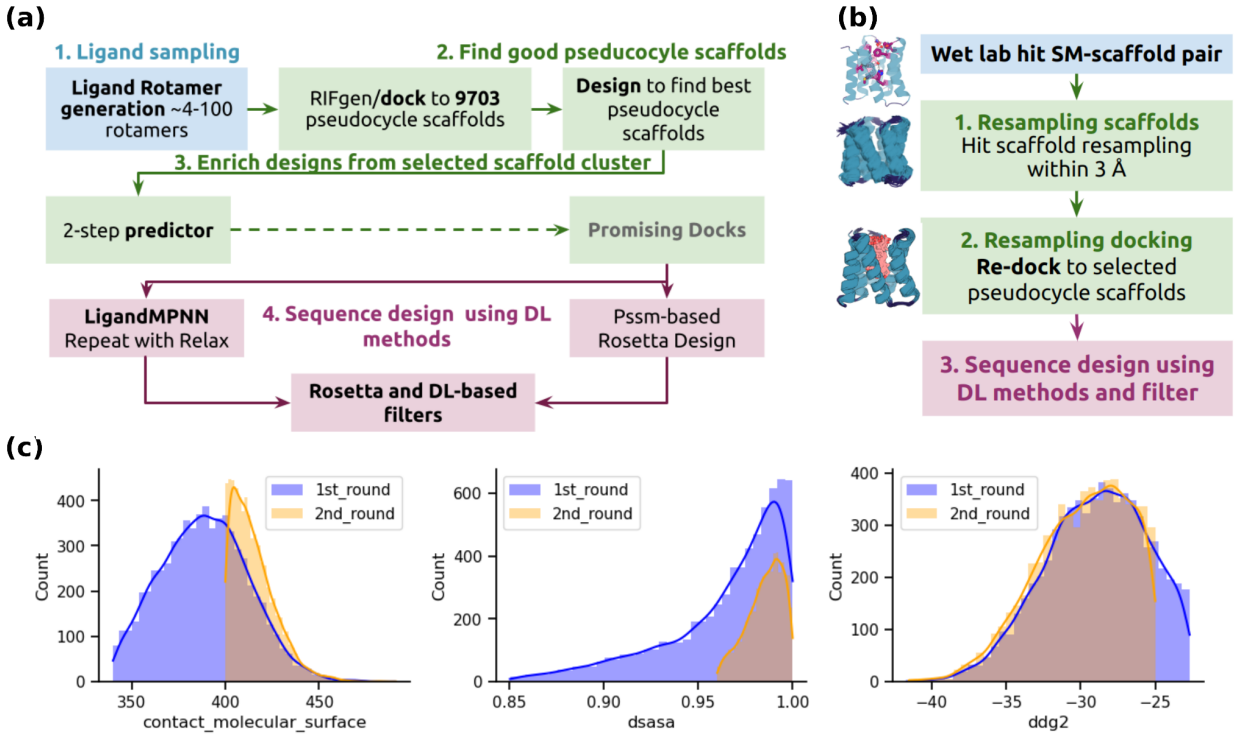

Figure S1 Computational design pipeline.

(a) The flowchart of the first computational step to identify scaffolds suitable with the each SM ligands; **(b)** The flowchart of the second computational step to focus on the verified suitable scaffold-ligand pair and sampling the docking and the interface packing of the designed binder proteins; **(c)** The histogram plots comparison of MTX binders for order from the first step sampling (1st_round) v.s second step sampling (2nd_round). The contacts between the ligand and the protein got significant improvement through stepwise sampling.

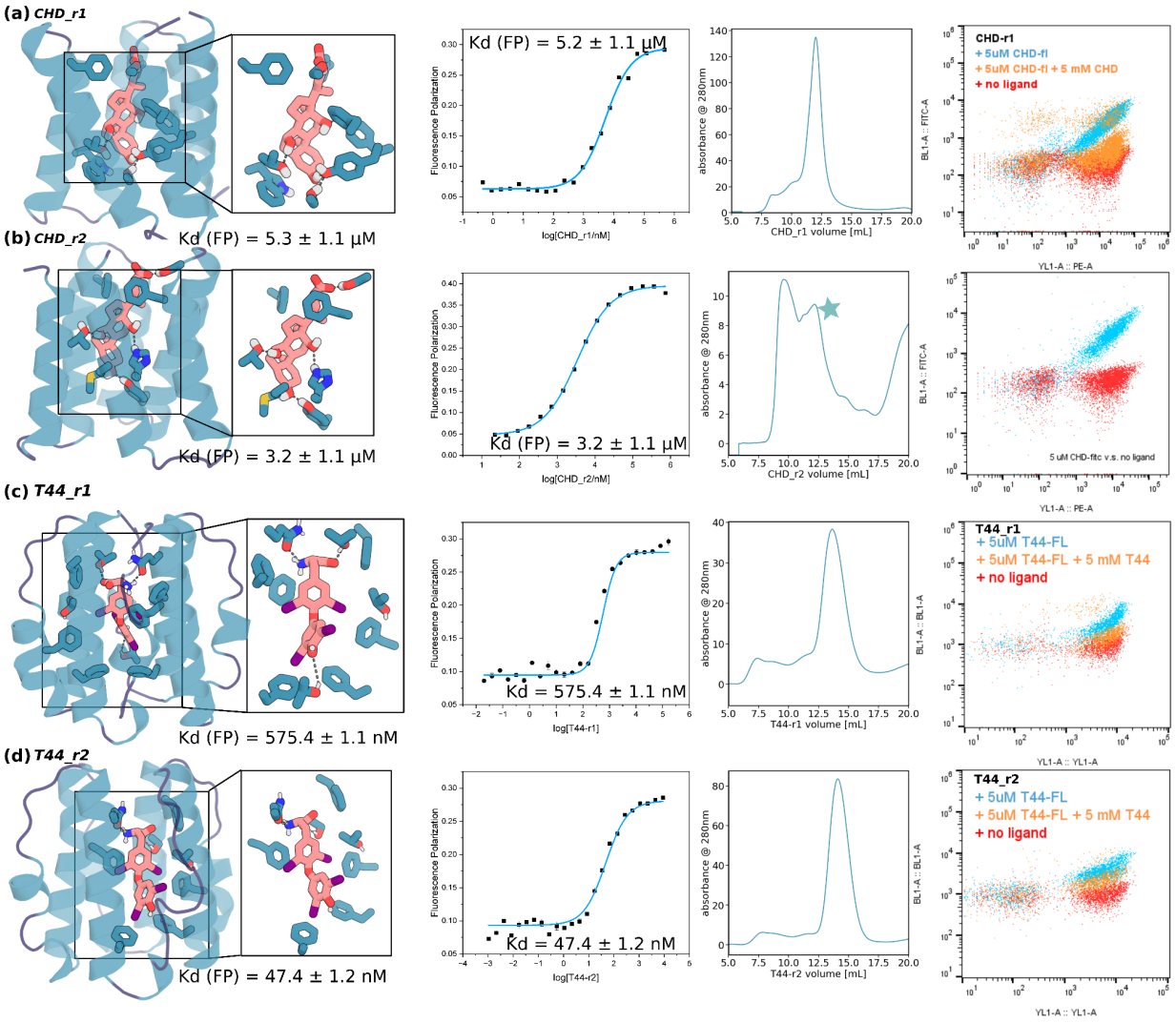

Figure S2 Characterization of the CHD and T44 binders from the first round of sampling.

The initial hits from the first step screening are shown in each panel, the design model, interface, binding and competition assays from FP studies of the Two binders for CHD **(a-b)** and T44 **(c-d)**. The monomeric fraction of CHD_r2 was collected and marked with a star on the third panel in **(b)**. The designs were shown in cartoons, the ligand and key interacting residues were shown in sticks. Oxygen, nitrogen, and sulfur were colored in red, blue, and yellow, respectively.

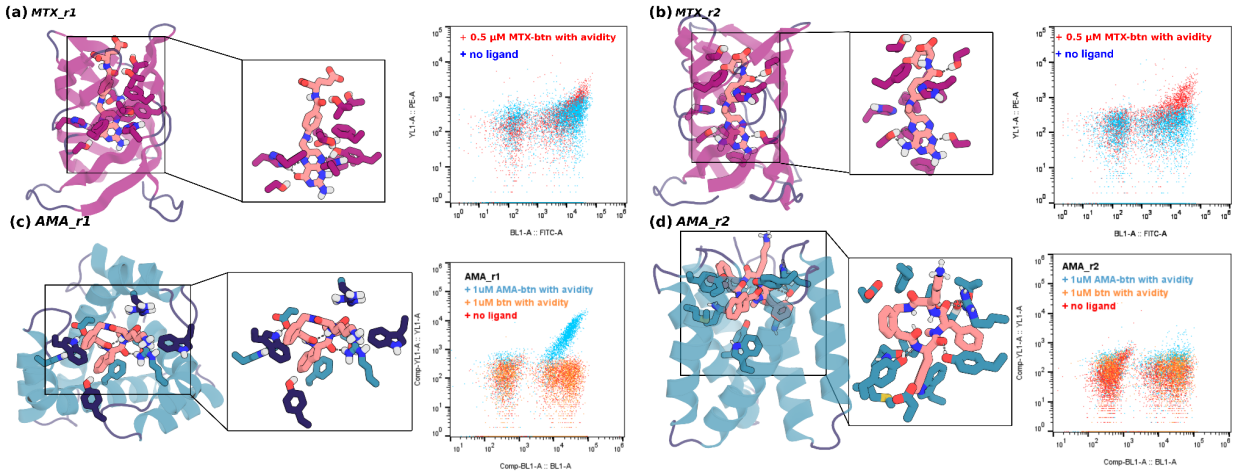

Figure S3 MTX and AMA binders from the first round of sampling.

The initial hits from the first step screening are shown in each panel, the design model, interface, binding and competition assays from yeast surface display studies of the MTX **(a-b)** and AMA **(c-d)**. The binder expressed yeast showed a positive FITC signal (BL1-A), and the binding yeast showed a positive PE signal (BL1-A). The binder-expressing yeast showed clear strong double positive signals (red for MTX, blue for AMA), and only FITC signals when no biotin-labeled ligand presence (blue for MTX, orange for AMA). The designs were shown in cartoons, the ligand and key interacting residues were shown in sticks. Oxygen, nitrogen, and sulfur were colored in red, blue, and yellow, respectively.

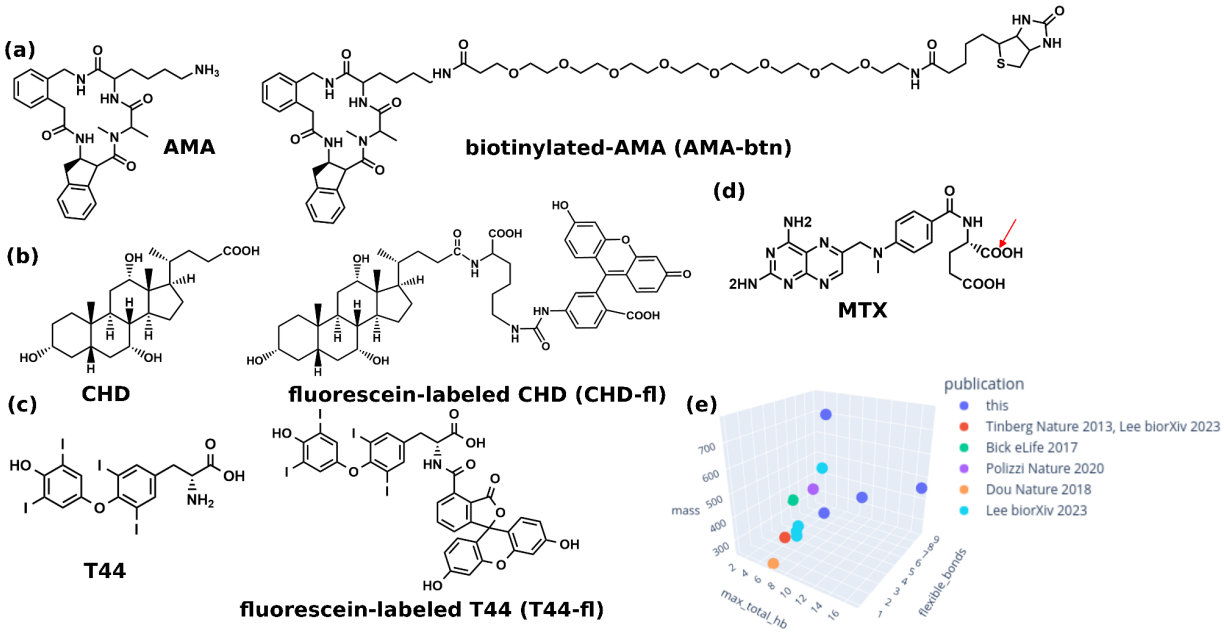

Figure S4. Target ligands.

(a-d) The structure and the tagged version of AMA (synthesized by WuXi Chemistry Service), CHD (both commercially available), T44 (tagged version synthesized in house, no-tagged version commercially available), and MTX (commercially available for tagged and untagged). While both the tagged structure of fluorescein-labeled MTX (MTX-fl), and biotinylated MTX (btn-MTX) is commercially available, the structure was not reported, but the tagged site was reported, marked in red. **(e)** The features of ligands designed against in this work and previous works. Only ligands with affinity (Kd) and validation (SSM, binding competition, or crystallography) data are included. More details of ligand features, see **Table S4**.

