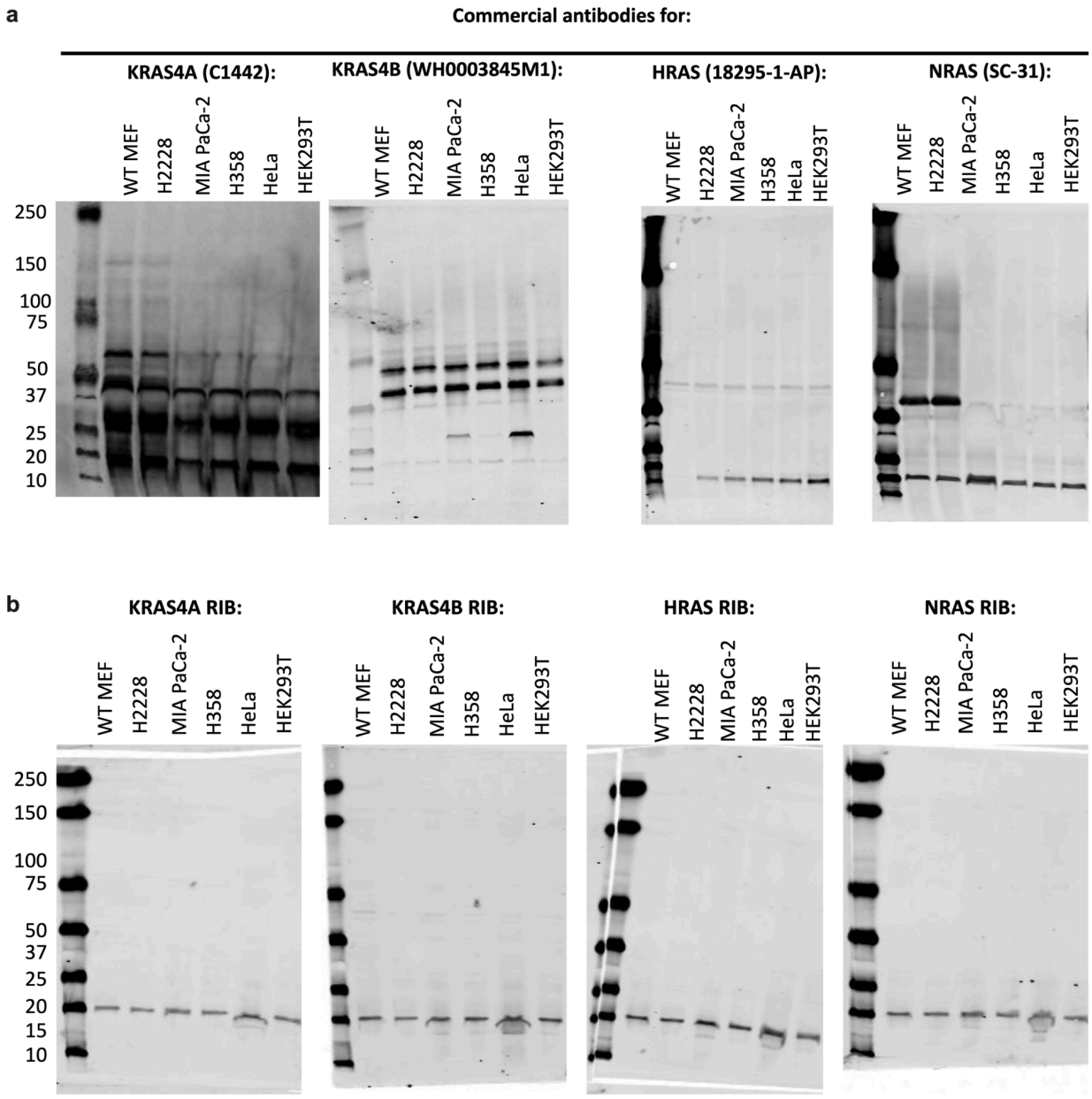


Figure S4: Comparison to commercially available antibodies.

a) Lysates from a panel of cell lines were lysed, ran on SDS-PAGE gels, and probed with antibodies that were shown previously to be the most specific among the commercially available reagents. Shown are representative full blots from 3 independent experiments.

b) Lysates from a panel of cell lines were ran on SDS-PAGE gels and probed with biotinylated RIBs. Shown are representative full blots from 3 independent experiments.

