Sequence-specific targeting of intrinsically disordered protein regions

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.

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 charge 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.

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).

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. 2B).

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. 2C).

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.

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. 4A-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) (Fig. 4D). 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. 3C; 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.

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). 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 <180 pM; Extended Data Figure 9, a variant of binder DYNA_2b1) (Fig. 4G). NMR confirmed that free dynorphin A is intrinsically disordered and becomes ordered upon binding except for the intentionally excluded regions (Fig. 3F). For both DYNA peptides the NMR data clearly support an extended bound-state conformation, as expected from the design models (Fig. 3F 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. 3F 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.

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. 4A, 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].

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).

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.

Acknowledgments

We thank J.M. Rogers for early discussions of disordered region binding mechanisms, P. Fordyce for inspiring ideas exchange, R.E. Stenkamp believed this was structurally possible. We thank A. Broerman, S. Honda, S. Kenny, and T. Schlichthaerle for valuable conversations to make this work better. We also thank K. VanWormer and L. Goldschmidt for their technical support. This work was funded by The Audacious Project at the Institute for Protein Design (D.H., A.B., C.L., A.K., S.G.), The Open Philanthropy Project Improving Protein Design Fund (H.J., B.C., A.B., S.G., M.A.K.), The Bill and Melinda Gates Foundation, #INV-010680 (A.K., A.K., S.G., X.L.), The Howard Hughes Medical Institute (D.B.), the National Science Foundation Molecular Foundations for Biotechnology No. CHE-2226466 (W.C.), the Defense Advanced Research Projects Agency Harnessing Enzymatic Activity for Lifesaving Remedies program Award HR0011-21-2-0012 (K.W., A.B., X.L.), the National Institutes of Health's National Institute on Aging U19AG065156 (D.H.), and the Defense Threat Reduction Agency Grant HDTRA1-21-1-0038 (W.Y.). NMR studies were supported by NIH Grant 5R35GM141818 (G.T.M.).

Author Contributions

K.W. and D.B. conceptualized the *de novo* platform. K.W. developed the computational pipeline *logos*, arbitrary unstructured sequence recognition. H.J. contributed ideas for early-stage di-peptide binder design, contributed the code for threading and parametric perturbation. D.R.H. participated in optimizing the final *logos* pipeline, contributed the code for customized MPNN, customized AF2 (including AF2-multimer, AF2-initial-multimer), and the ideas and code support of local resampling in threading, MPNN-AF2 iteration, organized the github repo. B.C. contributed the code for a customized RFdiffusion build used in the refinement protocol. K.W. with occasional assistance from D.R.H and others, designed, screened, and characterized almost all designs. C.L. helped with some protein purification, and participated in the design to target CSP_N and conducted screening. T.A.R., A.G., and G.T.M. designed, conducted, and analyzed NMR experiments. Y.L. designed, conducted, and analyzed LC-MS experiments in the CTN4 tryptic peptide enrichment detection. K.M. designed and performed the immunoprecipitation assays for the WASH complex, under the supervision of E.D. E.M. designed and performed the in vitro cAMP assay for dynorphin A binders. W.C. offered help and technical support in characterizing some binders. X.L. synthesized and purified peptides, and conducted LC-MS validation of proteins and peptides. S.G., M.Y.L., and A.M. handled large-scale protein purification for crystal sample preparation. A.K.B. and A.K. resolved crystal structures. M.A.K. kindly helped in manuscript writing and figure-making. W.Y. offered help in construct ordering. G.S. under the supervision of S.M.B. offered further downstream characterizations. E.D.

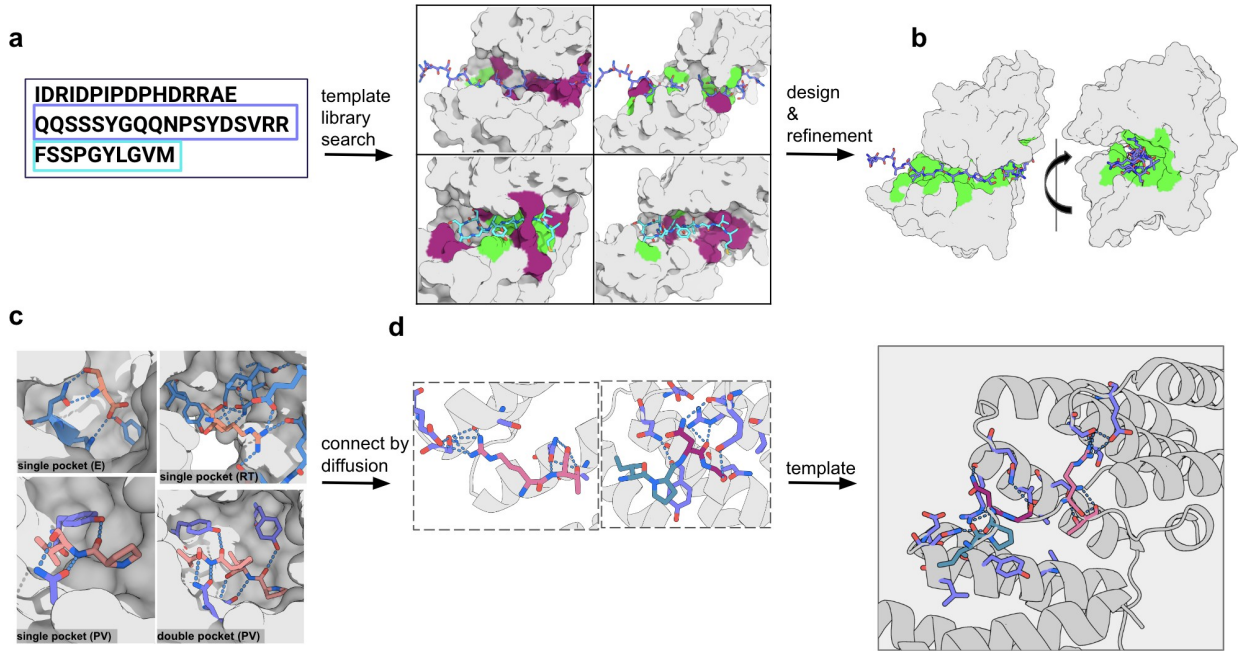
contributed conceptual support for the computation idea at the beginning. D.B. supervised the research. K.W. and D.B. drafted the manuscript. All authors reviewed and commented on the manuscript.

Competing Interests: A provisional patent application will be filed prior to publication, listing K.W., H.J., D.R.H., C.L., E.M., T.A.R., Y.L., K.M., M.H.G., G.T.M., E.D., and D.B. as inventors or contributors. G.T.M. is a founder of Nexomics Biosciences, Inc., which does not constitute a conflict with respect to this study.

Data and Materials Availability: All data are available in the main text or as supplementary materials. Design scripts and models are available through https://files.ipd.uw.edu/kejiawu/logos.zip and through https://github.com/drhicks/Kejia_peptide_binders/ . Crystallographic datasets have been deposited in the Protein Data Bank (PDB) with accession codes 9cce and 9ccf. NMR resonance assignments have been deposited in BioMagResDB.