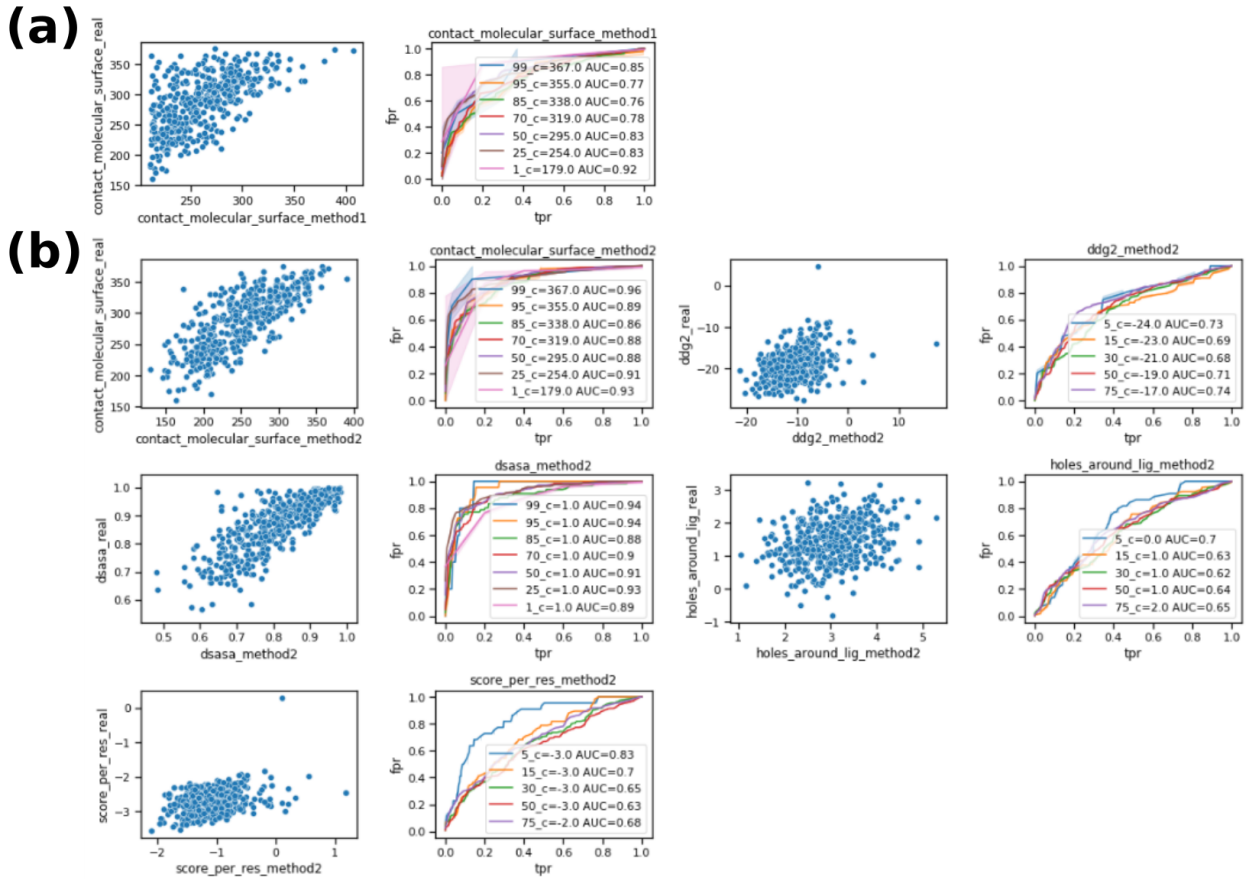

Figure S5. Two-step predictor scores are correlated with scores after full design

One example of evaluation on how predictive the two-step predictor is. 1000 docks to CHD were randomly selected to evaluable how the predictor works. For each set, the left panel is the scores generated from score from predictor (x-axis) v.s. final PSSM-based Rosetta design protocol (y-axis). The right panel is the receiver operating characteristic curve (ROC) curve at different quantiles to measure the predictiveness of the predictor script on the reported metric. The higher the false positive rate (fpr) with the increase of true positive rate (tpr), the more predictive the scripts are on the measured metric. The predictiveness at each quantile was also reported using area under curve (AUC), and the higher the AUC, the better the predictions are. **(a)** The CMS of input docks were clearly predicted well with predictor1; **(b)** The 'CMS', binding energy ('ddg2'), burial of the ligands ('dsasa' and 'holes_around_lig') and the protein quality ('score_per_res') were highly predicted using the second predictor.

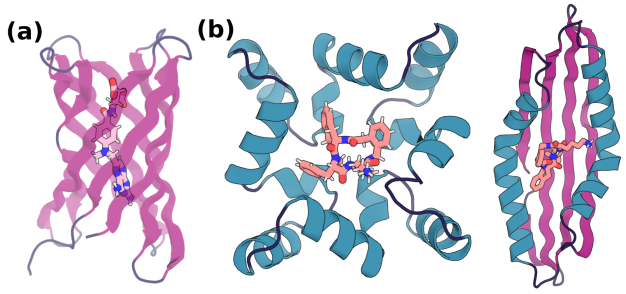

Figure S6. The other potential binding scaffold hits (not shown in Fig 1) from the first step sampling.

During first step HTP binder screening, for MTX and AMA, we also identified other scaffolds (apart from those shown in **Fig 1b**), which NGS enrichment suggested to be potential binders. These scaffolds were also identified as scaffolds in hit ligand-scaffold pairs and were used for second step computational sampling. Individual binders from these pseudocycle scaffolds were not verified individually for prioritization of other better performing binders. **(a)** One other potential binding pseudocycle scaffolds for MTX; **(b)** Two other potential binding pseudocycle scaffolds for AMA.

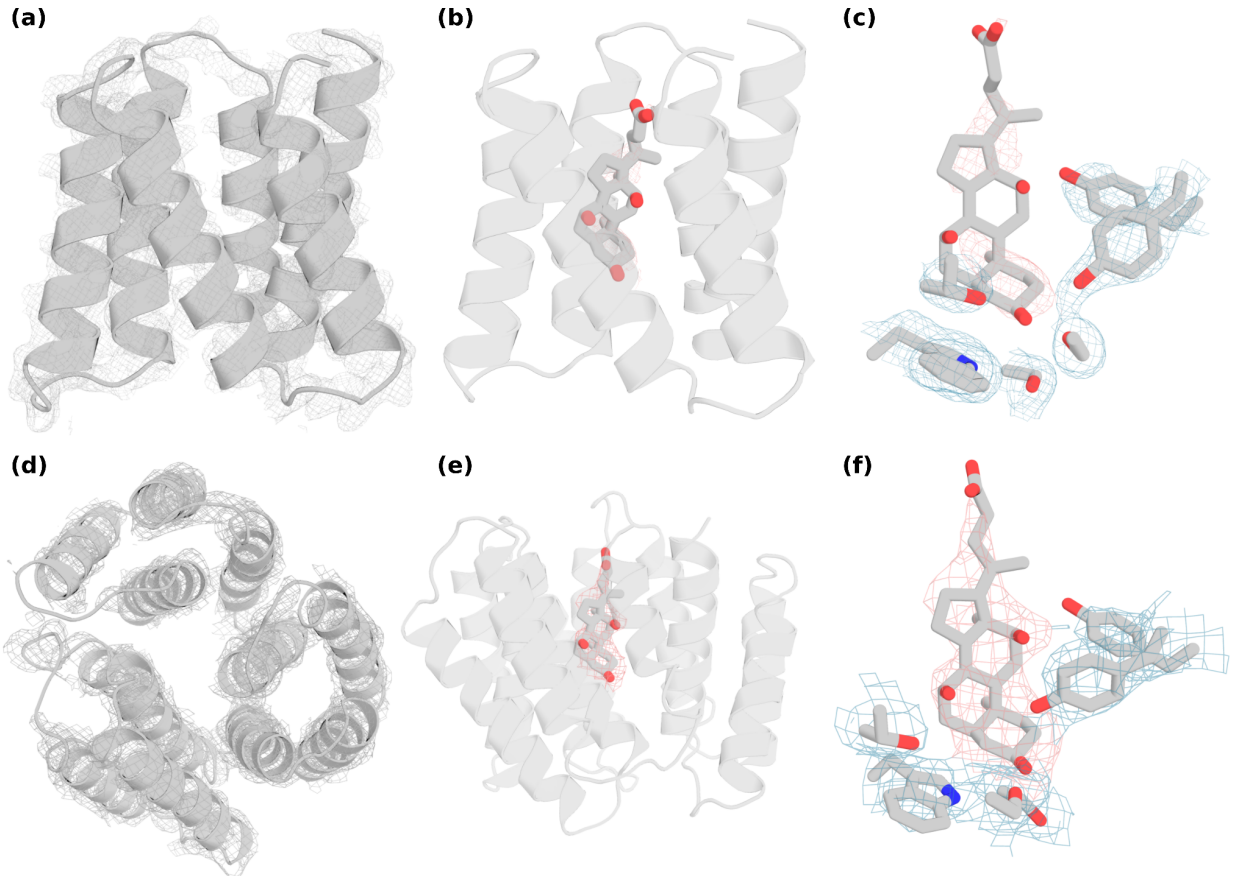

Figure S7. The comparison of crystal structures of CHD_d1 complex and CHD_buttress complex.

The crystal structure of CHD_d1 **(a-c)** and CHD_buttress **(d-f)**. Only the protein backbone densities are shown in **a** and **d** for clarity. The backbone electron density, the ligand density, and the key interacting rotamer electron densities are shown from left to right. The electron density is shown in mesh and the refined co-crystal structure is shown in gray. The ligand and the key interaction rotamers are shown in sticks, where oxygen, nitrogen are colored in red and blue, respectively.

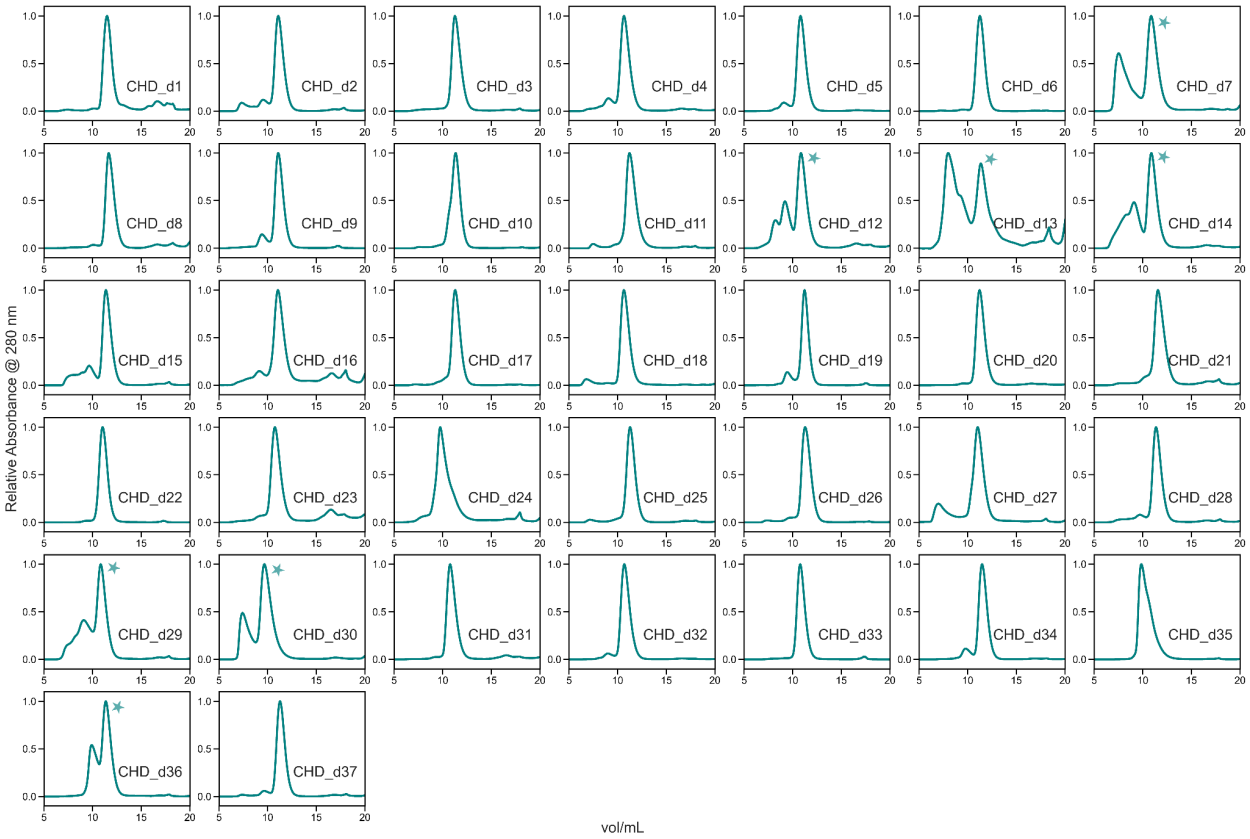

Figure S8. SEC profiles of 2nd round CHD binders.

SEC traces of all CHD binders chosen for off-yeast binding test are shown here. For traces where multiple oligomeric states are present, the monomeric peaks collected for FP studies are marked with stars.