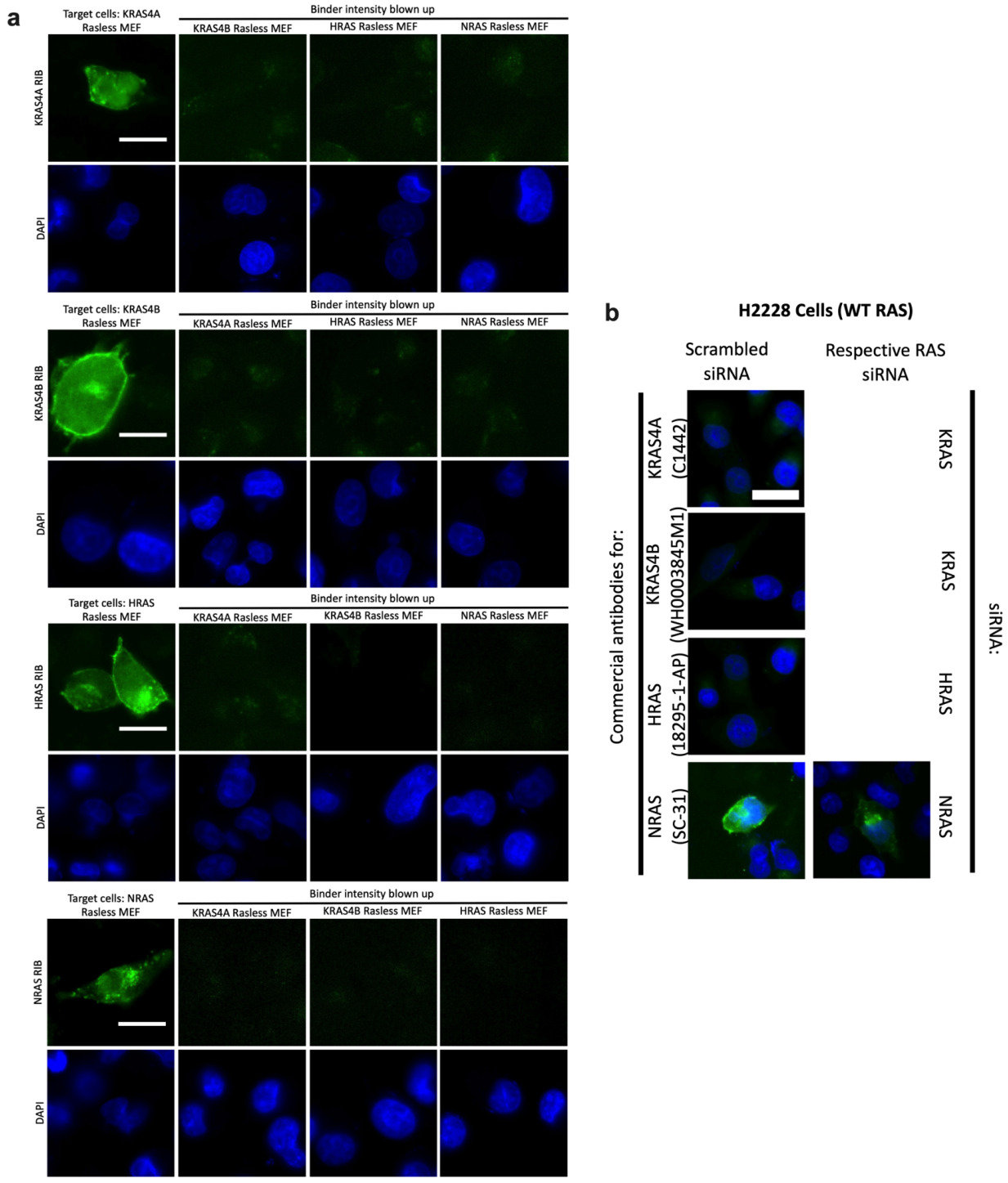


Figure S5: RIBs specifically bind to their target in mammalian cells.

a) Rasless MEFs were fixed, permeabilized, and probed with biotinylated RIBs and DAPI. Shown are the separated epifluorescence signals from RIBs and DAPI. RIBs incubated with their target cells are shown with a fixed brightness and contrast while RIBs incubated with their off-target cells are shown with the background signal amplified. Images are the same as **Fig 3f**. Results are representative of 3 independent experiments. Scale bar=10μm.

b) Representative epifluorescence images of Rasless MEFs either transfected with scrambled or on target siRNAs for 2 days and then immunostained with antibodies that were shown previously to be the most specific among the commercially available reagents. Results are representative of 3 independent experiments. Scale bar=10μm.

