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

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. 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 specific 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 to repeating peptides to a broader range of peptide conformations (previously we used it only for proline-rich peptide sequences with a three-residue repeat in polyproline II conformations). In this approach, repeating conformations of the peptide target sequence are generated, the superhelical parameters are determined, repeat proteins with matching superhelical parameters are generated, and the two are 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 both 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 (see 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 succeeds because polar interactions have considerably much more strict 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 as described above 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 with the designs described above, 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 it 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. 20 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] (Figure 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 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 which 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.

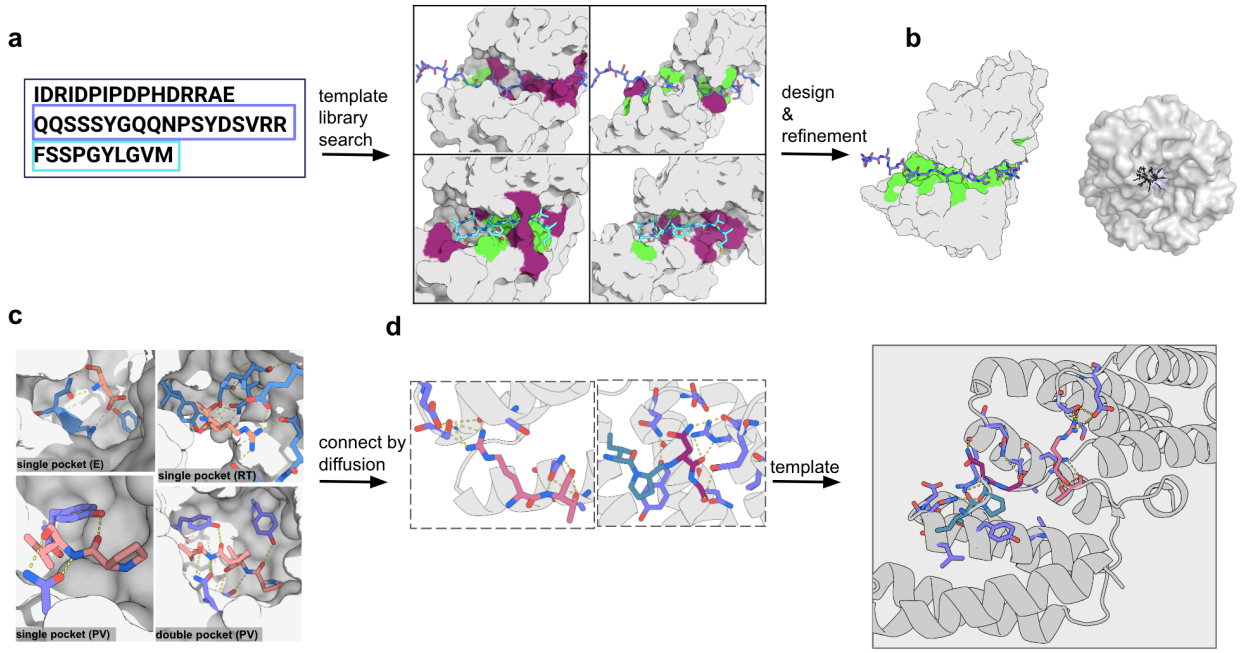
contributed conceptual support for the computation idea at the beginning. D.B. supervised the research. K.W. and D.B. drafted the manuscript. All authors reviewed and commented on the manuscript.

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 models can be found in supplementary file S2. Design scripts are available in supplementary file S3 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.