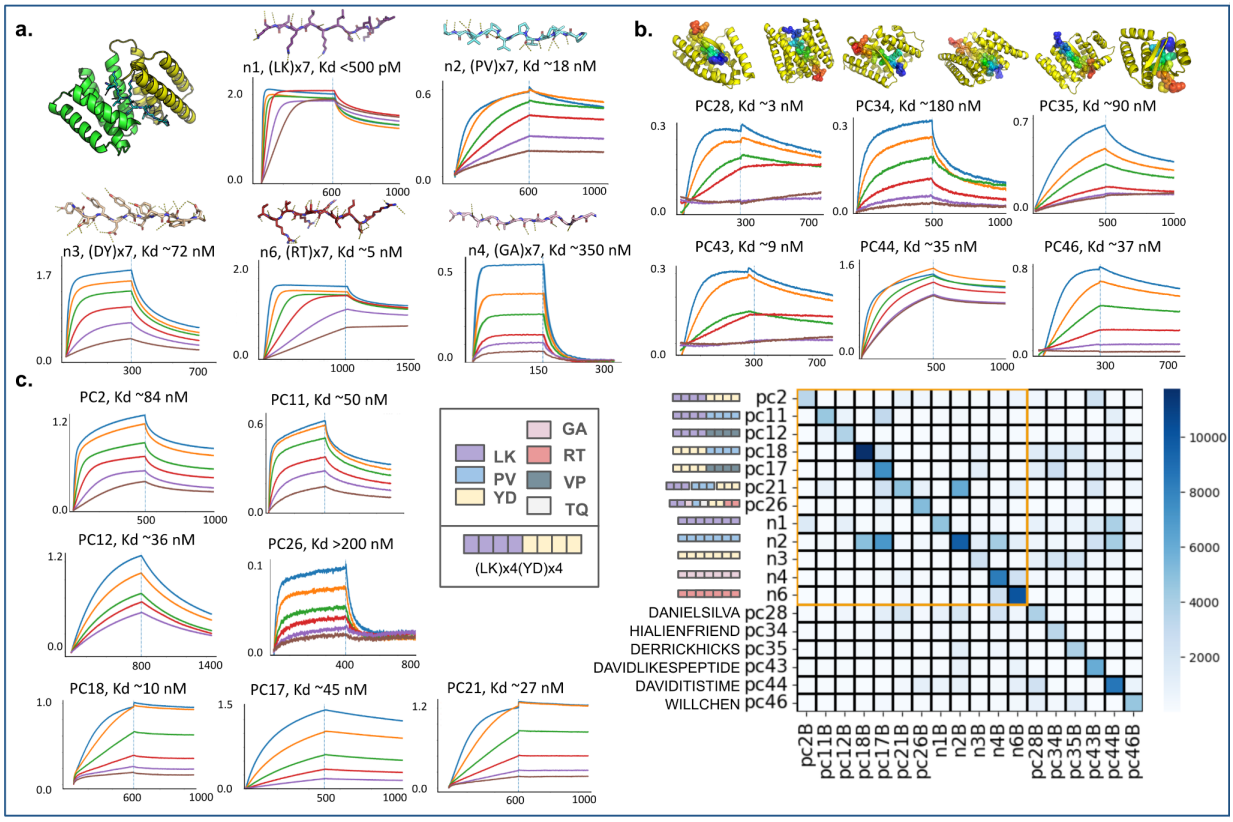
Figures

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**.

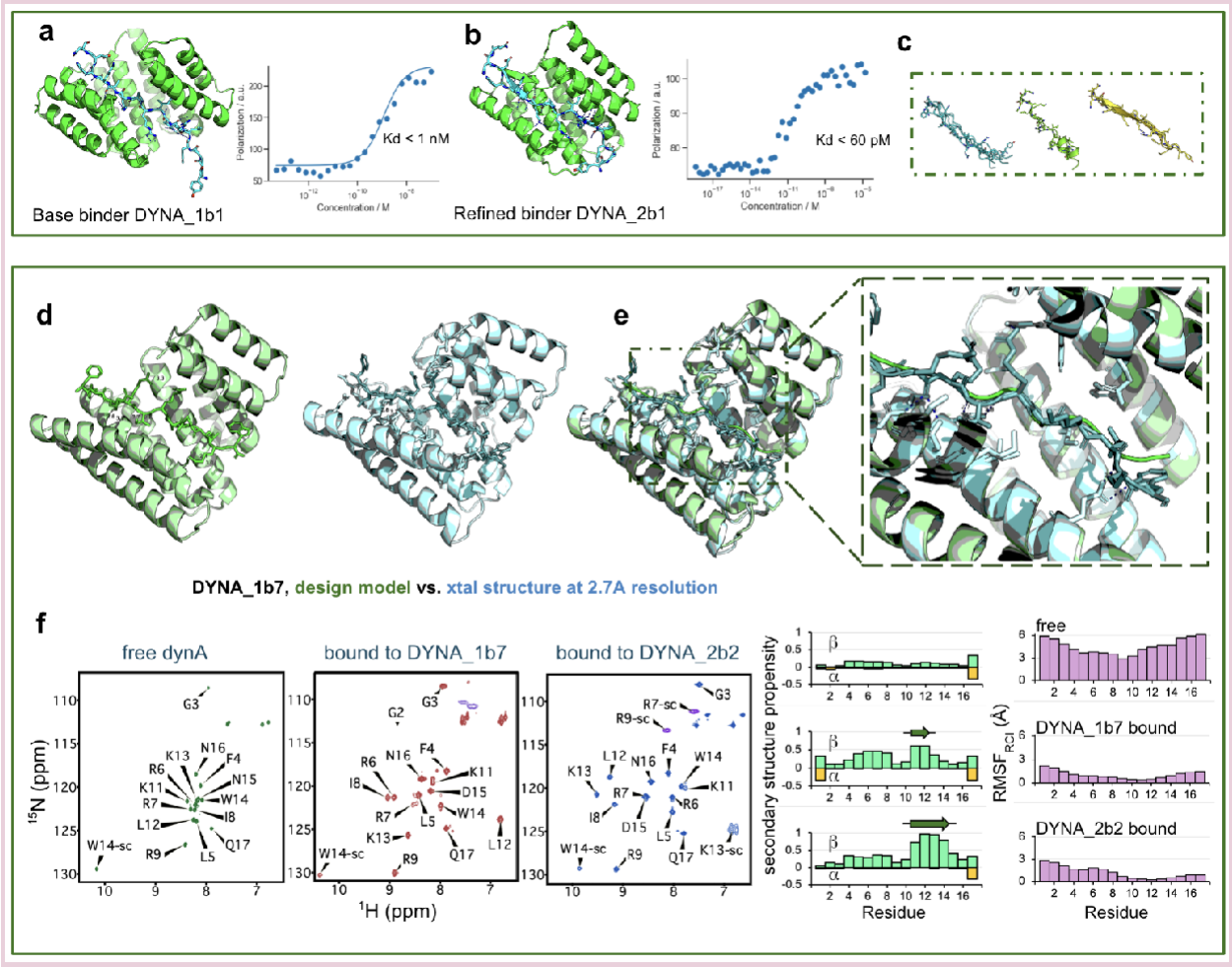
Fig. 2. Design of 18 synthetic peptide binding proteins.



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.