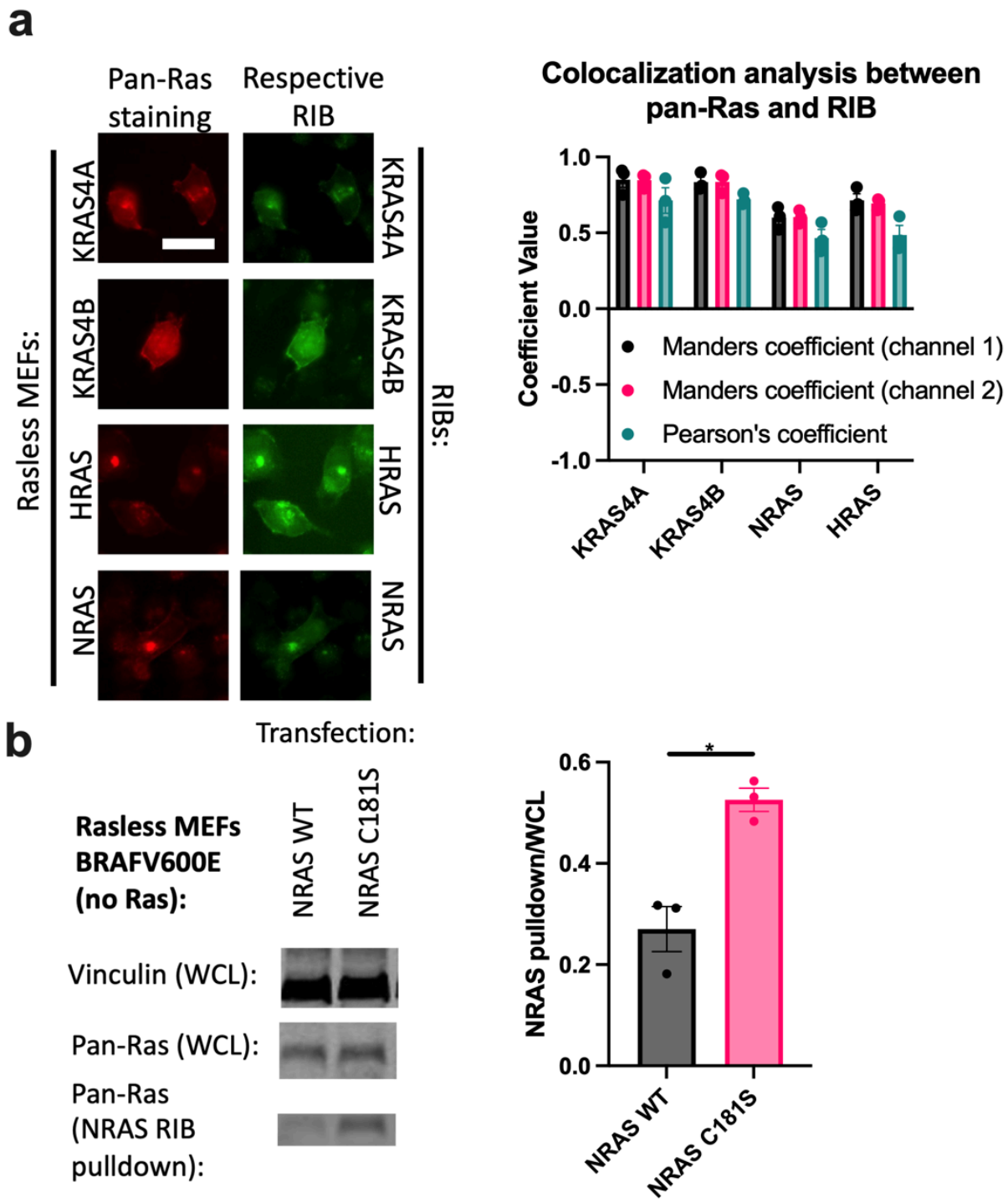


Figure S6: RIBs evenly bind to target throughout the mammalian cell.

a) Rasless MEFs were fixed, permeabilized, probed via immunostaining with the respective biotinylated RIBs and pan-Ras, and then imaged. Images were analyzed for co-localization

between pan-Ras and RIB signal (n=3 experiments). Images are representative of 3 independent experiments. Scale bar=10μm.

b) Rasless MEFs harboring BRAFV600E express none of the major Ras isoforms. These cells were transfected with either NRAS WT or NRAS C181S for 1 day, lysed, either pulled down with beads loaded with NRAS RIBs (see Methods for details) or had no pulldown (whole cell lysates=WCL), and then immunoprobed with the indicated antibodies. Left: representative immunoblot. Right: densitometry quantification of immunoblots (n=3 experiments). Statistics: Unpaired 2-tailed student's t-test.

