# Design of intrinsically disordered region binding proteins

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], Sebastian Kenny[1,2], Andrei Mihut[7], Amit Gaur[5], Brian Coventry[1,2,10], 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], Zihao Song[1,2], Gudrun Schober[8,9], Stuart M. Brierley[8,9], John O'Neill[7], 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

Intrinsically disordered proteins and peptides play key roles in biology, but the lack of defined structures and the high variability in sequence and conformational preferences has made targeting such systems challenging. We describe a general approach for designing proteins that bind intrinsically disordered protein regions in diverse extended conformations with side chains fitting into complementary binding pockets. We used the approach to design binders for 39 highly diverse unstructured targets and obtain designs with pM to 100 nM affinities in 34 cases, testing ~22 designs per target (including polar targets). The designs function in cells and as detection reagents, and are specific for their intended targets in all-by-all binding experiments. Our approach is a major step towards a general solution to the intrinsically disordered protein and peptide recognition problem.

## Main

Natural evolution has generated a variety of solutions to the challenge of binding unstructured regions of peptides and intrinsically disordered proteins (IDPs)(*1*–*6*), including natural antibodies(*1*–*3*), the major histocompatibility complexes(*4*), tetratricopeptide repeats(*5*), Armadillo repeat proteins(*6*), and lipocalins (Anticalins). Despite this diversity, engineering general peptide recognition remains challenging: peptide-specific antibodies have been obtained by immunization or by library selection, but this requires considerable effort, and disordered antigens are susceptible to degradation following injection. There has been progress in generalizing the binding modes of armadillo repeat and other natural peptide-binding proteins(*3*–*5*), but achieving completely new specificities has been challenging. De novo protein design methods have been used to design proteins that bind peptides in polyproline II, alpha-helical and beta-strand conformations(*7*–*9*), but 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 propensity for the same secondary structure throughout, or present suitable interfaces for binding in regular secondary structures. For example, amphipathic helices or strands can be recognized using designs with grooves that bind primarily to the non-polar side of the helix or strand, but if charged residues are distributed around the helix or strand axis, this binding mode would require energetically unfavorable charge burial. A method for achieving specific recognition of any target intrinsically disordered region (IDR)

sequence of interest would be broadly useful for applications in proteomics, targeting, sensing, and sequencing.

We reasoned that a general solution to the IDR binding problem might be achieved by combining the strengths of physical and deep learning design approaches. Rosetta design methods have been used to design binding proteins consisting of four to six tandemly repeated structural units that bind repeating proline-rich sequences in the polyproline II conformation, with each repeat unit in the designed binder interacting with a repeat unit in the peptide (Fig. 1a). Generative deep learning free diffusion methods(*8*, *10*) have been used to generate binding proteins with no such pre-imposed structural constraints on the binder structure (Fig. 1b), but because the model is trained on the PDB, this procedure generally folds the target sequence into the alpha-helical or beta-sheet conformations that dominate the PDB, which as noted above can be poorly compatible with binding. The repeat protein-repeat peptide approach can, in principle, be generalized beyond polyproline II conformations, but is limited by the requirement that each peptide unit have the same conformation, which again may be incompatible with heterogeneous target sequences. We reasoned that starting from different repeat protein architectures and recombining and specializing amino acid binding pockets in different repeat units for different amino acids and different conformations using diffusion could yield a family of templates enabling more general recognition of sequences with widely varying conformational preferences and sequences (Fig. 1c). In the following sections, we first describe the creation of such a scaffold library for general peptide recognition (Fig. 1d-f), and then the use of the library to design binders for a wide variety of non-repeating, both synthetic and native IDR targets (Fig. 1g-j).

## Generation of Template Library

We reasoned that the template library should have two properties. First, each template structure should “wrap” around extended peptide conformations with numerous opportunities for the hydrogen bonding and packing interactions with the target required for high specificity (Fig. 1d). Second, the structural variation in the template family should be sufficiently broad that for any target sequence at least one of the templates is able to induce it into a defined binding-competent conformation.

We developed a three step approach to generating such a library of template structures suitable for general recognition. In the first “scaffold generation” step (Fig. 1d; Methods I), we design repeat proteins that wrap around peptides in different repeating conformations, such that each repeat unit on the protein forms a binding pocket that interacts with a corresponding repeat unit on the peptide. 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. In the second “pocket specialization” step (Fig. 1e; Methods II), we fine-tune these pockets using diffusion to achieve more precise matching to specific target peptide sequences. We keep the four to nine amino acids surrounding each sidechain bidentate hydrogen bond from the repeat protein to the peptide backbone fixed (Fig. S1) while diversifying the hydrophobic interactions between designed binder; this is advantageous because hydrogen bonding interactions have more stringent geometric requirements than non-polar packing interactions and hence are more efficiently templated than repeatedly sampled from scratch (see Methods II-4). In the third “pocket assembly” step, we go beyond the limitations of repeating structures, which are optimal for repeating sequences but not more general sequence targets, by recombining pockets from different designs, using RFdiffusion(*11*) to generate interfaces between them where necessary to yield overall rigid structures (Fig. 1f). This generates a set of templates with diverse pockets arranged in different orders and geometries (Fig. 1f; Methods III).

For the first “scaffold generation” step, we chose to target peptides in a broad range of extended conformations rather than solely the polyproline II conformation as in our earlier study, as this is populated primarily by proline-rich peptides. In extended conformations, alternating side chains face in opposite directions, consistent with a two-residue sequence repeat. We used Rosetta design methods as described previously to generate designs targeting the dipeptide repeats LK, RT, YD, PV, and GA (single-letter amino acid codes) in a variety of extended conformations that wrap around these peptides such that each repeat unit interacts with one dipeptide unit (Fig. 2a; Methods I). Experimental characterization by fluorescence polarization of four-repeat versions of the designed binders revealed nanomolar binding for the LK and PV repeat peptides, but little binding for the more polar RT and highly flexible GA (Fig. S2; to avoid potentially unfavorable interactions with peptide termini, for experimental testing here and below, we pad all repeat target peptides with two additional repeats).

For the second “pocket specialization” step, we refined the designed binding pockets to improve the fit to the target sequences and extended the number of interacting repeat units from four to five to further increase affinity. This yielded designs with picomolar affinities for LK repeats, and low nanomolar affinities for RT and GA repeats (Fig. 2b-c; Fig. S3). The binding pockets in these designs have distinct geometries that are customized to the target being recognized (see example in Fig. S4).

For the third “pocket assembly” step, we enable more general recognition of non-repeating sequences by assembling the binding pockets into new backbones, keeping them positioned to interact with peptide targets in continuous extended

conformations (Fig. 2d-e). We assembled combinations of two to six binding pockets in silico (see below and Methods III), yielding models of chimeric proteins interacting with chimeric peptide targets. To do this, we positioned pockets parametrically (see Methods III-1), and connected them with RFdiffusion. We used this approach to generate 70 designs against seven chimeric targets. 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. Experimental characterization using nanoBiT split luciferase reconstitution(*12*) 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; Fig. S4).

To increase the size of the template library to cover a broader range of sequences, 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 silico (see Methods III-3; in the designs described above, the peptide sequence was always held constant). Together, this yielded a library of 1,000 templates, each consisting of a designed binding protein and a corresponding peptide backbone positioned such that the amino acids in the peptide fit into designed pockets in the binding protein (representative examples shown in Fig. S5).