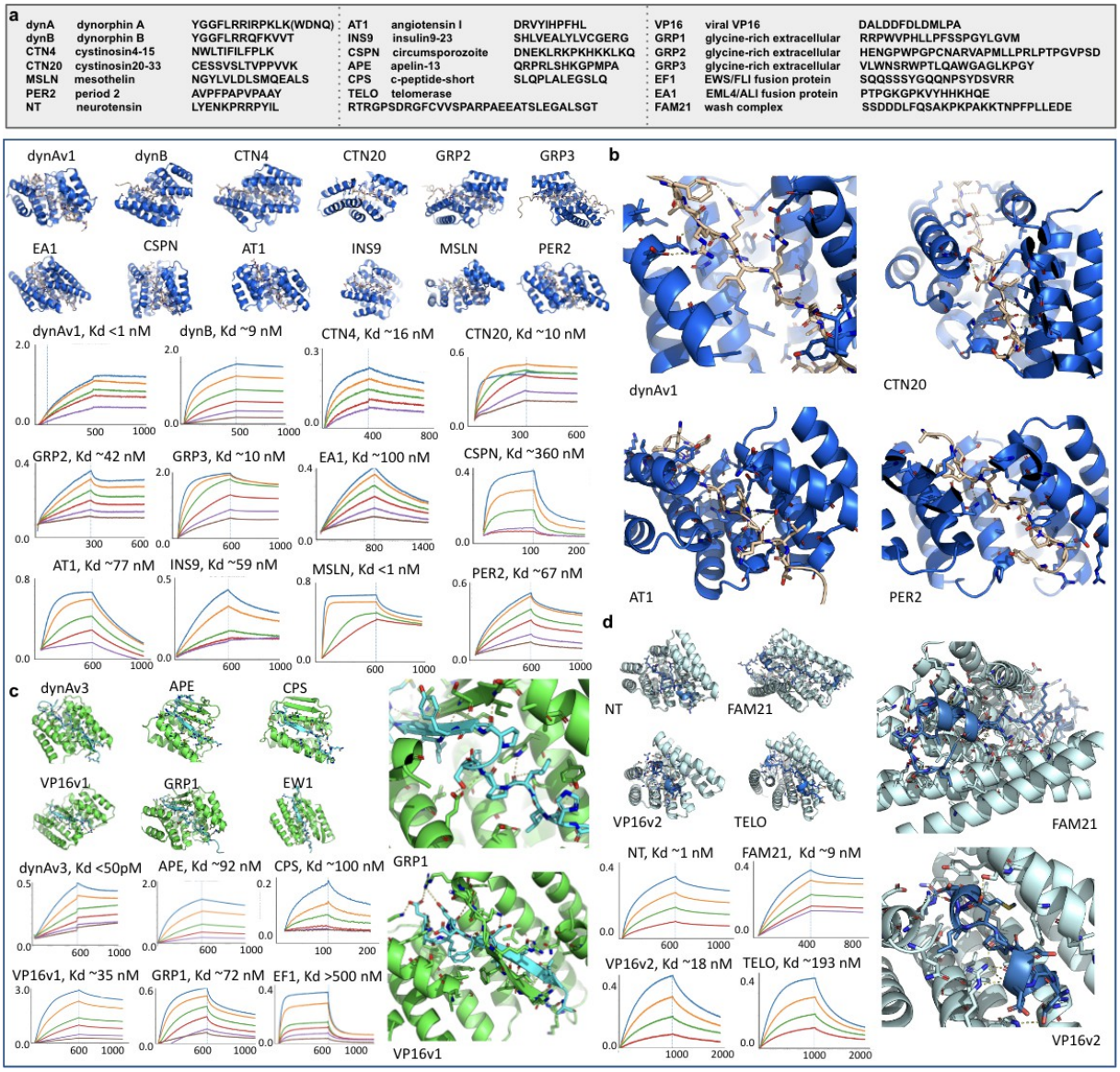
Fig. 3. Structural characterization of dynorphin A binder designs.



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 Cα positions ($RMSF_{RCI}$) based on backbone chemical

shift data. Sidechain amide resonance peaks in these HSQC spectra are not labeled.

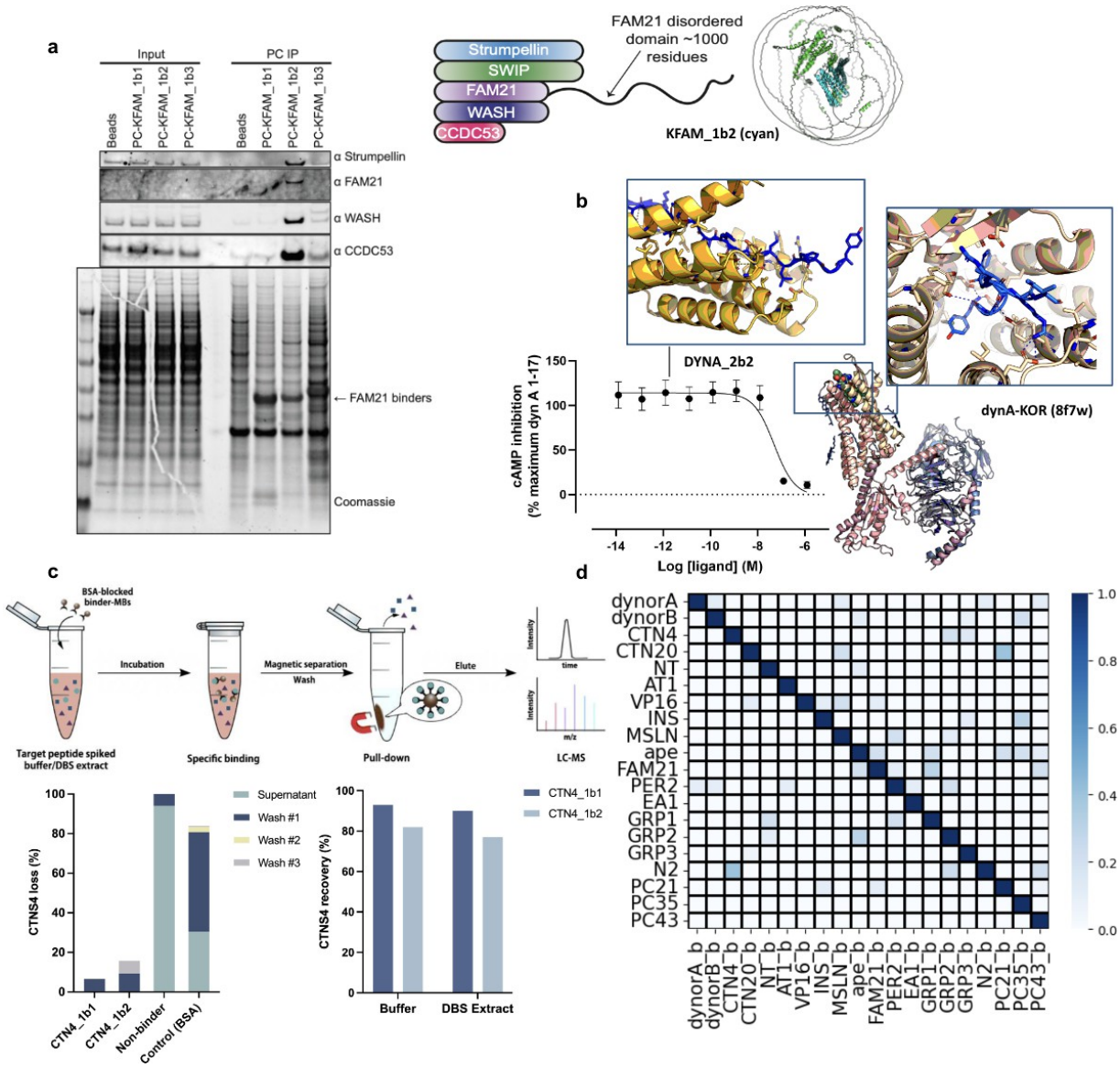
Fig. 4. Design of binders to 21 native protein disordered regions and peptides.



