Supplementary Information for

Non-proteinogenic deep mutational scanning of linear and cyclic peptides

Joseph M. Rogers, Toby Passioura and Hiroaki Suga*

correspondence to: hsuga@chem.s.u-tokyo.ac.jp

This PDF file includes:

Figs. S1 to S20
Tables S1 to S3
Supporting Methods
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1809901115

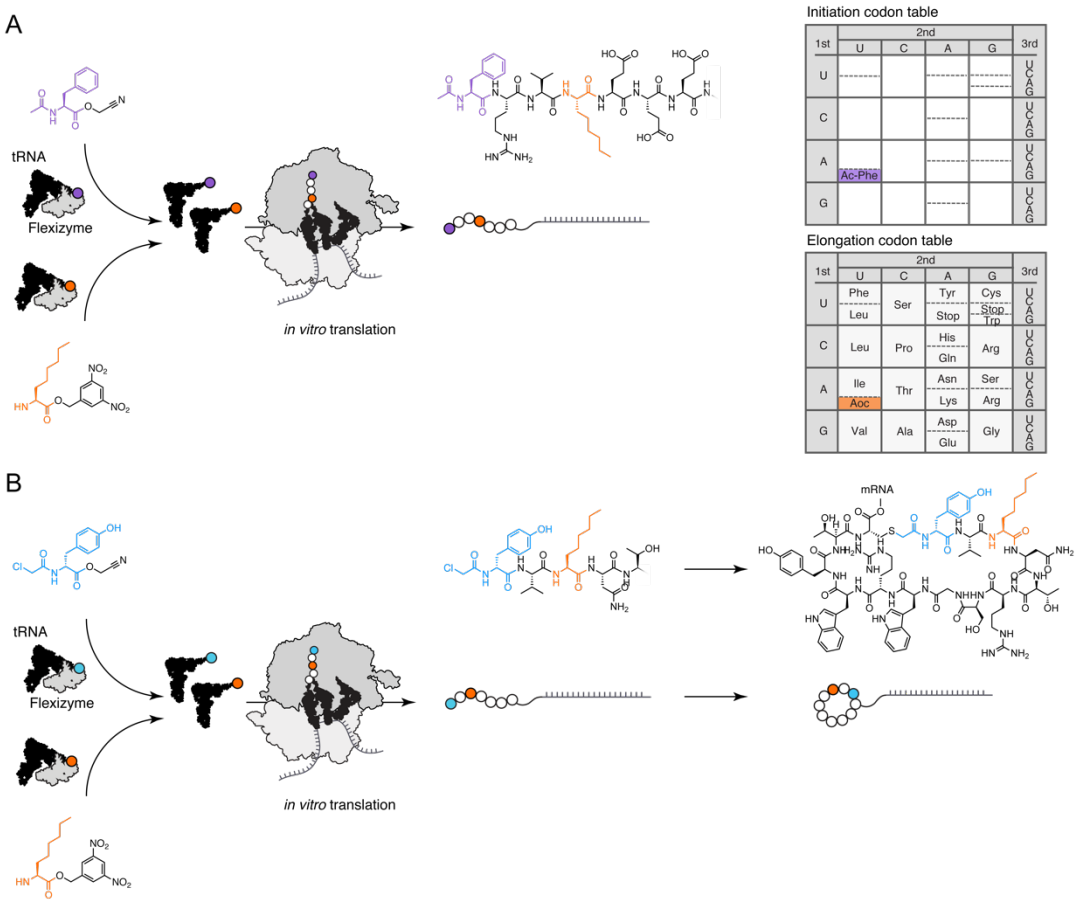


Fig. S1. Genetic code reprogramming using flexizyme to introduce non-proteinogenic amino acids at 'AUG' codons, with different non-proteinogenic amino acids at the initiation codon for CP2 and PUMA. (*A*) For translation of PUMA peptides, *N*-acetyl-$_{L}$-phenylalanine (cyan) was loaded onto $tRNA^{fMet}_{CAU}$ (black) using flexizyme, for use during initiation of peptide synthesis by the ribosome (grey). Additional non-proteinogenic amino acids (orange, e.g. Aoc shown) were loaded onto $tRNA^{EnGlu}_{CAU}$ using flexizyme, and supplied to the ribosome for use when any 'AUG' codons were encountered during peptide elongation. Multiple, separate genetic codes were constructed by supplying a different non-proteinogenic amino acid for elongation (i.e. replacing Aoc), and using the appropriate flexizyme (Table S3). (*B*) For translation of CP2 peptides, *N*-chloroacetyl-$_{D}$-tyrosine was used as the initiation non-proteinogenic amino acid. The chloroacetyl moiety allows for reaction with cysteine in the elongated peptide to produce a macrocycle. As for PUMA, different non-proteinogenic amino acids (Aoc shown) were used for elongation at 'AUG' codons.

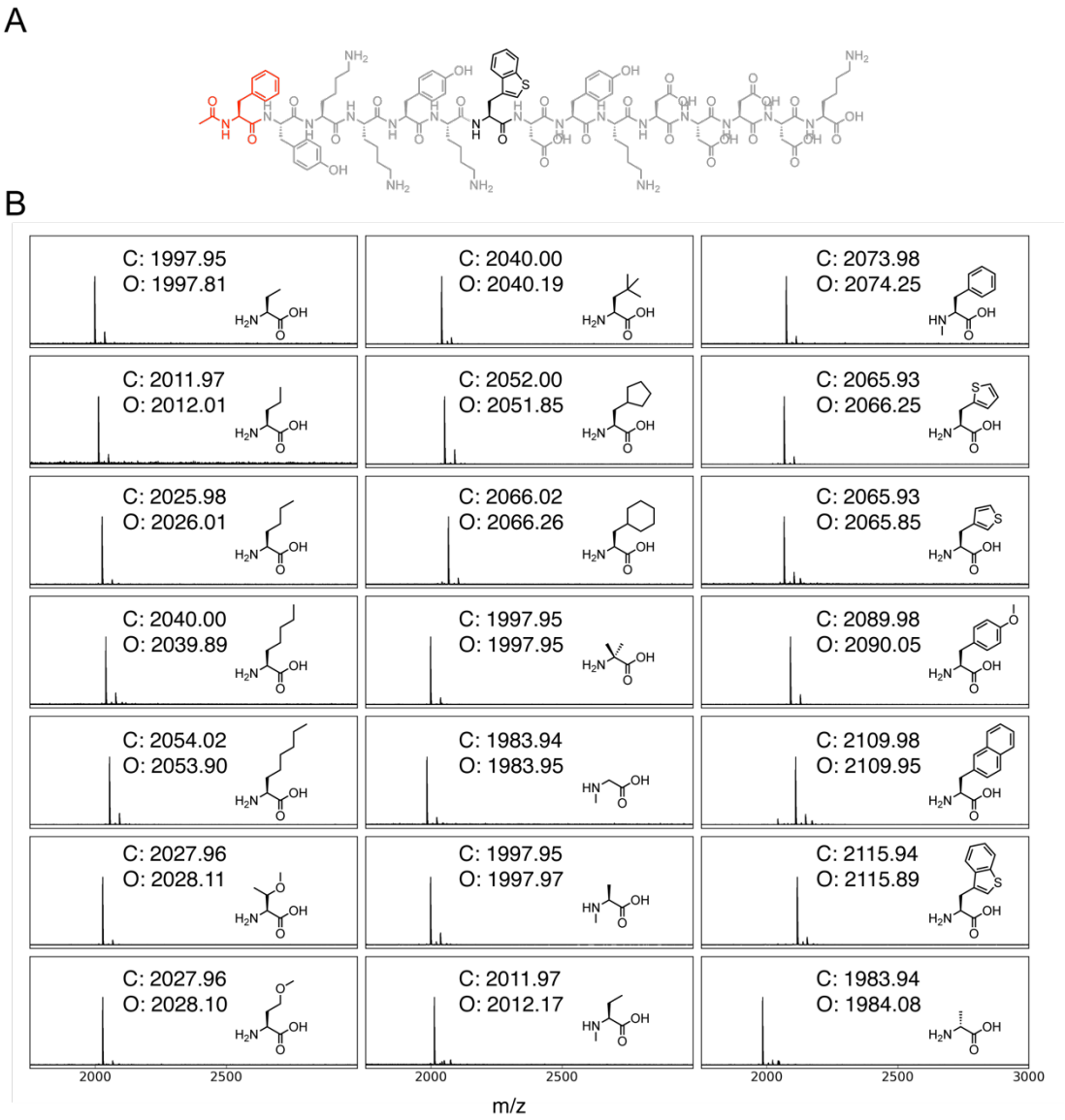


Fig. S2. Test of translation fidelity for the set of non-proteinogenic amino acids, each using a separated reprogrammed genetic code. (*A*) Expected peptide product upon successful translation using a test DNA/mRNA which codes for xYKKYKyDYKDDDDK, where x is the initiator non-proteinogenic amino acid (*N*-acetyl-$_{L}$-phenylalanine, red) and y is the non-proteinogenic amino acid of interest (e.g. Bzt, black, shown). (*B*) MALDI-TOF spectra of translation product for each reprogrammed genetic code with different non-proteinogenic amino acids (y) at the 'AUG' elongation codon. One major peak corresponding to the expected peptide mass was observed in each case, confirming the fidelity of genetic code reprogramming. C: calculated mass+H^+ O: observed mass.