## Threading Intrinsically Disordered Regions onto Template Library

We developed a two step approach for using the template library to generate binders for non-repeating synthetic sequences and arbitrary native unstructured targets. In the first “threading” step (Fig. 1g), the target sequence is threaded through the backbone of each template to identify the most compatible sequence segment-template pairs. In the second “refinement” step, the best matches are refined to increase the fit between the designed binder and target peptide (Fig. 1h-i).

For an IDP or 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 first discard segments with low sequence complexity and/or with multiple close matches in the proteome (Fig. 1g, S6, and S7), as binders to such targets would likely have some cross-reactivity. We map each of the remaining unique sequence segments of eight to 40 amino acids onto each of the target backbones in the library, carry out local backbone resampling, optimize the sequence of the binder using ProteinMPNN, and evaluate the designs based on the fit between the designed binder and the target sequence and the agreement between the AF2

prediction and the design model (see Methods IV). 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 V) to customize the backbone for the specific target.

We first tested this approach on synthetic sequences corresponding to six arbitrarily selected English words and names. We tested 45 designs against six targets, eight designs per target on average; the best binders for two out of the six targets had single-digit nanomolar affinities ($K_d$s of 9 nM), three had double-digit affinities ($K_d$ = 35 nM, 37 nM, 90 nM), and one a $K_d$ = 180 nM (Fig. 2b; Fig. S8). We investigated the selectivity of the designs for their peptide targets by carrying out all-by-all (18 by 18) nanoBiT interaction measurements, including the repeat sequence binders of the previous section. Although 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, the designs were quite orthogonal (Fig. 2c; Table S1).

We next used the threading approach to generate binders for 21 diverse therapeutically relevant IDPs, IDRs, and segments of IDPs ranging from eight to 40 amino acids. These include eight GPCR ligands, two insulin-related ligands, four disease detection-related disordered regions, four IDRs from cancer-related receptors, and three human scaffolding complexes for which there are no good monoclonal antibodies (target names and targeted sequences in Fig. 3a). For each target, three to 48 designs (on average 28) for which the AF2 predicted structures of the complexes were close to the computational design models (Fig. S8) were selected for experimental characterization. The designs have overall helical architectures reflecting their helical repeat protein origins (Fig. S9), with considerable structural variation in some regions introduced by the diffusion assembly and refinement steps, while the target sequences adopt a wide range of random coil and partially helical and strand conformations (Fig. 3b-c; Fig. S10).

The designed binders were expressed and purified, and binding to the targeted IDRs was measured by BLI with the peptide attached to the sensor chip and the binder in solution. The fraction of binders that showed a binding signal at 500 nM ranged from 2-67% (Fig. 3a). Together with the 18 synthetic targets, we obtained binders for 39 of 43 targets, testing on average 22 designs for each target (Table S2).

Polar targets have been long considered challenging in protein design. Twenty of the targets had more than 50% polar residues, and six had more than 75%; binders were obtained for the fusion fragment EF1 of EWS/FLI onco-fusion protein for Ewing sarcoma(*13–15*) (84% polar residues), and the N-terminal fusion fragment CSP-N of

Circumsporozoite protein (CSP) for malaria(*16*) (80% polar residues) (see Fig. 3d for target polarity distribution). The number of hydrogen bonds to the side chains of the target (per 10-amino-acid segment) was three-fold higher for the highly polar targets than the other targets (4.3 for target polarity <75%; 12.3 for target polarity >=75%). 77% of the targets had little predicted secondary structural propensity in isolation; 87% adopted predicted bound conformations lacking extensive secondary structure, and all adopted conformations were very different than in previously solved crystal structures in cases where these were available (Fig. 3e; Table S3). Binding affinities were determined by global fitting of BLI association and dissociation phases at a range of binder concentrations. There was little correlation between binding affinity and the polarity or intrinsic secondary structure of the target; Kds less than 10 nM were achieved for targets with a wide range of polarities and secondary structure propensities (Fig. 3d-f; Table S4), indicating the generality of the method.

To explore the optimization potential of the designs, we chose a binder, DYNA_1b1, with a Kd of ~1nM for dynorphin, a kappa opioid receptor (KOR) peptide ligand implicated in chronic pain(*17*, *18*). We used RFdiffusion refinement on top hits as described above for the synthetic targets (Fig. 4a, 2c, and Methods V). Forty-five out of 48 designs showed strong binding in the BLI screening assay at 5 nM, and six had $K_d$ <= 100 pM by BLI; fluorescence polarization measurements for two of these optimized designs, DYNA_2b1 and DYNA_2b2, indicated $K_d$s <60 pM and <200 pM respectively (Fig. 4b; Fig. S11). Over the set of original and optimized designs for dynorphin A, the peptide populated a wide diversity of random coil, partial strand, and partial helix conformations (Fig. 4c). The dynorphin A binder and dynorphin B binder were orthogonal, binding only to their intended targets (see below), despite having 62% sequence alignment.

## 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 3.15 Å resolution (Fig. 4d-e; Fig. S12). The backbones of both the protein and peptide in the crystal structure match the design model well, with an interface backbone RMSD of 1.2 Å for the complex and interface sidechain RMSD of 2.9 Å (Fig. 4d-e). The key interactions are in the central region of the peptide (Fig. 4e): during design, we excluded the N-terminal YGGF sequence, which is shared between dynorphin A and B and other neuropeptides(*19*), 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 and hence was not targeted. For this design, the two-sided diffusion refinement eliminated several of the asparagine-peptide backbone bidentate interactions, which could account for the decrease in binding affinity from ~1 nM for the starting design to 7 nM (Fig. S12; see

Fig. S13 for the peptide density maps). The crystal structure confirmed that all the hydrogen bonds in the design model of DYNA_1b7 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, and density was missing for the termini that were not included in the design calculations (YGG- and -DNQ).

To investigate changes in dynorphin structure upon binding, we examined the NMR spectra of isotope-labeled dynorphin A unbound in solution, bound to DYNA_1b7 ($K_d$ = 7 nM), and bound to the higher affinity design DYNA_2b2 ($K_d$ <200 pM; Fig. 4f and Fig. S11). NMR confirmed that free dynorphin A is intrinsically disordered and becomes ordered upon binding except for the regions not included in the design (Fig. 4f). For both bound complexes, the NMR data indicated an extended bound-state conformation, consistent with the design models (Fig. 4f and Fig. S14). The extent of ordering upon binding to DYNA_2b2 was greater than for DYNA_1b7 in both the C-terminal region and around the TRP14-ASN137 bidentate interaction, consistent with the more extensive sidechain-backbone hydrogen bonding in the former (Fig. 4f and Fig. S14).

The extended conformation of the dynorphin peptide in the designed complexes, confirmed by the X-ray and NMR data, is considerably different from any previously solved cryoEM or NMR structures of dynorphin with native KOR (Fig. S14), where it is bound in a compact, partial helix conformation (PDB ID 2n2f, 8f7w)(*20*, *21*). These data highlight the power of computational design for inducing disordered proteins and peptides into non-native conformations.

To assess the contributions of each target residue to binding, we carried out alanine scanning experiments on the dynorphin A binder DYNA_1b1 - dynorphin interaction. In nanoBiT binding experiments, each of nine alanine substitutions on the peptide that disrupt interactions with the binder considerably reduced the extent of binding compared to the wild-type dynorphin peptide (Fig. S15).