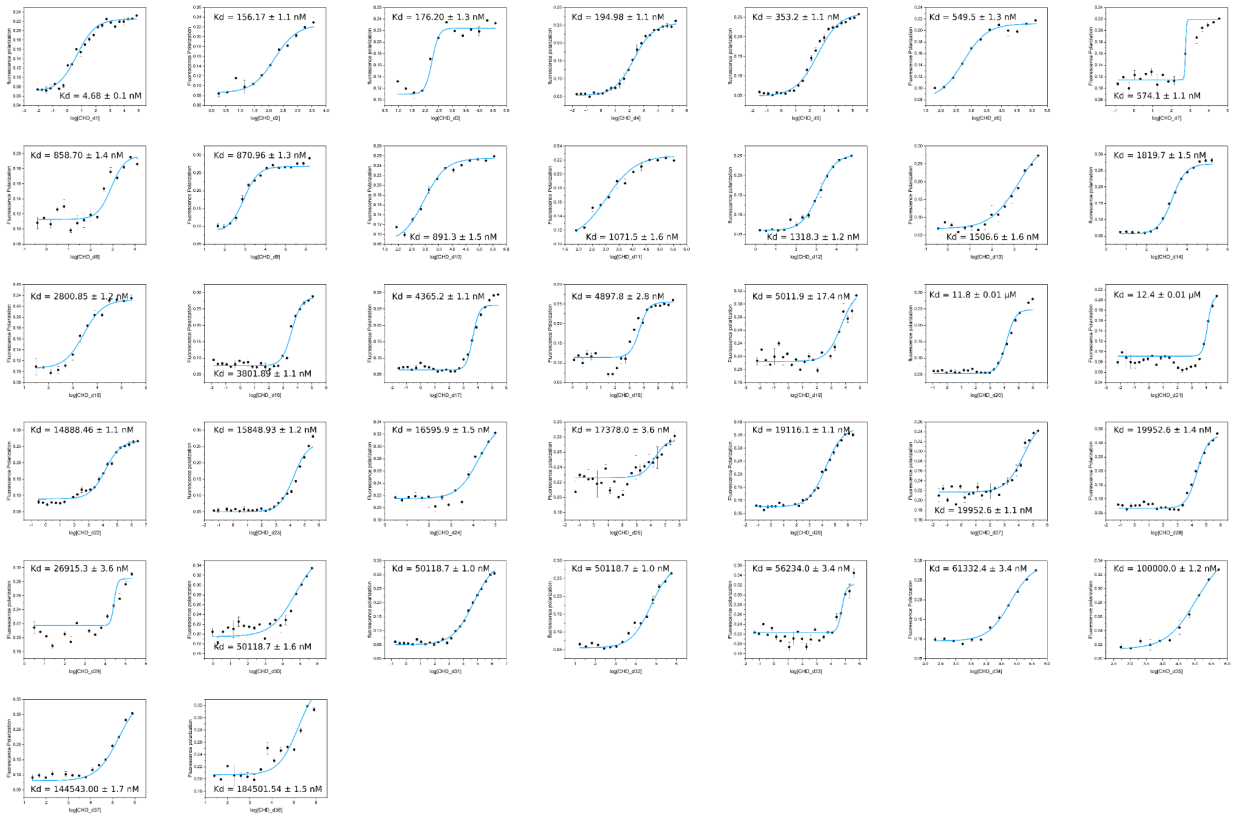

Figure S9. Fluorescence polarization titrations for all 2nd round CHD binders

FP profiles for all 37 purified 2nd round CHD binders; also see **Fig 3**.

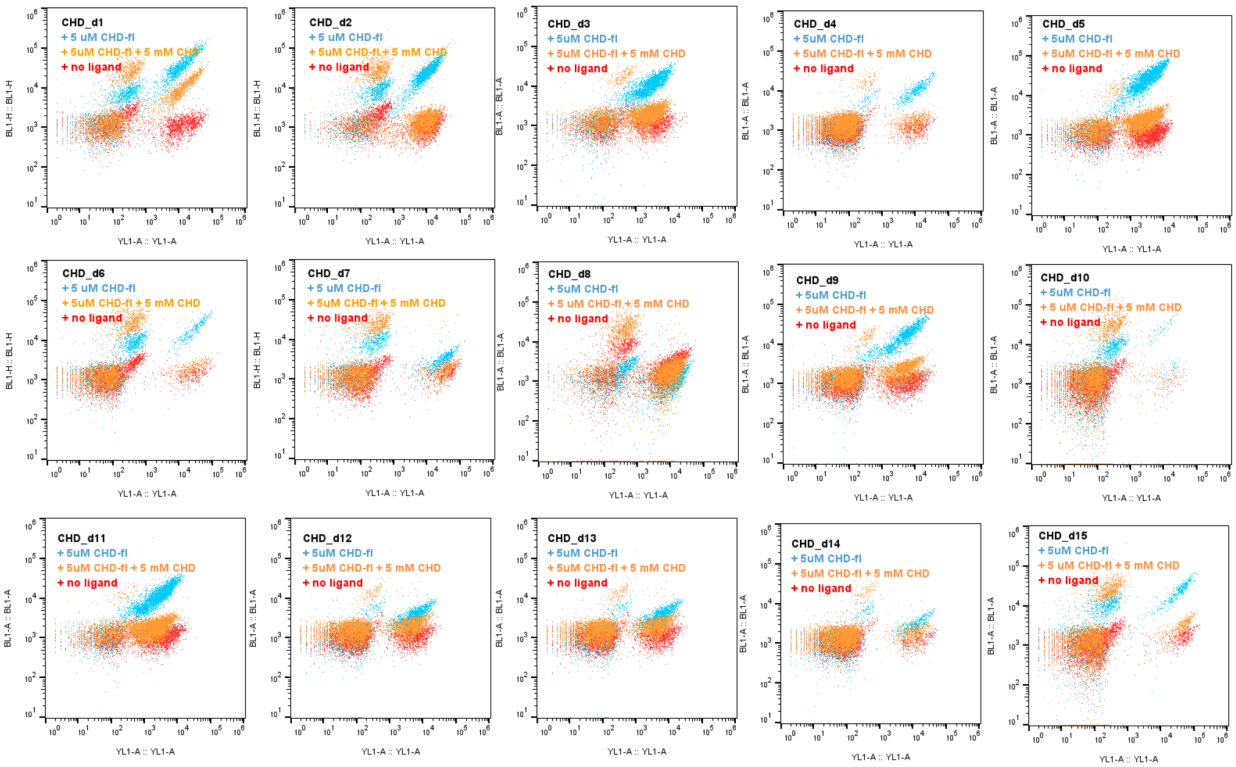

Figure S10. Competition of 2nd round CHD binders on yeast surface by free CHD

The designs were displayed on the surface of yeast, using 5 μM CHD-fl for binding and 5 mM free CHD for competition. The binder expressed yeast showed a positive PE signal (YL1-A), and the binding yeast showed a positive FITC signal (BL1-A). All binder expressed yeast showed clear strong double positive signals (blue), while weaker FITC signal upon free ligand competition (orange), and only PE signals when no FITC-labeThe binder expressed yeast showed a positive PE signal (YL1-A), and the binding yeast showed a positive FITC signal (BL1-A). All binder expressed yeast showed clear strong double positive signal (blue), while weaker FITC signal upon free ligand competition (orange), and only PE signal when no FITC-labeled ligand presence (red).

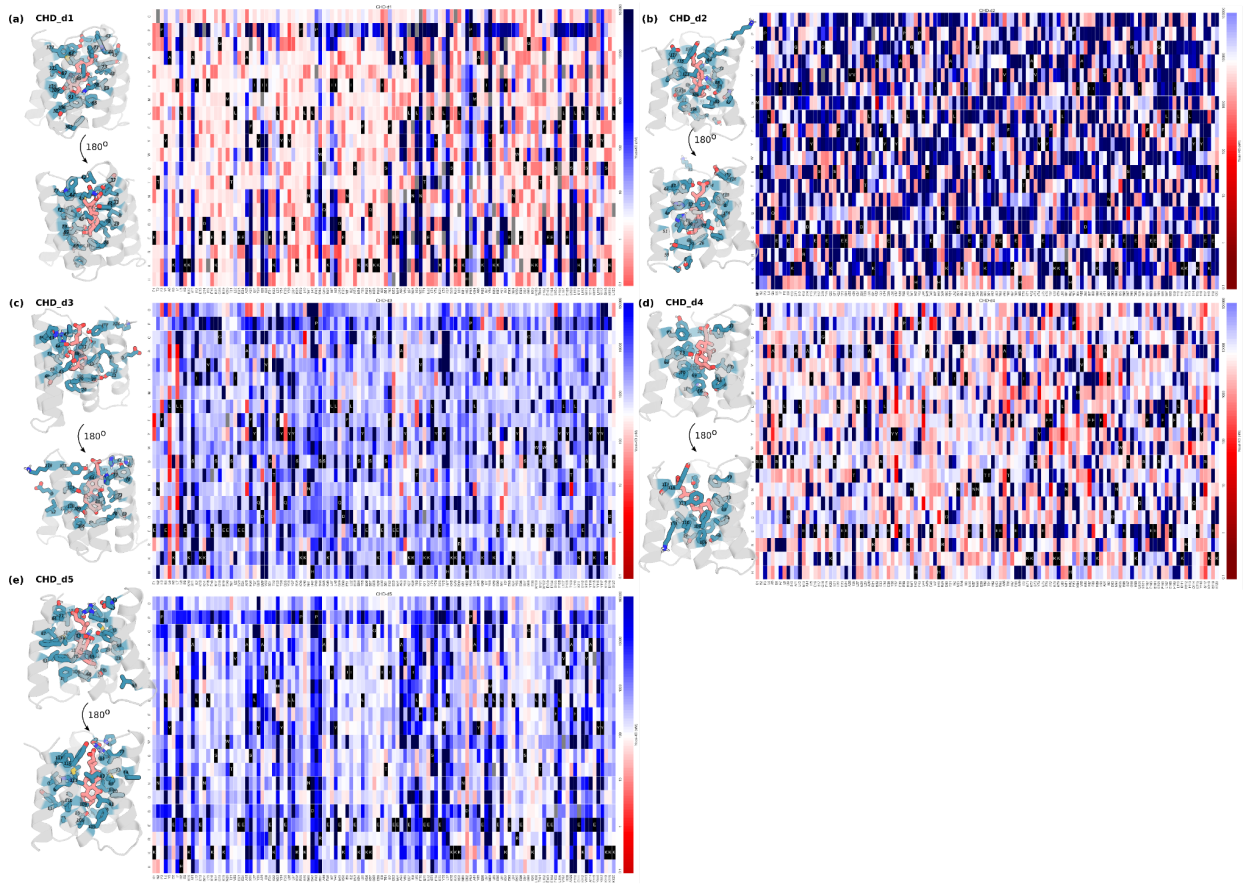

Figure S11. site saturation mutagenesis of CHD-d1 to CHD-d5.

SSM analyses were performed for CHD-d1 to CHD-d5 (**a-e**). The interface residues were analyzed, and those conserved were colored in teal. All SSMs showed high conservation are the residues having contact with the ligand, supporting the designed conformations. The native residues are listed on the x axial, and colored in black. The estimated binding affinity of each mutant is colored based on y axial; the residue choices which may improve binding are more red, while the residues jeopardizing the binding are more blue. The conserved interface residues are colored in teal.

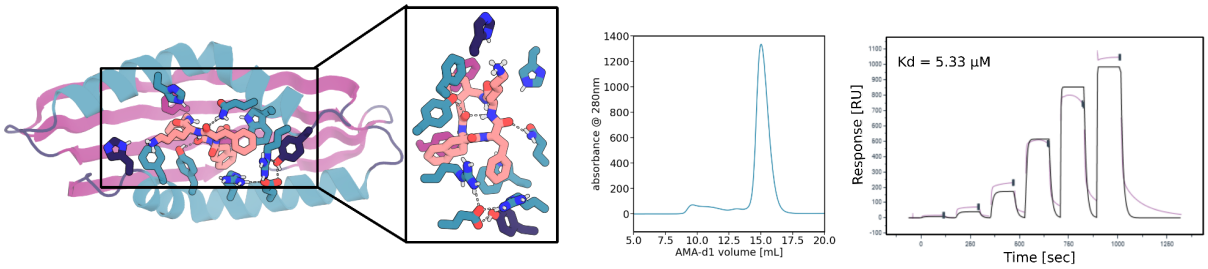

Figure S12. SPR analysis of additional 2nd round AMA binders not shown in Fig 3.

From left to right panel, the design model, SEC traces, and the SPR binding traces of AMA-d1 are shown in each panel. The design models are shown in cartoons, the ligand and the key interacting residues are shown in sticks. Oxygen, nitrogen are colored in red and blue, respectively. The cartoon of and the residues from helix, sheet, and loops are colored in teal, magenta, and dark blue, respectively.

