Protein design for cyclic peptide and small molecule binding

Stephanie Hanna

A dissertation

submitted in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

David Baker, chair

Jorge Marchand

Dustin Maly

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Protein design for cyclic peptide and small molecule binding

Stephanie Hanna

Chair of the Supervisory Committee:

David Baker

Biochemistry

Proteins can bind diverse classes of ligands and the presence of these ligands can induce a response in the cell. Particularly, for the design of systems where the interaction of a ligand with intracellular binding partners can elicit a response, the ligand must be able to cross the cell membrane. To achieve this aim, I explored two classes of ligand: cyclic peptides and small molecule HIV protease inhibitors. I designed protein binders for each of these ligand classes. For one cyclic peptide, I utilized a library of designed homodimeric scaffold proteins as the protein binding component and demonstrated their response to the cyclic peptide in mammalian cells. For HIV protease ligands, I explored the application of deep learning based protein design tools to generate backbones, design sequences and predict the ligand-protein complex. While two libraries of designs were screened for binding via yeast display with a biotinylated ligand, further development of the design pipeline is necessary to obtain protein binders with high affinity to these ligands.

Table of Contents

Acknowledgements

First, I would like to thank Dr. David Baker and my committee members, Dr. Jorge Marchand, Dr. Dusty Maly, and Dr. Champak Chatterjee for their guidance and input throughout the years.

My PhD experience would not have been the same without the people I met through the Baker Lab. I thank Dr. Jason Qian, Long Tran, Shajesh Sharma, Dr. Magnus Bauer, Dr. Thomas Schlichthaerle, Dr. Kathryn Shelley, Dr. Xinru Wang, Dr. Mohamad Abedi, and Dr. Basile Wicky for their support, mentorship and being great colleagues. I would also like to thank Laura Guerrero, Julia Bonzanini, Stephanie Adaniya, Riti Biswas, Dr. Preetham Venkatesh, Ayca Ersoy, Judy Ahr, and Cyrus Haas for being great friends and for all the adventures during my PhD.

Finally, I would like to thank my family including my parents Camille and Joe, my Aunt Clare, Uncle Frankie, Uncle Phil, and Aunt Karen as well as my cousins Lucy, Katherine, and Danielle for their support and encouragement throughout my life.

[1]Chapter 1: Design of a homodimeric protein that binds a cyclic peptide

This Chapter contains the contents of an accepted manuscript.

Background

Chemically induced dimerization (CID) systems in which protein-protein association is dependent on membrane permeable small molecule ligands are broadly useful in synthetic biology to enable external control of cellular processes. CID systems adapted from nature have been used to control biological processes ranging from apoptosis to targeted protein degradation to transcription[1, 2, 3]. The power and limitations of such systems are exemplified by the rapamycin-FKBP-FRB system which is widely used for cellular control. Rapamycin not only effectively induces FKBP-FRB association but also is an immunosuppressant that dimerizes endogenous proteins, leading in some cases to undesired off-target effects *in vivo*[4]. Alleviation of the toxicity issues associated with rapamycin CID systems have been accomplished either by derivatization of rapamycin and FK506 and the FKBP protein, or adapting CID systems native to plants into mammalian cells[5, 6, 7, 8]. Expansion to other ligand classes has relied upon directed evolution approaches to engineer a binding interface for a desired ligand or finding antibody fragment binding partners through phage display for a ligand or existing protein ligand binary complex[9, 10, 11, 12, 13]. Protein design methods can in principle yield a variety of CID systems; a farnesyl pyrophosphate sensor relying on a maltose binding and an ankyrin repeat protein, as well as a CID composed of NS3a protease, grazoprevir, danoprevir, and asunaprevir along with three designed proteins are two such examples[14, 15]. Developing CIDs with new chemical ligands, however, has been relatively unexplored due to a lack of protein domains that bind disparate or xenobiotic molecules. An attractive class of possible compounds for CID system construction are sets of designed, structured macrocycles (MCs) that are chemically diverse, easy to synthesize, and membrane permeable. These compounds have considerable exposed surface area available for binding, and an almost unlimited number of such molecules can now be designed[16].

We reason that designed proteins that dimerize in the presence of MCs could yield a wide variety of CID systems. To highlight the biochemical diversity that might be possible, we design both the MCs and the sensor protein from scratch. For the MC component of the system, desirable properties are 1) membrane permeability for intracellular applications, 2) sufficient exposed surface area to drive protein association, and 3) a well-ordered tertiary structure to favor on-target while disfavoring off-target interactions. Guided by previous studies indicating that structured MCs are generally cell permeable if all polar groups are satisfied in hydrogen bonding interactions, we develop a method to generate closed MCs with internal backbone-hydrogen

[1] This chapter contains collaborative work. P. J. Salveson designed macrocycles (MCs). D. R. Hicks designed homodimeric scaffold proteins. M. A. Kennedy and B. L. Stoddard crystallized protein-MC1 complex. A. K. Bera, P. J. Salveson, and A. Kang crystallized MCs. I designed, and experimentally screened homodimeric binders to MC1 as well performed data analysis of assays. C. Moller and X. Li ran and analyzed PAMPA data for MC1. S. Cheng ran cytotoxicity assay for MC1. The manuscript was written collaboratively.

bonding[16]. We generate large sets of loops of 5, 6, and 7 residues in which the terminal residues make beta-strand like hydrogen bonds and the remaining residues were N-methylated. Closed MCs are constructed by identifying pairs of loops with the terminal hydrogen bonding residues in nearly identical orientations. For computational efficiency, this identification is done using geometric hashing: for each loop we calculate the rigid body transform between the N- and C-terminal residues in both the N-to-C and C-to-N orientations, and closed MC backbones are then constructed by identifying pairs of loops for which the N-to-C transform of one is the same as the C-to-N transform of the other (Fig. 1A)[17]. These resulting MCs have two backbone-hydrogen bonds between the two residues at the fusion junctions, and the remainder of the backbone amides were N-methylated, leaving no unsatisfied backbone amides to compromise membrane permeability. For each such generated structure, we use RosettaFastDesign to generate low-energy amino acid sequences allowing both L and D amino acids at each position while maintaining the pattern of N-alkylation to maintain full NH satisfaction throughout the MC (Fig. 1B).

MC1, MC2, MC3 and MC4 are excellent candidates for induced dimerization systems as they are likely membrane permeable, with all backbone amides either engaged in hydrogen bonds or N-methylated and have well defined structures. An ordered structure reduces the chance of unwanted interactions with natural proteins and should increase affinity for designed binders since backbone entropy would not be lost in binding. The first three are asymmetric, and hence well suited to the design of induced heterodimerization CIDs. While we could apply these asymmetric macrocycles to the design of heterodimeric CIDs, this paper focuses on building induced homodimerization systems around the internally symmetric MC1 for two reasons: first many important processes in biology, such as caspase activation, involve homodimerization, and second, the design problem is simpler as only one protein chain need be designed[19].

Results

We solved crystal structures of four of these designed MCs, MC1, MC2, MC3, and MC4 (Fig. 1B). Three are asymmetric–MC2, MC3, and MC4– and have sequences apL*F*FPa*l*, fpL*a*APl*f*, and XPa*l*iX*Pa*P, respectively, where the asterisk represents N-methylation, X represents norleucine, capital letters are L amino acids and lower case D amino acids. As is evident in the structures, the two loops flanking the central backbone hydrogen bond pair have different conformations in different designs, and the structures are stabilized by torsional restraints from multiple proline residues as well as aromatic packing interactions (Fig. 1B). In proteins, backbone hydrogen bonds do not generally favor specific conformations as there are large numbers of alternative conformations in which these can be made, but in our designed MCs, extensive N-methylation and backbone cyclization strongly favors antiparallel versus parallel pairings. Thus, hydrogen bonding can effectively only occur between the intended residues and in the designed arrangement, likely also contributes to structural specification. The crystal structures of these designs are very close to the design models over the peptide backbone, but there are several evident flips of aromatic residues in some cases (Fig. 1C). The design scheme can be readily extended to incorporate loops which bind protein targets of interest as a

potential therapeutic strategy. The sequence of MC1 , apL*F*apL*F*, is internally duplicated, and both the design model and the crystal structure have C2 symmetry. Two of the eight amides are in the cis-conformation, which has been associated with increased membrane permeability[18]. An additional three crystal structures were solved for MC5 and MC6 as well as an L4F mutant of MC6 (Supplementary Note 1). MC5 and MC6 loops were not in agreement between the design and crystal structure, but the MC6-L4F mutant and its crystal structure align closely (Supplementary Fig. 1).

For the protein component of the new system, we chose to start from designed helical repeat homodimers containing central binding pockets with a wide range of shapes and sizes, including many that are in the size range of MCs[20]. An initial round of binder design with these scaffolds and MC1 yielded a co-crystal structure where MC1 was not in the correct orientation of the binding pocket with respect to the design model (Supplementary Fig. 2). The sequences of the original scaffolds were redesigned with ProteinMPNN to increase folding robustness and used for a successive round of binder design with MC1[21]. Homodimeric binders for MC1 were computationally designed by first generating a rotamer interaction field around MC1 using RIFgen. RIFgen generates large numbers of disembodied sidechains making favorable interactions with the proposed ligand[22]. RIFdock was then used to place MC1 into the central cavity of each homodimeric scaffold to maximize the number of these favorable sidechain interactions that could be realized, resulting in RIF sidechains with N-Cα-C backbone intersecting with backbone atoms of the scaffold (Fig. 1C). Docks were filtered on interface shape complementarity, binding energy (ΔΔG), contact molecular surface (CMS) area between the protein and the peptide, and contact patch. The peptide protein interface was then redesigned with Rosetta FastDesign, leaving the sequence of MC1 fixed. The resulting designs were filtered for the number of hydrogen bonds made to the ligand, CMS, ΔΔG, interface shape complementarity, and the number of buried unsaturated hydrogen bond donors. AlphaFold 2 (AF2) structure predictions were generated for each and predictions were then filtered on predicted local distance difference test (pLDDT), and predicted template modeling score (pTM), yielding 263 designs, seventeen of which were selected for experimental validation.

Synthetic genes encoding the designed proteins were obtained, expressed in *E. coli*, and evaluated for binding to MC1 by equilibrium dialysis. An initial binding screen by equilibrium dialysis suggested one out of eleven screened designs, CID7, bound MC1 at both protein concentration points assessed (Supplementary Fig. 3). CID7 exhibited a monodisperse size exclusion chromatography (SEC) trace. We then determined a 36 nM K_D of homodimeric CID7 to MC1 by Isothermal titration calorimetry (ITC) (Fig. 2A-B). Mass photometry experiments with maltose binding protein fused to CID7 revealed the approximate binding affinity of the CID7 homodimer is 0.4 nM (Supplementary Fig. 4).

The structure of MC1 in complex with CID7 was solved by X-ray crystallography. The complex has designed C2 symmetry, with the orientation of the peptide in the binding pocket very close to the design model and a 0.97 Å C-alpha RMSD between MC1 in the design and crystal structure (Fig. 2C). For the protein chains, the C-alpha RMSD between the design model

and crystal structure was 0.95 Å. The proline rings of MC1 pucker in the crystal structure, allowing for a backbone carbonyl to engage in hydrogen bonding with Y89. N53 and N218 also form hydrogen bonds with MC1 backbone carbonyls (Fig. 2D).

We investigated the membrane permeability of MC1 in an artificial membrane permeability assay (PAMPA). MC1 has an apparent permeability of $1.1x10^{-5}$ cm/s, above the 1.5×10^{-6} cm/s cutoff for a compound to be considered permeable[16] (Supplementary Fig. 5). We also assessed the cytotoxicity of MC1 in HEK293T cells using an LDH-Glo assay (Promega J2380). HEK293T cells treated with MC1 titrated 1 in 2 from 50 uM to 50 nM and aliquots of cell media were analyzed for LDH presence 48 h after treatment. We compared the LDH release to a vehicle only control (cell media with 1% DMSO) as well as a maximum LDH control where cells were treated with TritonX 15 minutes before LDH measurements were taken. We observed no significant cytotoxicity. Treatment with MC1 showed similar luminescence and percent cytotoxicity to that of vehicle-treated cells (Supplementary Fig. 6A-B).