## Applications of Designed Binders

Designed binders with high affinity could be useful as enrichment reagents for a broad range of low-abundance human proteins, particularly those involved in signaling pathways. We tested this using as a model the WASH complex, a pentameric complex responsible for the nucleation of branched actin on endosomes(*22*, *23*), and the PER complex involved in circadian clock function. The WASH complex includes WASH (WASHC1), FAM21 (WASHC2), CCDC53 (WASHC3), SWIP (WASHC4), and

Strumpellin (WASHC5)(*22*, *23*). FAM21 contains a C-terminal disordered region of about 1,000 residues in length involved in multiple protein:protein interactions (Fig. 5a). We designed three binders for a 27-amino acid disordered window in FAM21 (Fig. 3a). Immunoprecipitation studies showed that FAM21_1b1, retrieved the entire WASH complex from cell lysate (Fig. 5a; Fig. S16). Similarly, a designed binder (PER2_1b1) to a disordered region of PER2 (a phase separating circadian clock protein) enriched endogenous PER2 from cell lysates in pull down experiments (Fig. S17)(*24*, *25*). 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.

We explored using the designed binders in affinity enrichment coupled with LC-MS for detecting low-abundance peptides generated in proteolytic digests of the proteome (Fig. 5b), such as those from poorly folded mutant versions associated with disease, which can be very difficult to detect by mass spectrometry. To evaluate this, we chose a 12 amino acid tryptic peptide of a mutant form of the lysosomal cystine transporter cystinosin protein implicated in cystinosis, a lysosomal storage disease(*26*, *27*). Binder CTN4_1b1 targeting CTN4 was 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 (Fig. 5b; Fig. S18; Table S5), a higher recovery than achieved with previously described helical peptide binders(*8*).

Mesothelin is a cell surface glycoprotein upregulated in many cancers that is of considerable interest for tumor targeting (*28*). We investigated whether a designed binder (MSLN_1b1) to a juxtamembrane region of Mesothelin (MSLN) could specifically bind to cells expressing the target (proteolytic cleavage in this region makes more distal regions of the extracellular domain less useful for targeting). We incubated GFP-MSLN_1b1 fusions with cells expressing MSLN (HPAC) and cell lines not expressing MSLN (MCF7), along with a GFP-fusion to a control protein that does not bind MSLN (Fig. 5c). Fluorescence microscopy showed GFP localization of MSLN_1b1 at cell junctions, as expected for MSLN, on HPAC but not MCF7 control cells; no binding to HPAC cells was observed with the control binder (Fig. 5c). Thus MSLN_1b1 specifically binds MSLN on the cell surface.

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(*18*, *20*). To explore the potential of our binders to block KOR signaling mediated by dynorphin A (Fig. S19), 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. 5d; the IC50 is higher than the Kd because of competition with KOR binding). As noted above, to increase specificity, we excluded the N-terminal YGGF sequence during design to distinguish between dynorphin A and B; this region is critical for opioid receptor activation(*29*) (Fig. S20), and extension to include the YGGF- motif would likely increase potency.

## Binder Orthogonality

We 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, and observed little crossreactivity at concentrations up to 1 uM. Within the set of disordered targets considered here, each design thus only binds tightly to the target it was designed to bind (Fig. 5f and Fig. S21).

We investigated the specificity of interactions between the designed binders in cells using a mitochondria colocalization assay(*7*). We expressed six of the designed binders fused to sfGFP(*30*), and the six targeted disordered sequences fused to mCherry and a mitochondrial outer membrane targeting sequence. Each designed binder was expressed with each target one at a time, and binding was evaluated by localization of GFP fluorescence to the mitochondria. We observed localization of the GFP to the mitochondrial for each on-target pair (designed binder with its intended target), but not for any off-target pairs (Fig. 5e), indicating that the designed binders function in cells. As further in cell off-target controls, two sequence homologues of DYNA were also tested and found not to colocalize with the DYNA binder at all (see cartoons in Fig. 5e; and Experimental Methods).

## Discussion

We demonstrate that the conformational heterogeneity of IDPs and proteins can be exploited to make the binder design problem easier than for traditional stable folded targets. For each target, we sampled a wide variety of conformations and identified those compatible with high affinity binding—this contrasts with folded targets, whose fixed structures may admit few optimal binding solutions. Our designs induced the target disordered region to adopt bound structures completely different from those populated

in solution or present in previously solved native complexes. Induced fit is a general feature of disordered protein binding interactions in nature(*31–33*), and by taking advantage of induced fit, our approach of threading through a diverse extended scaffold set followed by diffusion-based refinement enables robust computational design of binders to a wide range of disordered sequences, including highly polar, challenging sequences. The design method, which we call *"logos*," has high computational efficiency and a high experimental success rate (22 designs were tested on average for each case described here), and should be broadly useful for making binders to arbitrary disordered targets of interest (Fig. S22 provides a step by step example of structural components progression of a design campaign). Beyond the examples presented here, the *logos* method has been used to generate binders to Hras, Kras-A, and Kras-B (three polar isoforms out of the four ras isoforms in total), which have high specificity in cells(*34*). More generally, the combination of physically based design approaches for generating a set of privileged starting scaffolds (such as the extended peptide binding protein structures we started with) and deep learning generative methods for introducing diversity and for case-specific refinement could be powerful for many other challenging design problems.

The range of possible applications of designed IDR binders is illustrated by the possible uses of the binders developed here. Cancer and disease-related cell receptors such as Mesothelin (MSLN)(*35*), onco-fusion proteins associated with childhood cancer such as EWS/FLI for ewing sarcoma and EML4-ALK for lung cancer(*36*), and CSP for malaria(*16*) 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(*37*). Soluble proteins and peptides ranging from neuropeptides in the brain to poorly studied noncanonical open reading frames (ORFs)(*38*, *39*), which are often intrinsically disordered, such as GREP1(*39*) implicated in breast cancer, could become accessible for imaging and sensing. The potential of our designs to be used in diagnostics is illustrated by the detection of tryptic peptide CTN4 from a mutant form of cystinosin associated with a newborn lysosomal function disorder. 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. For several of these applications, it will be important to extend the characterization of interaction specificity from beyond the sets of targets considered here to the entire proteome. An advantage of targeting long IDRs is that affinity and specificity can, in principle, both be increased if necessary with bispecific constructs targeting two distinct epitopes within the same target.

There are several 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 are advantageous properties for protease substrates. The binding pockets and conformations of peptides in most natural proteases resemble 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 general 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.