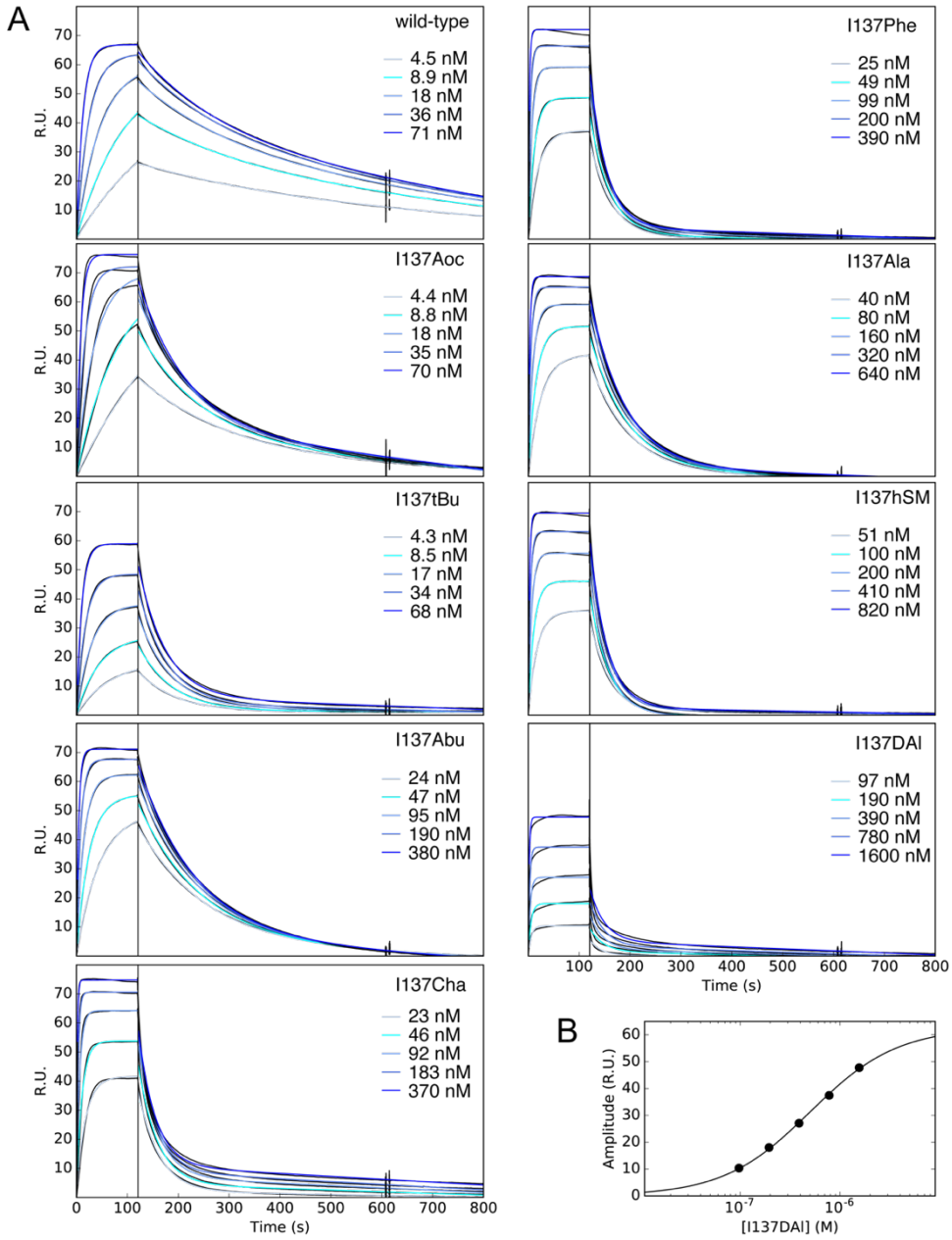


Fig. S3. K_D determination for a sample of proteinogenic and non-proteinogenic PUMA mutants. (*A*) BIACORE SPR association (first 120s) and dissociation (after 120s) kinetic traces for each mutant. Different concentrations of PUMA peptide were flowed over immobilized MCL1. (*B*) I137DAl showed complex kinetic behavior, so the amplitudes from the fit of the association phase were fit to give an apparent K_D.