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.

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.

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=htt ps:%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).

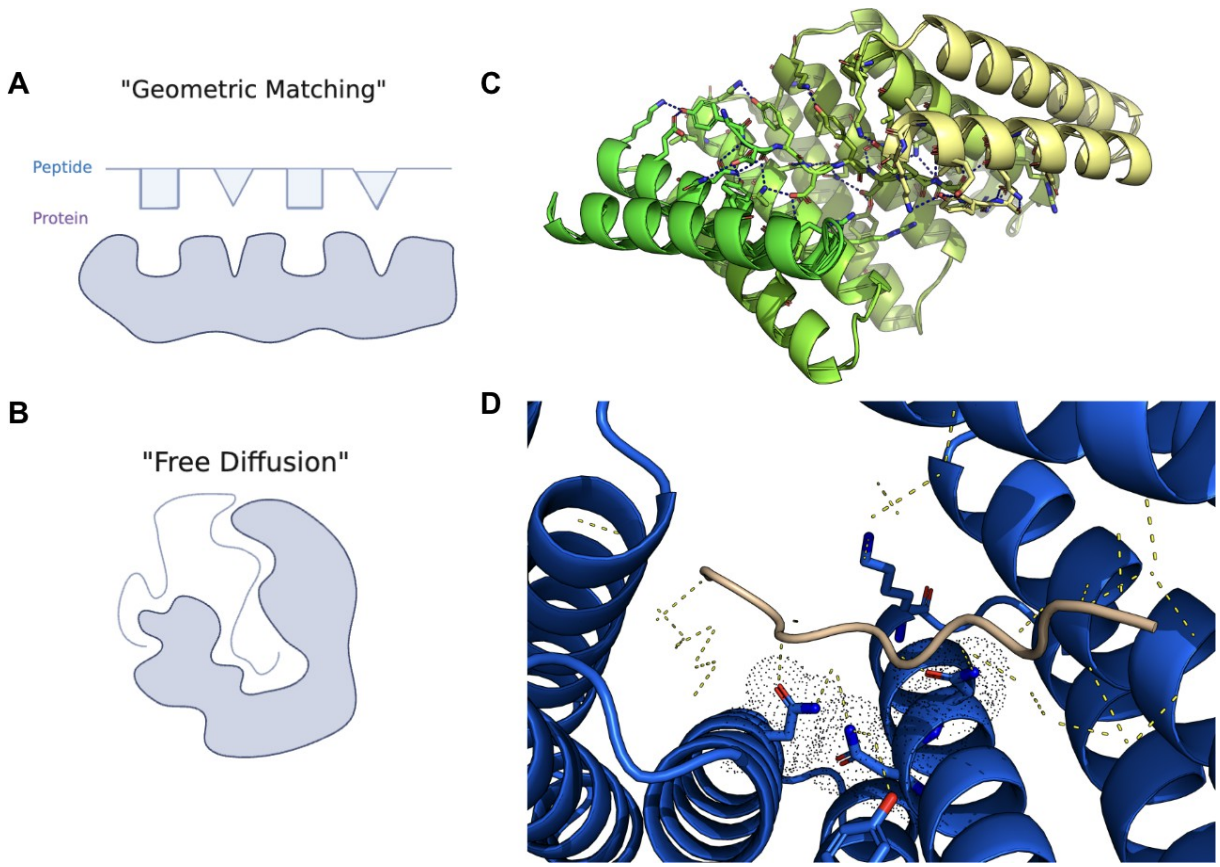
45.Sasaki T, Rodig SJ, Chirieac LR, Jänne PA. The biology and treatment of EML4-ALK non-small cell lung cancer. Eur J Cancer. 2010 Jul;46(10):1773-80.

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.

Supplementary Materials

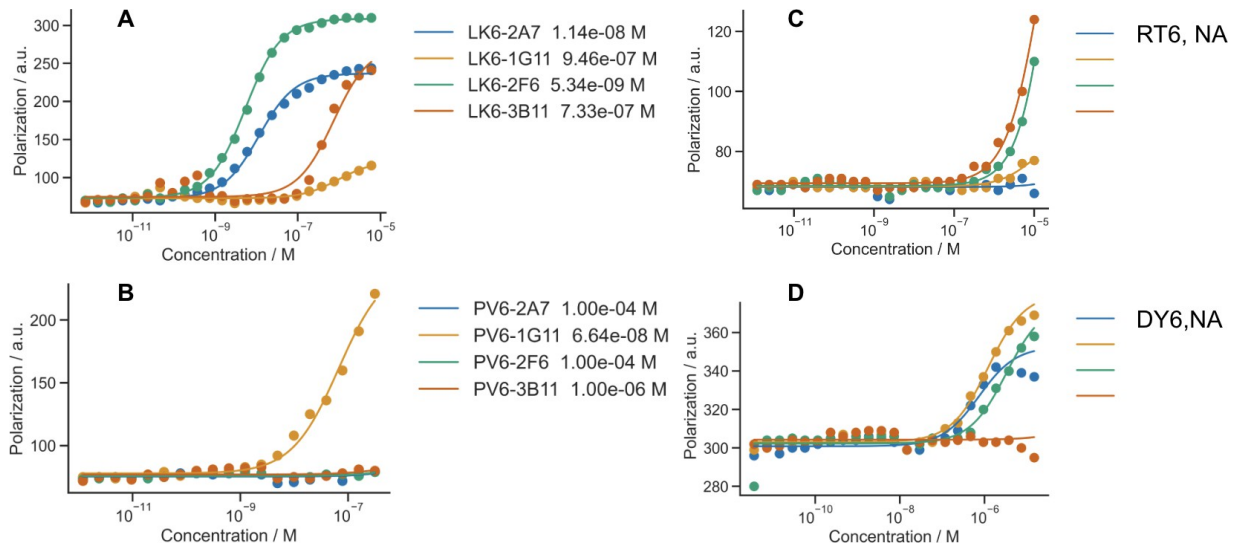


Extended Data Fig. 1: Template repeat di-peptide binding protein platform. A) Previous geometric matching approaches required a definite number of pre-built pockets as well as one-to-one parametric matching. B) Free diffusion approach folds relatively large disordered targets (30-40 amino acids) into 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