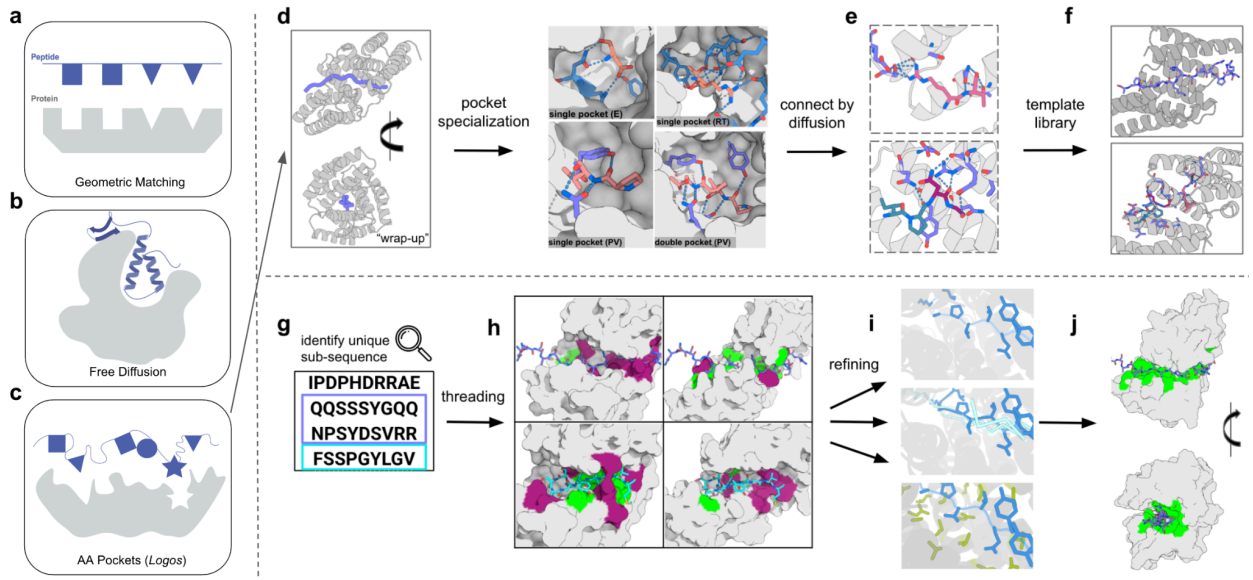
contributed conceptual support for the computation idea at the beginning. D.B. supervised the research. K.W. and D.B. drafted the manuscript. All authors reviewed and commented on the manuscript.

**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 FID data and resonance assignments have been deposited in BioMagResDB (BMRB) with accession codes 52752, 52755, and 52756.

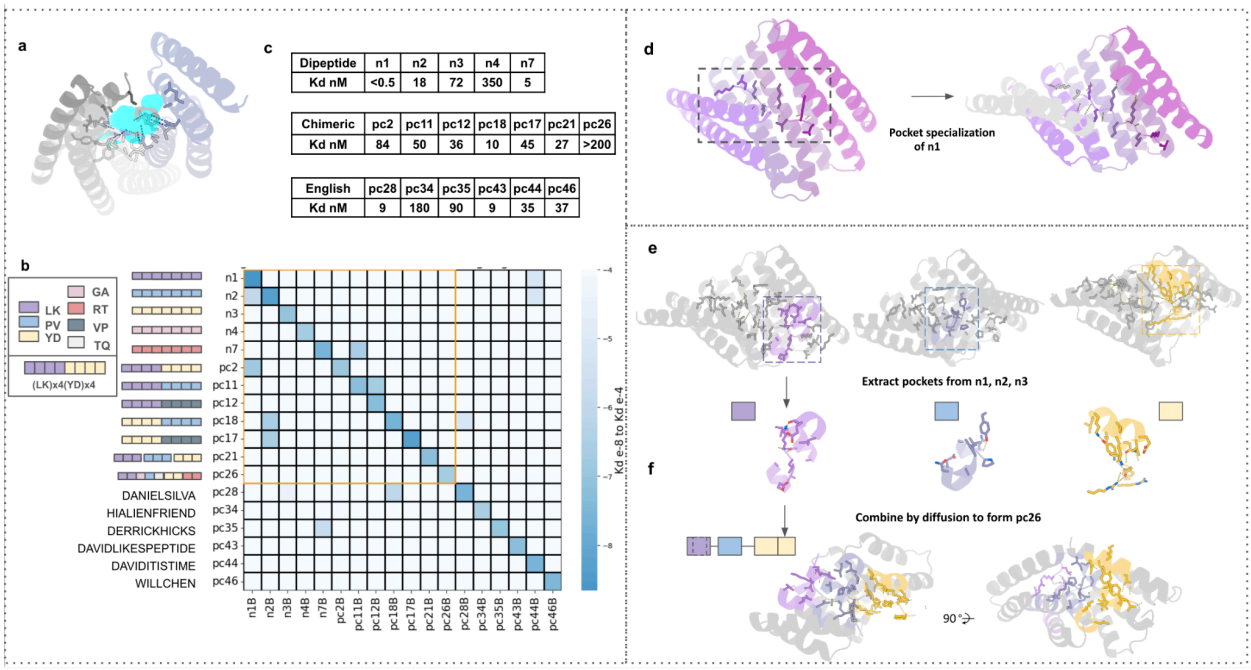
# Figures

## Fig. 1. Overview of IDR binder design protocol.




**a-c, Design methods. a,** Repeat protein based geometric matching approach requires one-to-one matching between identically spaced repeat units on designed binder and target peptide . **b,** Unconstrainted free diffusion approach folds targets into structures frequently observed in PDB training set, primarily helices and strands. **c,** The AA pockets approach explored here combines the designed pockets and extended scaffolds of the geometric matching approach with the ability of diffusion to recombine and diversify the pockets to achieve more general recognition of non-repeating sequences. **d-f, Template library construction. d,** Left, designed binder scaffolds wrap around extended peptide backbones, enabling contacts with each target amino acid. Right, example binding pockets. **e,** Binding pockets are connected using RFdiffusion (each peptide window colored differently) into templates for general sequence recognition. **f,** Examples of two of the 1000 generated templates. **g-j, IDR binding pipeline. g,** Unique sub-sequences (purple and cyan) are identified through a protein sequence database search, and **h,** threaded through the template library to identify optimal matches between amino acid segments and binder pockets. Pocket matches are green and mismatches are dark red on the protein surface. **i,** Matches are refined using “one-sided partial diffusion” (top), where only the binder is changed; “two-sided partial diffusion” (middle), where the target and the binder can be changed; “motif diffusion” (bottom), where key interacting motifs (target, blue; binder, green) are unchanged while the rest are noised, reconnected, diversified, and optimized. **j,** Examples of resulting designs.

Fig. 2. Design of binding to 18 synthetic peptides.