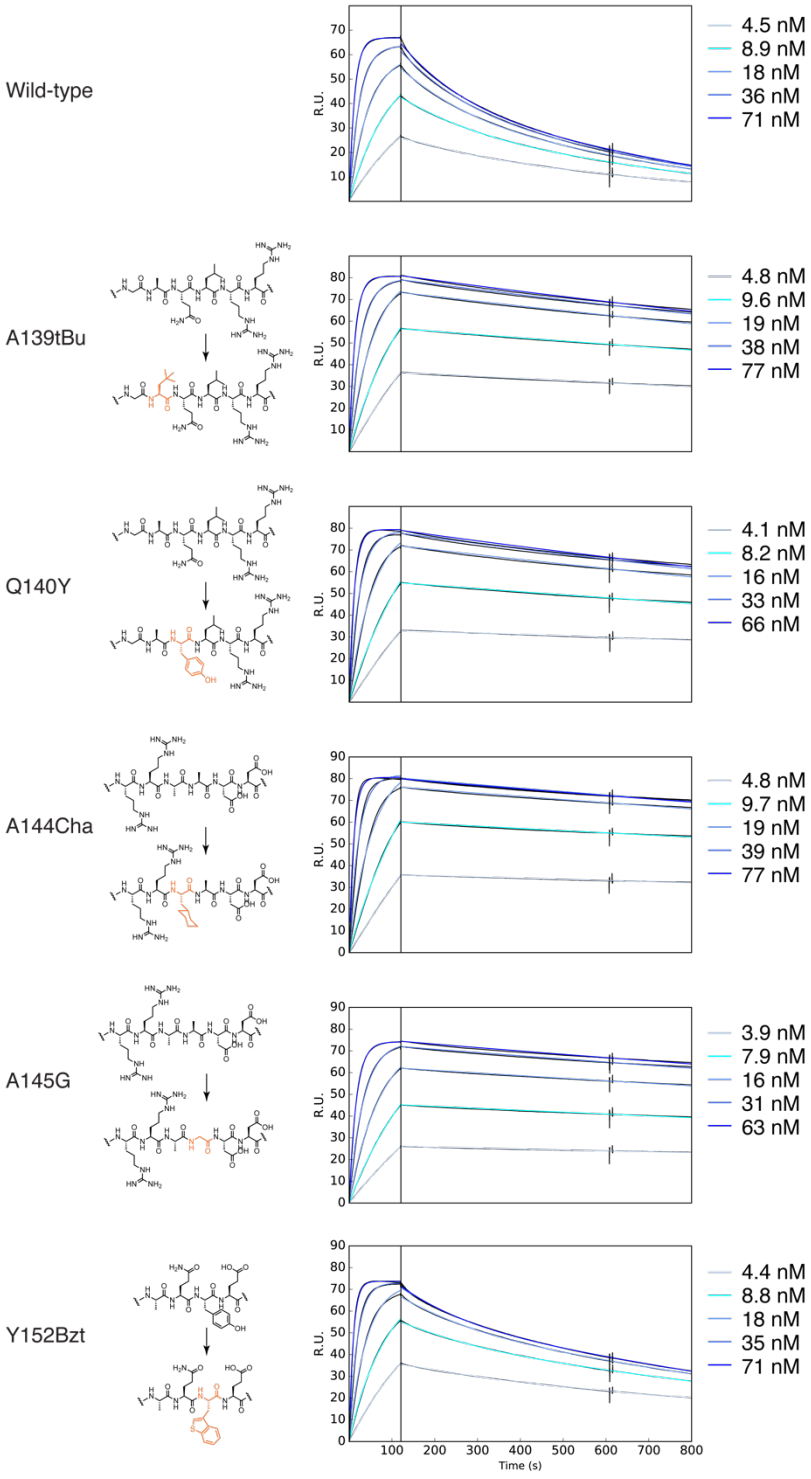
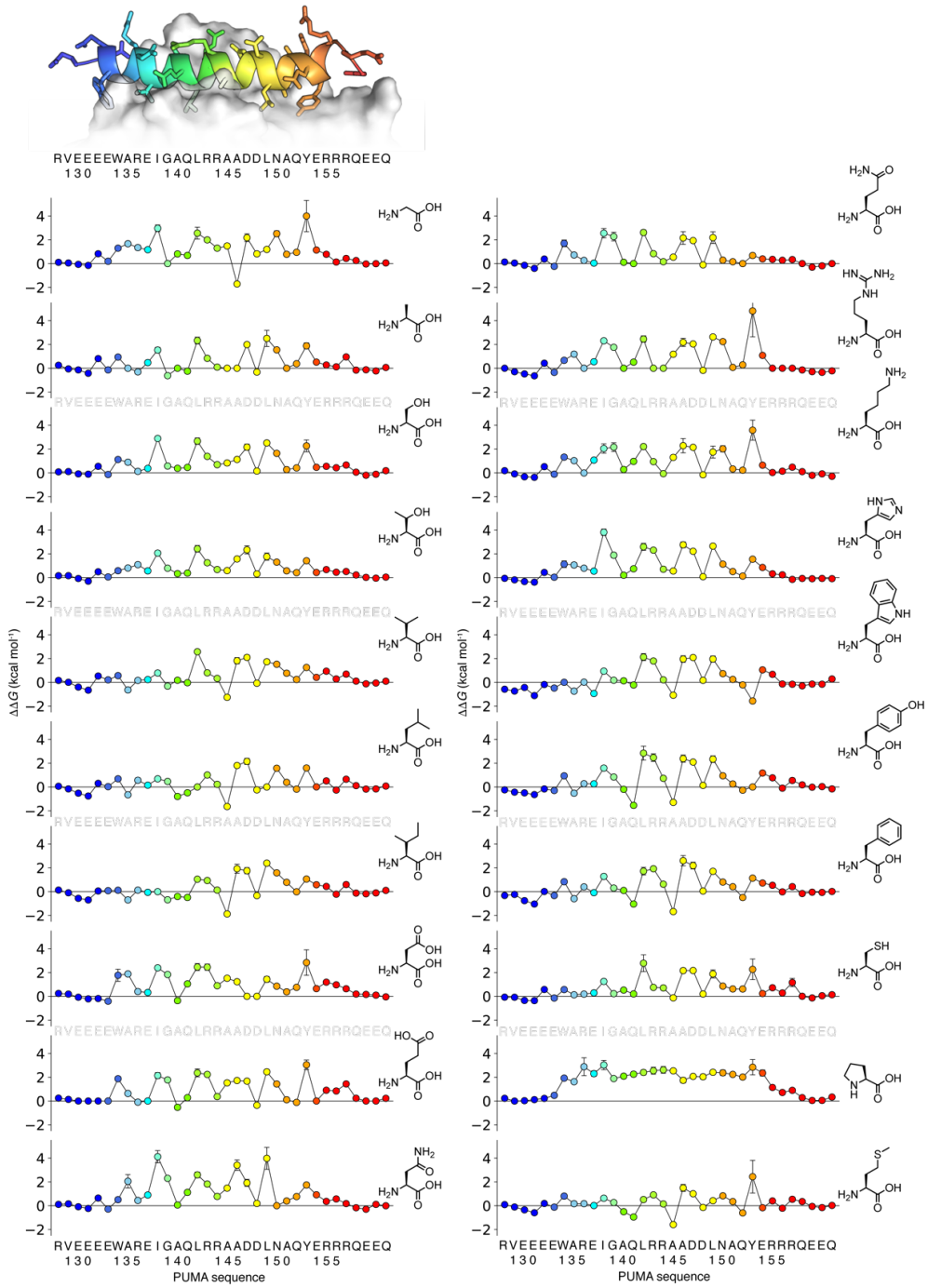


Fig. S4. K_D determination for a sample of PUMA mutants predicted to be stabilizing by deep mutational scanning. BIACORE SPR association (before 120s) and dissociation (after 120s) kinetic traces for each mutant. Different concentrations of PUMA peptide were flowed over immobilized MCL1. Segment of PUMA peptide chemical structure, wild-type and mutant, shown for each case.

A


RVEEEEWARE I GAQLRRAADDLNAQYERRRQEEQ
130 135 140 145 150 155
ΔΔG (kcal mol⁻¹)
PUMA sequence