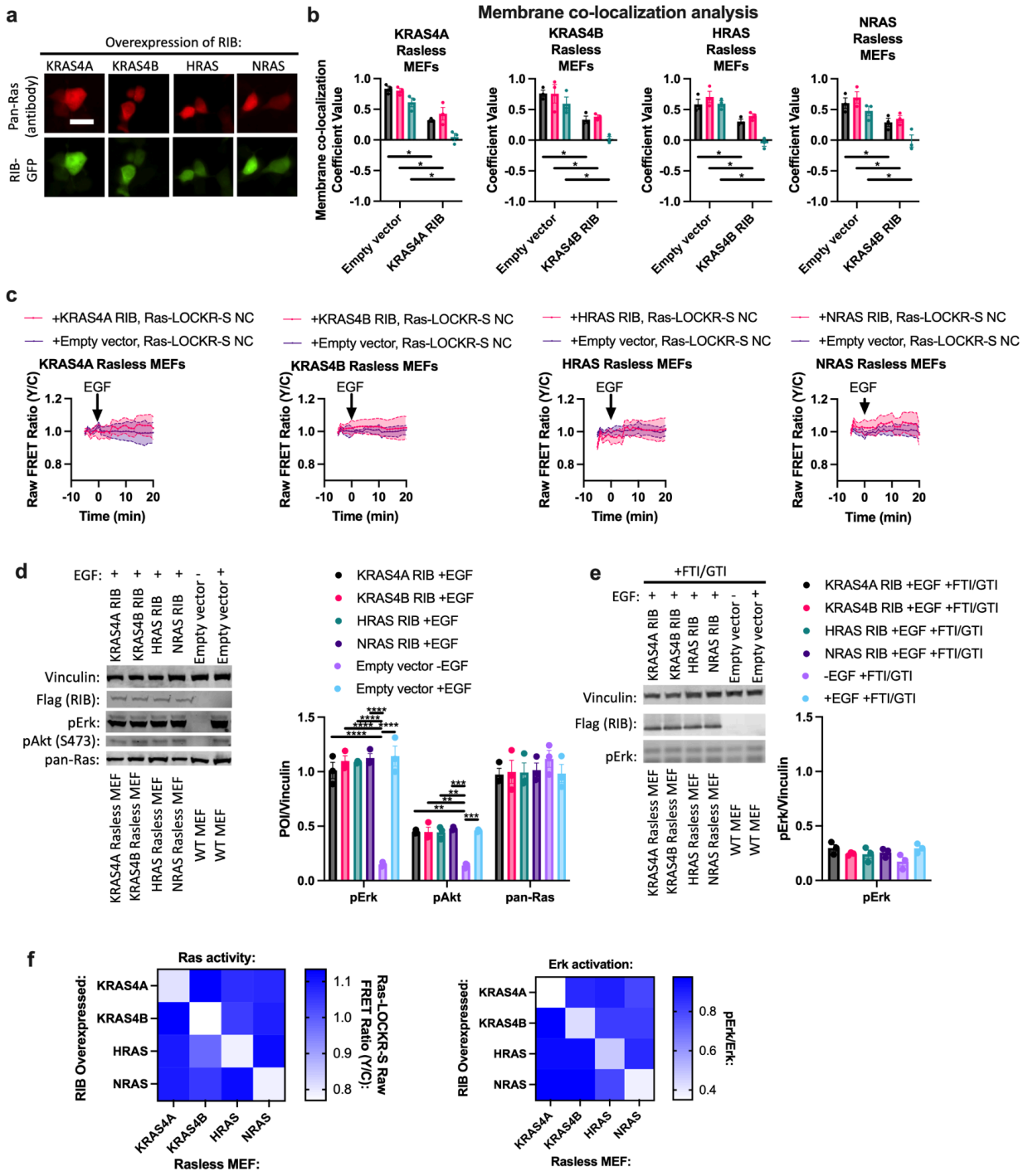


Figure S7: RIBs affect Ras signaling

a) Epifluorescence images of Rasless MEFs expressing RIB fused to GFP (RIB-GFP). These cells were then immunostained for pan-Ras. Images are representative of three biologically independent experiments. Scale bar=10μm.