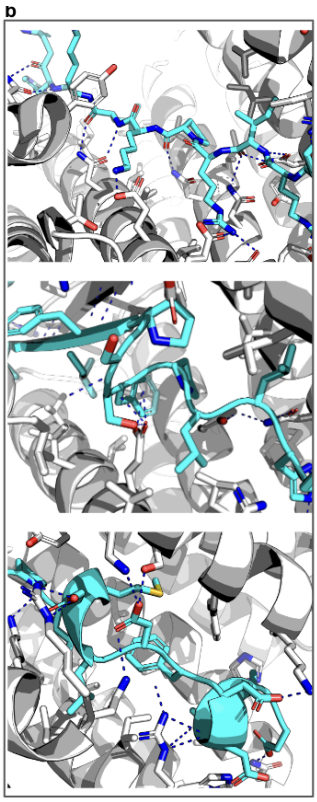


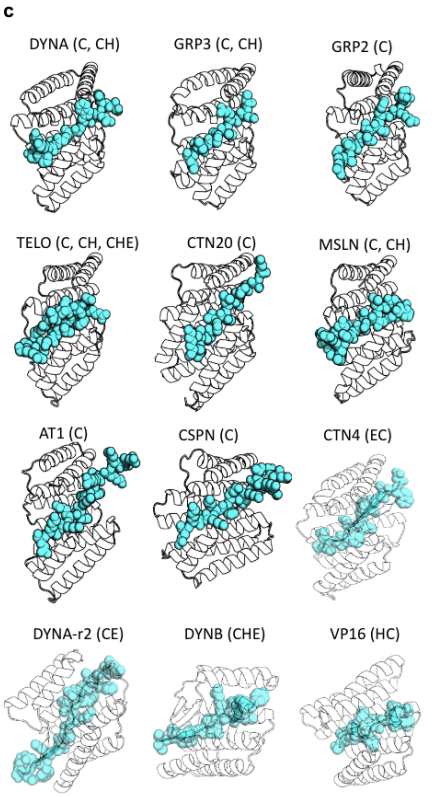

**a,** Representative binder design (cartoon) wrapping around peptide (cyan) in an extended conformation (hydrogen bonds shown as dotted lines). **b,** All-by-all binding Kd's obtained from nanoBiT binding titrations for 18 designed binder-synthetic peptide pairs. Target sequences are on the y-axis with each square representing one dipeptide motif using the color scheme in the legend (left). Heatmap intensities indicate Kd averages from N=2 titration experiments. Pairs within the orange square are composed of similar dipeptide repeats and hence have more cross-talk. **c,** Cognate designed binder-target Kd's measured by biolayer interferometry. Peptide label-identity pairings are on the y axis in b. Two to 35 binder designs were experimentally tested per target. **d-f,** Library construction example using pc26. The three recombined pockets are shown using color codes from panel **b**. **d,** "Pocket specialization," the repetitive pockets are optimized and specialized using motif diffusion from the old four-repeat scaffold (left) to the new five-repeat scaffold (right). The extended fifth repeat is shown in light grey. **e,** Extracting pockets containing two-sided interacting motifs from the base dipeptide repeats. **f,** Diffusion flexibly connects multi-motifs into coherent binding proteins with varied spacers and angles. An example of an assembled design is shown at the bottom.

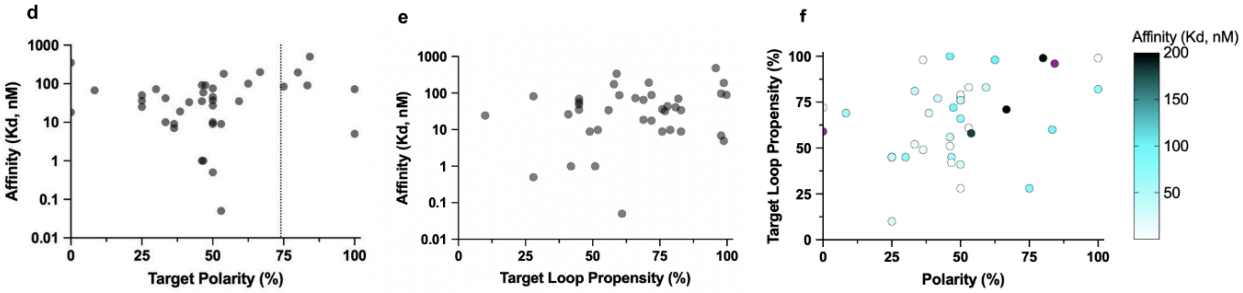
Fig. 3. Design of binding to 21 native protein disordered regions and peptides.

a

| Abbrev. | Full Name | Sequence | Hits % | Kd, nM (v1, v2) |
|---|---|---|---|---|
| DYNA-r2 | Dynorphin A-r2 | YGGF**LRRIRPKLKWDNQ** | 90* | <0.06* |
| DYNA | Dynorphin A | YGGF**LRRIRPKLK(WDNQ)** | 50 | <1, <1 |
| DYNB | Dynorphin B | YGGF**LRRQFKVVT** | 34 | <1 |
| MSLN | Mesothelin | NGY**LVLDLSMQEALS** | 21 | <1 |
| SP | Substance P | **RPKPQQFFGLM** | 13 | 7 |
| CTN20 | Cystinosin20-33 | C**ESSVSLTVPPVVK** | 40 | 9 |
| GRP3 | Glycine-rich extracellular | VLW**NSRWPTLQAWGAGLK PGY** | 9 | 10 |
| CTN4 | Cystinosin4-15 | **NWLTIFILFPLK** | 35 | 25 |
| CPS | C-peptide-short | **SLQPLALEGSLQ** | 8 | 33 |
| FAM21 | Wash complex | **SSDDDLFQSAKPKPAKKTN PFPLLEDE** | 67 | 35 |
| GRP2 | Glycine-rich extracellular | HENGPWPGPC**NARVAPMLL PRLPTPGVPSD** | 20 | 42 |
| VP16 | Viral VP16 | **DALDDFDLDMLPA** | 35 | 45, 19 |
| INS9 | Insulin9-23 | **SHLVEALYLVCGER**G | 19 | 59 |
| PER2 | Period 2 | **AVPFPAPVPAAY** | 42 | 67 |
| GRP1 | Glycine-rich extracellular | **RRPWVPHLLPFSSPGYL**GV M | 10 | 72 |
| AT1 | Angiotensin I | **DRVYIHPFHL** | 25 | 75 |
| TELO | Telomerase | **SWAHPGRTRGPSDRGFCVV SPARPAEEATSLEGALSGT** | 42 | 91 |
| APE | Apelin-13 | **QRPRLSHKGPMPA** | 9 | 92 |
| EA1 | EML4/ALI fusion protein | **PTPGKGPKVYHHKHQE** | 7 | 100 |
| CSPN | Circumsporozoite | **DNEKLRKPKHKKLKQ** | 15 | 195 |
| EF1 | EWS/FLI fusion protein | **SQQSSSYGQQNPSYDSVRR** | 2 | >500 |