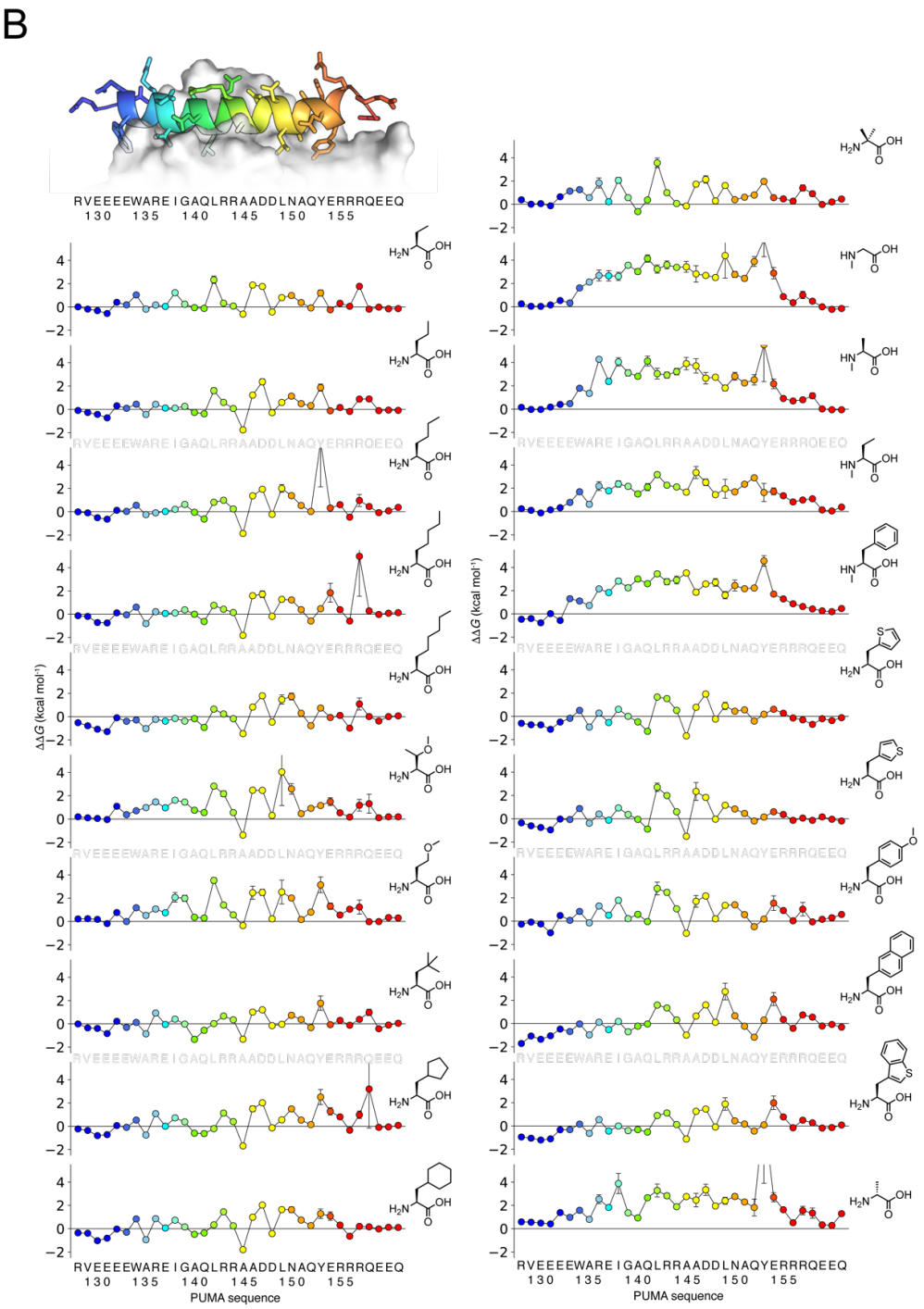
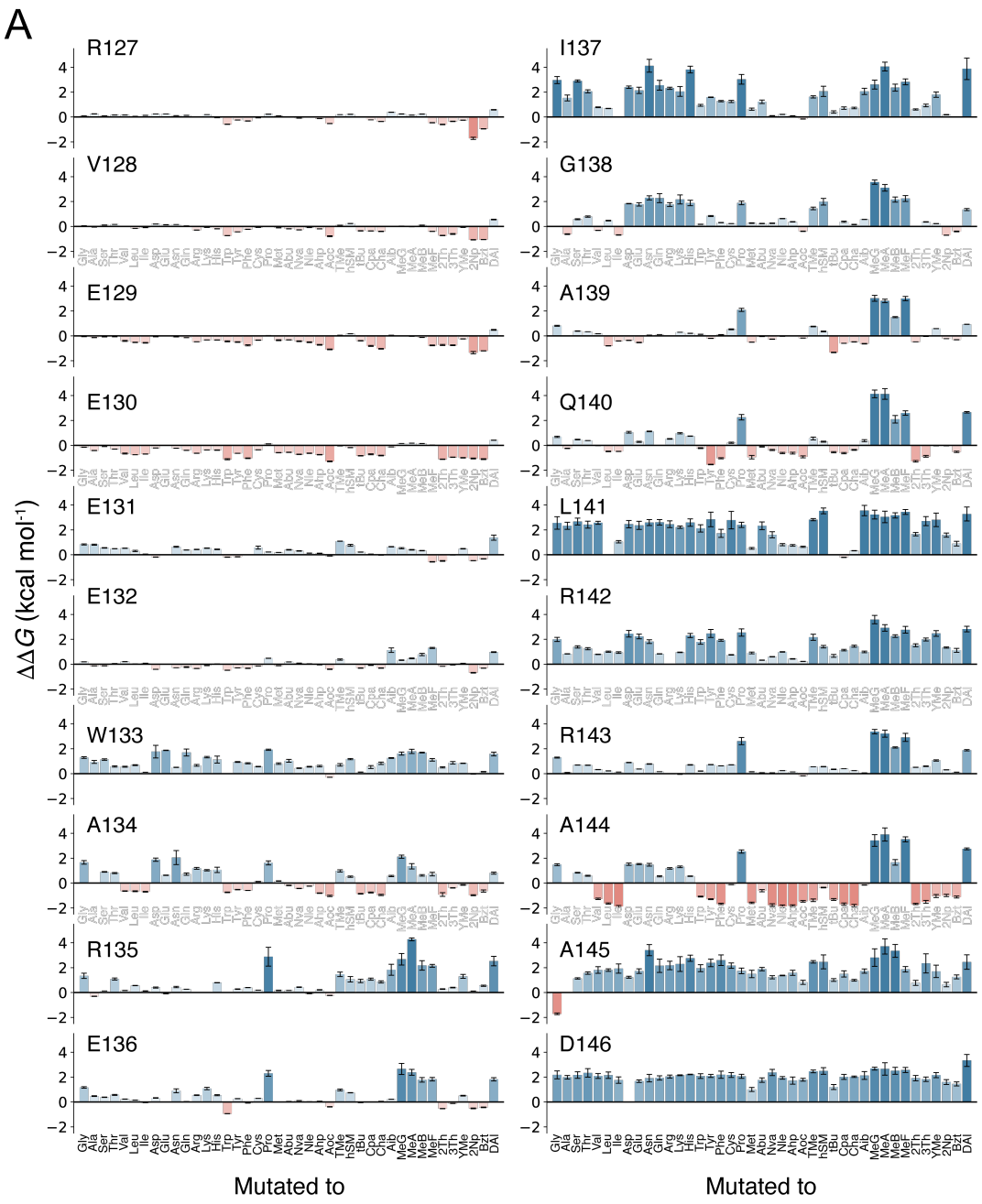


Fig. S5. Mutant scans of ΔΔ*G* across the PUMA sequence, for each proteinogenic and non-proteinogenic amino acid. (*A*) ΔΔ*G* scans for each proteinogenic amino acid mutant across the PUMA sequence. (*B*) ΔΔ*G* scans for each non-proteinogenic amino acid mutant across the PUMA sequence. Positive ΔΔ*G* corresponds to a mutant that destabilizes the interaction with MCL1. Errors in ΔΔ*G* shown propagated from standard deviation of $\log_2 E$, obtained from repeats of binding PUMA library to MCL1.


A
R127
V128
E129
E130
E131
E132
W133
A134
R135
E136
I137
G138
A139
Q140
L141
R142
R143
A144
A145
D146
ΔΔG (kcal mol-1)
Mutated to
Mutated to

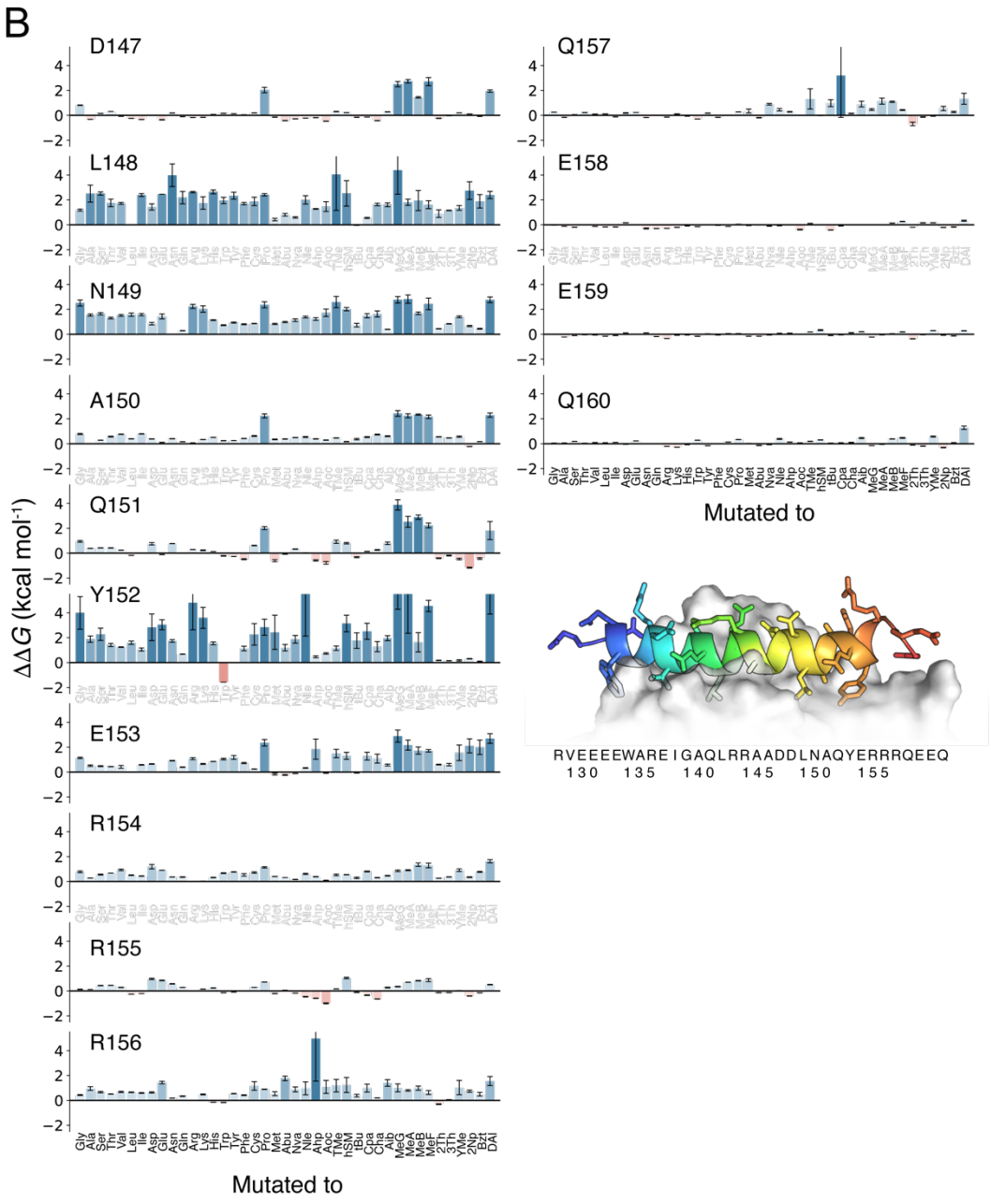


Fig. S6. ΔΔ*G* for each proteinogenic and non-proteinogenic amino acid mutant, at each position in the PUMA sequence. (Data from Fig. S5 transposed). (*A*) Each amino acid tried at PUMA positions R127-D146 (*B*) Each amino acid tried at positions D147-Q160 in PUMA. Positive ΔΔ*G* (blue) corresponds to a mutant that destabilizes the interaction with MCL1, negative ΔΔ*G* (red) corresponds to a mutant that stabilizes the interaction with MCL1. Errors in ΔΔ*G* shown propagated from $\log_2 E$ standard deviations, calculated from repeats of binding PUMA mutant library to MCL1.