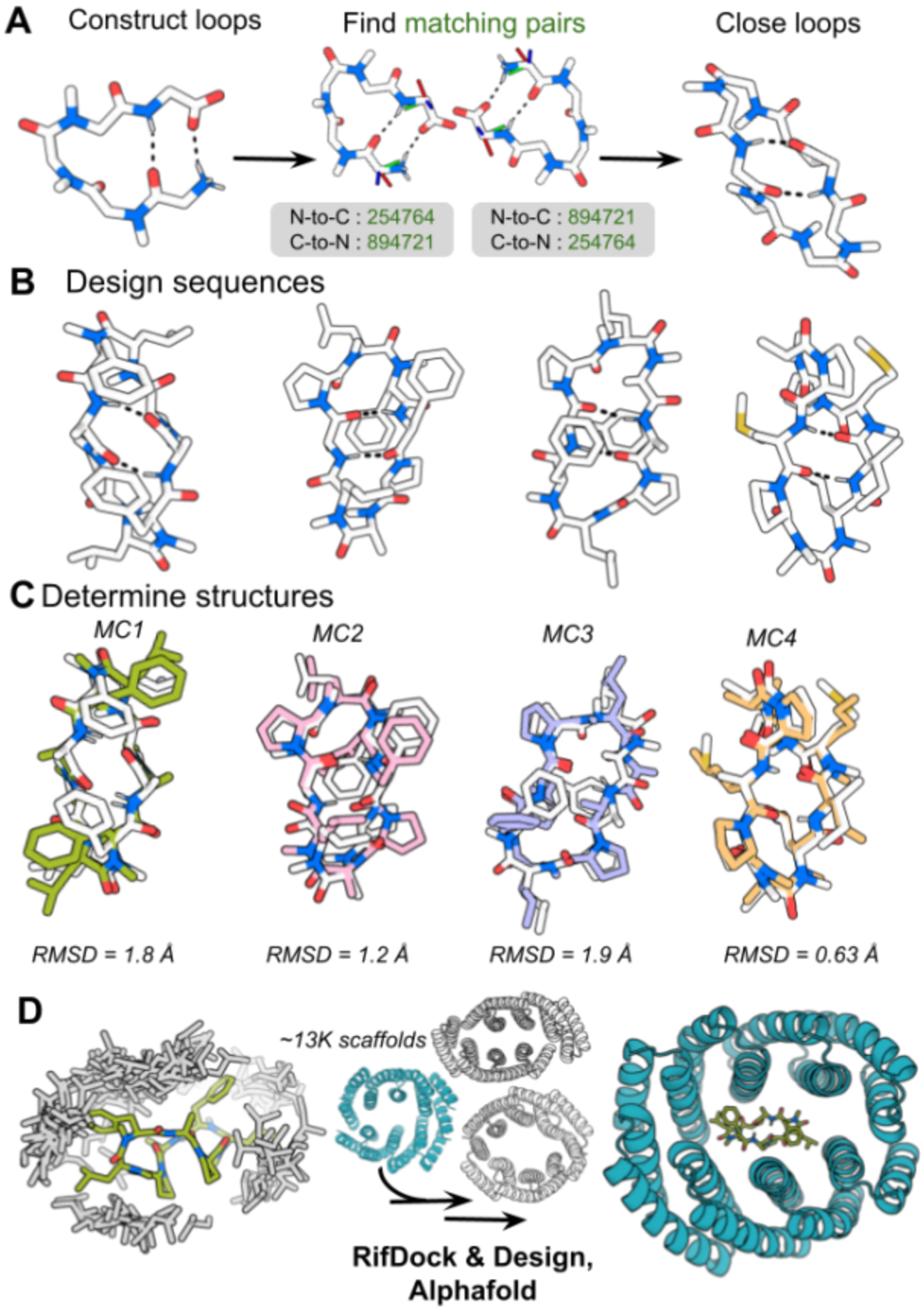
We evaluated the inducibility and reversibility of the CID7 system in an in-cell split luciferase (Nanobit) assay. CID7 proteins were expressed in HEK293T cells as N-terminal fusions to either smBiT (smBiT-CID) or lgBiT (lgBiT-CID7) NanoLuc® luciferase. The luminescence of samples expressing CID7 fusions was somewhat higher than that of a negative control consisting of Halo protein fused to smBit on the C-terminus and lgBiT-CID7, suggesting some background dimerization. Upon addition of MC1, approximately 3-fold increase in luminescence was observed. In contrast, the luminescence signal decreased in cells treated with vehicle rather than MC1 (Fig 3B). After washing out the ligand and adding additional substrate, the luminescent signal returned to close to the level before treatment with MC1 or vehicle, indicating reversibility (Fig 3B).

We next evaluated whether MC1 could induce dimerization of CID7 in a mammalian two-hybrid assay. CID7 was independently fused to either a zinc finger DNA binding domain (DBD) or a p65 activation domain (AD) and the two constructs were expressed in HEK293T cells along with an enhanced green fluorescent protein (eGFP) reporter which should be activated by reconstitution of the split transcription factor[23]. Upon addition of MC1, there was an increase in eGFP expression over control cells treated with cell media with 1% DMSO, likely reflecting dimerization of CID7 and reconstitution of the split transcription factor. Titration of MC1 yields a dose response curve with an EC_{50} of 9.4 μM and a 16-fold increase in geometric mean fluorescence intesity (gMFI) of eGFP at saturating ligand concentrations (Fig. 3A-B). Transfection efficiency was assessed by mscarlett-1 and mtagBFP2 fluorescence levels which are co-expressed with each construct through an internal ribosome entry site (IRES) in the AD and DBD plasmids, respectively. In the absence of MC1, there is some activation of eGFP expression due to basal homodimerization of CID7 in cells with high transfection efficiency (Supplementary Fig. 7). In control experiments, we generated similar constructs utilizing the FRB-FKBP rapamycin inducible system and observed 17-fold increase in eGFP gMFI between control cells and the highest concentration of rapamycin, 10 μM (Fig. 4C-D).

Discussion

The CID system described enables inducible control of gene expression using a designed macrocyclic peptide with a designed protein pair. Very little experimental optimization was required to generate our CID system as the sequences of both the MC ligand and the designed protein it dimerizes are directly from Rosetta design calculations. Cell based assays did not show detectable toxicity of MC1, and the MC1 permeability in the PAMPA assay is in line with other membrane permeable compounds. The CID7-MC1 system signal, however, could be limited by the ability of the designed proteins to form homodimeric species with each construct and we observe some background dimerization both in the transcriptional assay as well as the Nanobit assay. Thus, design optimization to improve EC_{50} and reduce background homodimerization could further increase the utility of the system.

More generally, the ability to jointly *de novo* design both sides of a CID system opens the door to an almost unlimited number of new systems. These systems could act in concert for simultaneous yet independent control over many aspects of cell function, for both mechanistic investigation and external guidance of adoptive cell therapies.


A
Construct loops
Find matching pairs
Close loops
N-to-C : 254764
C-to-N : 894721
N-to-C : 894721
C-to-N : 254764
B
Design sequences
C
Determine structures
MC1
MC2
MC3
MC4
RMSD = 1.8 Å
RMSD = 1.2 Å
RMSD = 1.9 Å
RMSD = 0.63 Å
D
~13K scaffolds
RifDock & Design,
Alphafold

F**igure 1. Computational design procedure. a** Five, six, or seven residue loops were constructed to mimic beta hairpins. Construction of macrocyclic peptide backbones is shown using a representative, five residue starting loop. 6D transforms were calculated and binned. Loops were closed together if their termini geometry for closure matched, as indicated by green bin numbers. **b** RosettaFastDesign was used to generate sequences for macrocyclic peptides, allowing both L and D hydrophobic residues. Design models of the MCs are shown in white with dashed black lines indicating hydrogen bonds. **c** X-ray crystal structures (color) overlaid on the corresponding design models (white) for four designs. The all atom RMSDs between the design models and crystal structures are depicted beneath the MCs. **d** A rotamer interaction field was generated from the MC1 crystal structure (green); leucine side chains in the RIF process are shown for illustration (gray). RifDock was then used to place the peptide into the central cavity of homodimeric scaffold proteins. Protein peptide interactions were optimized with Rosetta fastDesign, retaining the protein-protein interfaces, and the designs were selected for experimental characterization based on AF2 and Rosetta metrics.

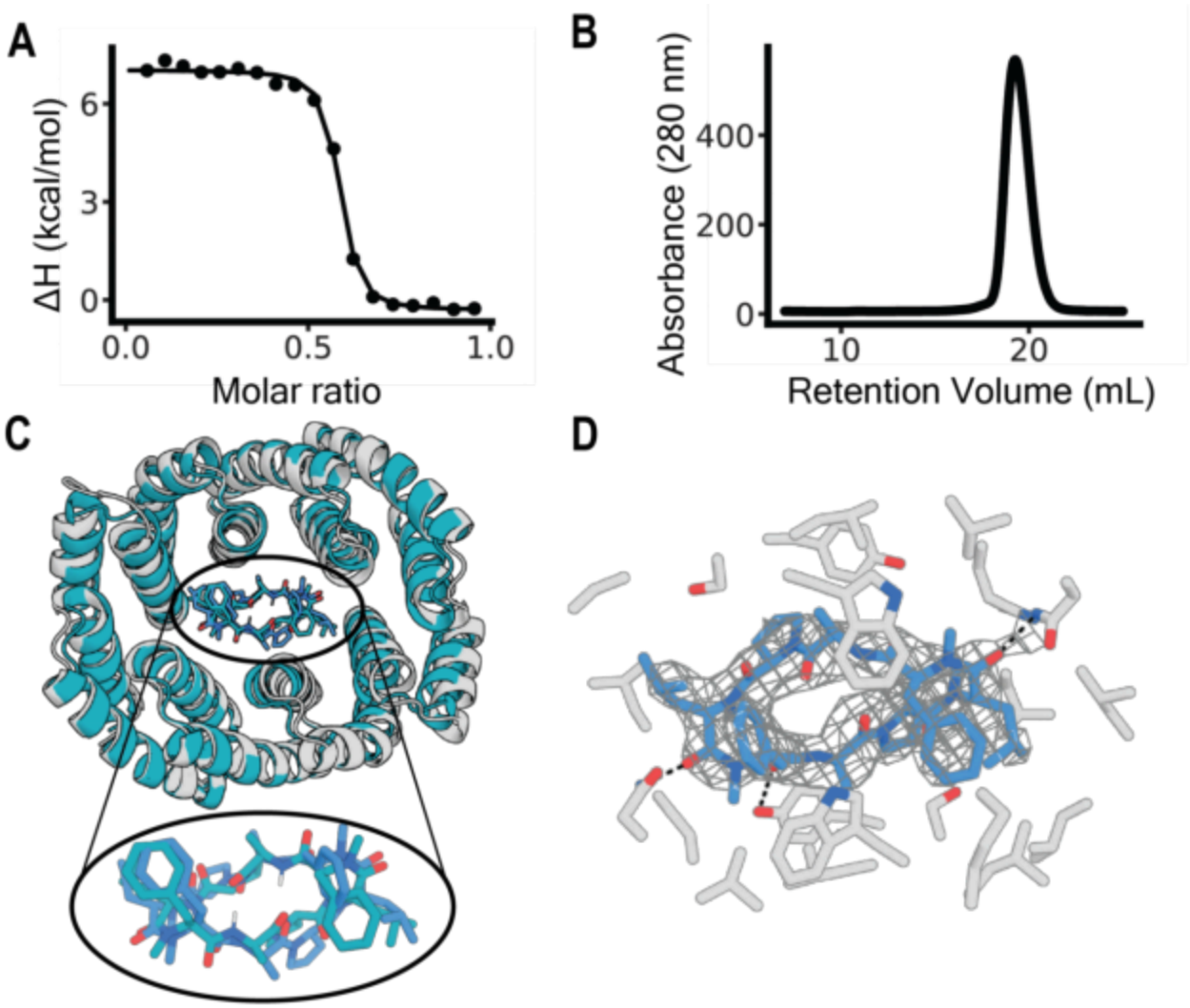


Figure 2. Structural and biophysical characterization of the designed macrocycle-protein interaction. a Titration curve starting with 250 μM of MC1 with CID7 obtained from ITC. K_D was measured at 36 nM. Data for one titration is shown. **b** CID7 SEC trace. Absorbance was measured at 280 nm. **c** Overlay of the design model (teal) with the crystal structure (gray). RMSD between the protein chains is 0.95 Å. Zoom of the binding pocket shows MC1 of the design model (teal) and MC1 of the crystal structure (blue). RMSD between the two structures is 0.97 Å. **d** Zoom in of the crystal structure binding pocket. Protein residues (gray) within 4 Å of MC1 (blue) are shown and hydrogen bonds are indicated with dashed black lines. Electron density around MC1 is shown in gray mesh with contour level 2.

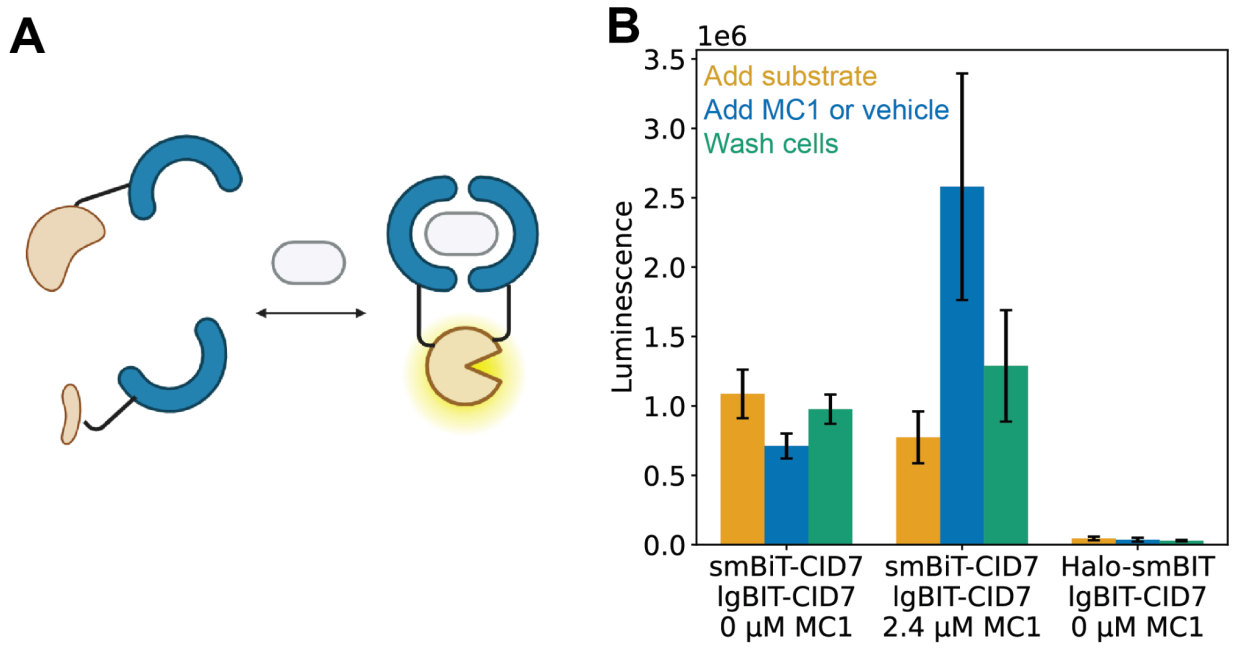


Figure 3. Inducibility and reversibility of MC1-dependent dimerization using split luciferase assay in HEK293T cells. a Schematic representation of the split luciferase assay. CID7 proteins (blue) were expressed in HEK293T cells as fusions with either small or large bit luciferase. Dimerization of the CID7 proteins reconstitutes the luciferase, which then can convert substrate to product, producing a luminescent signal. Created in BioRender.com. **b** HEK293T cells were transiently transfected with either LgBiT-CID7 and smBiT-CID7 or smBiT-CID7 and Halo-LgBiT. Substrate was added to cells and yellow bars represent the mean luminescence of each sample after substrate addition. Cells were then treated with either vehicle (0 µM, Opti-MEM with 2.5% DMSO) or 2.4 µM MC1, and mean luminescence was plotted (blue). Cells were washed with vehicle, additional substrate was added, and the mean luminescence was plotted (green). n=3 biological replicates from individual transfections were used for all conditions tested. Error bars represent standard deviation of the mean.