b) Rasless MEFs expressing RIB-GFP were stained for pan-Ras and BioTracker Cytosolic Membrane dyes. Co-localization between RIB-GFP signal and BioTracker signal was analyzed (n=3 experiments). Statistics: Unpaired 2-tailed student's t-test.

c) Time course imaging of Ras-LOCKR-S negative control (NC) expressed in Rasless MEFs either co-expressing a RIB or empty vector and stimulated with 100ng/mL EGF (n=10 cells for each experiment). These curves are representative of three biologically independent experiments.

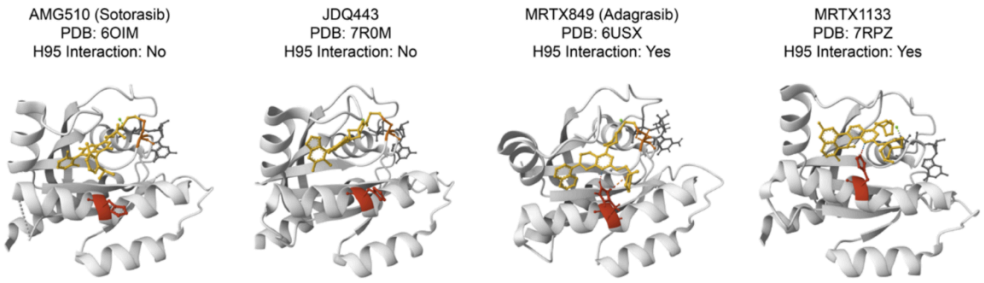
d-e) Rasless MEFs were transfected with their respective RIB and either treated without (**d**) or with (**e**) 10μM dual farnesyl transferase inhibitor (FTI) and geranylgeranyl transferase-1 inhibitor (GTI) FGTI-2734. 1 day later, cells were then stimulated with 100ng/mL EGF. These cells were compared to WT MEFs transfected with empty vector and stimulated with 100ng/mL EGF. These cells then underwent immunoblotting analysis. Right: densitometry quantification of immunoblots (n=3 experiments). Statistics: two-way ordinary ANOVA.

f) Rasless MEFs were transfected with RIBs for 1 day. Left: To measure Ras activity, Ras-LOCKR-S was co-transfected and imaged the next day. Right: To measure pErk levels, cells underwent immunoblotting analysis and pErk levels were compared to total Erk levels. All experiments were done in triplicates.