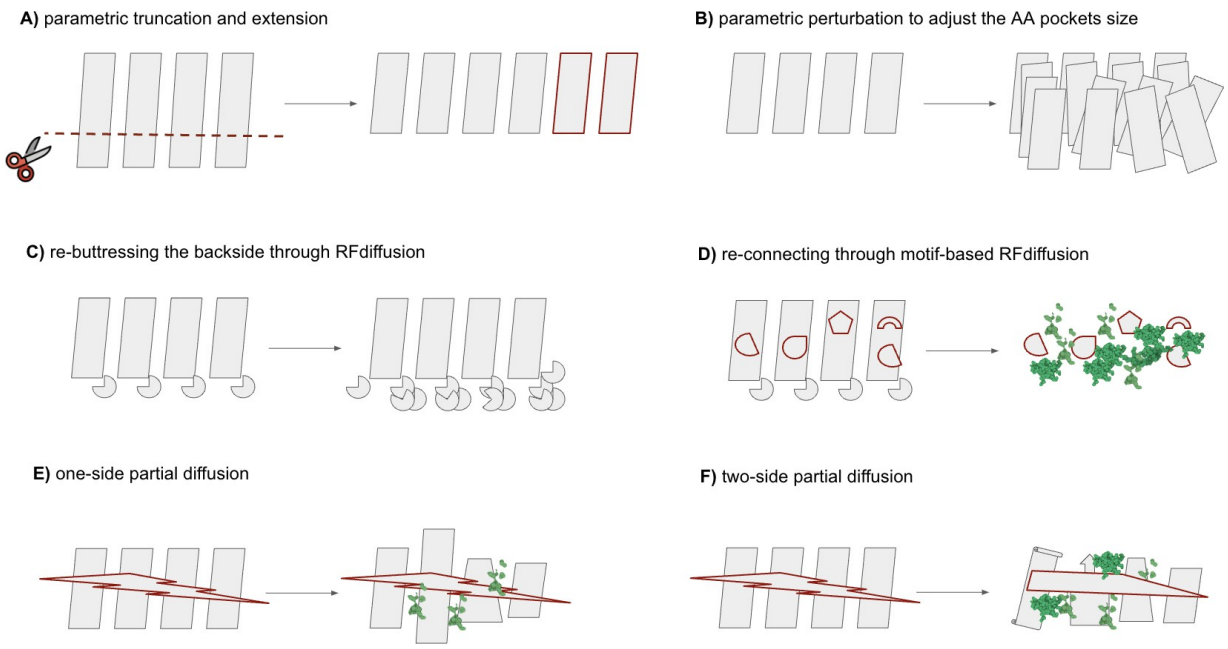
side chain - peptide side chain hydrogen bonding networks (shown in yellow dotted lines) for affinity and specificity.

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

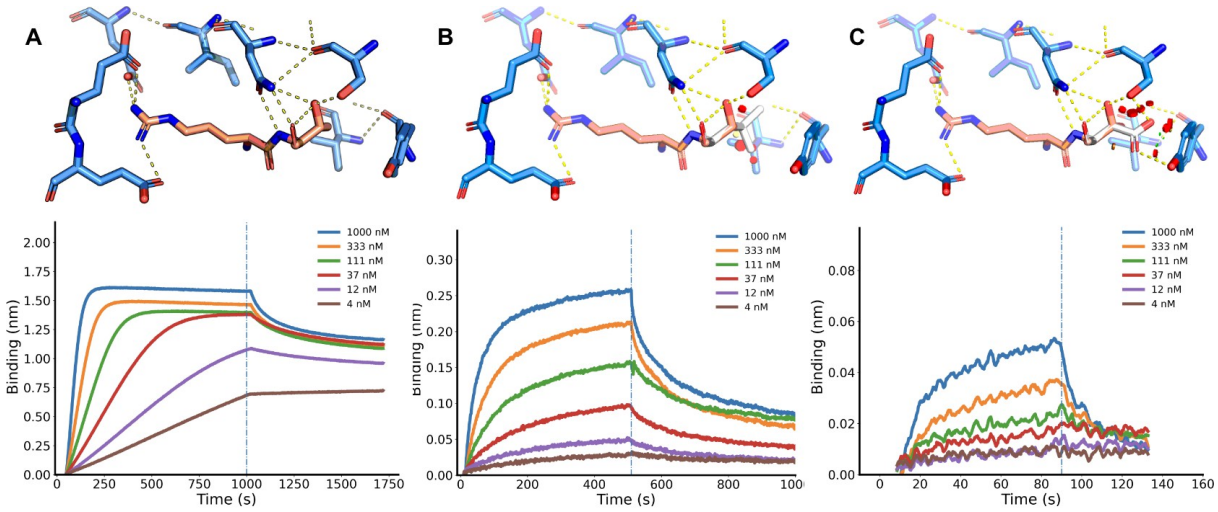
Extended Data Table. 2: Summary of computational and experimental protocols. To summarize, binders to 25/39 targets have not been experimentally or computationally optimized, and produced straight out of the computer through the base design pipeline. Binders to 2/39 targets have not been experimentally 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. 3: 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. 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.

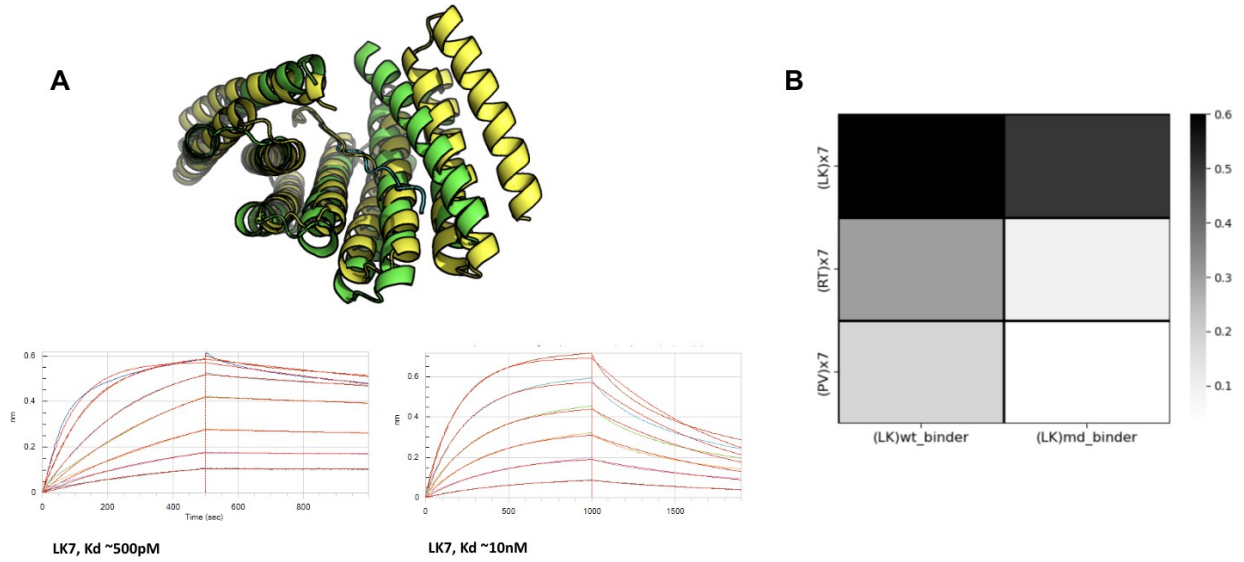


Extended Data Fig. 4: Template backbone optimization approaches. Multiple approaches were applied to generate the library templates during pocket design. **A)** Original four-repeat proteins with long helices were truncated by helical length and extended by repeating geometry parametrically to generate 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].