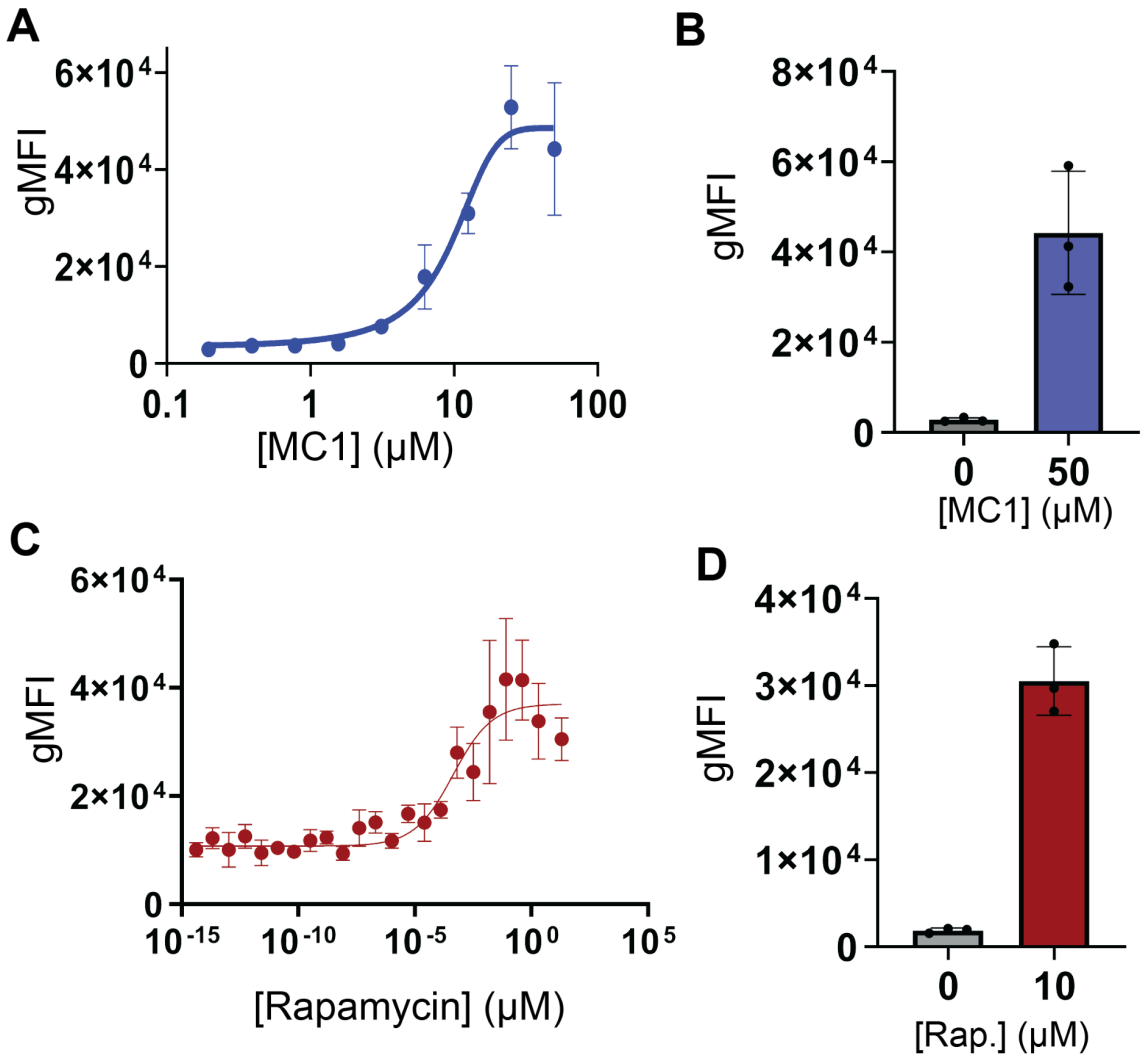


Figure 4. Designed CID system enables inducible transcription control. a HEK293T cells were transiently transfected with CID7-AD, CID7-DBD, and reporter. Cells were incubated with MC1 titrated in a 1 in 2, serial dilution with a starting concentration of 50 μM. Each concentration point plotted is the average of n=3 biological replicates from individual transfections. Error bars indicate the standard deviation from the mean of eGFP geometric mean fluorescence intensity (gMFI). **b** eGFP gMFI of cells transfected with CID7-DBD, CID7-AD, and reporter plasmids treated with either 0 μM MC1 (grey) or 50 μM (blue), Error bars represent standard deviation of the mean. n=3 biological replicates from individual transfections were plotted for both control and ligand treated samples. **c** HEK293T cells were transiently transfected with DBD-FRB, AD-FKBP, and reporter were incubated with rapamycin titrated in a 1 in 5, serial dilution with a starting concentration of 10 μM. Each concentration point plotted is the average eGFP gMFI of n=3 biological replicates from individual transfections. Error bars indicate the standard deviation from the mean. **d** eGFP gMFI of cells transfected with DBD-FRB, AD-FKBP, and reporter plasmids treated with 0 μM rapamycin (grey) or 10 μM

(red), Error bars represent standard deviation of the mean. n=3 biological replicates from individual transfections were plotted for both control and ligand treated samples.

Methods

Protein design

In brief, ProteinMPNN was used to generate ten new sequences for each previously designed homodimeric scaffold backbone. After MPNN sequence design, structures were created with AF2 and filtered for a Cα RMSD less than 2 Å to the input model, resulting in approximately thirteen thousand structures for binder design with MC1. Rotamer interaction fields were generated for MC1 and then used to dock the peptide into the scaffolds. Docks were filtered by rosetta metrics (interface area above 900 $Å^2$, interface shape complementarity greater than 0.7, ΔΔG less than -40, contact molecular surface above 480 $Å^2$, contact patch above 440 $Å^2$). The protein peptide interface of docks passing these metrics was then redesigned with Rosetta FastDesign. The resulting designs were filtered for contact molecular surface above 460 $Å^2$, ΔΔG less than -40, at least three hydrogen bonds between the ligand and protein, interface shape complementarity above 0.72, interface area above 959 $Å^2$, and for designs with less than one buried unsatisfied hydrogen bond. An AF2 structure prediction was generated and designs were filtered for pLDDT above 90, RMSD less than 1.1 Å between the design and prediction model, and a pTM score greater than 0.8, resulting in 263 designs.

Protein expression and purification

Plasmids were constructed by either (1) ordering synthetic genes with an Avi tag from Integrated DNA Technologies (IDT) in pET-29b (+), resulting in the expression product: GLNDIFEAQKIEWHEGHHHHHHGSGSGENLYFQSGSGSSS[POI] or (2) ordering synthetic genes as a gblock from IDT then cloning the synthetic gene fragment into a custom entry vector with golden gate assembly (GGA). An Echo acoustic liquid handler (Beckman Coulter) was used to dispense 1 µL reaction volumes (0.024 µL water, 0.100 µL T4 ligase buffer (New England Biolabs), 0.600 µL eblock fragment at 4 ng/µl, 0.060 µL BSaI-HFv2 (New England Biolabs), 0.100 µL T4 ligase (New England Biolabs), 0.116 µL entry vector at 115 ng/µl). Reaction was placed in thermocycler (Biorad T100) with the lid heated to 105 °C and subjected to 37℃ for 20 min followed by 60℃ for 5 min. The custom vector, LM627 (addgene 191551), yields the expression product MSG[POI]GSGSHHWGSTHHHHHH[24]. 1 µL of plasmid was transformed in BL21 *E. coli* cells (NEB) and cells were incubated on ice for 20 min then heat shocked for 10 s. Cells were recovered in 100 µL SOC media for 1 h at 37℃ then plated onto lysogeny broth (LB, MpBio) agar plates under kanamycin antibiotic selection (50 µg/mL, Teknova) and grown overnight at 37℃. A single colony was inoculated into Terrific broth (TB) II autoinduction media (MpBio) containing 50x5052 (glycerol 25% (v/w), glucose 2.5% (v/w), α-lactose 10% (v/w)), 2 mM $MgSO_4$, and trace metal for overnight protein expression at 37℃. After approximately 20 h cells were harvested by centrifugation (4000 g, 5 min), resuspended in cell lysis buffer (20 mM Tris, 300mM NaCl, 25 mM imidazole, pH 8.0, 0.1 mg/mL lysozyme, 0.01

mg mL-1, Deoxyribonuclease I (DNAse I) and 1 mM Phenylmethylsulfonyl fluoride (PMSF), and sonicated using Qsonica, Q500 with a 4-pronged horn and 10 s ON, 10 s OFF, 45% amplitude for 5 min total ON time. Cellular debris was pelleted by centrifugation (16000 g for 25 min) and the proteins were purified from lysate supernatants by immobilized metal affinity chromatography (IMAC). Lysates were incubated with 500 µL of nickel-nitrolacetic acid (Ni-NTA) beads (Qiagen) in a 50 ml column (Biorad) and then washed with 3 column volumes (CV) of wash buffer (20 mM Tris, 300mM NaCl, 25 mM imidazole, pH 8.0). Proteins were eluted in 1.2 mL of elution buffer (20 mM Tris, 300mM NaCl, 300 mM imidazole, pH 8.0).

Proteins were then purified by SEC using an Akta Pure with an autosampler (Cytiva) on a Superdex S200 Increase 10/300 GL column (Cytiva 28990944) at room temperature at a flow rate of 0.75 ml/min. Proteins were purified into 25 mM Tris, 100 mM NaCl, pH 8.0 (SEC buffer). Fractions of interest were pooled, protein identity validated by LC-MS. Intact mass spectra was obtained via reverse-phase LC/MS on an Agilent G6230B TOF on an AdvanceBio RP-Desalting column, and subsequently deconvoluted by way of Bioconfirm using a total entropy algorithm.

For X-ray crystallography, the SNACtag was cleaved in solution[25]. Following IMAC purification, protein was exchanged into SNAC cleavage buffer (100 mM CHES, 100 mM NaCl, 100 mM acetone oxime, 500 mM guanidine HCl, pH 8.6). 2 mM NiCl2 was added and the protein solution was incubated overnight at 37°C. Solutions were then incubated with Ni-NTA resin to bind uncleaved products and the flow through was collected.

MC synthesis

MCs were synthesized via solid phase peptide synthesis by WuXi Apptec using standard Fmoc-based solid-phase peptide synthesis followed by solution phase cyclization and purification by reverse-phase HPLC[17].

MC Crystallography

MC1 crystals were grown via vapor diffusion of Pentane into ethylacetate. Briefly, 1-3 mg of MC1 was dissolved in approximately 200 µL of ethyl acetate in a dram vial. The dram vial was placed into a scintillation vial containing approximately 3 mL of pentanes. All other MC crystals were grown via vapor diffusion of water into acetonitrile (ACN). Briefly, 1-3 mg of macrocycle was dissolved in approximately 200 µL of ACN in a dram vial. The dram vial was placed into a scintillation vial containing approximately 3 mL of water. The scintillation vials were sealed and monitored for crystal growth. Crystals diffraction data were collected from a single crystal at synchrotron (on APS 24ID-C) and at 100 K. All crystal diffraction data was collected to 0.83 Å resolution. Unit cell refinement, and data reduction were performed using XDS and CCP4 suites[26, 27]. The structure was identified by direct methods and refined by full-matrix least-squares on F^2 with anisotropic displacement parameters for the non-H atoms using SHELXL-2018/3[28, 29]. Structure analysis was aided by using Coot/Shelxle[30, 31]. The

hydrogen atoms on heavy atoms were calculated in ideal positions with isotropic displacement parameters set to $1.2 \times U_{eq}$ of the attached atoms. All designed MCs in this paper with solved crystal structures are summarized in Supplementary Table 1 and were deposited in Cambridge Crystallographic Data Centre (CCDC) with deposition numbers 2354604 (MC1, ethyl acetate conditions), 2354605 (MC1, ACN conditions), 2354606 (MC3), 2354607 (MC2), 2354608 (MC4), 2354609 (MC5), and 2360904 (MC6), and 2354610 (MC6-L4F).

X-ray crystallography of CID7-MC1 complex

Purified protein CID7 with and without its ligand, MC1, was initially tested for crystallization via sparse matrix screens in 96-well sitting drops (200 nL drop volumes versus 95 µL reservoir volumes) using a mosquito crystallization robot (TTP LabTech). Crystallization conditions were then optimized with constructs that proved capable of crystallizing in larger 24-well hanging drops corresponding to initial mixtures of 1 µL well solution and 1 µL protein solution equilibrated against 1000 µL reservoirs. Crystals were grown in the presence of the ligands.

The crystal of the designed protein CID7 was grown from 100 mM sodium acetate pH 4.5 and 8% PEG 4000 at a protein dimer concentration of 153 µM. The ligand MC1 was added to the protein CID7 at a concentration twice the protein dimer and 3% final DMSO and incubated for 20 minutes before being added to crystal trays. The crystal was transferred to a solution containing 25% PEG 4000 and flash-frozen in liquid nitrogen. Data were collected at ALS Beamline 5.0.1 and processed using the program HKL2000[32].

The structure of CID7 in complex with MC1 was solved by molecular replacement with Phaser via PHENIX using the coordinates of the computationally designed structure as a search model[33, 34]. The structure was then built and refined using Coot and PHENIX, respectively[35, 33]. The protein and solvent were modeled and refined first (while avoiding the modeling of any atoms within the binding site), and then the circular peptide ligand was built into the unambiguous density, with the last step being the addition of any missing waters. MC1 energies were calculated using eLBOW via PHENIX[34, 36]. Final Ramachandran statistics after refinement are listed in Supplementary Table 2. The co-crystal structure was deposited in the PDB with ascension code 8VX7.