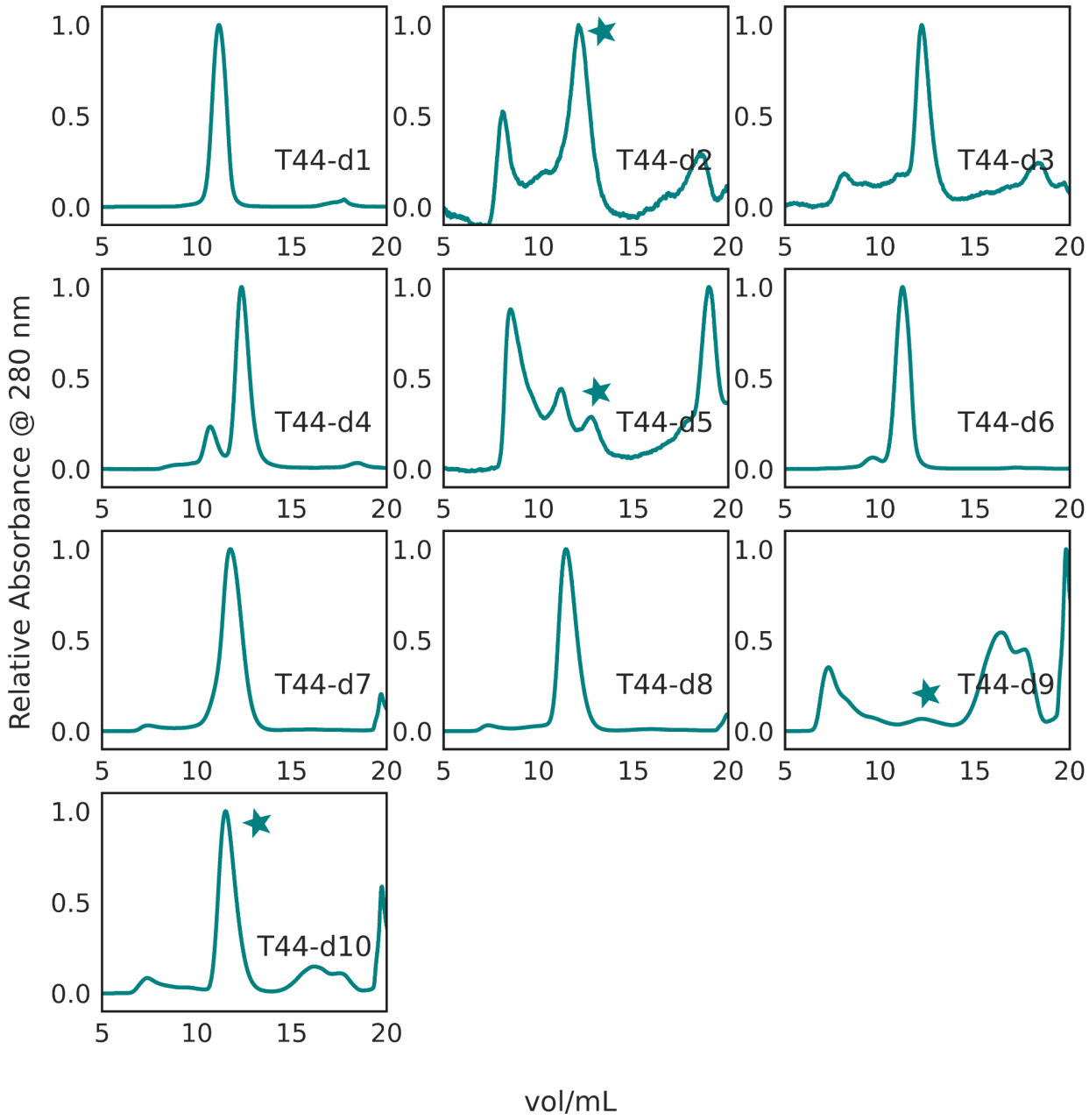

Figure S13. SEC analysis of second round T44 binders

SEC traces of T44 binders chosen for off-yeast binding test from the second round of sampling are shown here. For traces where multiple oligomeric states are present, the monomeric peaks collected for FP studies are marked with stars.

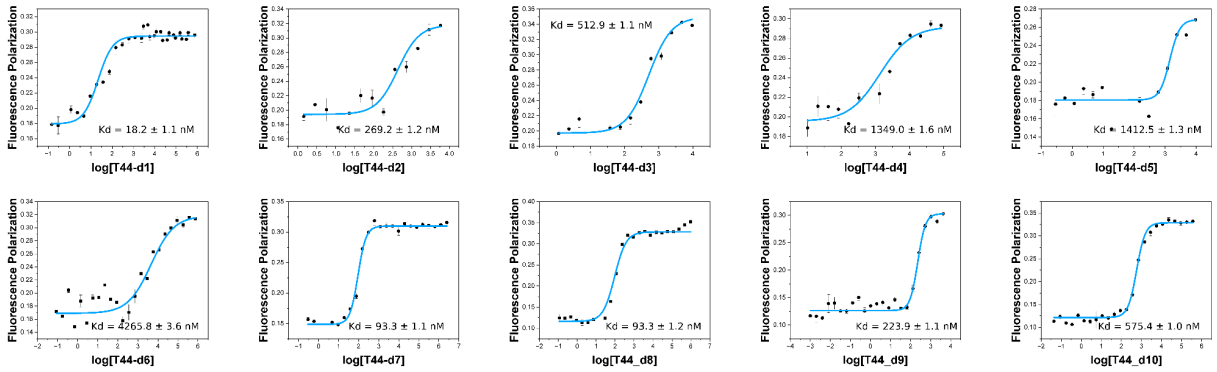

Figure S14. FP analysis of 2nd round T44 binders.

FP profiles for all 10 purified 2nd round T44 binders; also see **Fig 3**.

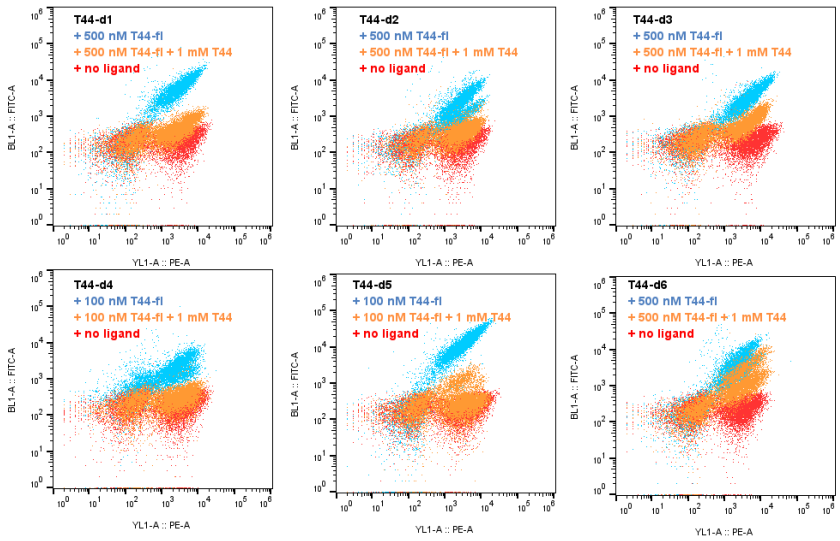

Figure S15. Competition of T44 binding on the yeast surface by free ligand

The designs were displayed on the surface of yeast, using T44-fl for binding and free T44 for competition; specific concenadditionaltrations used are marked on the figure. The binder expressed yeast showed a positive PE signal (YL1-A), and the binding yeast showed a positive FITC signal (BL1-A). All binder expressed yeast showed clear strong double positive signals (blue), while weaker FITC signal upon free ligand competition (orange), and only PE signals when no FITC-labeled ligand presence (red).

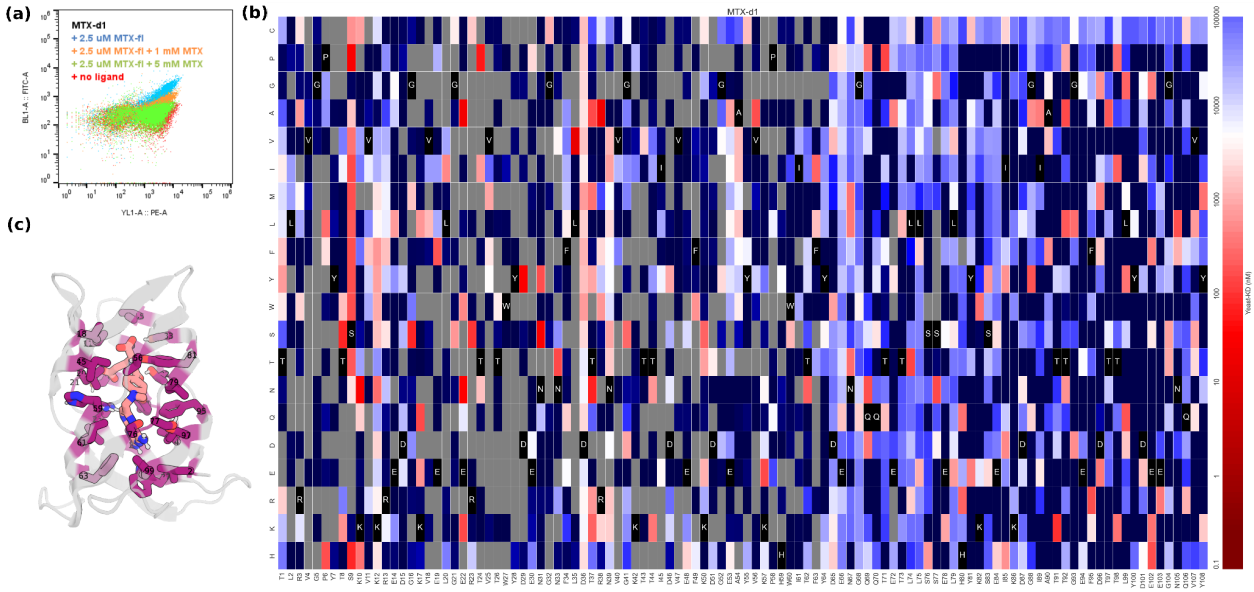

Figure S16. Binding verification of MTX_d1 by site saturation mutagenesis

MTX-d1 showed clear binding competed by free MTX **(a)** on yeast. SSM analysis of MTX supported the designed binding mode **(b-c)**. In the SSM matrix figure **(b)**, the native residues are listed on the x axial, and colored in black. The estimated binding affinity of each mutant is colored based on y axial; the residue choices which may improve binding are more red, while the residues that jeopardize the binding are more blue. For the sorting competition assay in **(a)**, the binder expressed yeast showed a positive PE signal (YL1-A), and the binding yeast showed a positive FITC signal (BL1-A). The binder expressed yeast clearly showed a strong double positive signal (blue), while weaker FITC signal upon free ligand competition (orange and green), and a PE-only signal when no FITC-labeled ligand presence (red). Based on the SSM matrix (**b**), conservative pocket residues are colored in magenta in (**c**).

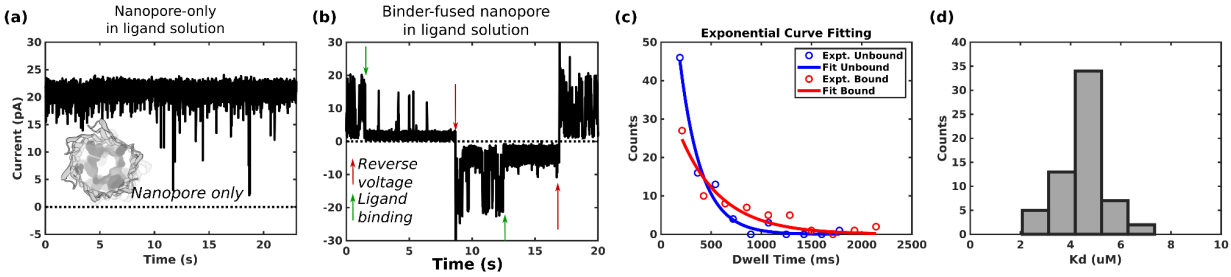

Figure S17. Characterization of binder fused nanopore.

The original nanopore did not show obvious blocking in presence of the ligand **(a)**. The binder fused nanopore can transition to off states and can be reversed by reversing the voltage **(b)**. Based on the ligand blocking dwell times (**c**, cutoff of > 100 ms shown here), the Kd of the binder-fused nanopore was estimated, which is in good agreement with reported Kd (CHD-r1, 5.3 ± 0.1 μM) **(d)**. See '***Estimation of dissociation constant (Kd) from nanopore conductance measurements***' in Methods).