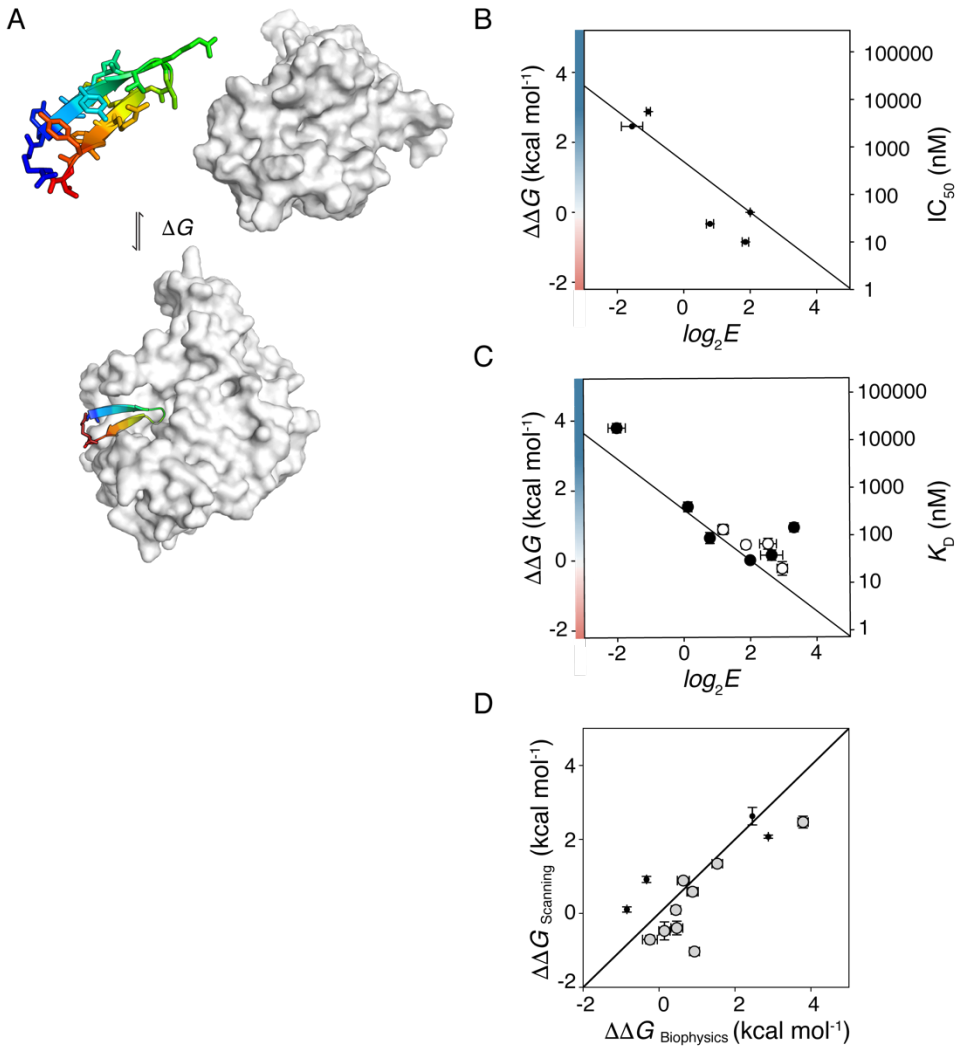


Fig. S7. Cyclic peptide CP2 binding to KDM4A and deep mutational scanning of CP2. (*A*) CP2 cyclic peptide (rainbow) can bind and inhibit the enzyme KDM4A (white, surface). CP2 forms a small β-sheet upon binding KDM4A. (*B*) $\log_2 E$ from deep mutational scanning correlate with effective $\Delta\Delta G$ values calculated using previously reported IC_{50} values as effective K_Ds (26). (*C*) A collection of CP2 mutants was synthesized (Fig. S8) and tested for binding to KDM4A. Proteinogenic (black) and non-proteinogenic (white) mutants overlay. Linear fit of $\Delta\Delta G$ and $\log_2 E$ were used to calibrate the deep mutational scanning data and calculate $\Delta\Delta G$ for all CP2 mutants. Linear fits in (*B*) and (*C*) produce similar gradients (-0.723 vs. -0.729). (*D*) Leave-one-out cross validation of empirical fit of $\Delta\Delta G$ and $\log_2 E$, showing agreement between calculated $\Delta\Delta G$ of the left out mutant ($\Delta\Delta G_{Scanning}$) and expected experimental value ($\Delta\Delta G_{Biophysics}$).

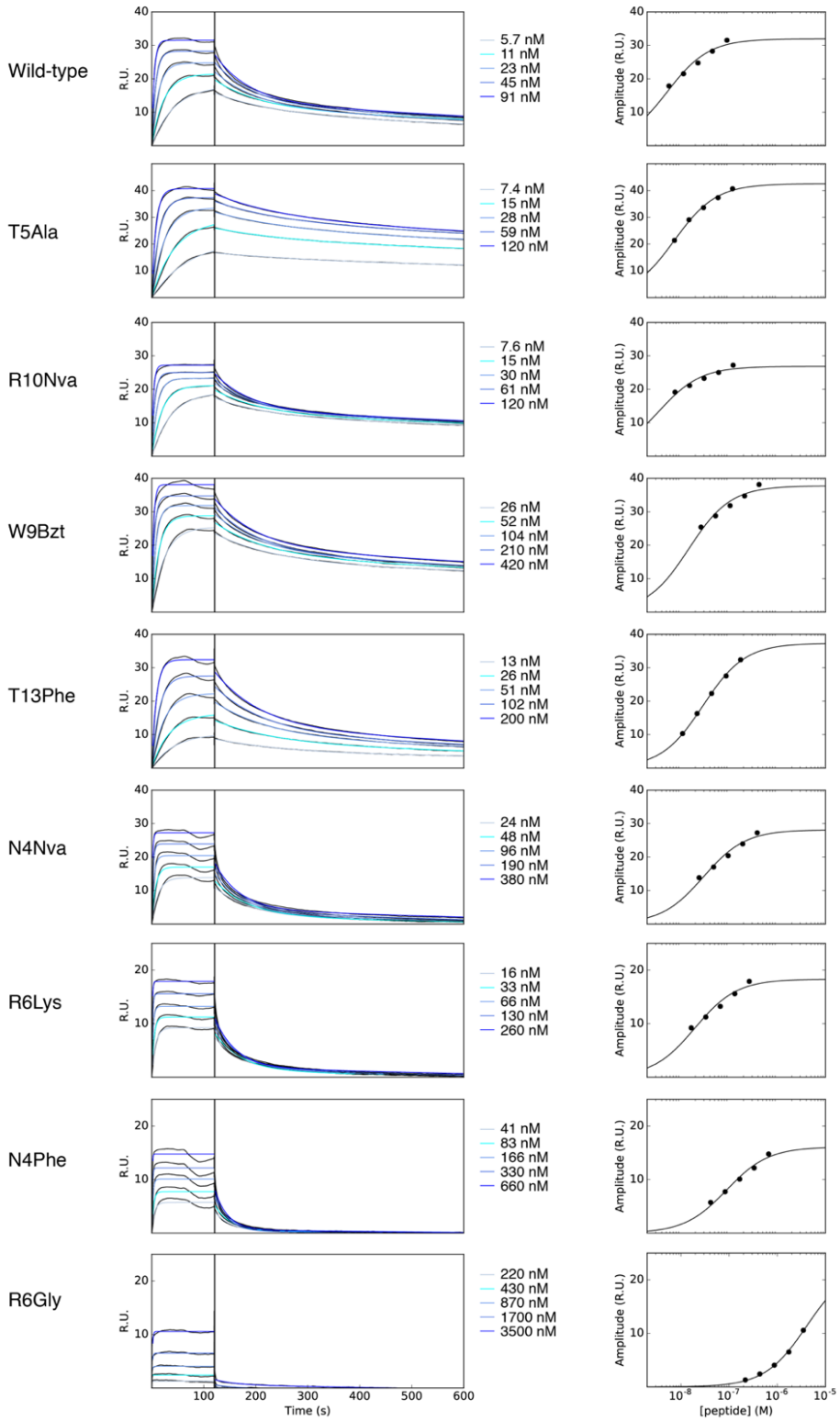


Fig. S8. **BIACORE SPR K_D determination for CP2 mutants.** Peptides were prepared using solid phase peptide synthesis (SPPS). Amplitudes for the association phase (right) where fit to calculate K_D for each mutant.