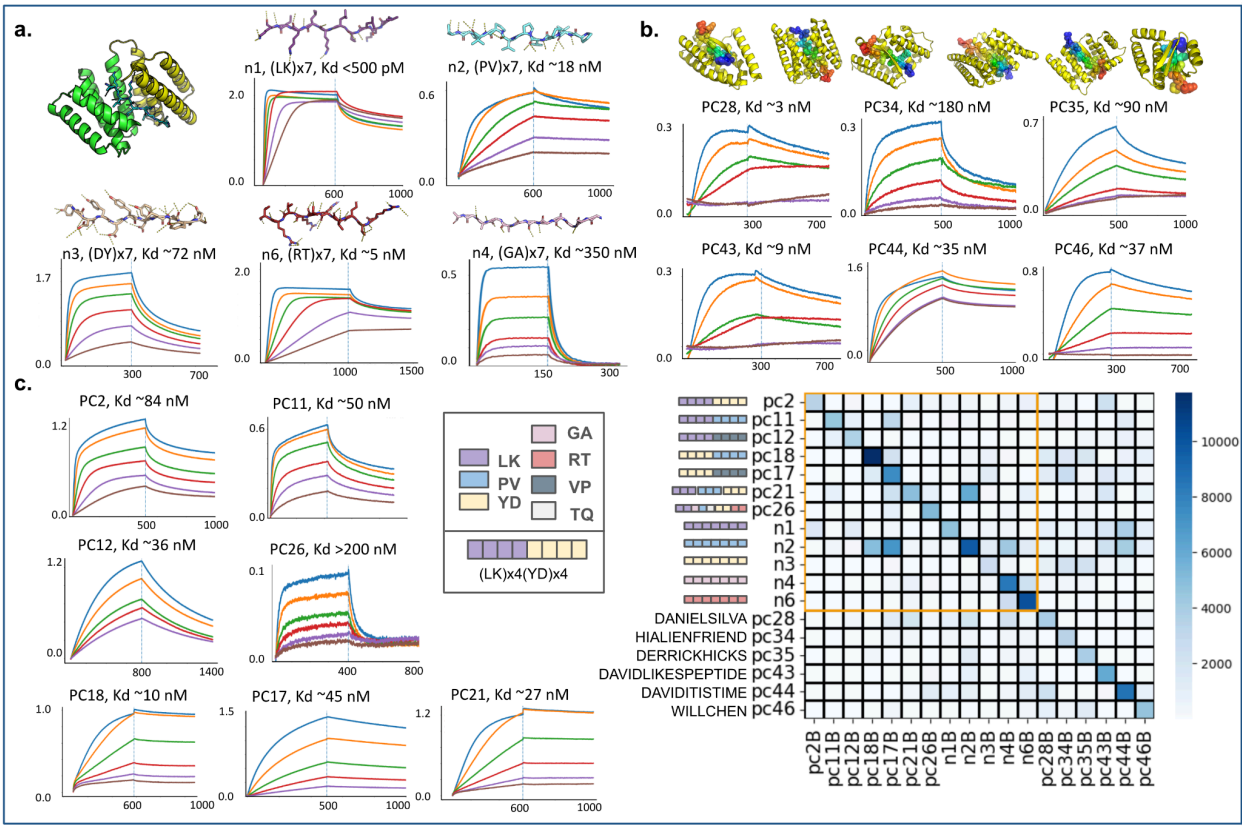
Figures

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 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**.

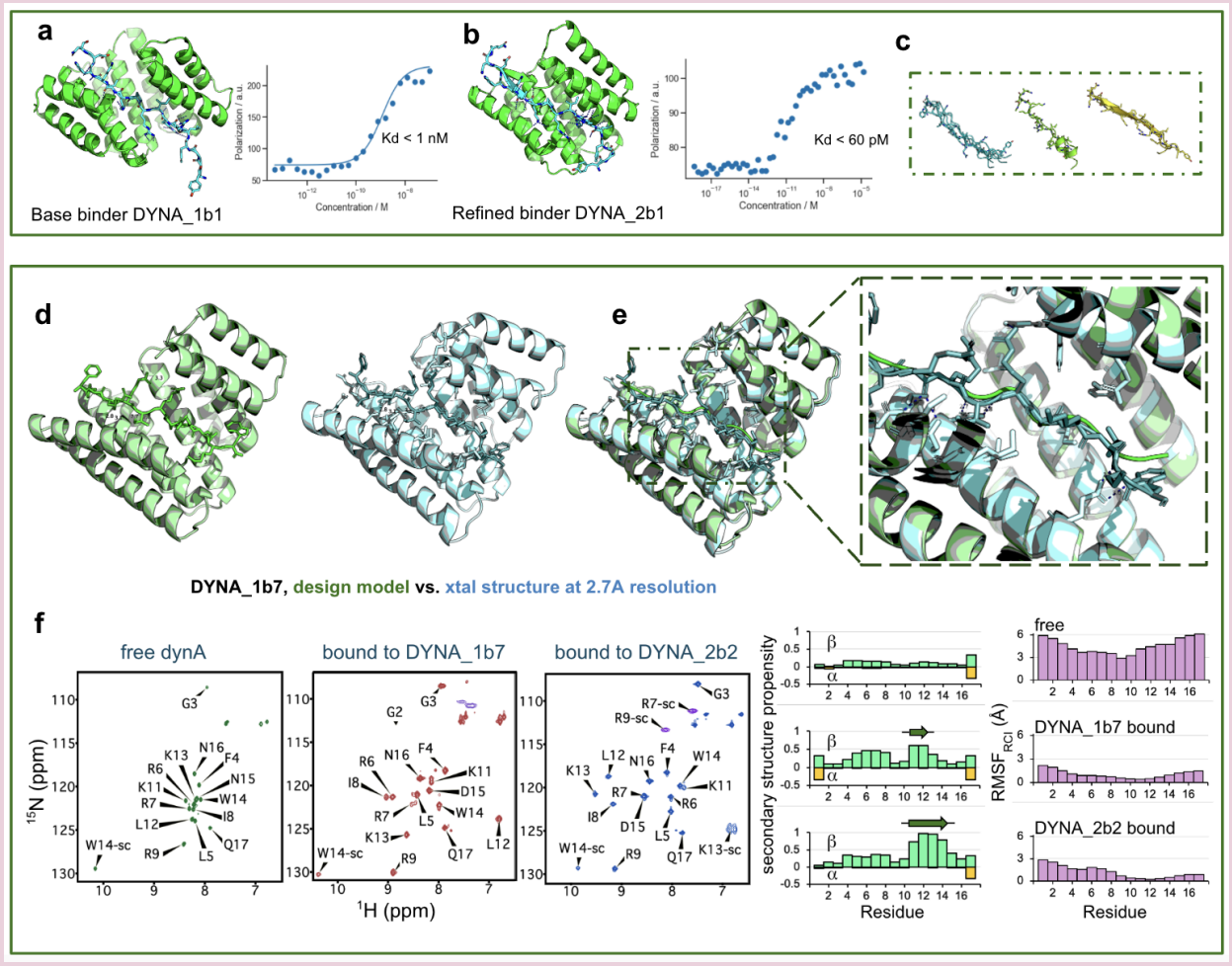
Fig. 2. Design of 18 synthetic peptide binding proteins.