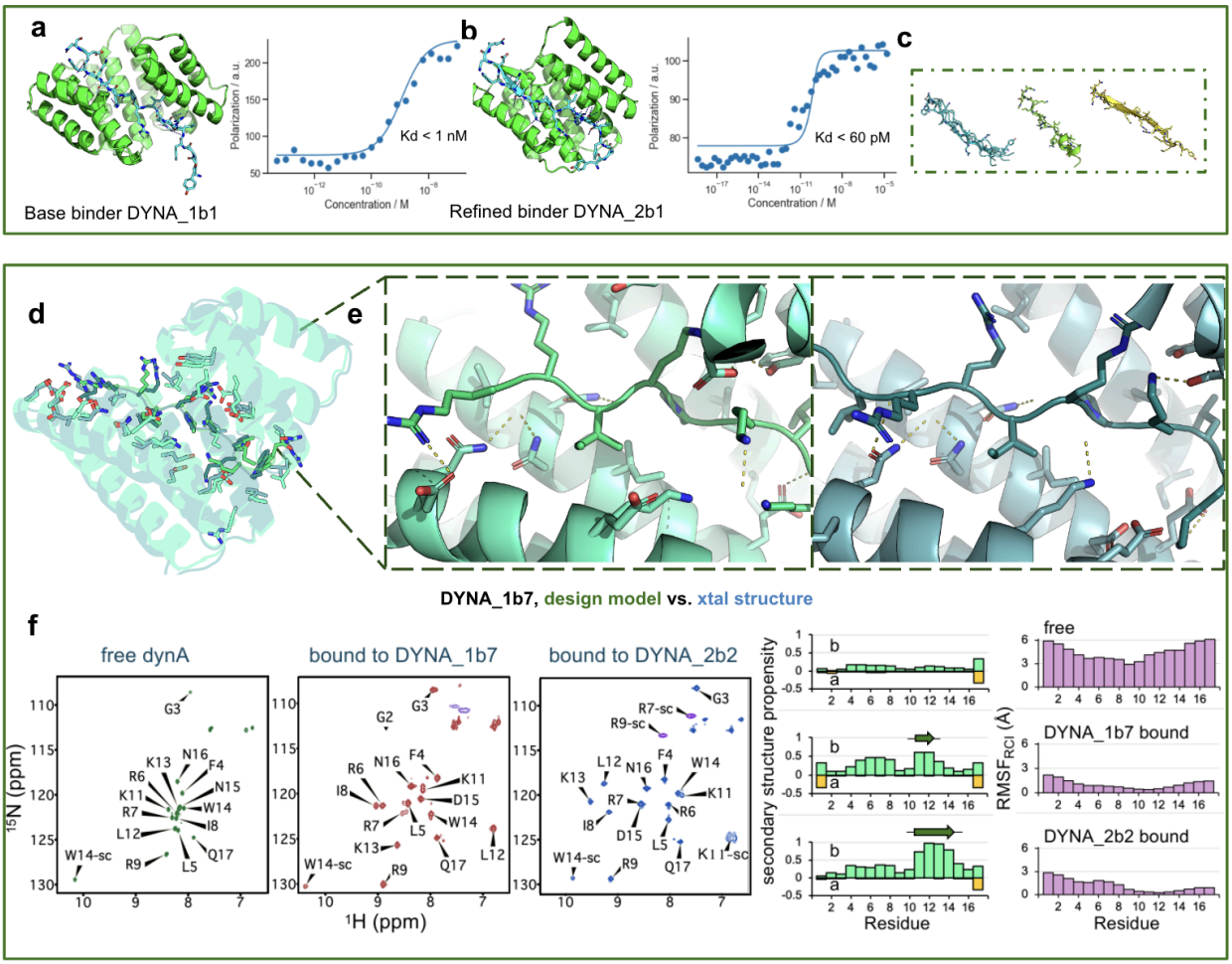




**a,** Targeted native bioactive peptides and IDRs; bold indicate the interacting sequences. Best obtained Kds are ranked from low to high; for the same target, the best two Kds from binders targeting two distinct target conformations are separated by commas. Hit rate was calculated based on the percentage of tested designs showing binding signals (> 0.1 AU) on BLI at 500nM. Asterisks indicate statistics calculated with the optimization campaign, instead of a one-shot campaign. Three to 48 designs were experimentally tested per target. **b,** Interactions of representative IDR binder conformations: random coil conformation (top); strand-containing conformation (middle); helical-containing conformation (bottom). **c,** Models of representative designed complexes. Abbreviations of the targets are above each model, with a bracket indicating the designed targeting conformations of the target. C = random coils, H = (partial) helix, E = (partial) strand. Commas separate multiple targeting conformations for the same target. **d,** Target

polarity vs. highest achieved affinities. Polarity calculated based on the percentage of polar and charged amino acids among the targeted window. **e,** Target loop propensity calculated with AIUPred3 algorithms vs. highest achieved affinities. **f,** Target polarity vs. target loop propensity with a color gradient showing the affinities (black, outliers; dark blue, 200 nM; white, low pM).

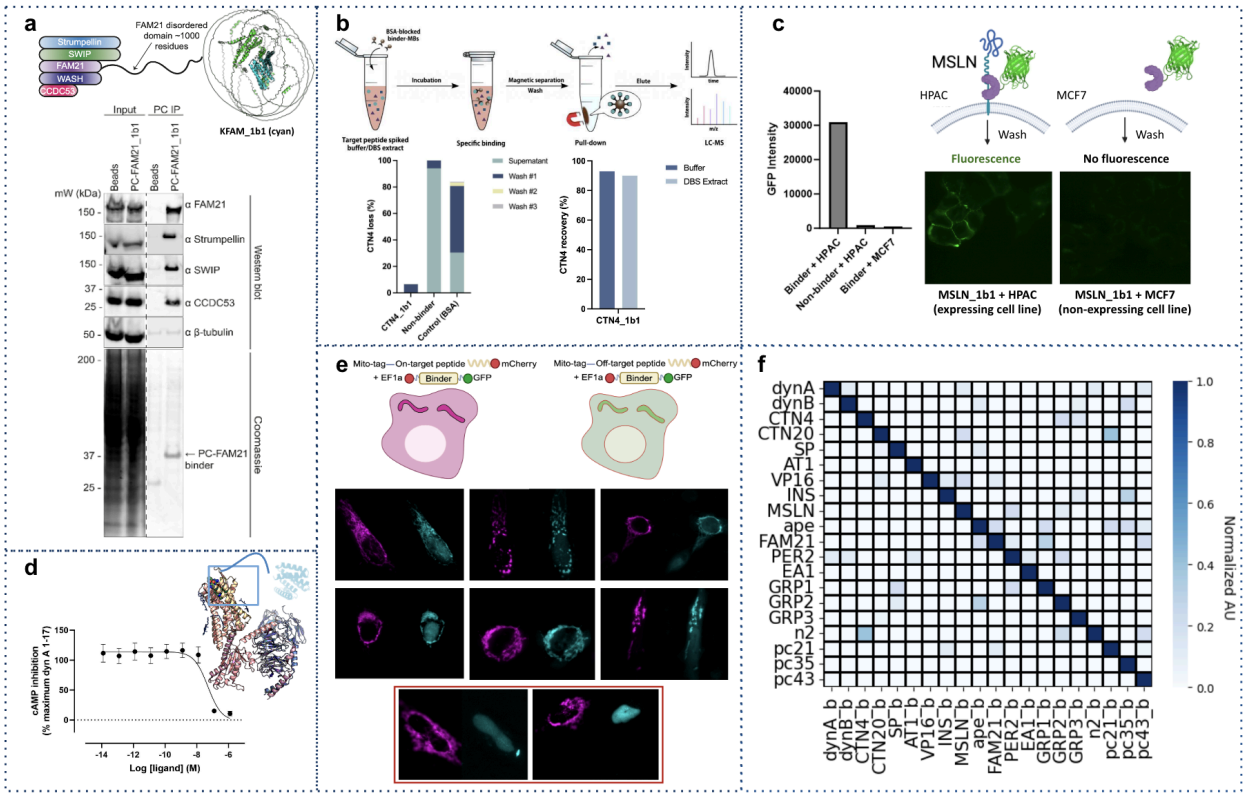
Fig. 4. Structural characterization of dynorphin A binder designs.




**a,** Design model of 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 1 nM TAMRA labeled peptide indicates a $K_d$ < 1 nM. **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,** Superposition of the zoom-out computational design model (green) and the 3.15 Å co-crystal structure (cyan) of dynorphin A bound with design DYNA_1b7, with interface residues shown in sticks. **e,** Zoom-in of the design model (green, left) and crystal structure (cyan, right) of the central designed interaction. **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 CαRMSF ($RMSF_{RCI}$) based on backbone chemical shift data. Sidechain Asn and Gln amide resonance peaks in these HSQC spectra are not labeled. When complexed with DYNA_2b2, dynorphin's Lys11

side chain (sc) amino group, folded from 72.8 ppm in the $^{15}N$ dimension, is observed only because it is stabilized by hydrogen bonding to 2b2 and buried within the complex.

Fig. 5. Designed binders are functional and orthogonal.