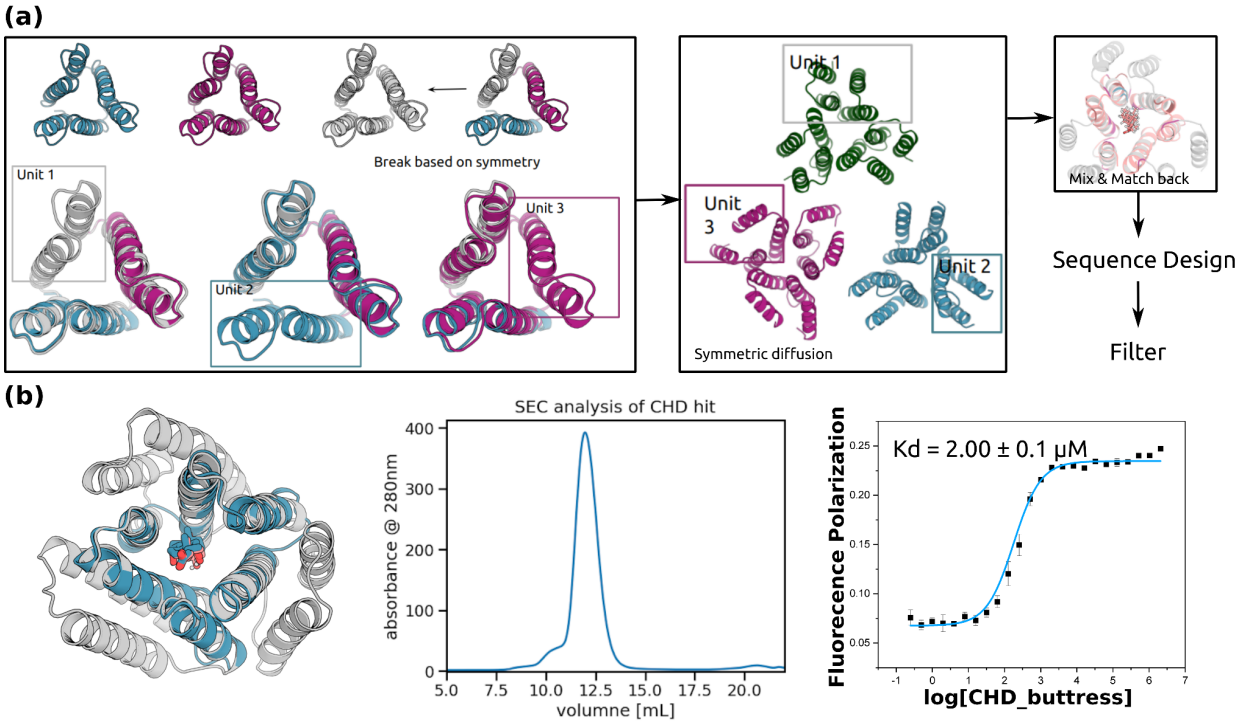

Figure S18. The design procedure for buttrased binder.

(a) Design strategy. Single ring designs are broken into repeat units based on the structural symmetry, and two helix buttresses were generated around each repeat protein using symmetric diffusion. The resulting four helical repeat units were pieced back together to generate the final buttrased scaffolds. ProteinMPNN was used to generate sequences. **(b)** The buttressed binder purified as a monodisperse peak and had tighter binding affinity than the original binder, (CHD-r1, 5.3 ± 0.1 μM). Also see **Fig S7** for the co-crystal structure of CHD_buttress with ligand.

Table S1. Computational statistics.

The numbers of the rotamer used, docks for ligandMPNN or Rosetta design, and final ordered designs from both rounds, and the potential binder, the selected verified binders were reported. 'Binders selected and verified using purified protein' meaning, based on predicted affinity and the resource, we choose a smaller number of binders to prioritize from all the NGS-verified binders to purify and verify their binding affinity.

items	CHD	MTX	T44	AMA
Rotamer number	112	28	21	1
Docks for long design protocols (1st step)	141,352	105,696	103,100	~100,000
Ordered design (1st step)	7,443	10,174	10,990	6,852
Docks for long design protocols (2nd step)	~100,000	~100,000	489,306	62,211
Ordered design (2nd step)	14,882 (5,574 from PSSM-based Rosetta design methods, 9,308 from ligandMPNN design method)	8,506	5,936	8,058
Binders based on NGS (2nd step)	231	1	84	2
Binders selected and verified using purified protein (2nd step)	37	1	10	2

Table S2. Summary of all reported binder in this paper.

The generation, characterization, sequence and structure similarity of each binder were reported here. Sequence similarities were calculated using BLASTP, and structural similarities were generated using Dali server(*35*).

number	Binder	Round	Biochemical characterization	Design method	Structural similarity to PDB50	Sequence similarity to Uniprot50	Sequence similarity to clostest parent
1	CHD_r1	1	FP, crystallography, Yeast surface display, SEC	PSSM-based Rosetta	10% to 2L6X	40%idt with 48% coverage to UniRef50 RepID=>UniRef50_A0A843FYG7	n.a
2	CHD_r2	1	FP, SEC	Repetitive ligandMPNN	10% to 2L6X	39%idt with 61% coverage to UniRef50 RepID=>UniRef50_N9S0L1	n.a
3	T44_r1	1	FP, SEC	PSSM-based Rosetta	10% to 6ait-A	22%idt with 55% coverage to UniRef50 RepID=>UniRef50_A0A8H8XJC6	n.a
4	T44_r2	1	FP, SEC	PSSM-based Rosetta	10% to 6ait-A	33%idt with 44% coverage to UniRef50 RepID=>UniRef50_A0A1B1E0J0	n.a

5	AMA_r1	1	Yeast surface display	PSSM-based Rosetta	14% to 5igh-A	37%idt with 49% coverage to UniRef50 RepID=>UniRef50_A0A1G5S3S9	n.a
6	AMA_r2	1	Yeast surface display	Repetitive ligandMPNN	12% to 6veo-A	35%idt with 51% coverage to UniRef50 RepID=>UniRef50_A0A517Y9K7	n.a
7	MTX_r1	1	Yeast surface display	Repetitive ligandMPNN	4% 70kn-A	36%idt with 59% coverage to UniRef50 RepID=>UniRef50_A0A1Y3NG75	n.a
8	MTX_r2	1	Yeast surface display	Repetitive ligandMPNN	9% to 70kn-A	32%idt with 52% coverage to UniRef50 RepID=>UniRef50_A0A524MJS0	n.a
9	CHD_d1	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	26%idt with 40% coverage to UniRef50 RepID=>UniRef50_A0A7X1B1Q9	30%idt with 54% coverage to CHD_r1
10	CHD_d2	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	29%idt with 42% coverage to UniRef50	28%idt with 55% coverage to CHD_r1

						RepID=>UniRef50_A0A7V5LIS7	
11	CHD_d3	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	33%idt with 41% coverage to UniRef50 RepID=>UniRef50_A0A5N6LWE3	35%idt with 65% coverage to CHD_r2
12	CHD_d4	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	32%idt with 45% coverage to UniRef50 RepID=A0A0B1RLI2_9NOCA	42%idt with 58% coverage to CHD_r1
13	CHD_d5	2	FP, SEC, yeast surface display	PSSM-based Rosetta	n.a	29%idt with 45% coverage to UniRef50 RepID=>UniRef50_A0A7W1U1H0	44%idt with 67% coverage to CHD_r1
14	CHD_d6	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	30%idt with 37% coverage to UniRef50 RepID=>UniRef50_A0A177KWZ7	37%idt with 58% coverage to CHD_r2
15	CHD_d7	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	26%idt with 51% coverage to UniRef50 RepID=>UniRef50_A0A350JEW0	26%idt with 41% coverage to CHD_r1

16	CHD_d8	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	26%idt with 32% coverage to UniRef50 RepID=>UniRef50_A0A256Y4E2	36%idt with 55% coverage to CHD_r2
17	CHD_d9	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	28%idt with 42% coverage to UniRef50 RepID=>UniRef50_A0A3Q3H3H9	40%idt with 72% coverage to CHD_r1
18	CHD_d10	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	32%idt with 37% coverage to UniRef50 RepID=>UniRef50_A0A5C3FD80	36%idt with 79% coverage to CHD_r2
19	CHD_d11	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	29%idt with 49% coverage to UniRef50 RepID=>UniRef50_A0A843BT52	38%idt with 60% coverage to CHD_r1
20	CHD_d12	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	28%idt with 41% coverage to UniRef50 RepID=>UniRef50_A0A2Z5T4F6	51%idt with 86% coverage to CHD_r1
21	CHD_d13	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	32%idt with 44% coverage to UniRef50	28%idt with 51% coverage to CHD_r1

						RepID=>UniRef50_A0A0M2PYL4	
22	CHD_d14	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	31%idt with 33% coverage to UniRef50 RepID=>UniRef50_A0A1I2G9I0	36%idt with 64% coverage to CHD_r2
23	CHD_d15	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	42%idt with 64% coverage to UniRef50 RepID=>UniRef50_A0A6P8C3U5	40%idt with 66% coverage to CHD_r2
24	CHD_d16	2	FP, SEC	Repetitive ligandMPNN	n.a	23%idt with 44% coverage to UniRef50 RepID=>UniRef50_I1C0H3	29%idt with 56% coverage to CHD_r1
25	CHD_d17	2	FP, SEC	Repetitive ligandMPNN	n.a	28%idt with 50% coverage to UniRef50 RepID=>UniRef50_K1PVQ4	33%idt with 67% coverage to CHD_r2
26	CHD_d18	2	FP, SEC	Repetitive ligandMPNN	n.a	42%idt with 52% coverage to UniRef50 RepID=>UniRef50_A0A4Y2LRW9	56%idt with 72% coverage to CHD_r1

27	CHD_d19	2	FP, SEC	Repetitive ligandMPNN	n.a	26%idt with 44% coverage to UniRef50 RepID=>UniRef50_A0A3Q3H3H9	35%idt with 50% coverage to CHD_r1
28	CHD_d20	2	FP, SEC	PSSM-based Rosetta	n.a	34%idt with 64% coverage to UniRef50 RepID=>UniRef50_A0A654GKB2	56%idt with 78% coverage to CHD_r2
29	CHD_d21	2	FP, SEC	Repetitive ligandMPNN	n.a	24%idt with 55% coverage to UniRef50 RepID=>UniRef50_A0A7R8ZID3	53%idt with 78% coverage to CHD_r1
30	CHD_d22	2	FP, SEC	Repetitive ligandMPNN	n.a	31%idt with 45% coverage to UniRef50 RepID=>UniRef50_A0A173XYP3	32%idt with 45% coverage to CHD_r1
31	CHD_d23	2	FP, SEC	PSSM-based Rosetta	n.a	22%idt with 28% coverage to UniRef50 RepID=>UniRef50_A0A806DV92	33%idt with 55% coverage to CHD_r1
32	CHD_d24	2	FP, SEC	Repetitive ligandMPNN	n.a	27%idt with 35% coverage to UniRef50	39%idt with 61% coverage to CHD_r1

						RepID=>UniRef50_UPI000E68460F	
33	CHD_d25	2	FP, SEC	PSSM-based Rosetta	n.a	30%idt with 36% coverage to UniRef50 RepID=>UniRef50_A0A6G1GIV4	29%idt with 56% coverage to CHD_r2
34	CHD_d26	2	FP, SEC	Repetitive ligandMPNN	n.a	41%idt with 57% coverage to UniRef50 RepID=>UniRef50_UPI0016673335	37%idt with 62% coverage to CHD_r1
35	CHD_d27	2	FP, SEC	PSSM-based Rosetta	n.a	34%idt with 46% coverage to UniRef50 RepID=>UniRef50_A0A3D4MUW3	36%idt with 57% coverage to CHD_r2
36	CHD_d28	2	FP, SEC	Repetitive ligandMPNN	n.a	28%idt with 32% coverage to UniRef50 RepID=>UniRef50_A0A7X7GV99	37%idt with 60% coverage to CHD_r2
37	CHD_d29	2	FP, SEC	PSSM-based Rosetta	n.a	28%idt with 36% coverage to UniRef50 RepID=>UniRef50_A0A6S7ITZ3	39%idt with 61% coverage to CHD_r1

38	CHD_d30	2	FP, SEC	PSSM-based Rosetta	n.a	27%idt with 39% coverage to UniRef50 RepID=>UniRef50_A0A518K765	48%idt with 64% coverage to CHD_r1
39	CHD_d31	2	FP, SEC	PSSM-based Rosetta	n.a	34%idt with 32% coverage to UniRef50 RepID=>UniRef50_A0A3M7S8C3	35%idt with 51% coverage to CHD_r2
40	CHD_d32	2	FP, SEC	Repetitive ligandMPNN	n.a	34%idt with 45% coverage to UniRef50 RepID=>UniRef50_UPI000711C2C5	52%idt with 76% coverage to CHD_r2
41	CHD_d33	2	FP, SEC	Repetitive ligandMPNN	n.a	30%idt with 60% coverage to UniRef50 RepID=>UniRef50_A0A5C5V838	47%idt with 72% coverage to CHD_r1
42	CHD_d34	2	FP, SEC	PSSM-based Rosetta	n.a	24%idt with 53% coverage to UniRef50 RepID=>UniRef50_A0A4U0VE94	54%idt with 77% coverage to CHD_r1
43	CHD_d35	2	FP, SEC	Repetitive ligandMPNN	n.a	37%idt with 32% coverage to UniRef50	41%idt with 63% coverage to CHD_r1

						RepID=>UniRef50_A0A1F9HRL8	
44	CHD_d36	2	FP, SEC	Repetitive ligandMPNN	n.a	36%idt with 47% coverage to UniRef50 RepID=>UniRef50_Q0UZJ0	38%idt with 66% coverage to CHD_r1
45	CHD_d37	2	FP, SEC	Repetitive ligandMPNN	n.a	32%idt with 36% coverage to UniRef50 RepID=>UniRef50_A0A7S3WRZ2	33%idt with 61% coverage to CHD_r1
46	MTX_d1	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	32%idt with 36% coverage to UniRef50 RepID=>UniRef50_E3HCG3	43%idt with 57% coverage to MTX_r1
47	AMA_d1	2	SPR, SEC, yeast surface display	PSSM-based Rosetta	n.a	37%idt with 49% coverage to UniRef50 RepID=>UniRef50_A0A1G5S3S9	44%idt with 67% coverage to MTX_r1
48	AMA_d2	2	SPR, SEC, yeast surface display	Repetitive ligandMPNN	n.a	35%idt with 51% coverage to UniRef50 RepID=>UniRef50_A0A517Y9K7	33%idt with 61% coverage to AMA_r2