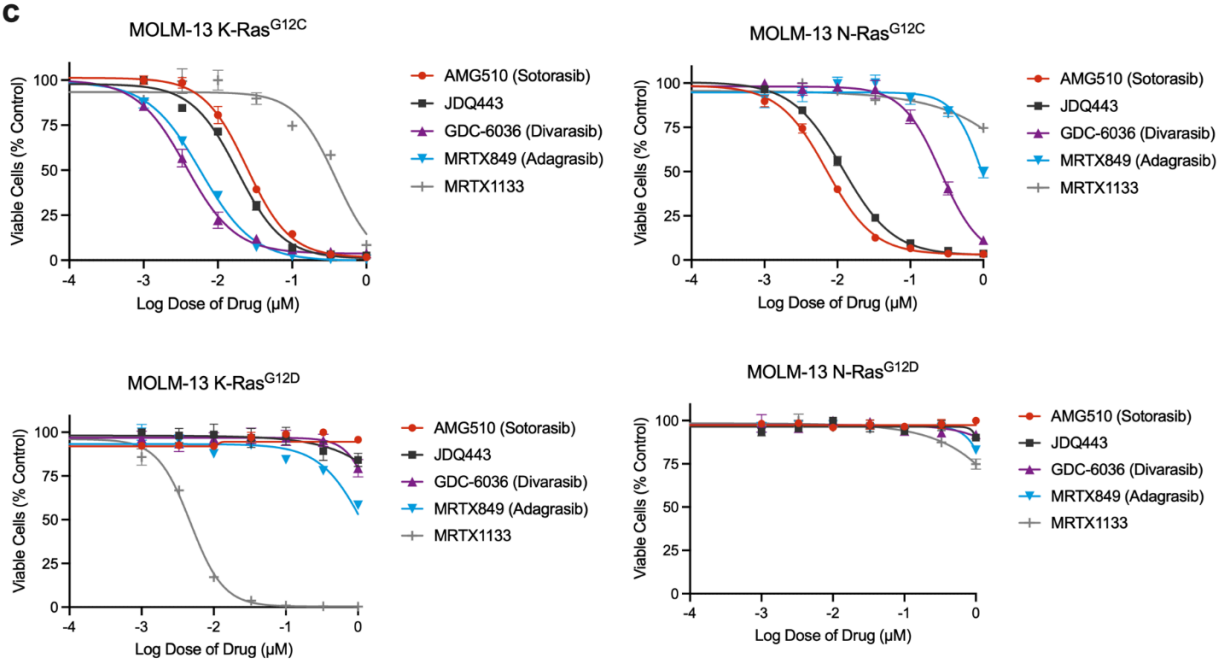
a

		Predicted and Experimental Response to Ras^{G12X} Inhibitors								
			K-Ras^{G12C}		N-Ras^{G12C}		K-Ras^{G12D}		N-Ras^{G12D}	
		K-Ras^{H95} Interaction	Predicted Response	Experimental IC50 (nM)	Predicted Response	Experimental IC50 (nM)	Predicted Response	Experimental IC50 (nM)	Predicted Response	Experimental IC50 (nM)
G12C Inhibitors										
(Sotorasib)	AMG510	No	Yes	24.6	Yes	7.1	No	N/A	No	N/A
	JDQ443	No	Yes	18.4	Yes	11.8	No	N/A	No	1365
(Divarasib)	GDC-6036	Yes	Yes	3.7	No	263.2	No	1895	No	N/A
(Adagrasib)	MRTX849	Yes	Yes	5.9	No	1035	No	1376	No	1876
G12D Inhibitors										
	MRTX1133	Yes	No	387.3	No	996.1	Yes	4.8	No	N/A

b



c



d

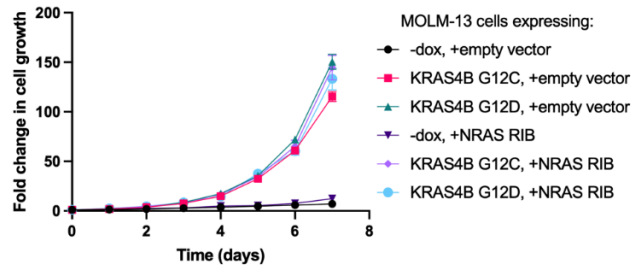


Figure S8: Characterization of MOLM-13 cells.

a) As previously described (PMID 37339170 and 38236605), the selectively of K-Ras^{G12C} and K-Ras^{G12D} inhibitors over N-Ras mutant counterparts depends upon whether a drug binds to K-Ras H95, a residue that is not conserved in N-Ras (L95). Experimental IC50s are based on (**c**).

b) Published crystal structures of K-Ras bound to G12C and G12D inhibitors showing the position of the drug in relation to H95 (in red). The structure for K-Ras bound to GDC-6036 (Divarasib) has not yet been published, but molecular dynamic simulations suggest that GDC-6036 binds to H95 (https://chemrxiv.org/engage/chemrxiv/article-details/64fecb59b6ab98a41c3d9c0f).

c) MOLM-13 KRAS or NRAS mutant cell lines treated with K-Ras^{G12C} or K-Ras^{G12D} inhibitors. Represented results shown from 3 independent experiments each performed in technical triplicate.

d) MOLM-13 cell lines expressing either KRAS4B G12C or G12D were treated with or without doxycycline to turn on mutant KRAS4B expression. Cells were also transfected with NRAS RIB expressing plasmid or empty vector and co-treated with AC220. Cells were then counted over a 7 day period (n=3 experiments). Statistics: one-way ordinary ANOVA.

Table S1: Characterization of all RIBs experimentally tested *in vitro* and in cells

The top 10 RIBs from yeast display for each target were further tested *in vitro* for binding affinity and in cells for their effects on Ras expression (via immunoblot), Ras activity (via Ras-LOCKR-S) before and after 100ng/mL EGF stimulation, pErk signaling (via immunoblot) before and after 100ng/mL EGF stimulation, localization of signal when used as staining probes, and number of bands when applied in blots.

Table S2: Reagents used within study

Reagents used throughout the study with catalog numbers.