Initial binding screen by equilibrium dialysis

6 µM MC1 stock solution was prepared in SEC buffer with 5% DMSO. Protein stock solutions were prepared in SEC buffer at either 20 or 100 µM. 50 µL protein solution was placed on one side of a Rapid Equilibrium Dialysis (RED) Device Inserts, 8K molecular weight cut off (MWCO) dialysis cassette (Thermo Fisher 89309) and dialyzed against 300 µL of MC1 stock solution. Dialysis cassettes were placed on 48 well plate and shaken (300 rpm, 30℃) for 5 h. 20 µL aliquots of each side of the cassette were diluted with 140 µL 90% acetonitrile (ACN), 0.1% formic acid (FA). Protein was then crashed out by centrifugation at 10,000 g for 10 min. Presence of MC1 on either side of the dialysis cassette for each protein sample was determined by injecting aliquots onto a Waters I-class ultra-performance liquid chromatography (UPLC)

equipped with an AB Sciex 5600 quadrupole time of flight (QTOF) mass spectrometer. A calibration curve was generated for MC1 and was used, along with the area integration values of the extracted ion current (EIC) of MC1 mass, to determine the concentration of MC1 on either side of the cassette for each experimental sample. EIC peaks were integrated with Analyst® TF 1.6 and MultiQuant™ Software. The area ratio of MC1 on each side of the dialysis cassette for each respective sample was used to assess whether MC1 was binding each respective protein. CID7 showed area ratios greater than one in both the 20 and 100 μM protein sample experiments, suggesting CID7 bound MC1 (Supplementary Fig. S3).

Characterization of binding affinity by ITC

SEC purified proteins were diluted to 50 μM in 20 mM Tris, 200 mM NaCl pH 8, 5% DMSO. Peptide solution was prepared at 250 μM in buffer matching that of the protein sample and was titrated into the protein sample. A control of 20 mM Tris, 200 mM NaCl pH 8, 5% DMSO was run for each experimental sample. ITC experiments were performed on an automated Microcal PEAQ-ITC. All titrations were performed with 19 total injections, where the first injection contained 0.4 μL of MC1 solution and the subsequent 18 injections contained 2 μL MC1 solution. The Microcal PEAQ-ITC Analysis Software was used to fit curves for dissociation constant determination.

Nanobit Assay in HEK293T cells

Plasmids were prepared by first digesting 1.5 μg of multiple cloning site vectors (Promega N2014) with NheI-HF and EcoRI-HF (New England Bio Labs). CID7 genes were ordered as gblocks from Twist Biosciences and digested with NheI-HF and EcoRI-HF prior to ligation with the respective linearized plasmid. Respective linearized vector and digested CID7 genes were ligated with T4 DNA Ligase (New England Bio Labs) at room temperature for approximately 20 minutes and the enzyme was heat inactivated at 65°C for 5 min. Approximately 2 μL of each ligation reaction was added to 50 μL 5α cells (New England Biosciences C2987), the cells rested on ice for 30 minutes, then were heat shocked at 42°C for 30 s. Cells were incubated with 200 μL SOC at 37°C for 1 h then plated on agar plates containing 50 μg/mL carbenicillin and allowed to incubate overnight at 37°C. Single colonies were grown in 5 mL LB with 50 μg/mL carbenicillin for approximately 20 h at 37°C. Colonies with the correctly assembled plasmid were purified by miniprep (Zymogen D4201) and sequence verified by whole plasmid sequencing (Plasmidsaurus).

HEK293T cells were plated onto 96-well TC-treated plates and allowed to reach 80% confluency prior to transfection. Plasmids stocks were diluted in Opti-MEM (Thermo Fisher 31985070) to a concentration of 6.25 ng/μl per construct. Viafect transfection reagent (Promega E4981) at 3:1 DNA : transfection reagent was added and the solution was mixed by pipetting up and down. Transfection solution sat for 10 min before it was added to each respective well.

Media containing transfection reagent was removed and 100 µL Opti-MEM with 2.5% DMSO was added to each well. Nano-Glo® Live Cell substrate was diluted 1 in 20 in Nano-Glo® LCS Dilution Buffer (Promega N2014), 25 µL was added to each well and luminescence was monitored for 20 min. 10 µL of either Opti-MEM with 2.5% DMSO (vehicle) or 2.4 uM MC1 in Opti-MEM with 2.5% DMSO was added to respective wells. Luminescence was measured continuously for another 1.5 h. To assess reversibility, cells were rinsed 3X with 100 µL vehicle, 110 µL vehicle was added to each well post rinsing, and then 25 µL of diluted substrate was added. Luminescence was recorded for another 30 min. All luminescence readings were recorded on a Synergy Neo2 plate reader with a LUM filter cube at 37°C. The maximum luminescence value occurring in each step of the experiment (initial addition of substrate, addition of compound or vehicle, compound washout and addition of more substrate) for each technical replicate was used to compute mean luminescence for plotting. All control samples received 10 µL vehicle at the compound addition stage. Fold difference in luminescence is calculated by dividing the mean luminescence signal of technical replicates after compound or vehicle addition by the mean luminescence signal of technical replicates before compound or vehicle addition.

Transcriptional assay in HEK293T cells

Plasmids were prepared by ordering synthetic genes encoding the designed proteins from IDT as eBlocks, and contained overhangs for Golden Gate Assembly (GGA) into custom-made entry vectors. These entry vectors contained an ampicillin resistance gene as well as a ccdb lethal gene between the SapI cut sites. Subcloning into MA0005 results in a [POI]-zinc finger DNA binding domain expression product, while subcloning into MA0006 results in [POI]-p65 transactivation domain expression product. An Echo acoustic liquid handler (Beckman Coulter) was used to dispense 1 µL reaction volumes (0.130 µL water, 0.100 µL T4 ligase buffer, 0.425 µL eblock fragment at 4 ng/µl, 0.150 µL of SapI (New England Biolabs), 0.125 µL T4 ligase, 0.070 µL entry vector at 200 ng/µl). 1 µL of reaction mixtures were then transformed in 5-alpha competent *E. coli* cells (New England Biolabs) and cells were incubated on ice for 20 min then heat shocked for 10 s. Cells were recovered in 100 µL SOC media for 1 h at 37℃ and then plated on LB agar plates under carbenicillin selection with 100 µg/mL carbenicillin. A single colony of each design was inoculated into LB and the plasmids were grown overnight at 37℃ under carbenicillin antibiotic selection (100 µg/mL). Plasmids were then purified by miniprep.

An echo acoustic liquid handler was used to aliquot plasmid solutions into the respective wells of a non-tissue culture treated 384-well plate. Each well was filled to a final volume of 5 µL with nuclease free water and mixed with transfection reagent using a 24-channel pipette.

HEK293T cells were cultured in tissue culture-treated 384-well plates. Each well contained 50 µL of HEK293T cells at a density of 600,000 cells/mL in cell media (Dulbecco's Modified Eagle Medium with Glutamax (DMEM, Thermo Fisher), 10% fetal bovine serum, heat inactivated (FBS, Thermo Fisher), 1X penicillin strep media (10,000 U/ml, Thermo Fisher)). At

approximately 80% confluency cells were then transfected with the respective vectors. Cells expressing CID7 were transfected with an equal ratio of plasmid (150 ng total plasmid) using 150 mM NaCl and 3X polyethylenimine (PEI) solution. After approximately 10 minutes the plasmid transfection reagent was added to cultured cells. Cells were incubated with plasmid and transfection reagent for approximately 24 h at 37℃. After incubation with the transfection reagent, media was removed and replaced with new cell media. MC1 was serially diluted 1 in 2 from 50 μM across nine wells in cell media with 1% DMSO. Control wells were made by adding DMSO to cell media to a final concentration of 1%. Cells were incubated with new media for approximately 24 h at 37℃ before analysis by flow cytometry.

Cells expressing rapamycin binding proteins were transfected using 150 mM NaCl and 3X Viafect transfection reagent. After approximately 10 minutes, the plasmid Viafect solution was pipetted onto HEK293T cultured cells. Following 48 h incubation with transfection reagent, the media was removed from cells and new media was added. Rapamycin (Millipore Sigma) dissolved in DMSO prior to addition to cell media and then serially diluted 1 in 5 from 10 μM across twenty-three wells in cell media with 1% DMSO. Control wells were made by adding DMSO to cell media to a final concentration of 1%. Cells were incubated with new media for approximately 24 h at 37℃ before analysis by flow cytometry.

To prepare cells for analysis with flow cytometry, media was removed from the cells and the cells detached by incubating with 40 μL Trypsin-EDTA (0.25%, Thermo Fisher) for approximately 10 min. 50 μL of cell media was added to each well and cells were then resuspended by pipetting. The gating strategy for flow cytometry data is outlined in supplementary Fig. S5.

Flow cytometry

Experiments were analyzed on an Thermo Fisher Attune NxT flow cytometer. For analysis of GFP expression, cells were gated on live singlets. To ensure only cells that were double transformed with both design-containing plasmids were included in downstream GFP quantification, cells were then gated for the expression of RFP and BFP. Finally, cells were gated for GFP fluorescence. The data were analyzed using FlowJo Software.

References

1. Di Stasi A et. al. Inducible apoptosis as a safety switch for adoptive cell therapy. N Engl J Med. 3, 1673-83 (2011).
2. Holland, A.J. et. al. Inducible, reversible system for the rapid and complete degradation of proteins in mammalian cells. Proc. Natl. Acad. Sci. U.S.A. 109, 3350–3357 (2012).
3. Zetsche, B. et. al. A split-Cas9 architecture for inducible genome editing and transcription modulation. Nat Biotechnol 33, 139–142 (2015).
4. Putyrski, M. et al. Protein translocation as a tool: The current rapamycin story. FEBS Letters. 586, 2097-2105 (2012).

5. Clackson, T. et. al. Redesigning an FKBP–ligand interface to generate chemical dimerizers with novel specificity. Proc. Natl. Acad. Sci. U.S.A. 95, 10437-10442 (1998).
6. Karin C. Straathof, et. al. An inducible caspase 9 safety switch for T-cell therapy. Blood 105 (11): 4247–4254 (2005).
7. Liang FS, et. al. Engineering the ABA plant stress pathway for regulation of induced proximity. Sci Signal. 4, (2011).
8. Miyamoto, T. et al. Rapid and orthogonal logic gating with a gibberellin-induced dimerization system Nat. Chem. Biol., 8, 465-470 (2012).
9. Ziegler, M.J., Yserentant, K., Dunsing, V. et al. Mandipropamid as a chemical inducer of proximity for in vivo applications. Nat Chem Biol 18, 64–69 (2022).
10. Hill, Z. B., Martinko, A. J., Nguyen, D. P. & Wells, J. A. Human antibody-based chemically induced dimerizers for cell therapeutic applications. Nat. Chem. Biol. 14, 112–117 (2018).
11. Chin, S.E., Schindler, C., Vinall, L. et al. A simeprevir-inducible molecular switch for the control of cell and gene therapies. Nat Commun 14, 7753 (2023).
12. Guo, Z., Smutok, O., Johnston, W.A. et al. Design of a methotrexate-controlled chemical dimerization system and its use in bio-electronic devices. Nat Commun 12, 7137 (2021).
13. Kang, S. et al. COMBINES-CID: an efficient method for de novo engineering of highly specific chemically induced protein dimerization systems. J. Am. Chem. Soc. 141, 10948–10952 (2019).
14. Anum A. Glasgow et al. ,Computational design of a modular protein sense-response system. Science 366,1024-1028 (2019).
15. Foight GW, et al. Multi-input chemical control of protein dimerization for programming graded cellular responses. Nat Biotechnol. 37, 1209-1216 (2019).
16. Bhardwaj G. et al. Accurate de novo design of membrane-traversing macrocycles. Cell 185, 3520–3532 (2022).
17. Salveson, P.J. et al. Expansive discovery of chemically diverse structured macrocyclic oligoamides. Science 384, 420-428 (2024).
18. Marelli, K.U. et al., cis-Peptide Bonds: A Key for intestinal Permeability of Peptides? Chemistry, 21, 15148-15152 (2015).
19. Renatus M., et al. Dimer formation drives the activation of the cell death protease caspase 9. Proc. Natl. Acad. Sci. U.S.A. 98, 14250–14255 (2001).
20. Hicks, D. R. et al. De novo design of protein homodimers containing tunable symmetric protein pockets. Proc. Natl Acad. Sci. USA 119, e2113400119 (2022).
21. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science 378, 49-56 (2022).
22. Dou, J., Vorobieva, A.A. et al. De novo design of a fluorescence-activating β-barrel. Nature 561, 485–491 (2018).
23. He R. et al. Mammalian two-hybrid assay for detecting protein-protein interactions in vivo. Methods Mol Biol. 439, 327-337 (2008).