49	T44_d1	2	FP, SEC, yeast surface display	PSSM-based Rosetta	n.a	34%idt with 43% coverage to UniRef50 RepID=>UniRef50_A0A4P7A274	67%idt with 83% coverage to T44_r1
50	T44_d2	2	FP, SEC, yeast surface display	PSSM-based Rosetta	n.a	24%idt with 31% coverage to UniRef50 RepID=>UniRef50_A0A8H8JH58	65%idt with 85% coverage to T44_r1
51	T44_d3	2	FP, SEC, yeast surface display	PSSM-based Rosetta	n.a	28%idt with 25% coverage to UniRef50 RepID=>UniRef50_R7RX94	68%idt with 84% coverage to T44_r1
52	T44_d4	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	31%idt with 42% coverage to UniRef50 RepID=>UniRef50_A0A7C2MB67	54%idt with 74% coverage to T44_r2
53	T44_d5	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	33%idt with 57% coverage to UniRef50 RepID=>UniRef50_A0A7S3ML05	48%idt with 65% coverage to T44_r2
54	T44_d6	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	21%idt with 49% coverage to UniRef50	48%idt with 71% coverage to T44_r1

						RepID=>UniRef50_A0A6A4XWT8	
55	T44_d7	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	31%idt with 37% coverage to UniRef50 RepID=>UniRef50_A0A6N6S117	69%idt with 88% coverage to T44_r1
56	T44_d8	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	31%idt with 35% coverage to UniRef50 RepID=>UniRef50_A0A6I7PU97	52%idt with 73% coverage to T44_r2
57	T44_d9	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	30%idt with 40% coverage to UniRef50 RepID=>UniRef50_A0A397TFF8	52%idt with 75% coverage to T44_r2
58	T44_d10	2	FP, SEC, yeast surface display	Repetitive ligandMPNN	n.a	26%idt with 34% coverage to UniRef50 RepID=>UniRef50_A0A1H8UAJ6	46%idt with 70% coverage to T44_r1

*Structural similarity described as %idt

Table S3. Data collection and refinement statistics for co-crystal structures.

	CHD_r1 (PDB: 8VEI)	**CHD_buttress (PDB:**8VEJ**)**
Wavelength	0.9791	0.9201
Resolution range	42.56 - 2.1 (2.26 - 2.1)	29.14 - 3.59 (3.71 - 3.59)
Space group	C 2 2 2_1	P 2_1
Unit cell	46.242, 57.819, 85.120; 90, 90, 90	72.607, 45.962, 73.664; 90, 104.045, 90
Unique reflections	6928 (1348))	4892 (483)
Multiplicity	12.6 (12.6)	2.1 (2.0)
Completeness (%)	99.86 (99.78)	85.78 (87.50)
Mean I/sigma(I)	12.51 (2.66)	4.55 (1.51)
Wilson B-factor	40.14	100.41
R-merge	0.143 (1.190)	0.143 (0.680)
R-pim	0.0415 (0.3465)	0.112 (0.551)
CC1/2	0.998 (0.892)	0.991 (0.510)
Reflections used in refinement	6928 (1348)	4894 (483)
Reflections used for R-free	693 (135)	484 (49)
R-work	0.2637 (0.3429)	0.2391 (0.2780)
R-free	0.3062 (0.4151)	0.2797 (0.3529)
Number of non-hydrogen atoms	1051	3713
macromolecules	1022	3655
ligands	29	58
solvent	0	0
Protein residues	120	482
RMS(bonds)	0.003	0.003
RMS(angles)	0.67	0.55
Ramachandran favored (%)	95.77	96.17

Ramachandran allowed (%)	2.54	3.62
Ramachandran outliers (%)	1.69	0.21
Average B-factor	51.79	94.42
macromolecules	51.52	94.97
ligands	61.42	59.55

Table S4. Ligand complexity characterization.

The features of Ligands from previous and current work (green) of SM binder designs are extracted from PubChem(*36*) or counted (for AMA) and listed here. Only ligands with binders with affinity reported and some binding pose validations (SSM, competition, or crystallization) are included here.

*max_total_hb equals addition of the number of hydrogen bond donors and the number of hydrogen bond acceptors.

Full_name	PDB_code	max_total_hb*	flexible_bonds	mass	XLogP3-AA	publication	scaffold
methotrexate	MTX	17	9	454.2	-1.8	this	barrel-like pseudocycle
Cholic acid	CHD	9	4	408.6	3.6	this	6 helical bundle-like pseudocycle
L-thyroxine	T44	8	5	776.9	2.4	this	4 helical bundle-like pseudocycle
AMA	AMA	14	>4	520.5	\	this	bowl-like pseudocycle
Digoxigenin	DIG	8	1	390.5	1.1	Tinberg, Nature, 2013; Lee, biorXiv, 2023.	NTF2
fentanyl	7v7	1	6	336.5	4	Bick, eLife, 2017	Native scaffold redesign
Apixaban	APX	6	5	459.5	2.2	Polizzi, Nature, 2020	four helical bundle
DFHBI	1TU	6	1	252.2	1.4	Dou, Nature, 2018	beta barrel
cortisol	COR	8	2	362.5	1.6	Lee, biorXiv, 2023	NTF2
Rocuronium	RBR	6	6	529.4	5	Lee, biorXiv, 2023	NTF2
Warfarin	SWF	5	4	308.3	2.7	Lee, biorXiv, 2023	NTF2
7-Ethyl-10-hydroxycamptothecin	SN-38	8	2	382.4	1.4	Lee, biorXiv, 2023	NTF2

Supplementary Materials

Materials and Methods

Stepwise SM binder design pipeline in detail

All scripts on stepwise SC-optimizing SM binder design pipeline are available at github (repository will be public after manuscript acceptance): https://github.com/LAnAlchemist/Pseudocycle_small_molecule_binder.

All binders were designed using the same computational pipeline unless specified otherwise.

Ligand preparation for computational processing

The 3D structure of the ligands were first taken from PDB or generated using ChemDraw 3D. The hydrogens were added using OpenBabel(*37*), and edited with VMD(*38*) for correct chemical structures and protonation states. Around 300-700 rotamers were generated for each ligand using RDKit(*23*). All rotamers were then relaxed using Rosetta with constraint, and scored for ddG. The scored rotamers were also aligned using RDKit, function 'Chem.rdMolAlign', and clustered based on r.m.s.d, using cutoff at 1.5 Å. The rotamer with lowest Rosetta energy from each cluster was selected as the starting point. The final selected rotamers are all physically allowed, but not necessarily the lowest-energy rotamers among all possible rotamers. The parameter file of each ligand based on 'ref2015', 'genpot'(*39*) score functions were generated using Rosetta. The only exception is AMA, where its single designed structure of AMA was used as the binding confirmation.

First step SC-optimizing sampling

The previously published 9838 pseucocles were reduced to 9,703 based on protein length; only proteins with length shorter than 155 aa were used for the following design work due to the limitation of the oligo library synthesis availability. The pocket residues of pseudocycles were annotated using in-house python script which identifies the largest internal cavity bounded by the protein after converting the protein to polyalanine and then identifies all side chain residues contacting this internal cavity.(*11*)

All rotamers of all four ligands were first docked to 9,703 pseudocycles using RIFgen/RIFdock suite(*3*) using the same docking protocol as the previous publication(*11*). Considering the aim of the first sampling step is to identify suitable scaffold for each ligand, to save resources, only 10 docks were requested between each ligand rotamer to each pseudocycle scaffold.

Predictor to select best docks

All docks obtained from the previous step were collected and the interfaces (a.k.a ligand and the pocket residues) were designed using a quick design and score Rosetta protocol ('FastPredictor_v2_talaris_ligand.xml'). The interface except for the previously seeded rifres was quickly packed with big hydrophobic residues (a.k.a Val, Leu, ILE, PHE, TYR) using score function 'talaris2014'. No Rosetta relaxation was performed, and only 'contact_molecular_surface' (CMS) was scored to save time. On average, 10 docks takes ~ 1 CPU second to score. The CMS scores were collected and ranked from highest to lowest. Docks at different CMS ranks were pulled out for manual inspection, and a cutoff CMS score was selected based on how well the ligand contacts with the scaffold. Usually 30-50% of the docks were dropped at this stage.

The selected docks were designed and scored again using another relatively quick design and score Rosetta protocol. The interface except for the previously seeded rif res were packed using layer design protocol ('LA_quick_design_select_dock_genpot.xml'). On average, 1 dock takes ~ 10 CPU seconds to score. 'CMS', 'dsasa', 'ddg' (a.k.a 'ddg_norepack' in the protocol), 'holes_around_lig' were used as major criterias to select final docks for design. The features were scored and ranked, and cutoffs were selected by manual inspection of docks at different ranks. The low cutoff, low cutoff, high cutoff, and high cutoff were selected for the abovementioned metrics, respectively.

All docks post first predictor were also taken and measured the potential ligand tail atom burial to avoid designs with too buried ligands.

All scores generated from predictor protocol scripts were found to be statistically highly correlated with the scores generated from the actual design protocols, which indicates the 2-step predictor is a good method to rank docks and select good docks to improve computational design efficiency (**Fig S5**).

PSSM-based Rosetta design protocol

One PSSM file was generated for each pseudocycle scaffold using ProteinMPNN(*13*). 100 sequences were generated for each pseudocycle scaffold using ProteinMPNN with temperature at 0.2 to increase sequence diversity. The PSSM score at each position of each amino acid type was calculated through comparing proteinMPNN generated frequency (p_observed) to that of BLOSUM62 amino acid background frequency (p_background). The calculation equation is:

Pssm_score = 2*log_2(p_observed/p_background)

Where p_observed is the frequency of the particular amino acid type observed at the specified location.

Two rounds of FastDesign were performed with the sequence biased using PSSM ('design_ligand_full_noHBNet_1.py'). For a 120-aa protein, usually it takes 15 min CPU time for each design and score. Because the quality of both the protein model (as assessed by AlphaFold2(*14*), AF2), and the ligand docks (selected by the two-step `**predictor**`) were high at this point, for both protocols, strong constraints were added to disallow significant movements of the protein backbone or the ligand docking conformation.

Iterative ligandMPNN design protocol

The output of the PSSM-based Rosetta designed proteins were used as input for repetitive ligandMPNN design protocol ('ligMPNN_FR_silent_in.py'). All short range interactions seeded from the previous design and RIFdock (polar interactions, π-π interactions) were kept in the first round of ligandMPNN design, while all residues were allowed for design during the second and third round of ligandMPNN design. Rosetta minimization with strong ligand-protein constraints were performed between ligandMPNN runs to remove potential unideal local interactions while preserving the ligand docking pose and protein backbone conformations. Temperature of 0.2 and polar/nonpolar bias were used to increase the sequence diversity and polar interface building bias.

Selection of designs for ordering

All designs were pooled together and filtered with the same criteria. Rosetta metrics including 'CMS', 'ddg2', 'hole_around_lig', 'dsasa', 'total_hb_to_lig' were used for design judgements. Designs with different scores of above-mentioned metrics were manually inspected for low cutoff, high cutoff, high cutoff, low cutoff, and low cutoff selection, respectively. The proteins of designs passed Rosetta filters were then first scored using AF2, and only designs with pLDDT (predicted local distance difference test) above 85, predicted C□-r.m.s.d smaller than 1 or 1.5 Å were ordered.

The second step SC-optimizing sampling

The pseudocycle scaffold-ligand pair identified from the first stem sampling and showed potential binding in the laboratory (**Fig 1c-f, S2, S3, S6**) were used for the second computational sampling step. The same second step of the sampling pipeline was used for all ligands. First, the hit scaffolds were taken and resampled through generating 1000 sequences using proteinMPNN(*13*) and folded using AF2(*14*). Only AF2 models (model 4) predicted to be within

Cα-r.m.s.d 3 Å, plDDT above 90, predicted template modeling score (pTM) above 0.65 were taken as the newly resampled models for next docking and design step.

The same docking procedure was used between each ligand rotamers and their hit scaffolds, except for 1000-1500 docks were requested. The same two-step predictor protocol was used to identify good docks for actual design protocols, which the selection criteria were changed to using the scores from the previous hits. The same sequence design protocols and filter steps were used, except for most of the selection criteria were taken from the previous hits.

Once optimal scaffold topologies were identified for each ligand in the first round, the success rate in the second design round was significantly higher than in previous studies using a single protein family for ligand binding(*3*–*5*, *7*, *8*). For easier targets, such as CHD, after NGS analysis, we confirmed 231 binders from 14,882 designs, with a 1.55% success rate. For the flexible T44, 84 binders were obtained based on NGS data from 5,936 designs, a 1.42% success rate. Both are significantly higher than previous studies where a single family of proteins were used for ligand binder design(*3*, *6*, *8*).

With further data on optimal pseudocycle geometries for different target ligands, it should become possible to identify optimal scaffolds for a given ligand computationally which could obviate the need for the initial design round, considerably streamlining the design process (the 'contact_molecular_surface', measuring protein-ligand contacts, is significantly higher for the round 1 designs with binding activity, but more data is needed to set thresholds; **Fig S1c**).

Preparation of the FITC-labeled SM target

Free CHD (1133503-200MG), T44 (T2376-1G), MTX (454126-100MG) were purchased from Sigma-Aldrich (St. Louis, MO). MTX-fl was purchased from Thermo Fisher (M1198MP, Bothwell, WA). MTX-btn (M260670) was purchased from Toronto Research Chemicals. CHD-fl was purchased from Life Sciences (#451041). Biotinylated AMA were synthesized, purified, and verified by WuXi Chemistry AppTec, Inc.