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.

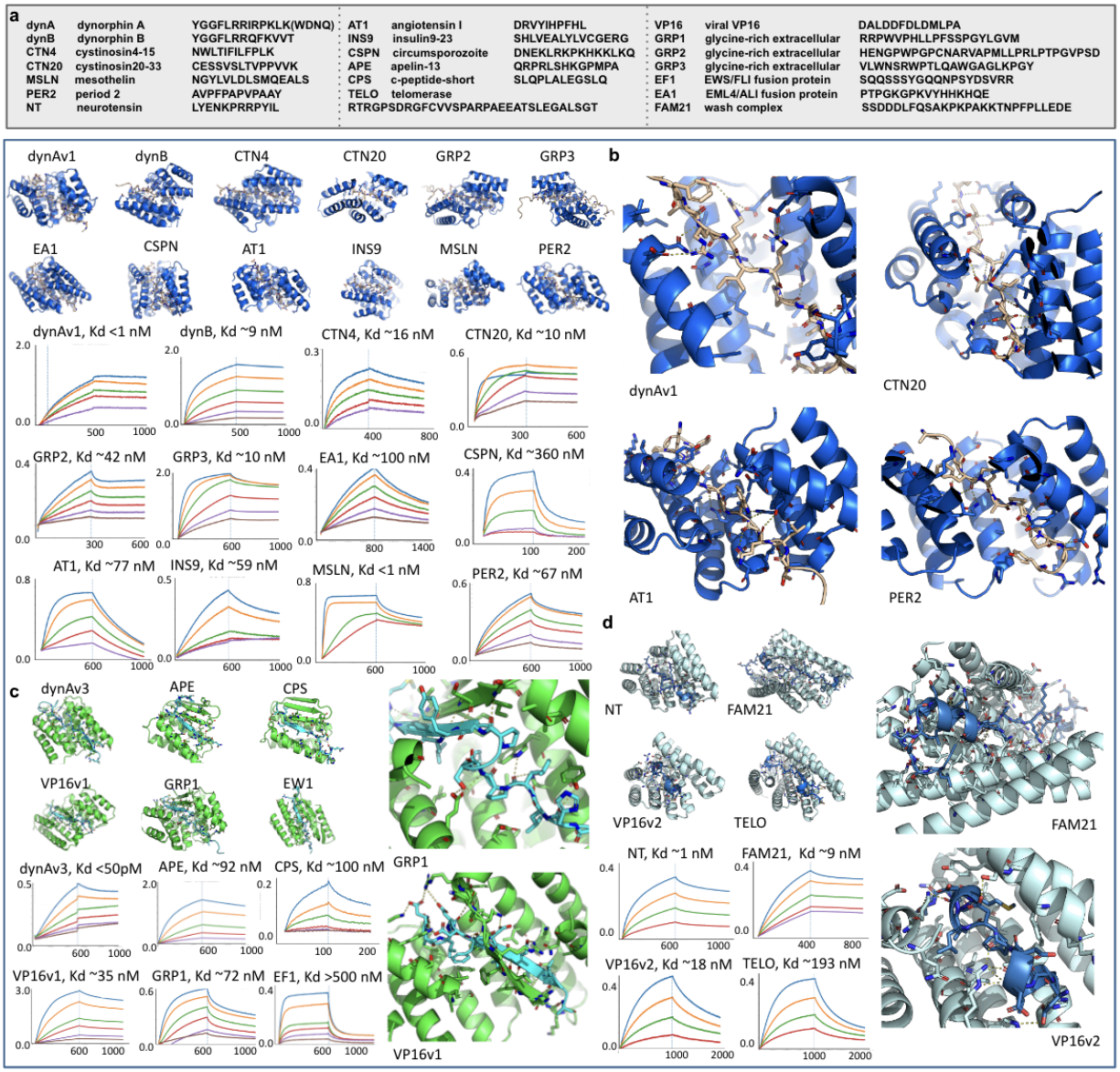
Fig. 3. Structural characterization of dynorphin A binder designs.