24. B. I. M. Wicky et al. Hallucinating symmetric protein assemblies. Science 378, 56-61 (2022).
25. Dang, B., Mravic, M., Hu, H. et al. SNAC-tag for sequence-specific chemical protein cleavage. Nat Methods 16, 319–322 (2019).
26. Kabsch, W. XDS. Acta Cryst. D66, 125-132 (2010).
27. Winn, M.D. et al. Acta. Cryst. D67, 235-242 (2011).
28. Sheldrick, G. M. SHELXT-Integrated space-group and crystal-structure determination. Acta Crystallographica Section A: Foundations and Advances 71, 3–8 (2015).
29. Sheldrick, G. M. Crystal structure refinement with SHELXL. Acta Crystallographica Section C: Structural Chemistry 71, 3–8 (2015).
30. Emsley, P. and Cowtan, K. Coot: Model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132 (2004).
31. Hübschle, C.B. , Sheldrick, G.M. and Dittrich, B. J. Appl. Cryst., 44, 1281-1284 (2011).
32. Otwinowski, Z. and Minor, W. Processing of X-ray diffraction data collected in oscillation mode. Macromolecular Crystallography, 276, 307-326 (1997).
33. McCoy, A.J. et al. Phaser crystallographic software. J Appl Crystallogr, 40(Pt 4), 658-674 (2007).
34. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr, 66(Pt 2): p. 213-21 (2010).
35. Emsley, P., et al., Features and development of Coot. Acta Crystallographica, 66, 486-501 (2007).
36. Moriarty, N.W., R.W. Grosse-Kunstleve, and P.D. Adams, electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. Acta Crystallogr D Biol Crystallogr, 65(Pt 10), 1074-1080 (2009)

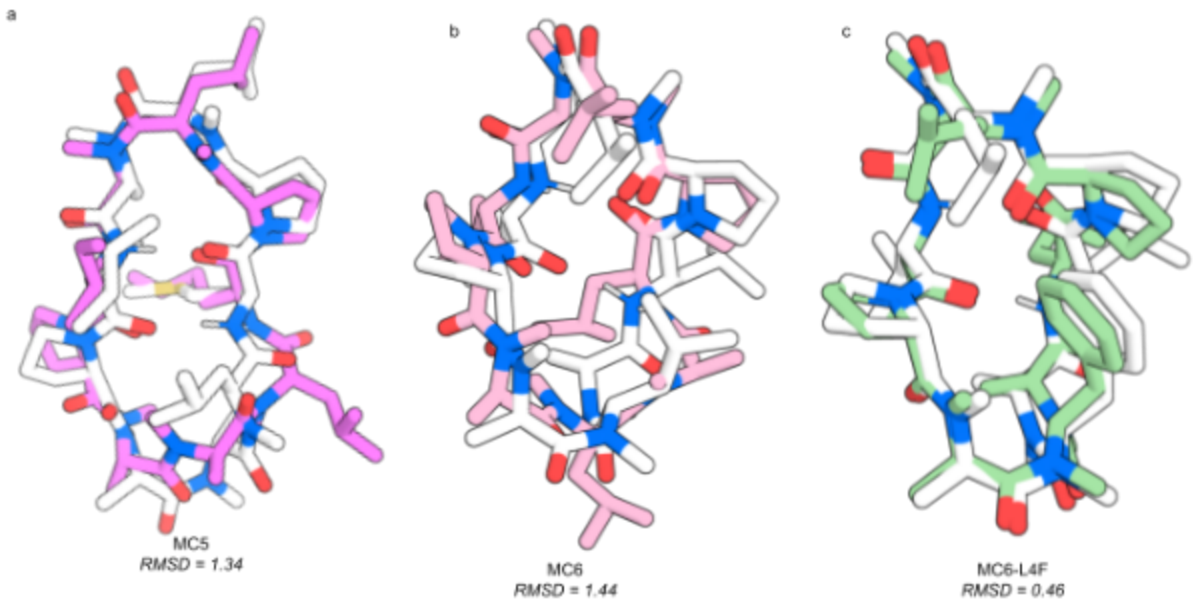
Supplementary Materials

Peptide	Design Sequence	Experimental Sequence	C-alpha RMSD (Å)	All atom RMSD (Å)	Calculated mass (g/mol)	Deposition number
MC1	apL*F*apL*F*	apL*F*apL*F*	0.63	1.89	912.6	2354605, 2354604
MC2	apL*F*FPa*l*	apL*F*FPa*l*	0.68	1.16	912.6	2354607
MC3	fpL*a*APl*p*	fpL*a*APl*f*	0.78	1.89	912.6	2354606
MC4	MPa*l*iM*Pa*P	XPa*l*iX*Pa*P	0.39	0.63	941.6	2354608

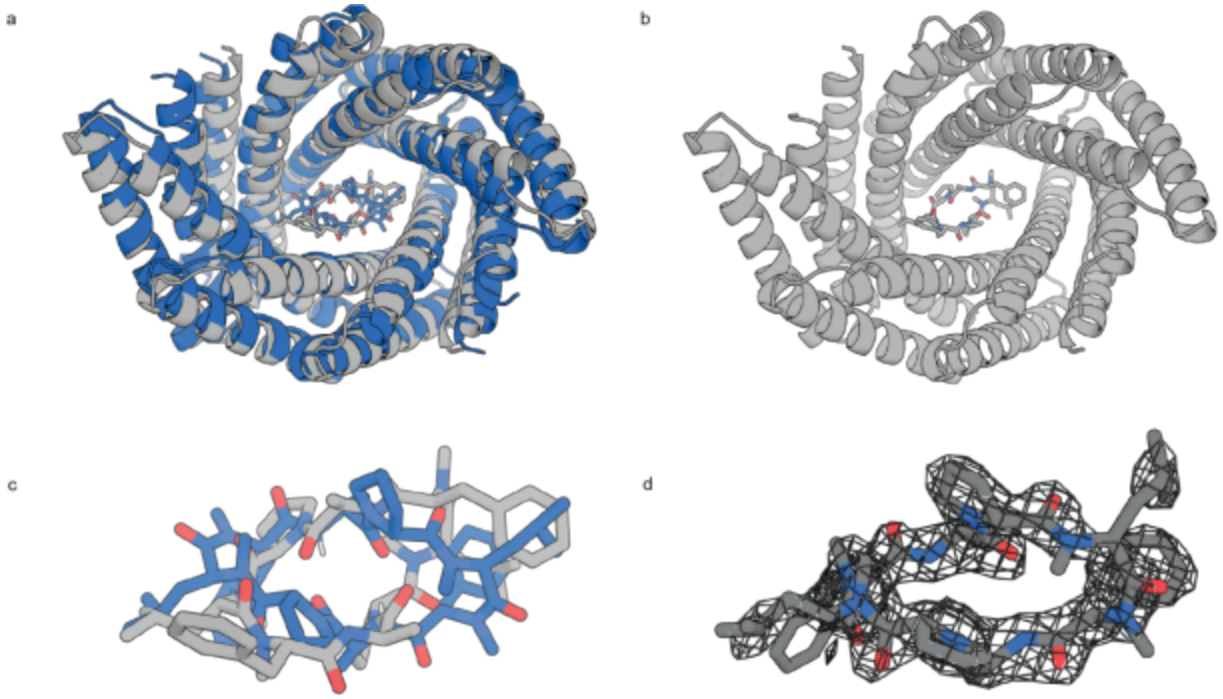
MC5	MPl*a*ipA*a*L*	XPl*a*ipA*a*L	1.34		929.6	2354609
MC6	VPa*L*aLpL*a*	VPa*L*aLpL*a*	1.44		915.6	2360904
MC6-L4F	VPa*F*a*LpL*a*	VPa*F*aLpL*a*	0.46		949.6	2354610

Supplementary Table 1. Summary of designed macrocycles with solved x-ray crystal structures. Crystal structures for seven designed peptides were solved. Lower case in sequence represents D-amino acids. An asterisk following an amino acid residue represents N-methylation of the amide. X represents norleucine. C-alpha RMSD was measured between the design model and crystal structure. All atom RMSD was measured between all heavy atoms of the design models and crystal structures. Molecular weight was calculated for the experimental sequences that were synthesized via solid phase peptide synthesis and crystallized. Cambridge Crystallographic Data Centre (CCDC) numbers are listed for each structure. Deposition numbers for MC1 are for acetonitrile and ethyl acetate crystallization conditions respectively.

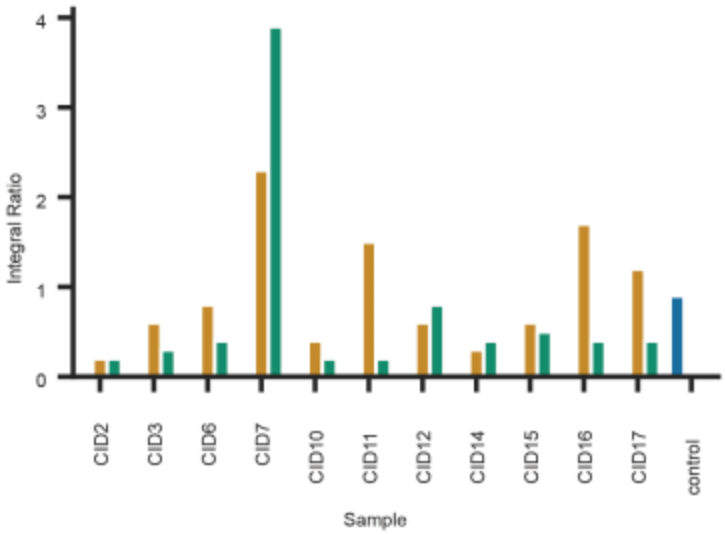
Supplementary Note 1:
While MC1, MC2, MC3, and MC4 showed close alignment to the design model, MC5 and MC6 crystal structures were slightly different from the design model. One loop of MC5 is in agreement between the design model and crystal structure, whereas the other loop is misfolded (Figure S1A). MC6 was initially designed with leucine residues at positions 4, 6, and 8, and the crystal structure of this compound was solved. Like MC5, the crystal structure showed that one loop of the MC matched the design model whereas the other loop was not aligned. A L4F mutant compound resulted in a forward funnel with a lower energy landscape than the original sequence. Thus, the mutated sequence was crystallized, and the backbone aligns with the design model (Figure S1B-C).



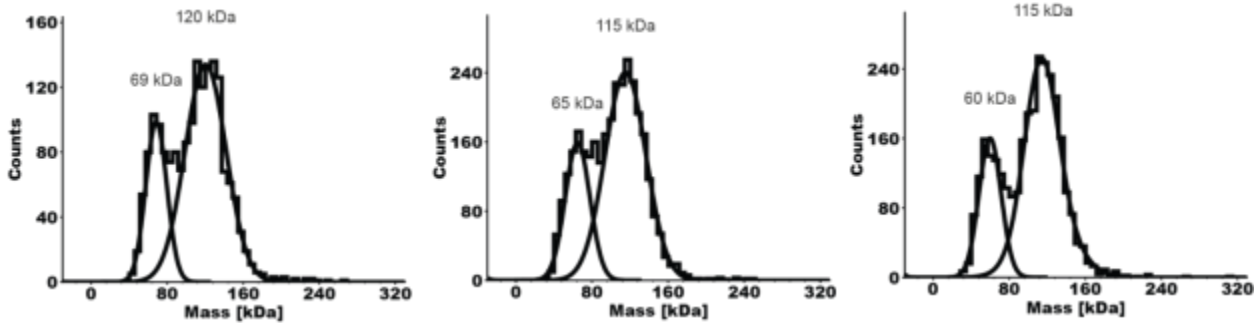
Supplementary Figure 1. Design models and crystal structures of MC5, MC6, and MC6-L4F. a Overlay of MC5 design model (white) with MC5 crystal structure (purple). **b** Overlay of MC6 design model (white) with MC6 crystal structure (pink). **c** Overlay of MC6-L4F design model (white) with MC6-L4F crystal structure (green). The C-alpha RMSDs between the design models and crystal structures are depicted beneath each MC.



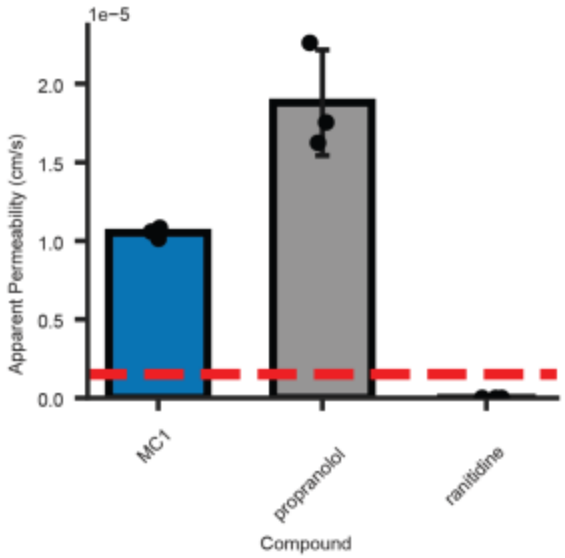
Supplementary Figure 2. Overlay of the design model for original MC1 binding protein, D_3_633_8x, with the crystal structure. a. Design model (blue) overlaid with the crystal strcture (grey) of the ligand bound protein. **b.** Crystal structure of the D_3_633 protein bound to MC1. **c.** Zoom in of MC1 of the design model (blue) overlaid with MC1 of the crystal structure (grey). **d.** Electron density map is shown in dark dray over MC1 of the crystal structure with contour level 1.



Supplementary Figure 3. Initial assessment of MC1 binding by equilibrium dialysis. Orange represents protein concentration of 20 µM, while green represents protein concentration of 100 µM dialyzed against 6 µM MC1. The control sample consisted of 25 mM Tris, 100 mM NaCl, pH 8.0 buffer with 0 µM protein dialyzed against 6 µM MC1. Integral ratio represents the ratio of MC1 concentration on the protein side of the cassette to MC1 concentration on the ligand side of the cassette. n=1 samples were used for all conditions tested.



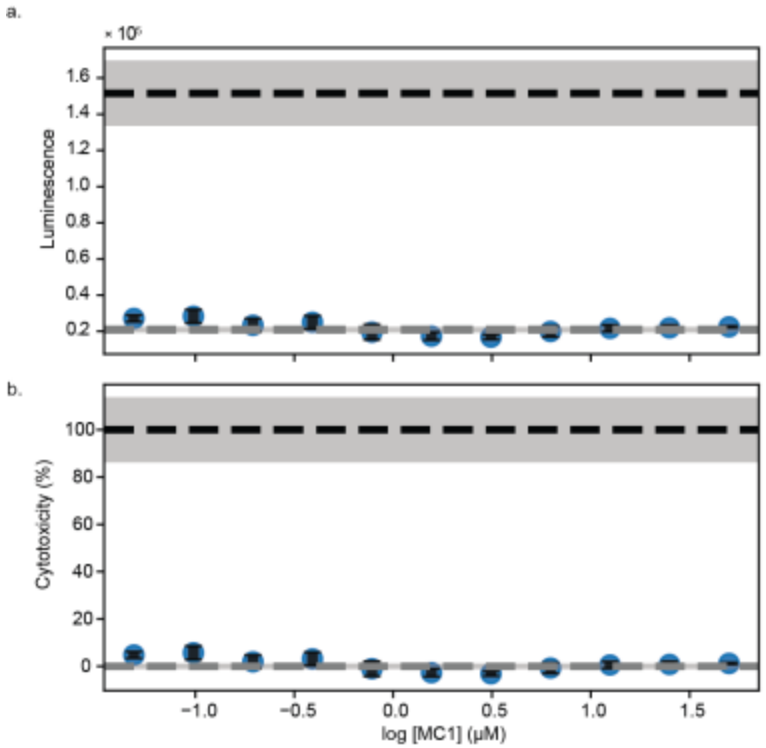
Supplementary Figure 4. MBP-CID7 mass photometry distributions. Mass distributions are shown for n =3 technical replicates containing 9.4 nM total protein. Masses obtained from converted contrast values are listed above each distribution.