(RT)x7, Kd ~5nM (RL)x7, Kd ~67nM (RD)x7, Kd NA

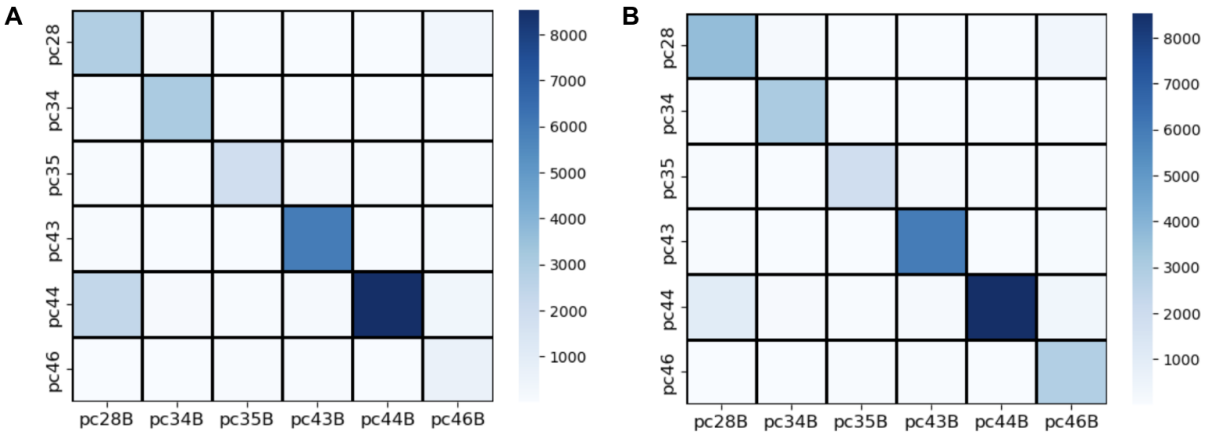
Extended Data Fig. 5: Repeat di-peptide (RT) binder showed selectivity between threonine, aspartate, and leucine by BLI. Bio-layer interferometry characterization for the designed (RT) binder titrating against three closely related peptide targets, (RT)x7, (RL)x7, (RD)x7. Three-fold serial dilutions 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	

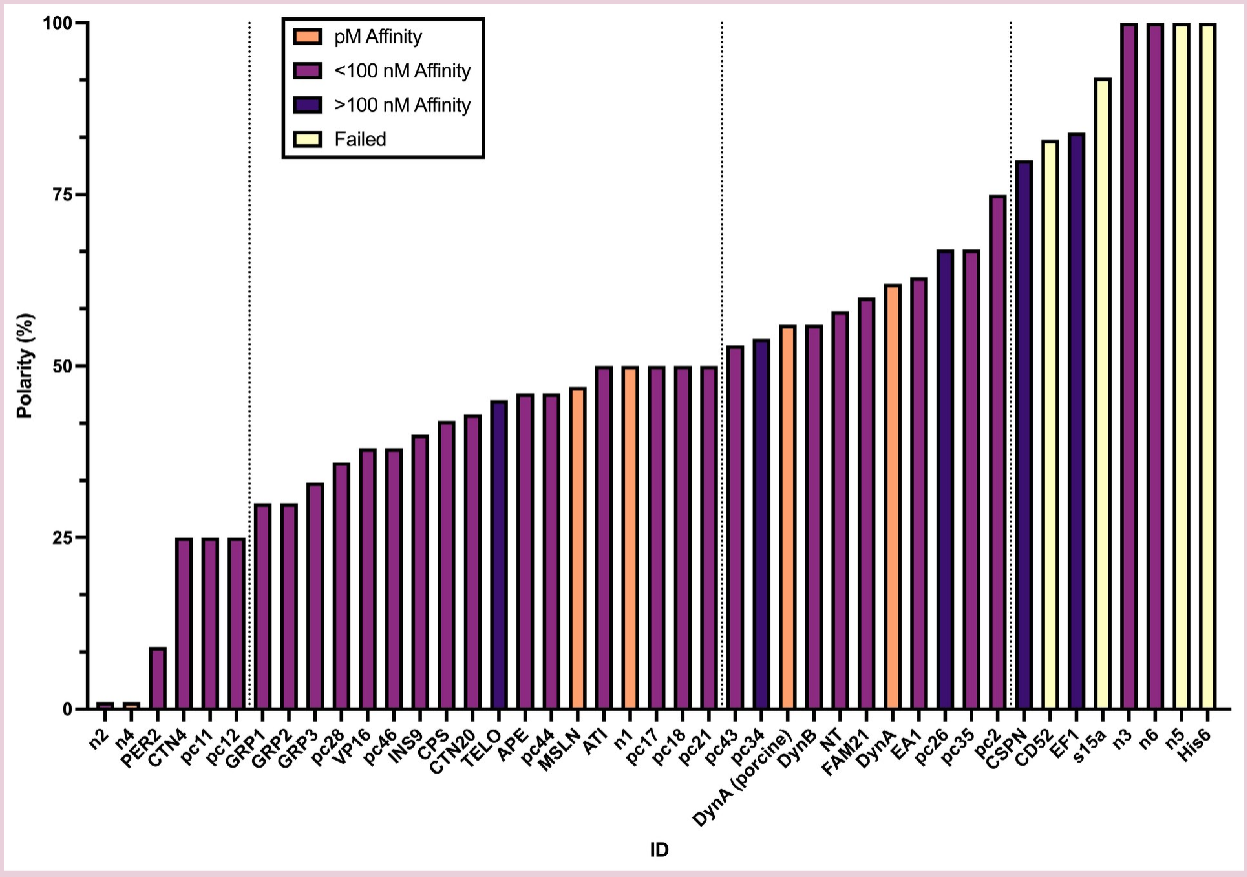
Extended Data Table 7: Recognition pockets (base and related) generated in this work. Collection of all pockets generated in this work, with either experimentally validated or computationally predicted confidence in context of disordered targets binding. The related pockets were structurally grouped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