Preparation of the FITC-labeled T44 was performed as previously described with some slight modifications(*40*). Briefly, L-thyroxine was incubated directly with N-hydroxysuccinimide-fluorescein (2 eq.) and N,N-diisopropoylethylamine (4 eq.) in dimethylformamide (DMF) for 4 hours. Following conversion to the fluorescent product, which was confirmed by mass spectrometry, the reaction was rotavapored to remove DMF, resuspended in acetonitrile and water, and purified by reverse-phase high-performance liquid chromatography to yield FITC-labeled T44.

All ligands were stored following commercial instructions, or in 30% (v/v) dimethyl sulfoxide (DMSO)/water at -20 ℃ for long-term storage.

High-throughput methods for binder identification

All library oligos were purchased from Twist Bioscience (South San Francisco, CA). The KAPA HiFi HotStart Uracil+ReadyMix (KK2802) and KAPA HiFi HotStart Kit (KK2101) were purchased from Roche (Basel, Switzerland). All gene fragments were purchased from Twist Bioscience (South San Francisco, CA) or IDT (Coralville, IA). The EvaGreen Dye, 20 x in water (#31000), was purchased from Biotium (Fremont, CA). The USER enzyme (NEB#M5508) and the NEBNext End Repair Module (E6050L) were purchased from New England BioLabs (Ipswich, MA). All DNA purification kits were purchased from QIAgen (Hilden, Germany). All chemicals and consumables were purchased from Thermo Fisher(Bothell, WA), unless specified otherwise. The anti-cMyc-R,Phycoerythrin (anti-cMyc-PE) was purchased from Cell Signaling Technologies (Danvers, MA). All tagged and free ligands were described in '***Preparation of the FITC-labeled SM target***' in Methods.

Yeast surface display library preparation

The designed proteins were reverse transcribed using DNAworks(*41*) with avoidance of common restriction enzyme cut sites. The nucleotide sequences were then broken into two parts (5' part, 3' part) with common complementary sequences in the middle for assembly using an in-house python script. The designs were separated into different groups of around 1000 based on length to avoid quantitative PCR (qPCR) bias caused by oligo length differences. The primers complementary to the pETCON3 yeast surface display vectors were added to the 5' of the 5' part and 3' of the 3' part. Inner primers with AA, TT in the middle were padded to the 3' of the 5' part and 5' of the 3' part to facilitate library PCR. The 250 and 300 nt oligo pools were ordered from Twist Bioscience depending on the protein length.

The purchased oligo pools were dissolved in water to prepare 100 ng/μL stock solution and stored at -20 ℃. The oligo pool stock was diluted to 2.5 ng/μL in water for qPCR experiments. The individual 5'-parts and 3'-parts were amplified using designed primers following commercially available protocol from KAPA HiFi HotStart Uracil+ReadyMix. DNA was purified using a PCR purification kit in between the qPCR runs. The inner primers of individual parts were removed using USER enzymes and the NEBNext End Repair Module following commercially available protocols. After DNA purification, the digested 5'-parts and 3'-parts were assembled using the same qPCR protocol, and the assembled oligo pools were checked using DNA electrophoresis. The designed DNA bands were cut from gels and purified for amplification. Roughly 4-6 μg of oligo pools were used for electrocompetent yeast

transformation for a library with complexity between 5,000-10,000. The oligo pool and the linearized pETCON3 vector were added to yeast at a ratio of 3:1 and the electro-transfer protocol for the yeast library was used for transformation following the previously published protocol(*42*).

Yeast surface display experiment for enrichment of the potential binders

The previously established yeast surface display protocol was followed with minor modification(*42*). PBSF, PBS with 0.1% (w/v) Bovine Serum Albumin (BSA), were used as all yeast sorting buffers. EBY1 were used for all yeast surface display experiments. All library fluorescence-activated cell sorting (FACS) experiments were performed on a Sony SH800S Cell Sorter using 100-µm chip, which is equipped with laser 488 nm and 561 nm.

For all yeast surface display libraries, one round of expression sort was first performed to enrich the binders with acceptable expression level. The binder-expressing yeast were stained at 1:50 (v/v) ratio with anti-cMyc-FITC for 30 min at 4 ℃ for expression sort; yeast with increased FITC signal, indicating expression, were collected for further binding assays. After expression sort, the collected yeast were grown up in C-Trp-Ura-2% glucose (CTUG) medium again, and expressed overnight at 30 ℃ in a shaker in SGCAA medium for binder expression.

If the ligands were tagged with FITC, the expressed yeast was first stained with 1:35 (v/v) anti-cMyc-PE for 45 min at 4 °C, followed by two washes using PBSF. Then the stained yeast was incubated with FITC-labeled ligands for 60-120 min at 4 °C. After two more washes, the stained yeast was sorted. The expressed-only yeast was used as negative control, while the yeast population with positive signals from both FITC and PE channels above the negative control samples were collected and grown up in CTUG medium. After 1-2 days of growth, the collected yeasts were expressed again in SGCAA medium and sorted again for binder enrichment through a similar above-described method. Usually it takes two to three sorts to have clear enrichment of the binding populations. If the binding population was strong (above 2%), a titration may be performed to select the strongest binders. Titration sorts were performed for CHD and T44 for both round1 and round2 binding assays. Usually 5 µM FITC-labeled ligands were used for sort1. The concentration would be dropped according to the positive signal intensity at sort2 and sort3. Usually 100 pM - 5 uM of FITC-labeled ligands were used for titration sorts.

If the ligands were tagged with biotin, the expressed yeast was always stained with anti-cMyc-FITC at 1:50 (v/v) for 35 min at 4 °C. The first and second sort were usually performed with avidity, meaning, 1 µM of biotinylated ligands were added to yeast with 0.25 µM anti-Streptavidin-R,Phycoerythrin (SAPE) at the same time for 120 min at 4 °C during the ligand binding step. If the binding population was strong enough, the second, or third round of sorting

would be performed without avidity, meaning 100 nM-10 μM of biotinylated ligands were added to design-expressing yeast for incubation for 120 min at 4 °C. After two washes using PBSF, SAPE would be added to stain the ligand-bound yeast at 1:50 (v/v) for 35 min at 4 °C.

Cultures from expression, and all sorting steps were collected for MiSeq analysis. All yeast sorting figures were prepared using FlowJo v10 (FlowJo, Inc).

Next-generation sequencing data analysis

All sequences were assembled and matched using an in-house script, and the git repository is at: https://github.com/feldman4/ngs_app.

For sequencing data with titration sorting, the data were analyzed based on previously published scripts to identify potential binders(*43*), and designs with predicted Kd with equal or lower than micromolar range were defined as potential binders, and some of them were selected for individual purification and binding assays, such as SEC, FP, and SPR.

For sequencing data without titration sortings, enrichment factors were calculated based on the reads from sort3 versus the reads from sort2, usually the binders with enrichment factor above 10 and reads from sort3 more than 1000 were ordered individually for binding test.

Yeast surface display for individual binder identification

A 120 μg of each gene fragment encoding individual design was mixed with 40 μg of linearized pETCON3 vectors and transferred to chemically competent yeast EBY1 following previously published protocols(*43*). The transformed yeast were first grown in CTUG media in a 96-deep well round bottom plate in plate shaker at 30 °C for 30 hours at 200 g, and expressed in SGCAA media for 14 hours under the same shaking conditions.

The expressed yeast were stained in the same fashion as previously described (see **'*High-throughput methods for binder identification*'** in Methods) and the binding and expression were checked on an Attune Flow Cytometer equipped with 488 nm and 561 nm lasers. For competition assays, free ligands were added with around 1000 folds higher than the FITC or biotin-labeled ligand at the same time, with the rest of the steps following the same like those described above.

All yeast sorting figures were prepared using FlowJo v10 (FlowJo, Inc).

Expression and purification of selected proteins

The major procedures were adapted from a previous publication(*11*). All chemicals and supplies were purchased from Thermo Fisher unless specified otherwise.

The selected designs were reverse-transcribed into DNA using DNAWorks(*41*). Eblocks (IDT or Twist Bioscience) were cloned into a pET29b-derived vector with C-terminal SNAC cleavable His-Tags using commercially available Golden Gate assembly kit (New England Biolabs, Ipswich, MA) and transformed into *Escherichia coli (E. coli)* BL21 strain. All designs were grown in 50 mL autoinduction Terrific Broth media in 250 mL baffled erlenmeyer flasks for production (6 hours at 37 ℃ followed by 24 hours at 18 ℃ shaking at around 200 g in New Brunswick Innova® 44 shakers). Cells for each design culture were harvested using centrifugation and resuspended in 25 mL of lysis buffer (25mM Tris 100 mM NaCl, pH 8, with protease inhibitor tablet) and sonicated to lyse (4.5 minutes sonication, 10 s pulse, 10 s pause, 65% amplitude). After centrifugation for 30 min at 14,000 g, soluble fractions were bound to 1 mL Ni-NTA resin (Qiagen) in a Econo-Pac® gravity column (BIO-RAD) for 1 hour with rotation. The resin was washed with 20 CV (column volume) low salt buffer (50 mM tris, 100 mM NaCl, 10 mM Imidazole, pH 8) and with 20 CV high salt buffer (50 mM tris, 1000 mM NaCl, 10 mM Imidazole, pH 8). Proteins were eluted with 2 CV of elution buffer (20 mM tris, 100 mM NaCl, 500 mM Imidazole, pH 8) and purified on a superdex 75 increase 10/300 GL column connected to ÄKTA protein purification systems in isocratic TBS buffer (20 mM Tris, 100 mM NaCl, pH 8).

For proteins for FP, SPR, and crystallization studies, the Histags were cleaved following on-bead SNAC tag cleavage protocol(*44*). The proteins were prepared using the same way described above until protein binding onto the resin. The column was washed twice with lysis buffer to remove imidazole, followed by equilibrium with 10 CV of SNAC buffer (100 mM CHES, 100 mM acetone oxime, 100 mM NaCl, 500 mM guanidine chloride, pH 8.5). Then the columns were incubated with 20 CV of fresh SNAC cleavage (2 mM NiCl2 in SNAC buffer) overnight at r.t. for SNAC tag on resin cleavage. The resins were then washed with 10 CV of weak elution buffer (20 mM Tris, 100 mM NaCl, 10 mM Imidazole, pH 8). Both SNAC cleavage solution and the weak elution solution were collected, and checked with protein SDS-Page for tag-free proteins. The tag-free proteins were pooled, concentrated, and purified on SEC using the same method described previously.

All purified proteins were verified using mass spectrometry.

Fluorescence polarization studies

The protein concentrations were judged using Bradford protein assay (Bio-Rad, Hercules, CA) following commercially available protocols.

All FP studies and analyses were performed following a previously established protocol(*45*) on a Biotek Synergy Neo2 Reader equipped with Dual PMT optical filter cube (part number: 1035108). The proteins were serial diluted and 5 or 20 nM of FITC-labeled ligands were added to the diluted protein. The mixtures were transferred to Corning 384 Well microplate (CLS4515-10EA). After 30 min incubation at 25 ℃ with shaking, the parallel and perpendicular light were read with scale to the highest-intensity well. Two independent FP were performed as replicates.

FP signal was calculated based on below equation:

$$FP = (I_{parallel}-I_{perpendicular})/(I_{parallel}+I_{perpendicular})$$

The data were fit using the Origin with GrowthCurve DropWise model to avoid ligand binding saturation.

SPR studies

The protein concentrations were judged using Bradford protein assay (Bio-Rad, Hercules, CA) following commercially available instructions. All SPR studies were performed on a Cytiva Biacore 8K SPR system, using a biotin CAPture chip, and in 1x HBS-EP+ (0.1 M HEPES, 1.5 M NaCl, 0.03 M EDTA and 0.5% v/v Surfactant P20, Cytiva) buffer.

All runs were performed using both flow cells, with one reference cell and one experimenting cell.

A start-up run was always first performed with running biotin CAPture reagent (Cytiva) through both cells for 300 seconds contacting, at flow rate of 2 µl/min, following 25 nM biotinylated ligand (biotinylated-AMA) contacting for 180 seconds at at a flow rate of 10 µl/min for flow cell 2. The buffer was then flowed through with 120 seconds contacting and 60 seconds dissociation at a flow rate of 30 µl/min, regeneration solution (3:1 (v/v) 8 M guanidine HCl pH 8, and 10 M NaOH) contacting for 120 seconds at flow rate of 10 µl/min, and lastly washed with buffer.

Blank runs were run before and after analysis runs to set up blanks, the only difference between analysis runs and blank runs was that buffers were used instead of binding proteins.

For analysis runs, the proteins were prepared at 6 concentrations, and tested from low concentration to high concentration sequentially following the same protocol. First, biotin CAPture reagent was flown through both cells for 300 seconds contacting, at a flow rate of 2 µl/min. The 25 nM biotinylated ligand (biotinylated-AMA) contacts for 180 seconds at a flow rate of 10 µl/min for only flow cell 2. Then, the protein solution was flow through for 120 seconds, with 300 seconds disassociation time, at a flow rate of 30 µl/min, to both cells. All 6 different concentrations were flown through sequentially from low to high concentrations. Lastly, the generation solution and wash cycle were performed similar to the start-up round.