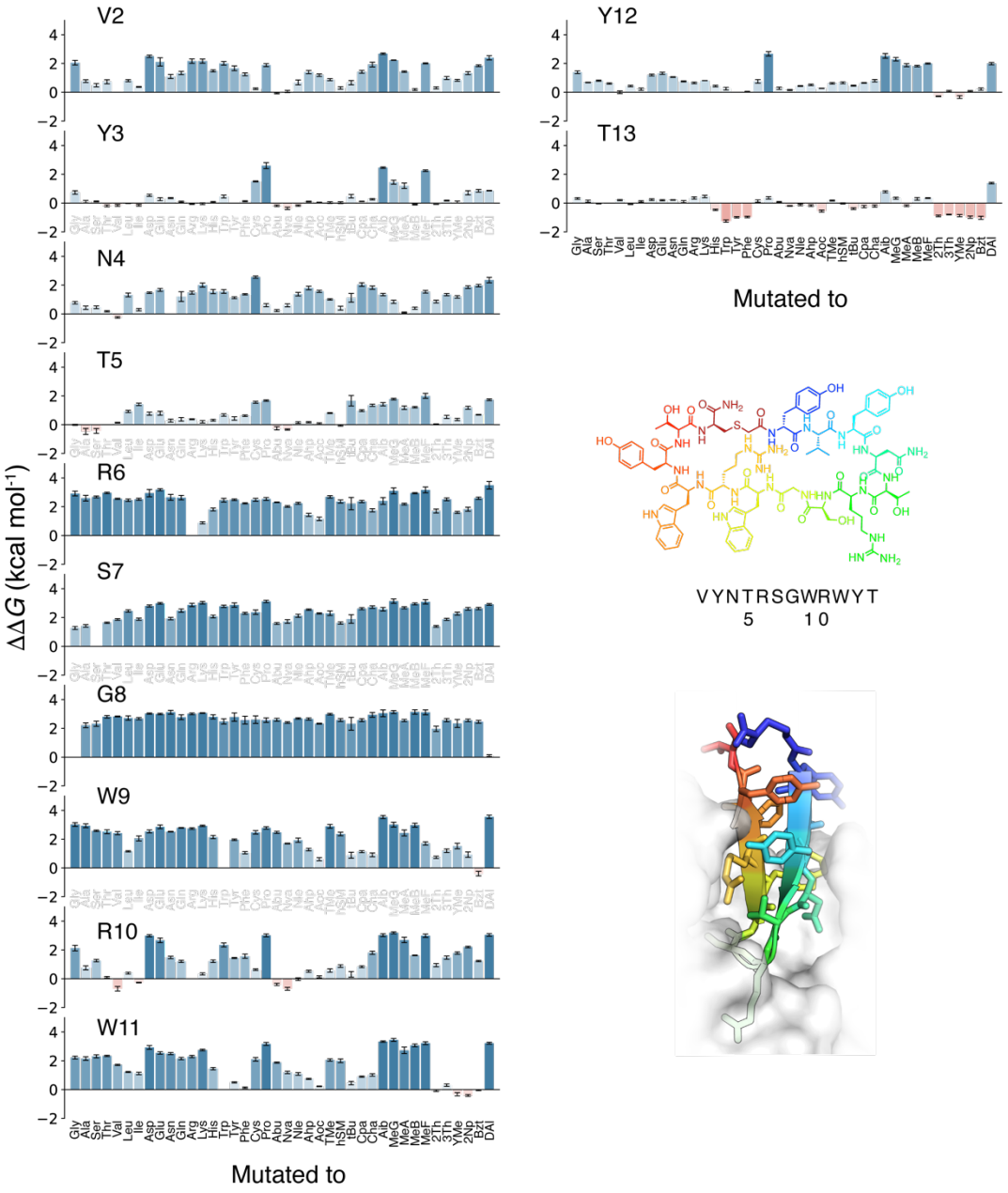
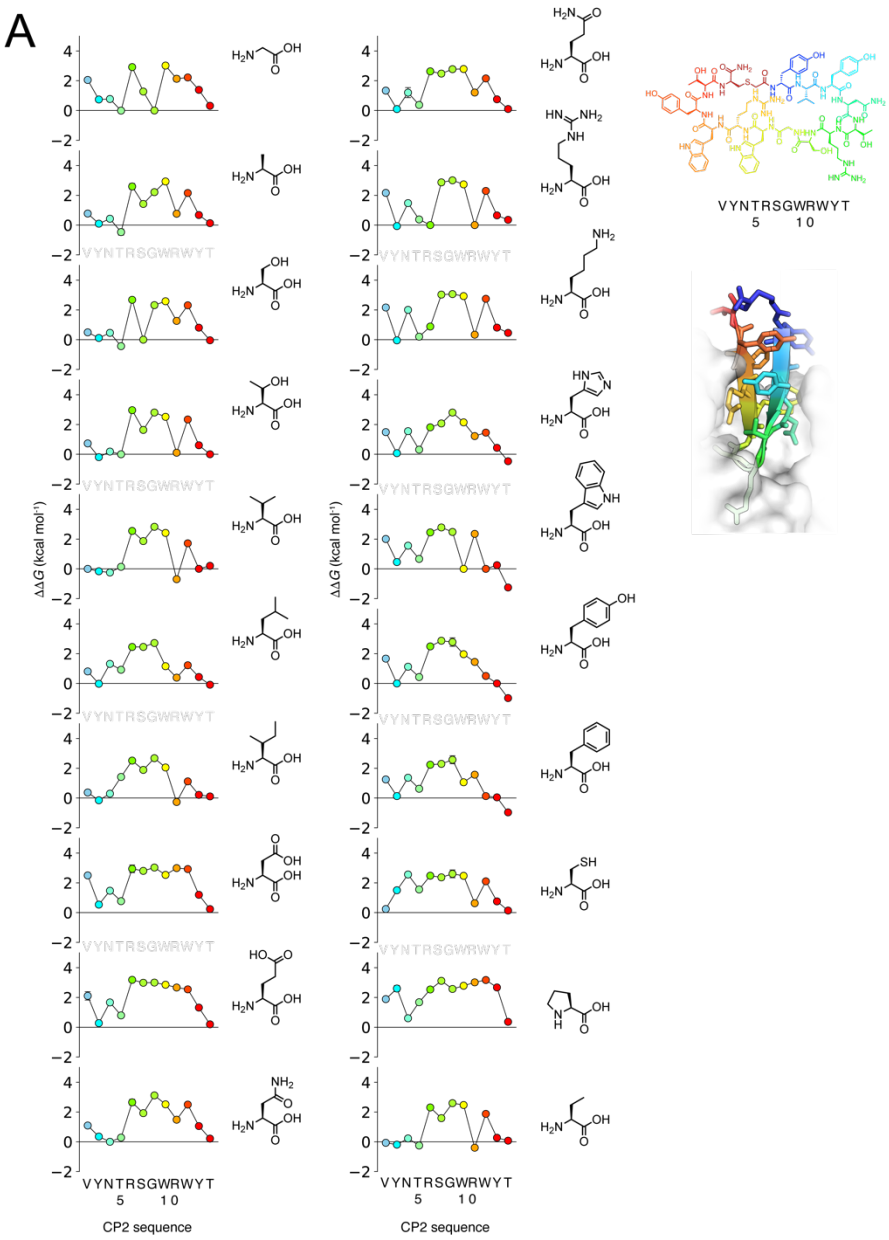


Fig. S9. ΔΔ*G* for each proteinogenic and non-proteinogenic amino acid mutant, at each position in the CP2 sequence. Each amino acid tried at all CP2 positions. Positive ΔΔ*G* (blue) corresponds to a mutant that destabilizes the interaction with KDM4A, negative ΔΔ*G* (red) corresponds to a mutant that stabilizes the interaction with KDM4A. N.B. Methionine does not appear in these scans, as this proteinogenic amino acid was replaced with a non-proteinogenic amino acid required for cyclization (*N*-chloroacetyl-D-tyrosine). Errors in ΔΔ*G* shown propagated from standard deviation of $\log_2 E$, obtained from repeats of binding CP2 library to KDM4A.