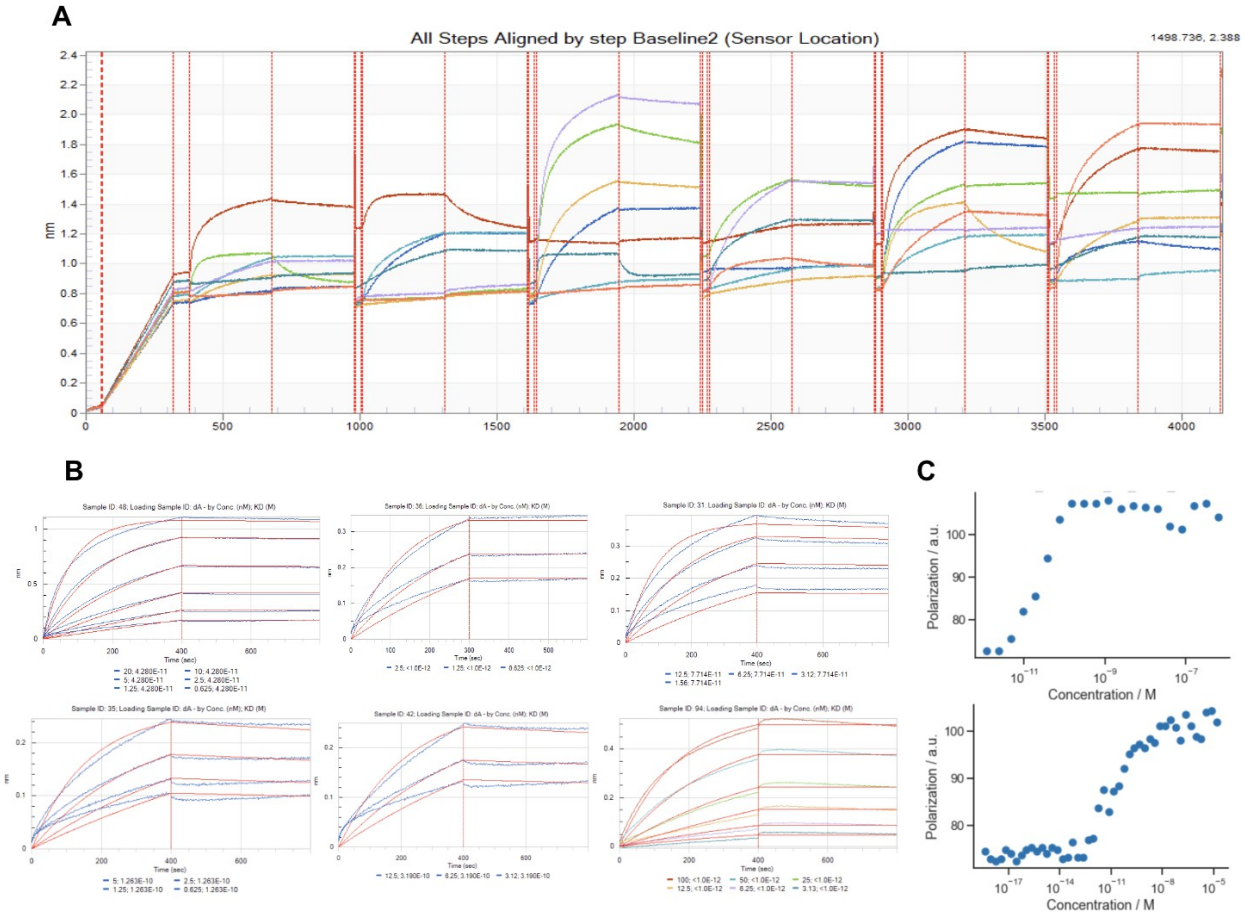
C

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

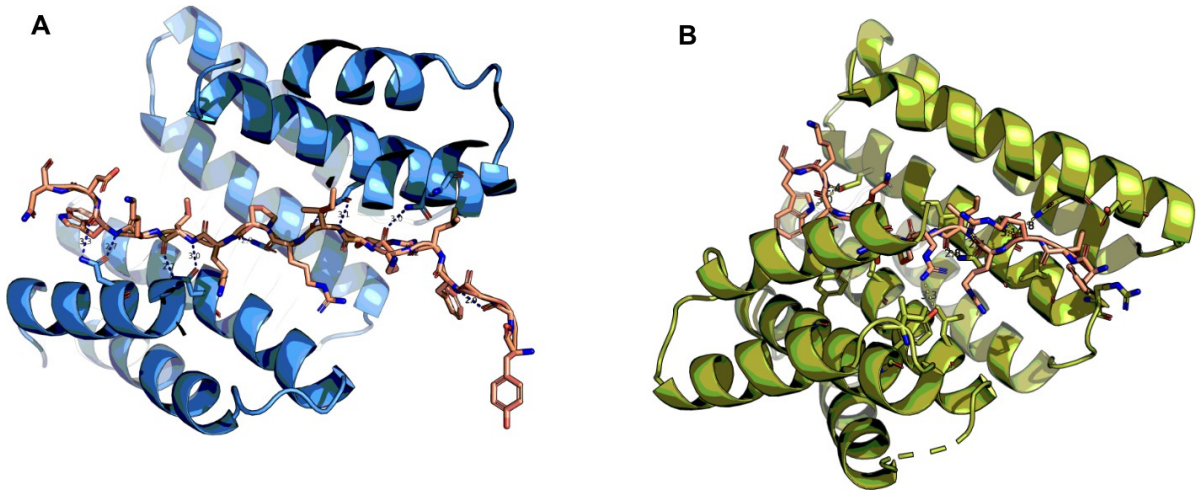
Extended Data Fig. 8: Orthogonality improved between targets of arbitrary English words through testing a slightly bigger set of binders (from 8 on average to 20 on average). Spells out of the English words are as follows: DANIELSILVA (pc28), HIALIENFRIEND (pc34), DERRICKHICKS (pc35), DAVIDLIKESPEPTIDE (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.



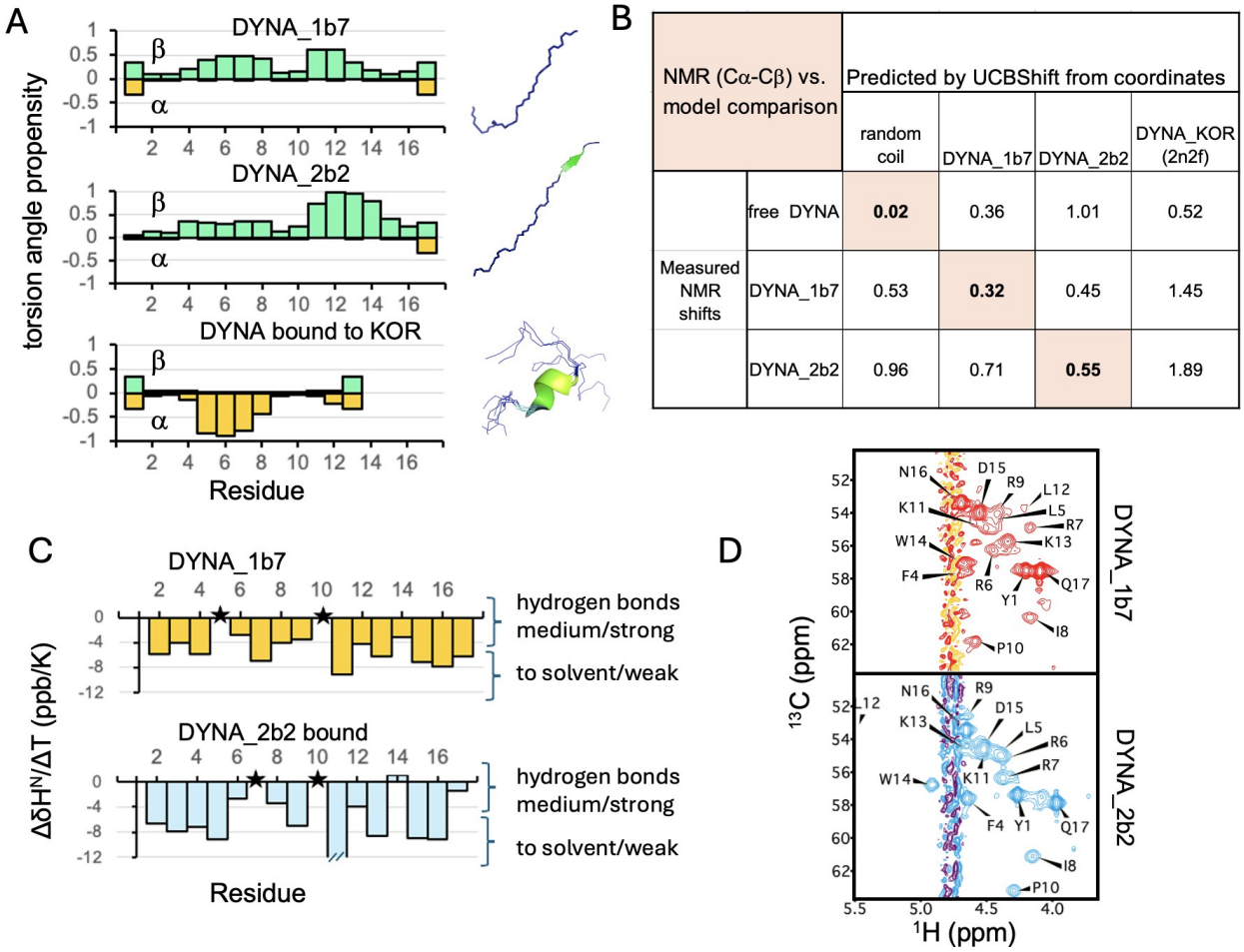
Extended Data Fig. 9: Target sequence polarity distribution on the 21 native targets and 18 synthetic targets. Polarity was calculated based on the percentage of polar and charged amino acids (D, E, H, K, N, Q, R, S, T, Y) among the targeted window of each target.