Supplementary Figure 5. Membrane Permeability measurement of MC1. Apparent permeability (Pe) was determined in PAMPA assay. Bar height represents mean Pe from n=3 technical replicates; error bars are the standard deviation of the mean calculated from the three technical replicates. Pe values were calculated from calibration curve data fit to quadratic regression. Ranitidine and propranolol were used as negative and positive controls for membrane permeability, respectively. The dashed, red line represents the Pe cutoff for a compound to be considered permeable. n=3 technical replicates were used for all conditions tested.

Analyte	**Transitions**
Ranitidine	315 >130 / 315 > 176 / 315 > 315
Propranolol	260 > 127 / 260 > 144 / 260 > 183 / 4: 260 > 260
MC1	913.53 > 913.53 /930.53 > 930.53 (NH_4^+ adduct) / 935.51 > 935.51 (Na^+ adduct)

Supplementary Table 2. Calculated MS transitions for each analyte in the PAMPA expreiment.



Supplementary Figure 6. Percent cytotoxicity of HEK293T cells treated with MC1. a Luminescence values were plotted against log MC1 concentration. The dashed horizontal line represents the mean of the max LDH release control samples consisting of 3 technical replicates. Data after 48 h incubation was MC1 is plotted. Error bars represent standard deviation of the mean. n=3 technical replicates were used for all conditions tested. **b** Percent cytotoxicity was calculated and plotted against log MC1 concentration. Percent cytotoxicity was calculated for each of three technical replicates after 48 h incubation with MC1. Error bars represent standard deviation of the mean. n=3 technical replicates were used for all conditions tested.

Protein	**Expression Sequence**	**Observed mass (Da)**	**Calculated mass (Da)**
CID7	MSGMLEEIERLVLSGLLTGDKELLKKASELLKEE MEKLLEEGDLDALKKALQLAVNVADHNGDKEL LAHAAEVIKRALDLALEAKDLQSAKYLASLALW IAKRAGDKELYAYLEEKIKKIIELAEEAGDRESLK ILILLGIFIARDAGSEEVKAFVAEQLERLLGSGSH HWGSTHHHHHH	20012	20012
MBP-CID7	SHHHHHHGKIEEGKLVIWINGDKGYNGLAEVGK KFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDII FWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFT	59567	59567

	WDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPK TWEEIPALDKELKAKGKSALMFNLQEPYFTWPLI AADGGYAFKYENGKYDIKDVGVDNAGAKAGLT FLVDLIKNKHMNADTDYSIAEAAFNKGETAMTI NGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFV GVLSAGINAASPNKELAKEFLENYLLTDEGLEAV NKDKPLGAVALKSYEEELAKDPRIAATMENAQK GEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEA LKDAQTGSSGGMLEEIERLVLSGLLTGDKELLKK ASELLKEEMEKLLEEGDLDALKKALQLAVNVAD HNGDKELLAHAAEVIKRALDLALEAKDLQSAK YLASLALWIAKRAGDKELYAYLEEKIKKIIELAEE AGDRESLKILILLGIFIARDAGSEEVKAFVAEQLE RLLGS		

Supplementary Table 3. Summary and characterization of proteins expressed and purified from e. coli. Observed mass represents the mass observed after deconvolution of the mass spectrum.

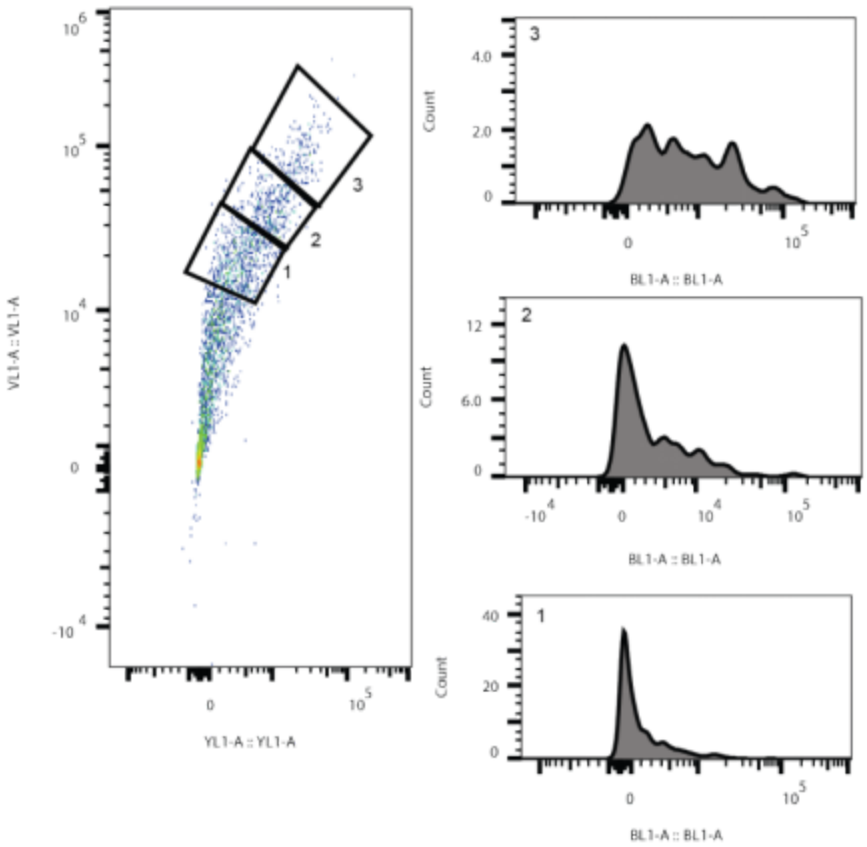
	8TLP	8TM9	8VX7
Data collection			
Space group	P 2_1 2_1 2_1	P 1 2_1 1	P 2_1 2_1 2_1
Cell dimensions			
a, *b*, *c* (Å)	54.38, 55.03, 146.53	54.72, 54.56, 80.48	33.43, 79.80, 105.94
a, b, g (°)	90, 90, 90	90, 96.1, 90	90, 90, 90
Resolution (Å)	44.00 – 2.0 (2.071 – 2.0)	47.4 – 2.05 (2.126 – 2.05)	39.9 – 2.75 (2.848 – 2.75)
R_{merge}	0.056 (1.546)	0.074 (0.42)	0.113 (0.497)
I / sI	31.36 (0.6)	22.05 (2.07)	23.44 (2.5)
Completeness (%)	99.62 (96.41)	99.18 (94.18)	98.75 (91.03)
Redundancy	11.8 (6.4)	6.3 (4.8)	11.6 (9.2)
Refinement			
Resolution (Å)	2.0	2.05	2.75
No. reflections	30453 (2906)	29557 (2789)	7762 (700)
R_{work} / R_{free}	23.3 / 28.1 (34.5 / 38.4)	20.5 / 24.3 (25.2 / 29.2)	25.5 / 30.4 (33.9 / 35.9)

No. atoms			
Protein	450	452	322
Ligand/ion	5	93	66
Water	450	101	3
B-factors			
Protein	52.78	47.52	70.02
Ligand/ion	80.28	66.89	56.20
Water	53.06	46.83	52.18
R.m.s. deviations			
Bond lengths (Å)	0.007	0.002	0.001
Bond angles (°)	0.81	0.40	0.36

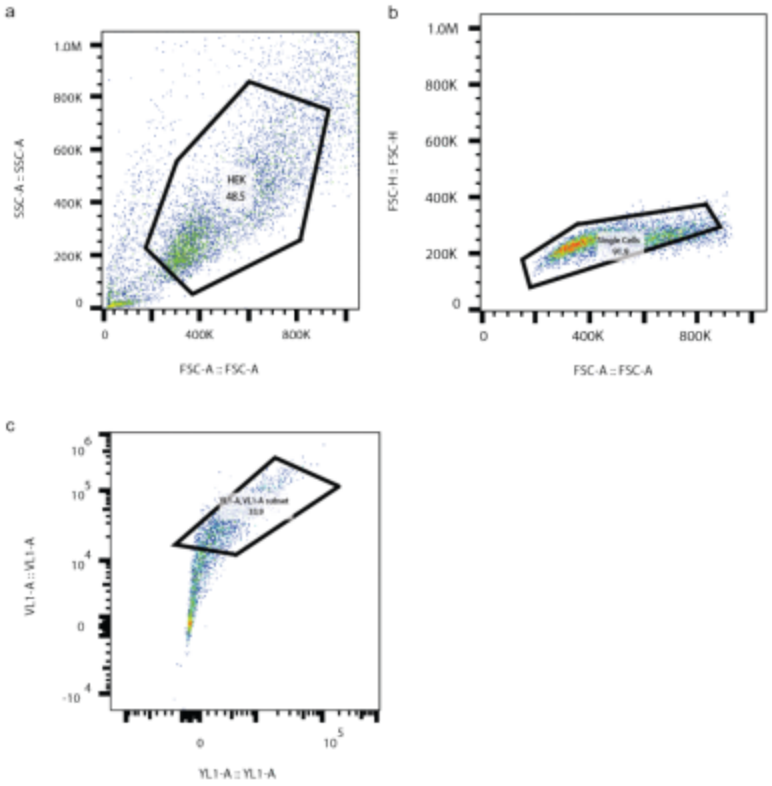
Supplementary Table 4. Final Ramachandran statistics after refinement for protein crystal structures deposited in the PDB.

Insert DNA sequence ordered	Entry Vector
GGACTGCTACGAATTCAATGCTGGAGGAGATCGAGAGG CTGGTGCTTAGCGGCCTGCTGACAGGCGACAAGGAACT GCTCAAGAAAGCCAGCGAGCTGCTGAAGGAGGAAATG GAGAAACTGCTGGAAGAGGGCGACCTGGATGCCCTGA AGAAGGCTCTGCAGTTGGCCGTGAACGTGGCTGACCA CAACGGCGATAAGGAGCTGCTGGCTCACGCTGCCGAA GTGATTAAAAGAGCCCTGGACCTGGCCCTGGAGGCCAA GGACCTTCAGAGCGCTAAGTACTTGGCCAGCCTGGCTC TGTGGATTGCCAAGAGAGCCGGAGACAAAGAGCTGTA CGCCTACCTCGAGGAGAAGATCAAGAAGATTATCGAGC TGGCCGAGGAGGCCGGCGACAGGGAGTCTCTGAAGAT CCTGATTCTGCTGGGCATCTTCATCGCCAGAGACGCCG GAAGCGAAGAGGTGAAGGCCTTCGTGGCCGAACAGTT GGAGAGGCTGCTGTGAGCTAGCGGACGCTAC	pBiT1.1-N [TK/LgBiT] and pBiT2.1-N [TK/SmBiT]

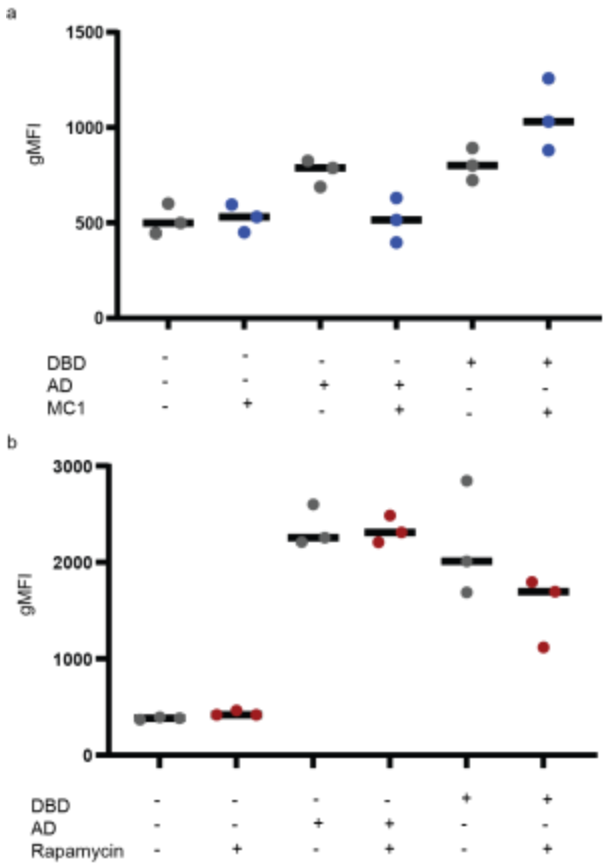
Supplementary Table 7. DNA sequences ordered for split luciferase assay plasmid construction.