**a,** The WASH complex contains the FAM21 protein with a disordered tail (green cartoon, FAM21; cyan, designed binder). Cell lysates were immunoprecipitated with designed binders, and the bound proteins were assessed by Coomassie stain and Western blot. Designed binder PC-FAM21_1b1 immunoprecipitated FAM21 and other WASH complex subunits from the cell lysate. β-tubulin was blotted as loading control. **b,** Illustration of the use of BSA-blocked designed binder-conjugated magnetic beads (MBs) to capture trypsin-generated target peptides for the case of cystinosin. The amount of peptide recovered from elution was quantified by LC-MS. Left, percentage of unbound peptide from each step normalized to the peak area of peptide standards. Right, the percentage of peptide recovery by LC-MS and normalized to the peak area of peptide standards. Non-binder MBs and BSA-blocked unfunctionalized MBs were used as negative controls. **c**, Designed binder–GFP fusions specifically recognize MSLN targets on cells. MSLN_1b1-GFP staining is observed on the cell surface in MSLN-expressing (HPAC) but not non-MSLN-expressing (MCF7) cell lines following incubation at 1uM. No signals observed when incubating HPAC cells with a non-binder-GFP fusion. **d,** Antagonism of dynorphin A-stimulated KOR signaling by DYNA_2b2 binder competition 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. See Fig. S19-20 for activation mechanism. **e**, HeLa (CCL-2) cells co-expressing the targets fused to mCherry and a mitochondria-targeting sequence (Mito-Tag) and the binders fused to GFP. GFP signal is only relocalized to the mitochondria for designs with their

cognate targets. **f,** 20x20 orthogonality binding matrix determined using BLI. Biotinylated target peptide was loaded onto streptavidin biosensors, and incubated with the designed cognate binder and non-cognate binders. Heat map shows the maximum response signal for each binder-target pair normalized by the maximum response signal of the cognate at 1 uM.

# References


1. M. L. Chiu, D. R. Goulet, A. Teplyakov, G. L. Gilliland, Antibody structure and function: The basis for engineering therapeutics. *Antibodies (Basel)* **8**, 55 (2019).

2. J. H. Lee, R. Yin, G. Ofek, B. G. Pierce, Structural features of antibody-peptide recognition. *Front. Immunol.* **13**, 910367 (2022).

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

4. M. A. S. Perez, M. A. Cuendet, U. F. Röhrig, O. Michielin, V. Zoete, Structural prediction of peptide-MHC binding modes. *Methods Mol. Biol.* **2405**, 245–282 (2022).

5. N. Zeytuni, R. Zarivach, Structural and functional discussion of the tetra-trico-peptide repeat, a protein interaction module. *Structure* **20**, 397–405 (2012).

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

7. K. Wu, H. Bai, Y.-T. Chang, R. Redler, K. E. McNally, W. Sheffler, T. J. Brunette, D. R. Hicks, T. E. Morgan, T. J. Stevens, A. Broerman, I. Goreshnik, M. DeWitt, C. M. Chow, Y. Shen, L. Stewart, E. Derivery, D. A. Silva, G. Bhabha, D. C. Ekiert, D. Baker, De novo design of modular peptide-binding proteins by superhelical matching. *Nature* **616**, 581–589 (2023).

8. S. Vázquez Torres, P. J. Y. Leung, P. Venkatesh, I. D. Lutz, F. Hink, H.-H. Huynh, J. Becker, A. H.-W. Yeh, D. Juergens, N. R. Bennett, A. N. Hoofnagle, E. Huang, M. J. MacCoss, M. Expòsit, G. R. Lee, A. K. Bera, A. Kang, J. De La Cruz, P. M. Levine, X. Li, M. Lamb, S. R. Gerben, A. Murray, P. Heine, E. N. Korkmaz, J. Nivala, L. Stewart, J. L. Watson, J. M. Rogers, D. Baker, De novo design of high-affinity binders of bioactive helical peptides. *Nature* **626**, 435–442 (2024).

9. D. D. Sahtoe, E. A. Andrzejewska, H. L. Han, E. Rennella, M. M. Schneider, G. Meisl, M. Ahlrichs, J. Decarreau, H. Nguyen, A. Kang, P. Levine, M. Lamb, X. Li, A. K. Bera, L. E. Kay, T. P. J. Knowles, D. Baker, Design of amyloidogenic peptide traps. *Nat. Chem. Biol.* **20**, 981–990 (2024).

10. C. Liu, K. Wu, H. Choi, H. Han, X. Zhang, J. L. Watson, S. Shijo, A. K. Bera, A. Kang, E. Brackenbrough, B. Coventry, D. R. Hick, A. N. Hoofnagle, P. Zhu, X. Li, J. Decarreau, S. R. Gerben, W. Yang, X. Wang, M. Lamp, A. Murray, M. Bauer, D. Baker, Diffusing protein binders to intrinsically disordered proteins, *biorxiv* (2024). https://www.biorxiv.org/content/10.1101/2024.07.16.603789v1.

11. J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

12. R. Rozbeh, K. Forchhammer, 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).

13. X. Li, M. E. McGee-Lawrence, M. Decker, J. J. Westendorf, The Ewing's sarcoma fusion protein, EWS-FLI, binds Runx2 and blocks osteoblast differentiation. *J. Cell. Biochem.* **111**, 933–943 (2010).

14. S. P. Selvanathan, G. T. Graham, H. V. Erkizan, U. Dirksen, T. G. Natarajan, A. Dakic, S. Yu, X. Liu, M. T. Paulsen, M. E. Ljungman, C. H. Wu, E. R. Lawlor, A. Üren, J. A. Toretsky, Oncogenic fusion protein EWS-FLI1 is a network hub that regulates alternative splicing. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1307–16 (2015).

15. G. Flores, P. J. Grohar, One oncogene, several vulnerabilities: EWS/FLI targeted therapies for Ewing sarcoma. *J. Bone Oncol.* **31**, 100404 (2021).

16. J. Tan, H. Cho, T. Pholcharee, L. S. Pereira, S. Doumbo, D. Doumtabe, B. J. Flynn, A. Schön, S. Kanatani, S. O. Aylor, D. Oyen, R. Vistein, L. Wang, M. Dillon, J. Skinner, M. Peterson, S. Li, A. H. Idris, A. Molina-Cruz, M. Zhao, L. R. Olano, P. J. Lee, A. Roth, P. Sinnis, C. Barillas-Mury, K. Kayentao, A. Ongoiba, J. R. Francica, B. Traore, I. A. Wilson, R. A. Seder, P. D. Crompton, Functional human IgA targets a conserved site on malaria sporozoites. *Sci. Transl. Med.* **13**, eabg2344 (2021).

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

18. S. Podvin, T. Yaksh, V. Hook, The emerging role of spinal dynorphin in chronic pain: A therapeutic perspective. *Annu. Rev. Pharmacol. Toxicol.* **56**, 511–533 (2016).

19. E. B. Margolis, M. G. Moulton, P. S. Lambeth, M. J. O'Meara, 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).

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

21. Y. Wang, Y. Zhuang, J. F. DiBerto, X. E. Zhou, G. P. Schmitz, Q. Yuan, M. K. Jain, W. Liu, K. Melcher, Y. Jiang, B. L. Roth, H. E. Xu, Structures of the entire human opioid receptor family. *Cell* **186**, 413–427.e17 (2023).

22. E. Derivery, C. Sousa, J. J. Gautier, B. Lombard, D. Loew, A. Gautreau, The Arp2/3 activator WASH controls the fission of endosomes through a large multiprotein complex. *Dev. Cell* **17**, 712–723 (2009).