The SPR data were analyzed and plotted using Cytiva Biacore Insight Evaluation software.

Site-specific mutagenesis SSM studies

SSM studies were performed following the previous protocol(*43*). The SSM library was prepared and sorted following previous protocol (see '***Yeast surface display library preparation***') with minor modification. Each protein residue was mutated to all 20 possible residues, including Cystine. 1 to 5 binders to the same target were pooled together as 1 single SSM library for yeast surface display studies. The only difference during sorting was that the same gate was used for sort 1, 2. Various concentrations of ligands were used for sort2 for titration studies following previously published protocol(*42*). The site-specific Shannon entropy was calculated based on the counts and enrichment following the previously published protocol, and the SSM matrix plots were generated based on the Shannon entropy(*43*).

Design of SM binder fused nanopore

The nanopore TMB12_3 from our de novo nanopore design(46) was chosen for the design of the cholic acid binder-based nanopore sensor. This nanopore was shown to have a stable conductance at ~230 pS in a 500 mM NaCl buffer. Also, since the SM binder for cholic acid has 6 helices, a 12 stranded beta-barrel nanopore seemed appropriate for the fusion where 4 strands from the beta-barrel could be fused to 2 helices from the binder leading to a maximum of 6 loop connections between the two proteins. 6 loops with the right lengths would possibly allow for complete blockage of the nanopore opening from a direction perpendicular to the pore and SM binder axis. This is a desirable feature for a binary nanopore sensor where a binding event can be associated with significant current blocking leading to nanopore currents very close to 0 pA. Therefore, the SM binder was placed on one side of the nanopore and roughly aligned to match the closest helical and strand termini that would generate the most optimal fusion without significantly changing the structure of either protein. Complete freedom was allowed for different loop closure solutions based on the orientations of the binder and the nanopore. Further sampling of the binder position was carried out by rotating the binder around the pore axis within

an angular range that would allow for different termini distances between the fusion points of the binder and the nanopore. The upper limit on termini distance (between a helix of the binder and a strand of the nanopore) was chosen as 12 Å based on initial sampling which showed that further distances would lead to a different loop closure solution as compared to the starting configuration. Also, the binder was translated along the pore axis within 5-8 Å of the pore opening to allow for different loop conformations upon fusion. The loop closure was carried out using a deep learning model(25) for a range of starting termini configurations and their corresponding conformations under different rotations and translations of the binder as described above. Different loop lengths were sampled and final selection for the generated backbones was carried out by filtering out structures with chain breaks. Sequence design was carried out using ProteinMPNN(13) and Rosetta total scores were used for final filtering as AF2 was incapable of predicting full structures from single sequences for these types of fusion proteins.

Expression and purification of designed binder-based nanopore sensors

Fusion constructs as designed above were ordered as Eblocks from IDT and cloned into a pET29b vector with T7 promoter and kanamycin resistance gene. The proteins were expressed and purified using an auto-induction media as previously described(*47*). All proteins were refolded from inclusion bodies using dodecyl-phosphatidylcholine (DPC) detergent and were subsequently run through a SEC column (Superdex S200 from Cytiva). Appropriate fractions were diluted to nanomolar to picomolar concentrations and tested for membrane insertion and conductance as described below.

Conductance measurement in planar lipid bilayers

The conductance measurements were performed as previously described(*47*). All ion-conductance measurements were carried out using the Nanion Orbit 16TC instrument (https://www.nanion.de/products/orbit-16-tc/) on MECA chips. Lipid stock solutions were freshly made in dodecane at a final concentration of 5mg/mL. Di-phytanoyl-phosphatidylcholine (DPhPC) lipids were used for all experiments. Designed proteins were diluted in a buffer containing 0.05% DPC (~ 1 critical micelle concentration), 25 mM Tris-Cl pH 8.0 and 150 mM NaCl to a final concentration of ~100 nM. Subsequently, 0.5 μL or less of this stock was added to the cis chamber of the chip containing 200 μL of buffer while simultaneously making lipid bilayers using the in-built rotating stir-bar setup. All measurements were carried out at 25 ℃ and with a positive potential bias of 100 mV. Spontaneous insertions were recorded over multiple rounds of bilayer formation. All chips were washed with multiple rounds of ethanol and water and completely dried before testing subsequent designs. A 500 mM NaCl buffer was used on

both sides of the membrane for all current recordings. Raw signals were recorded at a sampling frequency of 5 kHz. Only current recordings from bilayers whose capacitances were in the range 15-25 pF were used for subsequent analysis. The raw signals at 5 kHz were downsampled to 100 Hz using an 8-pole bessel filter. Estimation of current jumps were carried out using a custom script with appropriate thresholds.

Estimation of dissociation constant (Kd) from nanopore conductance measurements

The analysis of nanopore data was generally following previously published protocols(*48*). The following assumptions were made for estimating the Kd of ligand binding to the CHD nanopore sensor. First, despite the presence of three distinct current states of the nanopore in the presence of ligand, only the lowest current level was considered to be indicative of ligand binding. Therefore current level jumps from the lowest state to either of the higher current states were treated identically and were considered to be indicative of unbinding. Second, ligand binding was assumed to stabilize the binder domain in a closed conformation and likely prolong the lowest current state of the nanopore. However, since the lowest current state is also observed for the control condition with no ligand, it was assumed that the dwell time of this current state in the presence of ligand would be a combination of inherent stochasticity of the binder domain closing and potential ligand binding.

With the above assumptions, a hill equation model with hill coefficient as one was sufficient to capture the binder-domain transitions from an unbound to a ligand-bound state:

$$f = [L]/(Kd + [L])$$

Here f denotes the fraction of the ligand bound state of the nanopore and [L] is the concentration of the ligand in solution which was fixed at 10 uM for this experiment.

Dwell time distribution for events resulting in the lowest current state of the nanopore were calculated for both cases (absence and presence of ligand) and were fitted to an exponential curve to estimate the mean dwell times (inverse of the exponential decay constants), Tc (nanopore only) and Tl (nanopore with ligand) respectively (**Figure S17c**). The mean dwell times were normalized by the duration of the recordings to represent the probabilities of the unbound and bound states. To prevent user bias in determining a cutoff for discarding very small dwell times resulting from noise or initial signal-filtering artifacts, a range of dwell time cutoffs from 50 ms to 400 ms were used. This resulted in a range of mean dwell times post exponential curve-fitting for each experimental condition. For each pair of Tc and Tl, the fraction of ligand-bound receptors was estimated as:

$$f = (Tl - Tc)/Tc$$

Corresponding Kd values were calculated with [L] = 10 uM. The range of calculated Kd is shown as a histogram in **Figure S17d**.

Engineering/Buttressing pseudocycle binders

The overall steps are shown in **Fig S18a**. The starting pseudocycle binder was first broken into individual repeat units based on their structural symmetry order and expanded to multiple perfect symmetric structures. We mutated the surface residues to apolar residues to help diffuse buttrased structures out. We also found use of `adjacent_matrices` helps placement of buttressed structures. Then, we used symmetry diffusion to generate buttress around the individual symmetric structures(*10*). Symmetric diffusion was used because it is significantly faster and easier to converge to high quality protein scaffolds than free diffusion. We then take individual parts of the buttressed unit, mixed and matched them together to make buttressed binders. The binding interface residues were not designed while the rest of the residues were redesigned using proteinMPNN(*13*), scored with Rosetta, and folded using AF2(*14*). The 68 buttrased designs passed Rosetta, AF2 filters were ordered using Twist gene fragments. The designs were first transformed into EBY1 with pETCON3 vector for yeast surface display test to identify initial binders (see method '***Yeast surface display experiment for enrichment of the potential binders***'). The designs with potential binding signal on yeast surface display were purified following protocol '***Expression and purification of selected proteins***', and the binding affinity was determined using FP following protocol '***Fluorescence polarization studies***'.

CID design and testing

Computational design of CID

The scripts for CID design were uploaded to github: https://github.com/iamlongtran/pseudocycle_paper (will change from private to public upon BiorXiv release).

The design of CID starts with verified buttressed binders, which were splitted into two halves (A and B). The main objective is to keep the ligand binding in presence of the SM ligand, while preventing the A-B interactions without the ligand. We first used ligandMPNN(*16*) for sequence design. Specifically, we kept the verified ligand-protein interface, and used tied sequence design functions, which allows simultaneous sequence design on A-ligand complex, B-ligand complex, and A-ligand-B complex, and used tied weight to reward A-ligand-B complex design and weak the rest conformation design.

Once we have designed CIDs, we predicted the A-B dimer formation using superfold, an in-house AF2 parser (https://github.com/rdkibler/superfold) to select designs which were predicted not to form hetero/homo-dimers but form stable individual monomeric proteins.

For A-ligand-B complexes, we use RosettaFold All Atom (RFAA)(*9*) to predict the complex structure. We selected designs with Predicted Aligned Error (PAE) lower than 5 and predicted model with low Cα-r.m.s.d to designs, and ordered as gene fragments from Twist Bioscience.

Chromatography test for identification of CID pairs

The expression and purification protocol of all proteins are closely followed to the section '***Expression and purification of selected proteins***', with a few minor adjustments. For the volume of autoinduction TB media, for CID screening, designs were grown in 5 mL TB autoinduction media in Falcon™ Round-Bottom Polypropylene Test Tubes with Cap in replicates. After growing for 6 hours at 37 ℃ followed by 24 hours at 18 ℃ with shaking at 225 rpm, cultures of each design were transferred to a separate well from a 24-well plate. Design cultures were harvested using centrifugation at 4000 g for 15 minutes, then resuspended in a 10 mL lysis buffer in the 24-well plate. The suspended cells were sonicated to lyse for 7.5 minutes, using 10 second pulse, 10 second pause, at 65% amplitude. The lyzed mixtures in 24-well collection plates were then centrifuged at 2000 g for 20 minutes to separate cell debris and lysate. The soluble fraction collected post centrifugation were bound to 1 mL of Ni-NTA resin in a 24-well filter plate, followed by wash and elution facilitated by a vacuum manifesto. The individual eluted proteins were first purified by SEC, and the CID pairs were analyzed on an Agilent 1260 Infinity I HPLC system with DAD using an analytical SEC column, superdex 75 increase 3.2/300 GL column. An 4-minute isocratic run was performed for each CID part and CID combinations to identify CID pairs.

Mass photometry test for identification of CID pairs

The mass photometry assays were performed following previously published protocol(*49*). All mass photometry measurements were carried out in an instrument TwoMP (Refeyn) Mass photometer. The same buffer (20 mM Tris 100 mM NaCl, pH 8) was used for all mass photometry experiments. Due to the small size of the CID monomers (A is 9.8 kDa, B is 18.9 kDa), we fused N-terminus gfp to each monomer to increase their molecular weight so that we can observe them under mass photometry. CID pairs were prepared as 10 nM protein gfp-A, gfp-B, with and without 10 μM ligand, and were used for all measurements after 1 hour incubation at r.t. A 24-well gasket was placed on a clean glass slide as sample holders. A 5 μL buffer was first added to one well of this gasket and used to bring the camera into focus after orienting the laser to the center of the sample well. A 5 μL sample prepared as stated-above was

added to this droplet and 1-minute videos were collected with a large field of view in AcquireMP. Ratiometric contrast values for individual particles were measured and processed into mass distributions with DiscoverMP based on calibration curves generated with 20 nM β-amylase (consists of monomer 56 kDa, dimer 112 kDa, and tetramer 224 kDa). The dimerized peak of gfp-A and gfp-B in presence of ligand was clearly observed with a clear right shift compared to the monomer-only peaks and a good mass fitting based on the standard curve. The monomer peak of gfp-A, gfp-B was observed, clearly smaller than the dimerized peak, but cannot be fitted to report mass value due to their small size (gfp-A: 37 kDa, gfp-B: 46 kDa).

Crystallization condition and analysis

Crystallization experiment for the designed protein was conducted using the sitting drop vapor diffusion method. Crystallization trials were set up in 200 nL drops using the 96-well plate format at 20°C. Crystallization plates were set up using a Mosquito LCP from SPT Labtech, then imaged using UVEX microscopes from JAN Scientific. Diffraction quality crystals formed in 0.1 M Tris/Biocine pH 8.5, 25% v/v 2-methyl-2,4-pentanediol (MPD); 25% PEG1000; 25% w/v PEG 3350 and 0.093 MSodium fluoride; 0.3 M Sodium bromide; 0.3 M Sodium iodide for CHD_r1. CHD_r1_buttress crystal formed in 0.2 M Sodium chloride 20% (w/v) PEG 3350.

Diffraction data were collected at Advanced Photon Source on Beamline NECAT 24IDC for CHD_r1 and CHD_r1_buttress. X-ray intensities and data reduction were evaluated and integrated using XDS(*50*) and merged/scaled using Pointless/Aimless in the CCP4 program suite(*51*). Structure determination and refinement starting phases were obtained by molecular replacement using Phaser(*52*) using the designed model structure. Following molecular replacement, the models were improved using phenix.autobuild(*53*). Structures were refined in Phenix(*53*). Model building was performed using COOT(*54*). The final model was evaluated using MolProbity(*55*). Data collection and refinement statistics are recorded in the Supplementary **Table S3**. Data deposition, atomic coordinates, and structure factors reported for the protein in this paper have been deposited in the Protein Data Bank (PDB), http://www.rcsb.org/ with accession code 8VEI and 8VEJ.