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).

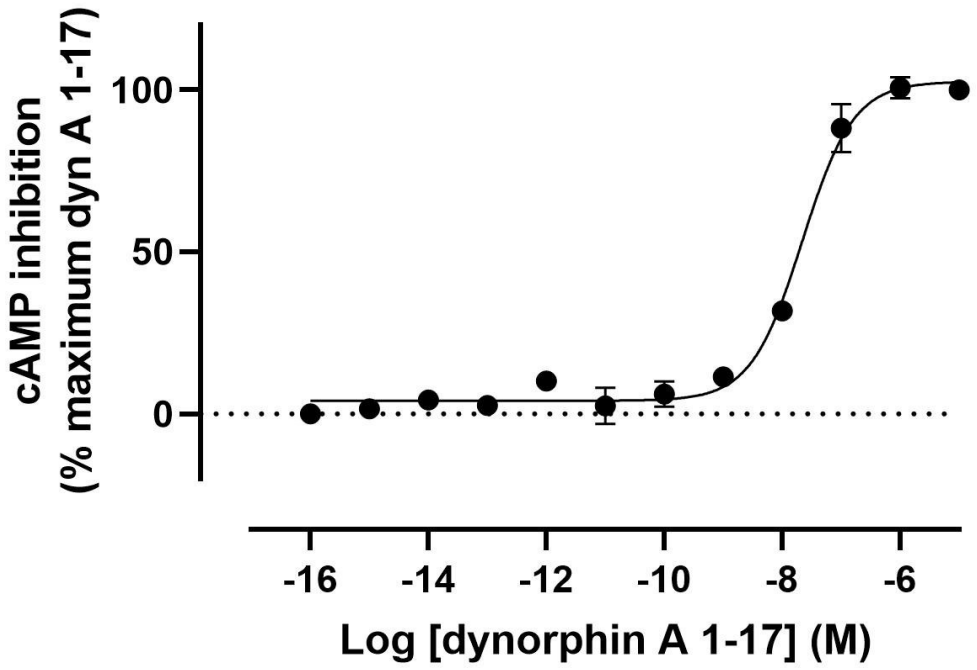


Extended Data Fig. 11: Comparison of the structure of A) initial design DYNA_1b1 to B) two-sided partial diffusion refined design DYNA_1b7. During partial diffusion, the binding mode was further changed and diversified on both the binder and the target.

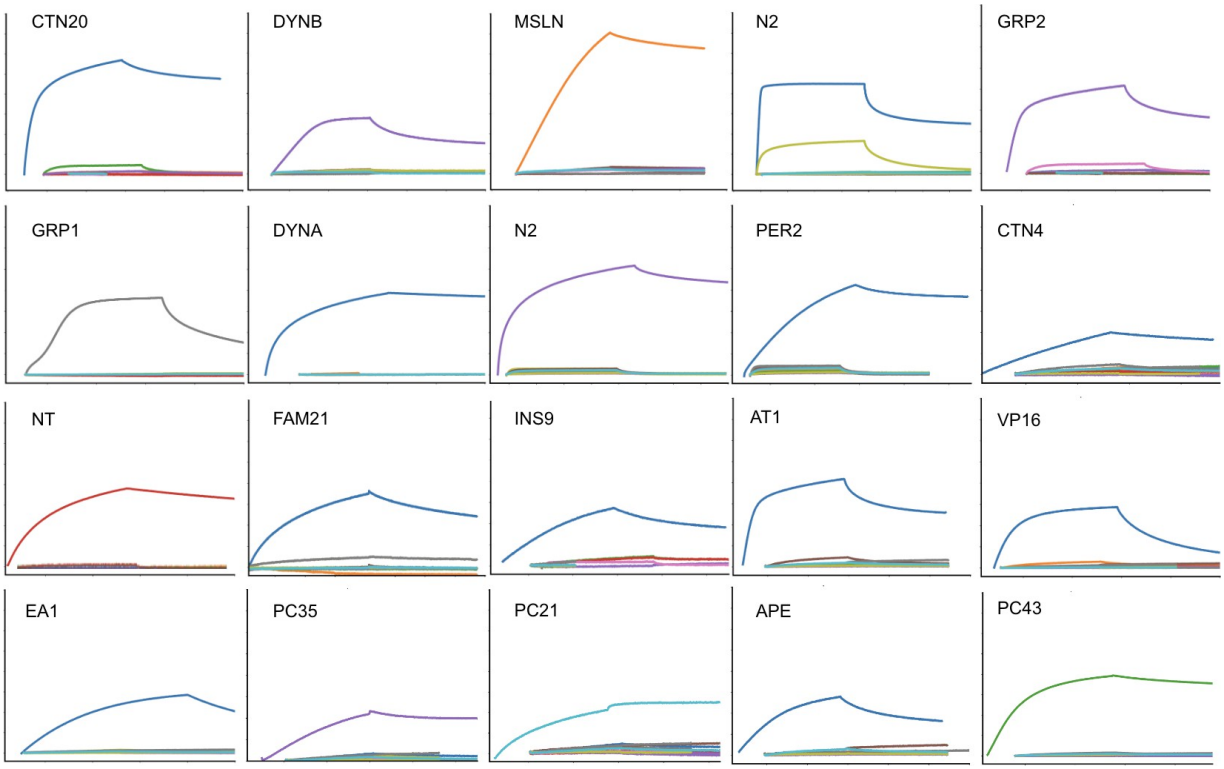


Extended Data Fig. 12: NMR analysis on dynorphin A complexes. Dynorphin A binds as extended and partial beta strands as designed. **A)** NMR chemical shift-based analysis using Talos-N of isotope-enriched dynorphins bound to DYNA_1b7 and DYNA_2b2 distinguish helical (α) from extended (β) conformations 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. 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.

Methods 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**

10) Purification of KFAM binders
11) Immunoprecipitation with KFAM binders
12) Dynorphin A inhibition cAMP assay
13) Affinity enrichment of CTN4 analyzed by LC-MS
14) NMR sample preparation and data collection of dynorphin-binder complex data
15) 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 [ref]. 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 central 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 to supplementary file S3 and through https://github.com/drhicks/Kejia_peptide_binders , expected release upon journal acceptance).

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
path_to/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. path_to/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 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.

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 were 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°C, 200 RPM. Bacterial culture was induced with 0.25 mM of IPTG at 18°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°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 was 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).