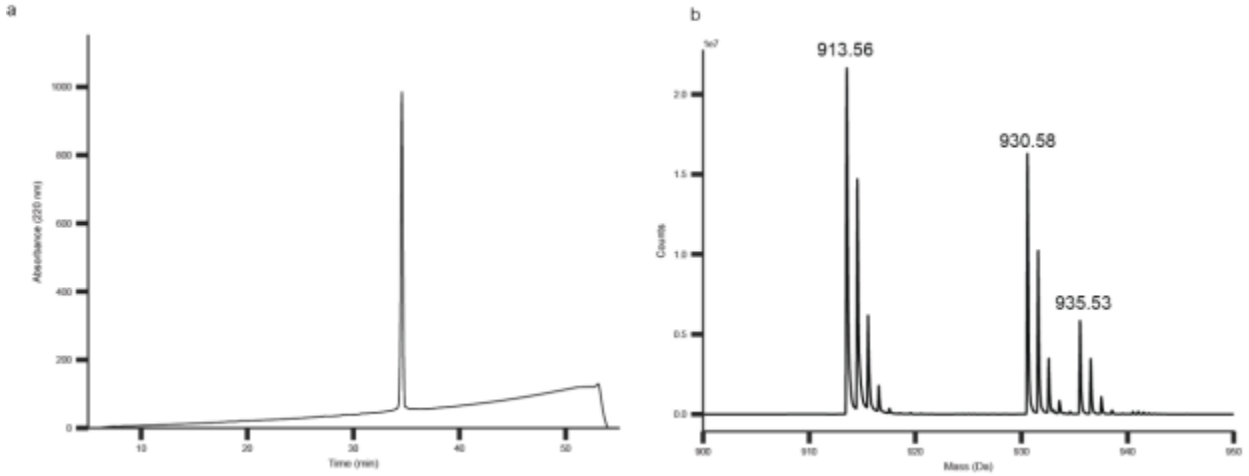
Supplementary Figure 8. Construct expression vs GFP fluorescence of CID7. a HEK293T cells were transfected with CID7-DBD, CID7-AD, and reporter plasmids and then treated with 1% DMSO. mscarlett-1 (assessed in YL1-A channel) vs mtagBFP2 (assesed in VL1-A channel) were plotted to show transfected population. Plots are composed of data from n=1 representative replicate. **b** eGFP fluorescence (assessed in channel BL1-A) distributions from gates 1, 2, and 3 in a.



Supplementary Figure 9. Gating strategy for HEK cells expressing construct proteins. a Gates for live cells. **b** Live cells were then gated for singlet cells. **c** Of the singlet cells, the entire population expressing CID7 construct proteins as measured by RFP (channel YL1-A) and BFP (channel; VL1-A) expression markers. Plots are shown for one well of HEK cells on 384 tissue culture plate that were transfected with all three plasmids and treated with 1% DMSO in cell media. Gates were drawn in and plots were exported from flowjo software. Data represented is for cells expressing CID7 construct proteins.



Supplementary Figure 10. Construct controls for HEK cell experiments with either MC1 or rapamycin CID systems. a HEK cells were transiently transfected with either 150 ng total plasmid of reporter and puc, CID7-DNA binding domain (DBD), reporter, and puc, or CID7-activation domain (AD), reporter and puc. Samples were treated with either 1% DMSO in cell media (grey) or 50 μM MC1 in cell media with 1% DMSO (blue). n=3 technical replicates were used for all conditions tested. is shown for each sample. **b** HEK cells were transiently transfected with either 150 ng total plasmid of reporter and puc, DBD-FRB, reporter, and puc, or AD-FKBP, reporter and puc. Samples were treated with either 1% DMSO in cell media (grey) or 1 μM MC1 (red). Data for three individual transfections is shown for each sample. Geometric mean of the GFP fluorescence for each sample was computed in flowjo and plotted in prism. n=3 technical replicates were used for all conditions tested.



Supplementary Figure 11. Purity trace and mass spectrum of MC1. a Purity of the MC1 compound was 99% by HPLC. **b** The observed mass from the mass spectrum corresponds to the calculated mass of 912.6 Da. 930.6 Da MS product corresponds to the loss of ammonia during ionization.

Supplementary Note 2

The following provides more background on the three step Rifgen, RifDock and Interface design process:

RifGen

RifGen was used to build disembodied hotspot side chains around the C2 symmetric peptide MC1, which had its symmetry axis aligned to the Z-axis and centered at the origin.

RifDock

We modified the inputs to RIFdock to only sample C2 docks. Then RifDock was used to dock the homodimeric scaffolds onto the MC1 hotspot interactions. To do this, we initially aligned the scaffolds symmetry axis to the Z-axis and centered them at the origin to match MC1. Instead of using RifDock's default hierarchical docking and scoring system that docks in 6D space, we provided C2 symmetric transforms to RifDock that moved the scaffolds up and down the Z-axis, while rotating them around the Z-axis, and performed a 180° flip perpendicular to the Z-axis to enumerate C2 symmetric docking orientations. These docks were scored with RifDock's default hash based scoring system and the top scoring docks were output for subsequent design.

Interface Design

The C2-symmetric MC1-scaffold complexes were then designed using the c2_peptide_design.xml script. Briefly, the script extracts an asymmetric unit from the input complex, applies rosetta C2 symmetry, add bonds to cyclize the peptide, designs the MC1-scaffold interface with FastDesign, performs FastRelax, and then repeats FastDesign, and

FastRelax. The designed complexes were scored and various energy metrics were calculated by rosetta.

[2]Chapter 2: Design of protein binders to HIV protease inhibitors

Introduction

The ability to precisely control intracellular processes using chemically induced dimerization (CID) systems has important medical applications. For example, rapamycin or rimiducid-controlled Cas9–induced apoptosis has been used to eliminate CAR T cells in patients experiencing cytokine release syndrome, reversing the adverse effects[1]. While the homodimeric CID systems previously described demonstrate the utility of protein design as a proof-of-concept strategy, further engineering of the system to limit background homodimerization would be needed to increase their broader utility.

Heterodimeric CID systems offer a potential solution to these limitations, and protein design has been successfully applied in proof-of-concept studies to generate such systems. However, the small-molecule inputs used to date have largely been restricted to endogenous ligands or research-grade compounds such as fluorescent dyes. Even the clinically explored rimiducid-based system has drawbacks, as rimiducid is not FDA-approved and rapamycin has off-target effects, activating endogenous mTOR[1]. CID systems consisting of FDA-approved ligands have been made, one such example consists of hepatitis protease inhibitors as input ligands and the NS3a protease along with designed reader proteins as the protein components[2]. Thus, there is interest in developing new CID systems that are orthogonal to one another and utilize ligands with well characterized pharmacological profiles.

With these constraints in mind, we chose amprenavir as a starting point for binder design. In addition to FDA approval, this ligand also has a site for biotinylation, enabling us to screen for binding via high throughput experimental methods, such as yeast surface display. Any hits could be verified for binding via ITC or SPR and eventually the fully characterized monomer could be engineered into a heterodimeric CID system.

[2] This chapter contains collaborative work. L. Tran designed the first library of amprenavir binders. I designed the second library and experimentally screened both libraries by yeast surface display.

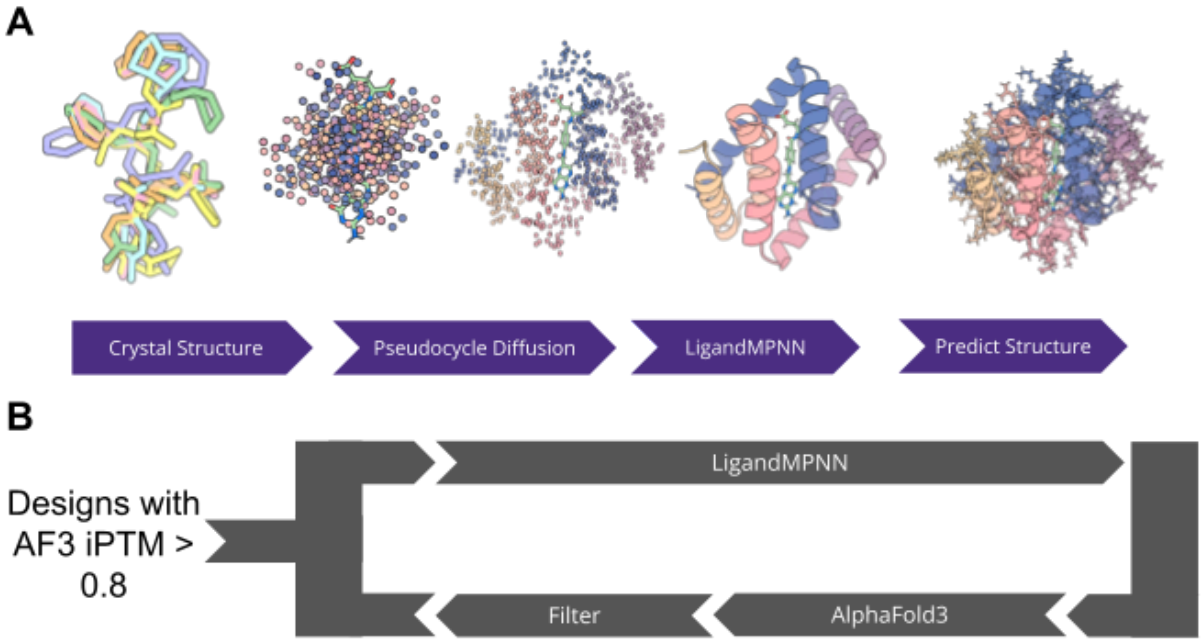


Figure 1. Computational workflow for the design of monomeric binders to HIV protease inhibitors using pseudocycle diffusion. A Initial round of designs were generated by using Pseudocycle diffusion to obtain Ca carbons around a number of conformers of the ligand. A refinement step fills in remaining backbone atoms and the sequence is designed with LigandMPNN. Designed sequences are then predicted with either Chai or AlphaFold3. **B** Designs with AF3 iPTM > 0.8 were subject to an AF3-LigandMPNN resampling method where additional sequences for the backbones were generated using the AF3 predicted structure.

Results

We reasoned that a pseudocycle diffusion CA-RFDiffusion blended method would be useful for this protein design task. This blended diffusion method has previously yielded binders to a number of small molecule dyes. The CA-RFDiffusion component of this method allows for the backbone to be directly generated around the ligand, circumventing a ligand docking that was necessary in previous small molecule design methods[3, 4]. The pseudocycle component of the method allows for the creation of a ligand binding cavity. The monomeric binders could later be split and engineered into two distinct heterodimeric chains downstream.

An initial library of designs were made for amprenavir by diffusing backbones around the ligand, designing sequence with LigandMPNN and generating a chai structure prediction of the protein ligand complex (Figure 1A). Ordered designs for the initial amprenavir library were cloned and transformed[3]. Next generation sequencing (NGS) of the naive library revealed 77% of the designs ordered were transformed. They were evaluated for binding to PEG4-biotin labeled amprenavir. Results were compared to a control sample consisting of library expressing

cells incubated with Alexafluor 594 + 0 μM labeled ligand. No binding signal was observed in the PE and FITC channels above the control (Figure 2).

We next tested whether AF3 structure prediction could aid in a redesign process using the designed proteins from the initial amprenavir library. We utilized a AF3-LigandMPNN resampling design method on designs from the original amprenavir library which had an AF3 iPTM > 0.8 (Figure 1b). The ordered library of designs was transformed and expressed on yeast as described previously. NGS of the naive library revealed 80% of the designs ordered were transformed. They were evaluated for binding to both PEG4-biotin labeled amprenavir. Again, we observed no binding signal in the PE and FITC channels above the control for yeast incubated with either PEG4-biotin-amprenavir or PEG4-biotin-darunavir (Figure 3A-B). A library designed with the same pipeline as in Figure 1 was also made for darunavir and was screened for binding to amprenavir since these ligands are structurally similar. No binding was observed when PEG4-biotin-amprenavir was incubated with the library (Figure 3C-D)

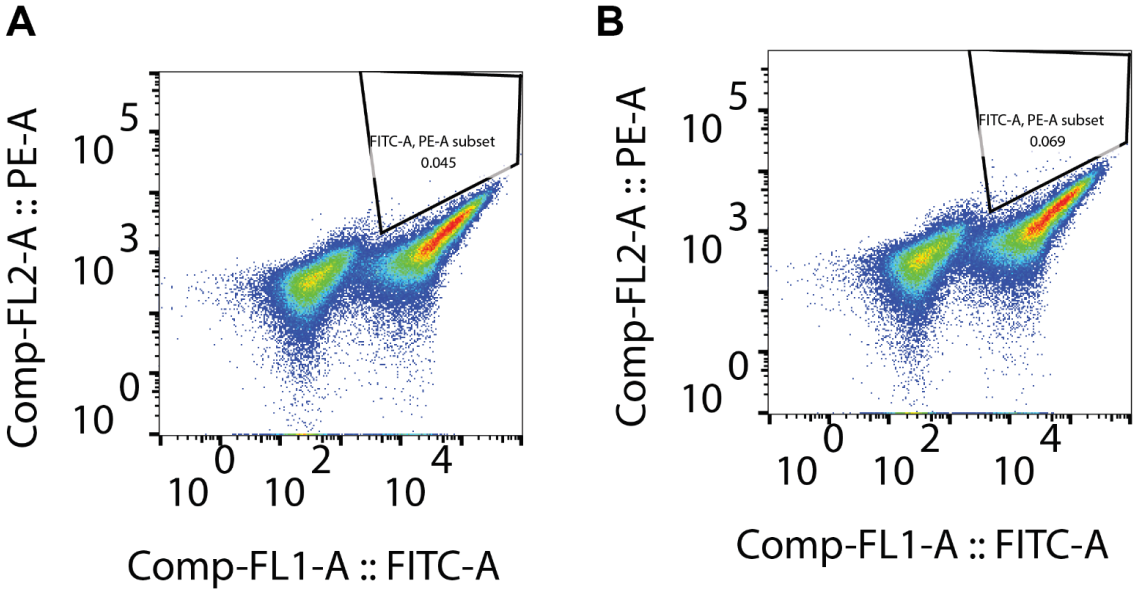


Figure 2. Binding sort results for the initial library of designs ordered for amprenavir. A Control sample consisting of 0.25 μM Alexafluor 594 incubated with the library designed for amprenavir. **B** 0.25 μM Alexafluor 594 and 1 μM amprenavir-PEG4-biotin incubated with the library. Plots and gates were rendered in and exported from FlowJo.