Chapter 6 – Other relevant work

Thanks to the great collaborative environment at the Institute of Protein Design and University of Washington, I have been able to always work with amazing scientists and fellow researcher friends, learning new perspectives from them.

Here I would like to take this chance (but not too much space) to point out two other published work from my earlier years, where I was a co-first author and working with wonderful people inside and outside of the lab; not as closely related to the topic of this thesis, but helped me grow and learn from the memorable collaborations.

The work on buttressed loops helped me understand new, delicate, precise rules of controlling the loop-based hydrogen bonding networks; the concept of de novo “DARPin” to some degree inspired the main concept of building logos platform in Chapter 2 (de novo IDR binding version of “DARPin”- one set of rules to bind any arbitrary given protein sequences computationally, despite being independent of any native protein families or borrowed native principles). The work on COVID neutralizing minibinders paved my understanding of protein therapeutics, also guided me to learn and build my own de novo IDR binding domain from the help of our de novo protein target binding pioneers Dr. Longxing Cao, Dr. Andrew Hunt, etc. The audience who found common interests can feel free to refer to the online papers.

De novo design of buttressed loops for sculpting protein functions

Hanlun Jiang[1,2#], Kevin M. Jude[3,#], Kejia Wu[1,2,4#], Jorge Fallas[1,2], George Ueda[1,2], TJ Brunette[1,2], Derrick Hicks[1,2], Harley Pyles[1,2], Aerin Yang[5], Lauren Carter[1,2], Mila Lamb[1,2], Xinting Li[1,2], Paul M. Levine[1,2], Lance Stewart[1,2], K. Christopher Garcia[3,5,6*], David Baker[1,2,3,7*]

[#]Equal contribution

*e-mail: kcgarcia@stanford.edu; dabaker@uw.edu

[1]Department of Biochemistry, University of Washington.

[2]Institute for Protein Design, University of Washington.

[3]Howard Hughes Medical Institute, Stanford University School of Medicine

[4]Biological Physics, Structure and Design Graduate Program, University of Washington.

[5]Department of Molecular and Cellular Physiology, Stanford University School of Medicine.

[6]Department of Structural Biology, Stanford University School of Medicine.

[7]Howard Hughes Medical Institute, University of Washington

Abstract

In natural proteins, structured loops play central roles in molecular recognition, signal transduction and enzyme catalysis. However, because of the intrinsic flexibility and irregularity of loop regions, organizing multiple structured loops at protein functional sites has been very difficult to achieve by de novo protein design. Here we describe a solution to this problem that designs tandem repeat proteins with structured loops (9-14 residues) buttressed by extensive hydrogen bonding interactions. Experimental characterization shows the designs are monodisperse, highly soluble, folded and thermally stable. Crystal structures are in close

agreement with the design models, with the loops structured and buttressed as designed. We demonstrate the functionality afforded by loop buttressing by designing and characterizing binders for extended peptides in which the loops form one side of an extended binding pocket. The ability to design multiple structured loops should contribute generally to efforts to design new protein functions.

Title: Multivalent designed proteins neutralize SARS-CoV-2 variants of concern and confer protection against infection in mice

Authors: Andrew C. Hunt[1,2†], James Brett Case[3†], Young-Jun Park[4†], Longxing Cao[4,5†], Kejia Wu[4,5†], Alexandra C. Walls[4,6†], Zhuoming Liu[7], John E. Bowen[4], Hsien-Wei Yeh[4,5], Shally Saini[4,8], Louisa Helms[8,9,10,24], Yan Ting Zhao[4,8,11], Tien-Ying Hsiang[12], Tyler N. Starr[13], Inna Goreshnik[4,5], Lisa Kozodoy[4,5], Lauren Carter[4,5], Rashmi Ravichandran[4,5], Lydia B. Green[14], Wadim L. Matochko[14], Christy A. Thomson[14], Bastian Vögeli[1,2,15], Antje Krüger[1,2], Laura A. VanBlargan[3], Rita E. Chen[3,16], Baoling Ying[3], Adam L. Bailey[16,17], Natasha M. Kafai[3,16], Scott E. Boyken[4,5], Ajasja Ljubetič[4,5,18], Natasha Edman[4,5,19,20], George Ueda[4,5], Cameron M. Chow[4,5,21], Max Johnson[4,5], Amin Addetia[4,22], Mary Jane Navarro[4], Nuttada Panpradist[23], Michael Gale Jr.[12], Benjamin S. Freedman[8,9,10,23,24], Jesse D. Bloom[13,6,25], Hannele Ruohola-Baker[4,8,11,23], Sean P. J. Whelan[7], Lance Stewart[4,5], Michael S. Diamond[3,7,16,26*], David Veesler[4,6*], Michael C. Jewett[1,2,27,28*], David Baker[4,5,6*]