A
ΔΔG (kcal mol⁻¹)
VYNTRSGWRWYT
5 10
CP2 sequence
VYNTRSGWRWYT
5 10
CP2 sequence
VYNTRSGWRWYT
5 10

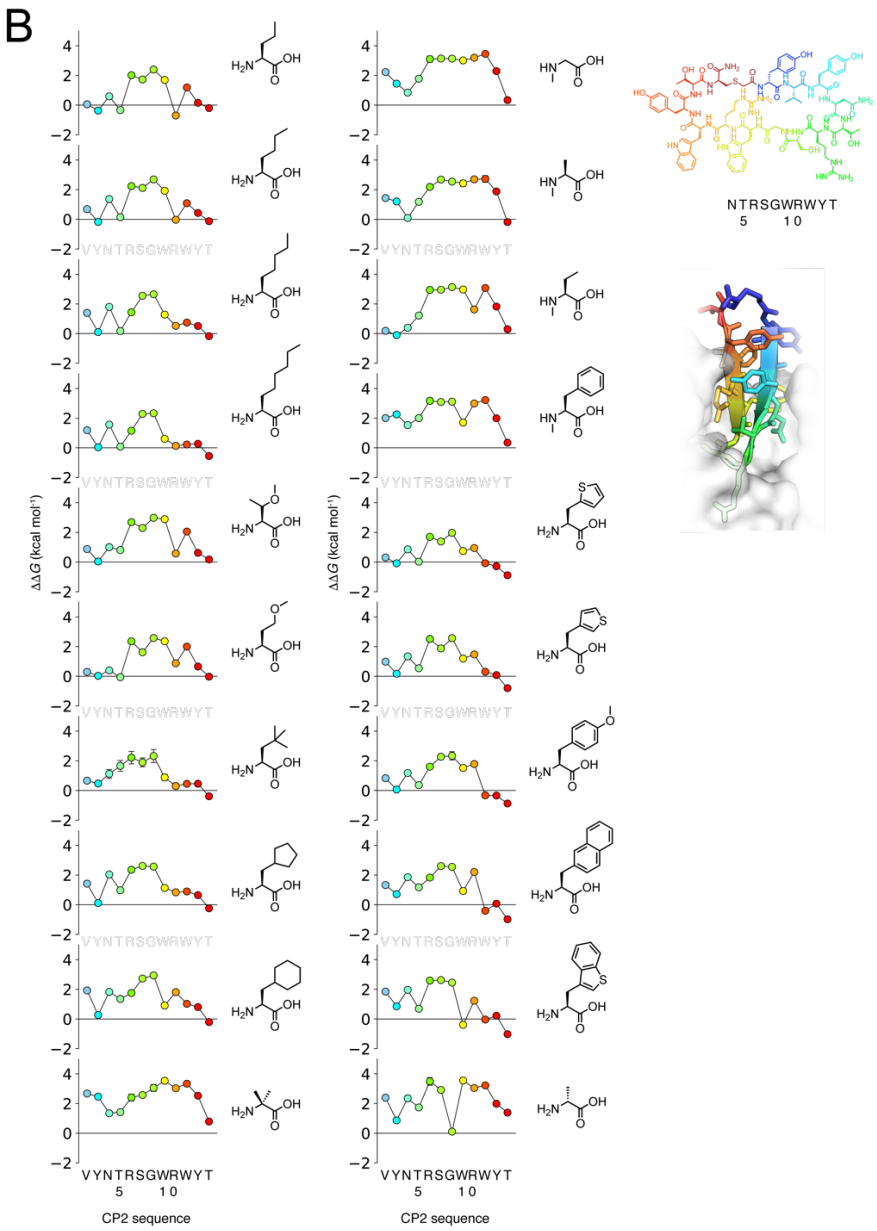


Fig. S10. Mutant scans of ΔΔ*G* across the CP2 sequence, for each proteinogenic and non-proteinogenic amino acid. (Data from Fig. S9 transposed). (*A*) ΔΔ*G* scans for each proteinogenic amino acid mutant across the CP2 sequence. (*B*) ΔΔ*G* scans for each non-proteinogenic amino acid mutant across the CP2 sequence. N.B. Methionine does not appear in these scans, as this proteinogenic amino acid was replaced with a non-proteinogenic amino acid required for cyclization (*N*-chloroacetyl-$_D$-tyrosine). Positive ΔΔ*G* corresponds to a mutant that destabilizes the interaction with KDM4A. Errors in ΔΔ*G* shown propagated from standard deviation of $\log_2 E$, obtained from repeats of binding CP2 library to KDM4A.

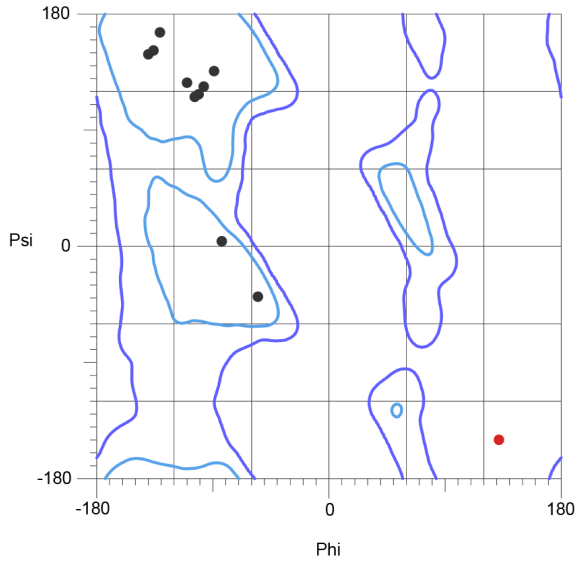


Fig. S11. Ramachandran plot for CP2 torsional angles when bound to KDM4A. Gly8 shown in red is in a disallowed area for L-amino acids but is in an acceptable area for D-amino acids. All other CP2 amino acid phi-psi angles show in black. From pdb 5LY1