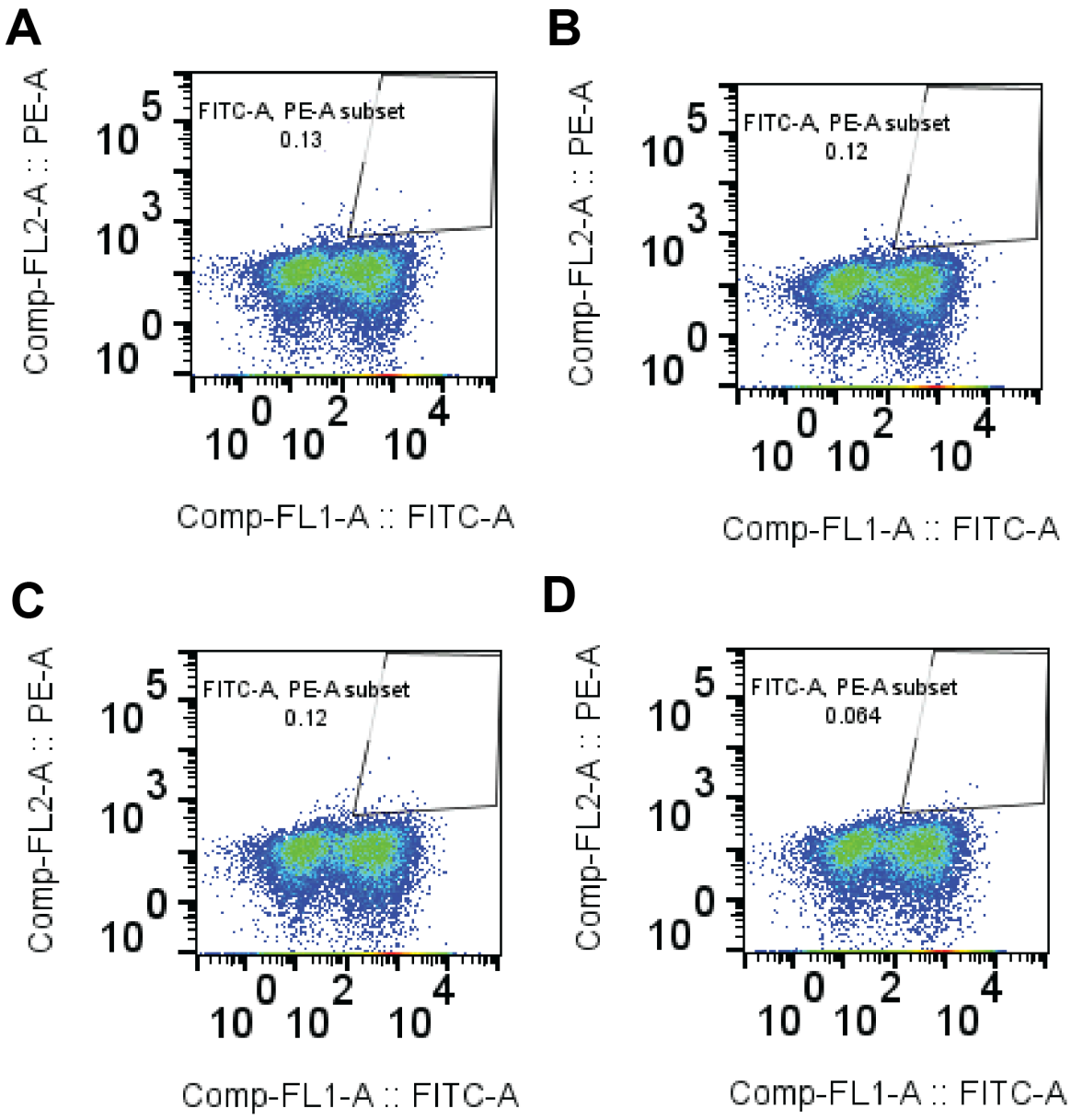


Figure 3. Libraries designed for amprenavir and darunvir with the AF3-ligandMPNN resampling method screened for binding to amprenavir. A 0.25 μM Alexafluor 594 incubated with the library designed for amprenavir. **B** 0.25 μM Alexafluor 594 and 1 μM amprenavir-biotin incubated with the library designed for amprenavir. Gates shown represent where binding signals would appear. **C** 0.25 μM Alexafluor 594 incubated with the library designed for darunavir. **D** 0.25 μM Alexafluor 594 and 1 μM PEG4-biotin-amprenavir incubated with the library designed for darunavir. Plots and gates were rendered in and exported from FlowJo.

Methods

Ligand Conformer generation

A known crystal structure of amprenavir (extracted from PDB 3NU6) as well as RDKit and CREST generated conformers to design with[5] .

Pseudocycle-CARFDiffusion

For the initial library ordered, Cα carbons were generated around amprenavir. The remaining backbone atoms were added during a refinement step and backbones were filtered for distance of center of mass between the ligand and the backbone (d_COM >X), number of neighboring Cα carbons that are within 10 Å of the ligand, and the number of backbones with Cα carbons that are less than 2 Å away from the ligand (ligand_clash)[3, 6]. Sequences were then designed for these backbones with LigandMPNN and included X cycles of FastRelax during the design process. Rosetta energy metrics were calculated during this step for downstream filtering. A final structure prediction was made with Chai (Figure 1A). The designs were filtered with the following metrics and logic statement, resulting in 6057 designs for experimental testing:

Chai iptm $\geq$ 0.65 & chai interaction pae $\leq$ 8 | $\Delta\Delta G \leq$ -39 & protein length $\leq$ 165 amino acids & chai iptm !=0.75

Structures for the first library of amprenavir designs were re-predicted with AlphaFold3 and the designs were filtered for iptm greater than 0.8[7]. Any designs passing this filter were then subjected to redesign with LigandMPNN and the structures repredicted with AlphaFold3. A second library of designs was ordered with designs showing AF3 iptm > 0.8, AF3 ptm >0.8, AF3 ranking score > 0.8, protein plddt > 90, ligand plddt >90, protein length < 165 amino acids, number of neighboring Cα carbons calculated from the AF3 prediction model > 75, & distance of the center of mass between the ligand and protein in the AF3 model < 7 Å.

Biotinylation of HIV protease ligands

Amprenavir (Millpore Sigma) was biotinylated at the aniline with an NHS ester PEG4-biotin (Thermo Fisher). Amprenavir and biotin solutions were mixed in 1 : 3 molar ratio in DMSO. Reaction pH was adjusted to ~9 with NaOH and the reaction was placed on a stir plate for approximately 2 h at room temperature. Reaction was purified on AdvanceBio RP-Desalting column, collected fractions were assessed by mass spectrometry, and the product expected mass matches the observed mass (Figure 4). Mass spectrometry analysis was run via reverse-phase LC/MS on an Agilent G6230B TOF with an AdvanceBio RP-Desalting column.

Darunavir with PEG4-bition at the aniline was synthesized by and ordered from Wuxi.

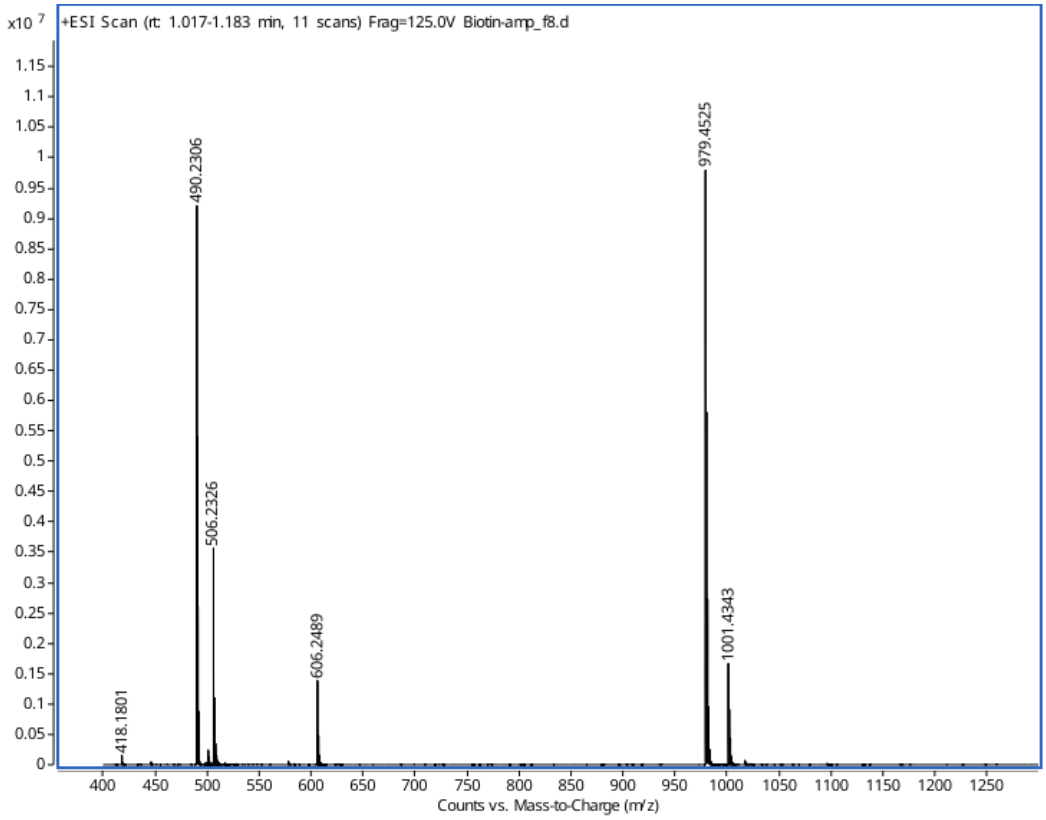


Figure 4. Mass spectrum of biotinylated amprenavir. The observed mass of the PEG4-amprenavir product was 979.5 Da. The calculated mass was 979.2 Da.

Yeast display for screening designed protein binders

Cells were first stained for expression of the designed proteins by measuring OD600 and calculating cell density of the yeast cultures. 50 million cells were then washed 2 times with 1 mL PBS with 0.1% BSA and washed pellet was resuspended in 490 uL PBS with 0.1% BSA. 10 uL of anti-c-myc-FITC (Immunology Consultants Laboratory, CMYC-45F) was added and the cells incubated with antibody for 30 min on rotator at 4℃. The cells were then washed 2 times with 1 mL PBS with 0.1% BSA and the washed pellet was resuspended in 250 uL PBS with 0.1% BSA. Sorting was performed by Fluorescence-activated cell sorting (FACS) on a Sony SH800S Cell Sorter fitted with a 100 µm chip and 488 nm/561 nm/638 nm lasers. Cells expressing the designed proteins were collected and regrown in CTUG overnight. OD600 was calculated for the expression sort culture and used to calculate cell density of the culture. 50 million cells were stained for those expressing library proteins as described above. 2.5 million of the 50 million cells were then diverted to a separate tube for a control. The control consisted of 0 μM of ligand, 1.5 μL of Alexa Fluor 594 (Abcam, ab223894) in a final volume of 50 μL PBS with 0.1% BSA. Experimental samples consisted of 22.5 million cells, 1 μM of 500 μM ligand stock solution in DMSO, 13.5 μM Alexa Fluor 594 at a final volume of 450 μL PBS in 0.1% BSA. The final concentration of Alexa Fluor 594 in both the control and experimental samples

was 8.3 μM. Samples were incubated for 30 min on rotator at 4℃. The cells were then washed 2 times with 500 mL PBS with 0.1% BSA and the washed pellet was resuspended at a concentration of 15 million cells/mL in PBS in 0.1% BSA for sorting on FACS. Yeast-sorting data were processed and visualized using FlowJo v10 (FlowJo, Inc.).

References

1. Adams, E.L., McGovern, A.C., So, V. *et al.* Small-molecule control of CAR T cells. *Nat Rev Chem* 9, 809–825 (2025).
2. Foight GW, et al. Multi-input chemical control of protein dimerization for programming graded cellular responses. Nat Biotechnol. 37, 1209-1216 (2019).
3. Tran L., et al. Design of Orthogonal Far-Red, Orange and Green Fluorophore-binding Proteins for Multiplex Imaging. Biorxiv (2025).
4. An L., Said M., Tran L., Majumder S., Goreshnik I., Lee G. R., Juergens D., Dauparas J., Anishchenko I., Coventry B., Bera A. K., Kang A., Levine P. M., Alvarez V., Pillai A., Norn C., Feldman D., Zorine D., Hicks D. R., Li X., Sanchez M. G., Vafeados D. K., Salveson P. J., Vorobieva A. A., Baker D., Binding and sensing diverse small molecules using shape-complementary pseudocycles. Science 385, 276–282 (2024).
5. P. Pracht, F. Bohle, S. Grimme, Automated exploration of the low-energy chemical space with fast quantum chemical methods. Phys. Chem. Chem. Phys. PCCP 22, 7169–7192(2020)
6. Anna Lauko *et al.* Computational design of serine hydrolases. *Science* **388**, eadu2454 (2025).
7. Abramson, J., Adler, J., Dunger, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500 (2024).

Chapter 3: Conclusions

The computational design methods in Chapter 1 show the use of rosetta, proteinMPNN, and Alphafold2 for creation of a CID system. In HEK293T protein-protein interaction screening assays, we demonstrated the conditional response of the designed proteins to the cyclic peptide. While this method yielded a homodimeric system, there are still some limitations to the current system. Engineering the protein protein interface to tune the amount of background dimerization from the proteins or even engineering the system into a heterodimeric CID could increase the utility, preventing any signal from being lost due to the background dimerization.

The design methods outlined in Chapter 2 explore the utility of novel machine learning tools to generate CID systems. We utilized a pseudocycle diffusion model to generate backbones followed by LigandMPNN for sequence design. After screening two libraries via yeast display for binding to the HIV protease inhibitor amprenavir, we did not obtain any binders. Since the pseudocycle diffusion method used for backbone generation is meant to fully encapsulate the ligand, future efforts involve using newer diffusion models to generate backbones with more diverse binding pockets. Incorporating more sequence and structure predicting cycling along

with partial diffusion could help optimize the design pipeline and yield binders in future design rounds,

Over the course of my PhD, protein design tools have expanded from physics based design methods to the incorporation of deep learning based methods for backbone generation, sequence design, and structure prediction. Looking forward, structure prediction models that incorporate more experimental data could potentially increase the ability to discriminate between binder and non-binder designs *in silico*. With more advanced protein design tools, protein designers could, for example, expand into the design of cellular logic gates to include systems that turn off, rather than turn on, a cellular response in the presence of a ligand.