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. Inset on 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.

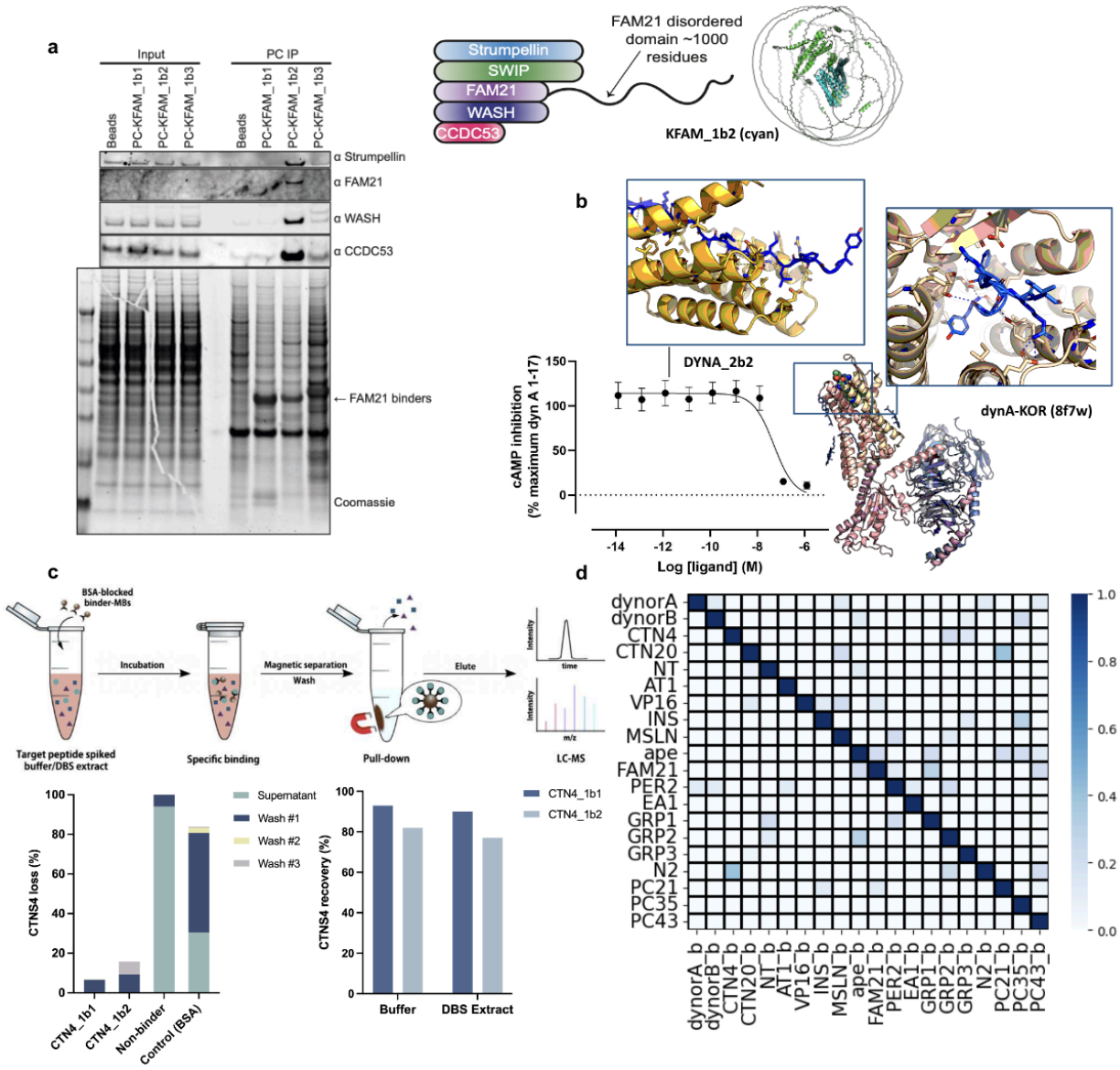
Fig. 4. Design of binders to 21 native protein disordered regions and peptides.



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=https:%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2212137421000592%3Fshowall%3Dtrue&referrer=.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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