23. T. S. Gomez, D. D. Billadeau, A FAM21-containing WASH complex regulates retromer-dependent sorting. *Dev. Cell* **17**, 699–711 (2009).

24. M. Putker, D. C. S. Wong, E. Seinkmane, N. M. Rzechorzek, A. Zeng, N. P. Hoyle, J. E. Chesham, M. D. Edwards, K. A. Feeney, R. Fischer, N. Peschel, K.-F. Chen, M. Vanden

Oever, R. S. Edgar, C. P. Selby, A. Sancar, J. S. O'Neill, CRYPTOCHROMES confer robustness, not rhythmicity, to circadian timekeeping. *EMBO J.* **40**, e106745 (2021).

25. A. Mihut, J. S. O'Neill, C. L. Partch, P. Crosby, PERspectives on circadian cell biology. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **380**, 20230483 (2025).

26. M. Al-Haggar, Cystinosis as a lysosomal storage disease with multiple mutant alleles: Phenotypic-genotypic correlations. *World J. Nephrol.* **2**, 94–102 (2013).

27. M. A. Elmonem, K. R. Veys, N. A. Soliman, M. van Dyck, L. P. van den Heuvel, E. Levtchenko, Cystinosis: a review. *Orphanet J. Rare Dis.* **11**, 47 (2016).

28. X. Liu, A. Chan, C.-H. Tai, T. Andresson, I. Pastan, Multiple proteases are involved in mesothelin shedding by cancer cells. *Commun. Biol.* **3**, 728 (2020).

29. P. Ernst, F. Zosel, C. Reichen, D. Nettels, B. Schuler, A. Plückthun, Structure-guided design of a peptide lock for modular peptide binders. *ACS Chem. Biol.* **15**, 457–468 (2020).

30. X. Wang, Z. Xu, Z. Tian, X. Zhang, D. Xu, Q. Li, J. Zhang, T. Wang, The EF-1α promoter maintains high-level transgene expression from episomal vectors in transfected CHO-K1 cells. *J. Cell. Mol. Med.* **21**, 3044–3054 (2017).

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

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

33. S. Sipeki, K. Koprivanacz, T. Takács, A. Kurilla, L. László, V. Vas, L. Buday, Novel roles of SH2 and SH3 domains in lipid binding. *Cells* **10**, 1191 (2021).

34. J. Z. Zhang, X. Li, C. Liu, H. Jiang, K. Wu, D. Baker, De novo design of Ras isoform selective binders, *Bioengineering* (2024). https://www.biorxiv.org/content/10.1101/2024.08.29.610300v3.

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

36. T. Sasaki, S. J. Rodig, L. R. Chirieac, P. A. Jänne, The biology and treatment of EML4-ALK non-small cell lung cancer. *Eur. J. Cancer* **46**, 1773–1780 (2010).

37. S. Aiyer, G. V. T. Swapna, L.-C. Ma, G. Liu, J. Hao, G. Chalmers, B. C. Jacobs, G. T. Montelione, M. J. Roth, A common binding motif in the ET domain of BRD3 forms polymorphic structural interfaces with host and viral proteins. *Structure* **29**, 886–898.e6 (2021).

38. C. Erady, K. Amin, T. O. A. E. Onilogbo, J. Tomasik, R. Jukes-Jones, Y. Umrania, S. Bahn, S. Prabakaran, 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).

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

40. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

41. L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

42. N. R. Bennett, B. Coventry, I. Goreshnik, B. Huang, A. Allen, D. Vafeados, Y. P. Peng, J. Dauparas, M. Baek, L. Stewart, F. DiMaio, S. De Munck, S. N. Savvides, D. Baker, Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).

43. J. Umper, R. Evans, A. Pritzel, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

44. C. Reichen, S. Hansen, C. Forzani, A. Honegger, S. J. Fleishman, T. Zhou, F. Parmeggiani, P. Ernst, C. Madhurantakam, C. Ewald, P. R. E. Mittl, O. Zerbe, D. Baker, A. Caflisch, A. Plückthun, Computationally designed armadillo repeat proteins for modular peptide recognition. *J. Mol. Biol.* **428**, 4467–4489 (2016).

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

46. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

47. H. Jiang, K. M. Jude, K. Wu, J. Fallas, G. Ueda, T. J. Brunette, D. R. Hicks, H. Pyles, A. Yang, L. Carter, M. Lamb, X. Li, P. M. Levine, L. Stewart, K. C. Garcia, D. Baker, De novo design of buttressed loops for sculpting protein functions. *Nat. Chem. Biol.* **20**, 974–980 (2024).

48. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer, *bioRxiv* (2021). https://doi.org/10.1101/2021.10.04.463034.

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

50. N. London, B. Raveh, E. Cohen, G. Fathi, O. Schueler-Furman, Rosetta FlexPepDock web server-high resolution modeling of peptide-protein interactions. *Nucleic Acids Res* **39**,

W249–W253 (2011).

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

52. A. D. Beale, N. M. Rzechorzek, A. Mihut, A. Zeng, N. J. Smyllie, V. Pilorz, R. Richardson, M. F. Bertlesen, N. R. James, S. V. Fazal, Z. Voysey, J. Pelletier, P. Crosby, S. Y. Peak-Chew, M. A. Lancaster, R. A. Hut, J. S. O'Neill, Thermosensitivity of translation underlies the mammalian nocturnal-diurnal switch, *bioRxiv* (2023). https://doi.org/10.1101/2023.06.22.546020.

53. E. Muratspahić, N. Tomašević, J. Koehbach, L. Duerrauer, S. Hadžić, J. Castro, G. Schober, S. Sideromenos, R. J. Clark, S. M. Brierley, D. J. Craik, C. W. Gruber, Design of a stable cyclic peptide analgesic derived from sunflower seeds that targets the κ-opioid receptor for the treatment of chronic abdominal pain. *J. Med. Chem.* **64**, 9042–9055 (2021).

54. E. Muratspahić, A. M. White, C. I. Ciotu, N. Hochrainer, N. Tomašević, J. Koehbach, R. J. Lewis, M. Spetea, M. J. M. Fischer, D. J. Craik, C. W. Gruber, Development of a selective peptide κ-opioid receptor antagonist by late-stage functionalization with cysteine staples. *J. Med. Chem.* **66**, 11843–11854 (2023).

55. E. Muratspahić, K. Deibler, J. Han, N. Tomašević, K. B. Jadhav, A.-L. Olivé-Marti, N. Hochrainer, R. Hellinger, J. Koehbach, J. F. Fay, M. H. Rahman, L. Hegazy, T. W. Craven, B. R. Varga, G. Bhardwaj, K. Appourchaux, S. Majumdar, M. Muttenthaler, P. Hosseinzadeh, D. J. Craik, M. Spetea, T. Che, D. Baker, C. W. Gruber, Design and structural validation of peptide-drug conjugate ligands of the kappa-opioid receptor. *Nat. Commun.* **14**, 8064 (2023).

56. F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, A. Bax, NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).

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

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

59. Y. Hen, A. Bax, Protein Structural Information Derived from NMR Chemical Shift with the Neural Network Program TALOS-N. *Methods Mol. Biol. Clifton NJ*.

60. M. V. Berjanskii, D. S. Wishart, A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.* **127**, 14970–14971 (2005).

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

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

63. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

64. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

65. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

66. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall 3rd, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).