Affiliations:

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, 60208, USA

[2]Center for Synthetic Biology, Northwestern University, Evanston, IL, 60208, USA, [3]Department of Medicine, Washington University School of Medicine, St. Louis, MO, 63110, USA

[4]Department of Biochemistry, University of Washington, Seattle, WA, 98195, USA

[5]Institute for Protein Design, University of Washington, Seattle, WA, 98195, USA

[6]Howard Hughes Medical Institute, University of Washington, Seattle, WA, 98195, USA

[7]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, 63110, USA

[8]Institute for Stem Cell and Regenerative Medicine, University of Washington School of Medicine, Seattle, WA, 98109, USA

[9]Division of Nephrology, Department of Medicine, University of Washington School of Medicine, Seattle, WA, 98109, USA

[10]Kidney Research Institute, University of Washington School of Medicine, Seattle, WA, 98109, USA

[11]Oral Health Sciences, School of Dentistry, University of Washington, Seattle, WA, 98195, USA

[12]Department of Immunology, Center for Innate Immunity and Immune Disease, University of Washington, Seattle, WA, 98195, USA

[13]Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA

[14]Amgen Research, Biologic Discovery, Burnaby, V5A 1V7, BC, Canada

[15]Invizyne Technologies Inc., Monrovia, CA, 91016, USA

[16]Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO, 63110, USA

[17]Department of Pathology & Laboratory Medicine, University of Wisconsin – Madison, Madison, WI, 53705, USA

[18]Department for Synthetic Biology and Immunology, National Institute of Chemistry, Ljubljana, SI-1000, Slovenia

[19]Molecular and Cellular Biology Graduate Program, University of Washington, Seattle, WA, 98195, USA

[20]USA Medical Scientist Training Program, University of Washington, Seattle, WA, 98195, USA

[21]Neolukin Therapeutics Inc., Seattle, WA, 98102, USA

[22]The Molecular and Cellular Biology Program, University of Washington, Seattle, WA, 98195, USA

[23]Department of Bioengineering, University of Washington, Seattle, WA, 98195, USA

[24]Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, WA, 98109, USA

[25]Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA

[26]Andrew M. and Jane M. Bursky Center for Human Immunology and Immunotherapy Programs, Washington University School of Medicine, St. Louis, MO, 63110, USA

[27]Chemistry of Life Processes Institute, Northwestern University, Evanston, IL, 60208, USA

[28]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, 60611, USA

[†]These authors contributed equally

***Corresponding authors:** Michael S. Diamond (mdiamond@wustl.edu), David Veesler (dveesler@uw.edu), Michael C. Jewett (m-jewett@northwestern.edu), David Baker (dabaker@uw.edu)

ONE SENTENCE SUMMARY

Computationally designed trivalent minibinders provide therapeutic protection in mice against emerging SARS-CoV-2 variants of concern.

ABSTRACT

New variants of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) continue to arise and prolong the coronavirus disease 2019 (COVID-19) pandemic. Here we used a cell-free expression workflow to rapidly screen and optimize constructs containing multiple computationally designed miniprotein inhibitors of SARS-CoV-2. We found the broadest efficacy with a homo-trimeric version of the 75-residue angiotensin converting enzyme 2 (ACE2) mimic AHB2 (TRI2-2) designed to geometrically match the trimeric spike architecture. In the cryo-electron microscopy structure, TRI2 formed a tripod on top of the spike protein which

engaged all three receptor binding domains (RBDs) simultaneously as in the design model. TRI2-2 neutralized Omicron (B.1.1.529), Delta (B.1.617.2), and all other variants tested with greater potency than that of monoclonal antibodies used clinically for the treatment of COVID-19. TRI2-2 also conferred prophylactic and therapeutic protection against SARS-CoV-2 challenge when administered intranasally in mice. Designed miniprotein receptor mimics geometrically arrayed to match pathogen receptor binding sites could be a widely applicable antiviral therapeutic strategy with advantages over antibodies and native receptor traps. By comparison, the designed proteins have resistance to viral escape and antigenic drift by construction, precisely tuned avidity, and greatly reduced chance of autoimmune responses.