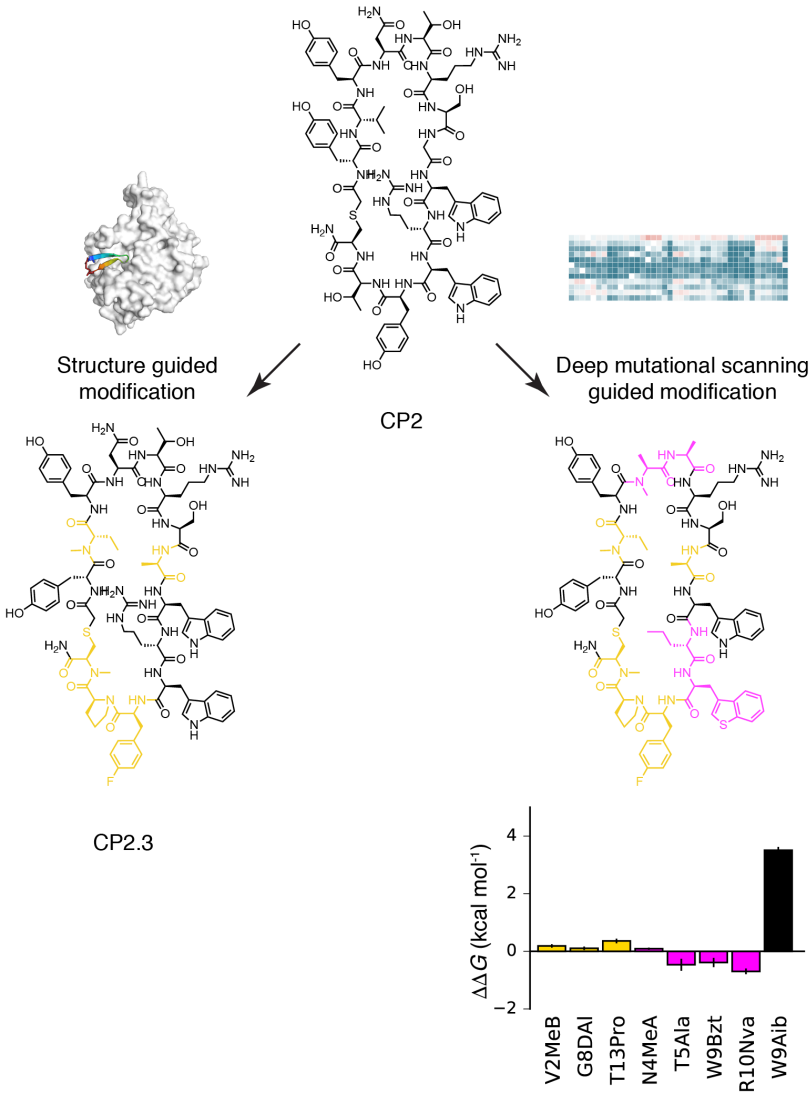


Fig. S12. Deep mutational scanning recapitulates optimization of CP2 by structure-guided design and suggests further modifications. In a previous study (26) the solved structure of CP2 bound to its partner KDM4A was used to mutate CP2 for improved *in vivo* activity, reducing polarity for membrane permeability and including non-proteinogenic amino acids to protect against degradation, resulting in the peptide CP2.3 (mutations shown in yellow). Deep mutational scanning also predicted that these CP2.3 mutations (yellow) are tolerated without affecting binding (using MeB as the closest amino acid to *N*-methyl valine used in CP2.3). However, deep mutational scanning suggests additional modifications at different positions in the CP2 peptide. For example, the mutants N4MeA, T5Ala, W9Bzt, R10Nva (purple) are tolerated in isolation, with low values of $\Delta\Delta G$. One destabilizing, positive $\Delta\Delta G$, mutant W9Aib (black) is shown for comparison. Mutants N4MeA, T5Ala, W9Bzt, R10Nva are also accepted in combination (Fig. 4*B*).

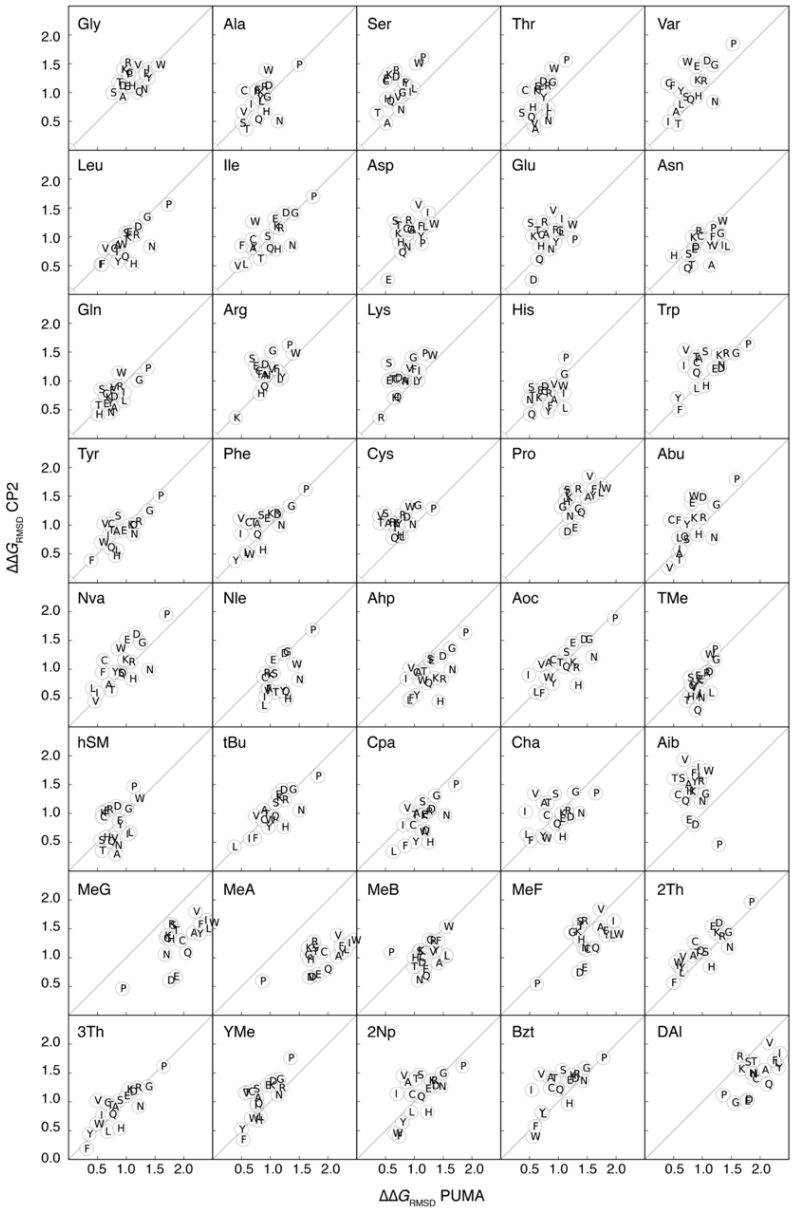


Fig. S13. Similarity of each amino acid to the proteinogenic amino acids in the α-helical PUMA and β-sheet CP2, assessed using deep mutational scanning ΔΔ*G*. ΔΔ*G* values for pairs of amino acids were compared using root mean squared deviation in ΔΔ*G* (ΔΔG_{RMSD}) over all positions in each peptide sequence. For Arg, ΔΔG_{RMSD} is low for K (Lys) in both PUMA and CP2. This is expected, given that they are both positively charged amino acids. Similarly, for Asp, ΔΔG_{RMSD} is low for the other negatively charged amino acid E (Glu). The non-proteinogenic amino acid Bzt has relatively low ΔΔG_{RMSD} values with W (Trp) - these two amino acids only differ by a functional group substitution. In contract, DAl does not show any low ΔΔG_{RMSD}, mutations to this $_D$-amino acid do not resemble mutations to any of the proteinogenic amino acids.

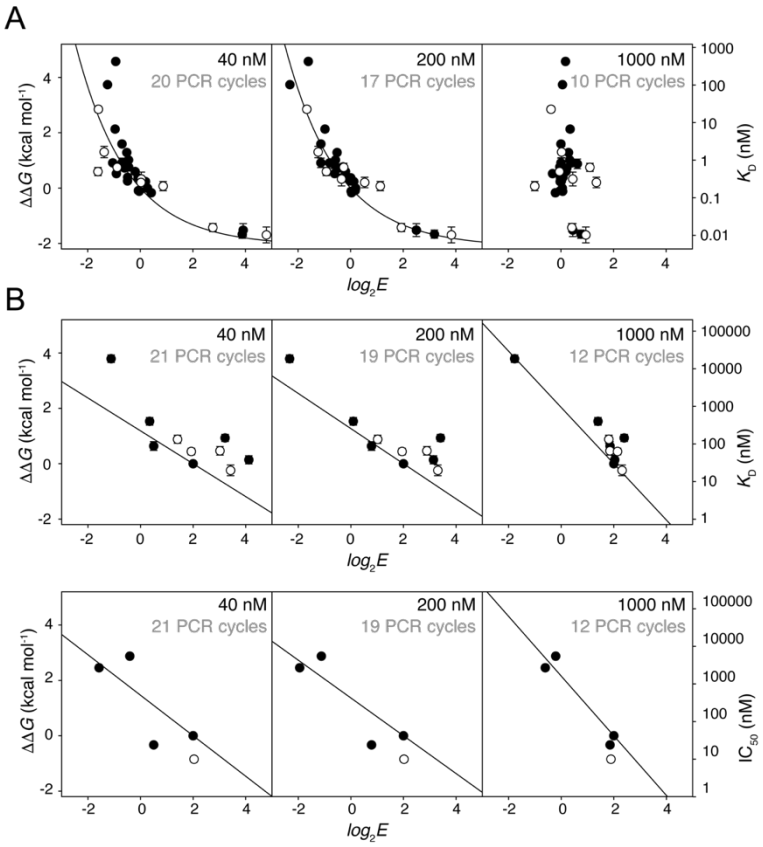


Fig. S14. Effect of partner protein concentration on enrichment values.
Preliminary tests revealed that using higher concentration of partner protein resulted in poorer discrimination between the binding ability of mutants, producing a narrower range of $\log_2 E$. Here, $\log_2 E$ values shown where experimental $\Delta\Delta G$ (or IC_{50}) were available, for PUMA mutants binding partner MCL1 at concentrations of 40, 200 and 1000 nM (*A*) and CP2 mutants binding partner KDM4A at concentrations of 40, 200 and 1000 nM (*B*). Proteinogenic mutants shown as black circles, non-proteinogenic mutants shown as white circles. Lowering partner protein concentration increased the range of $\log_2 E$ values but also reduced the amount of recovered DNA. 200 nM partner protein concentration was chosen for its broad $\log_2 E$ while minimizing PCR cycles (to obtain sufficient DNA concentration for sequencing). The concentration of input cDNA-peptide mutant library was approximately 200 nM in each case, with an unknown concentration of peptides without cDNA.

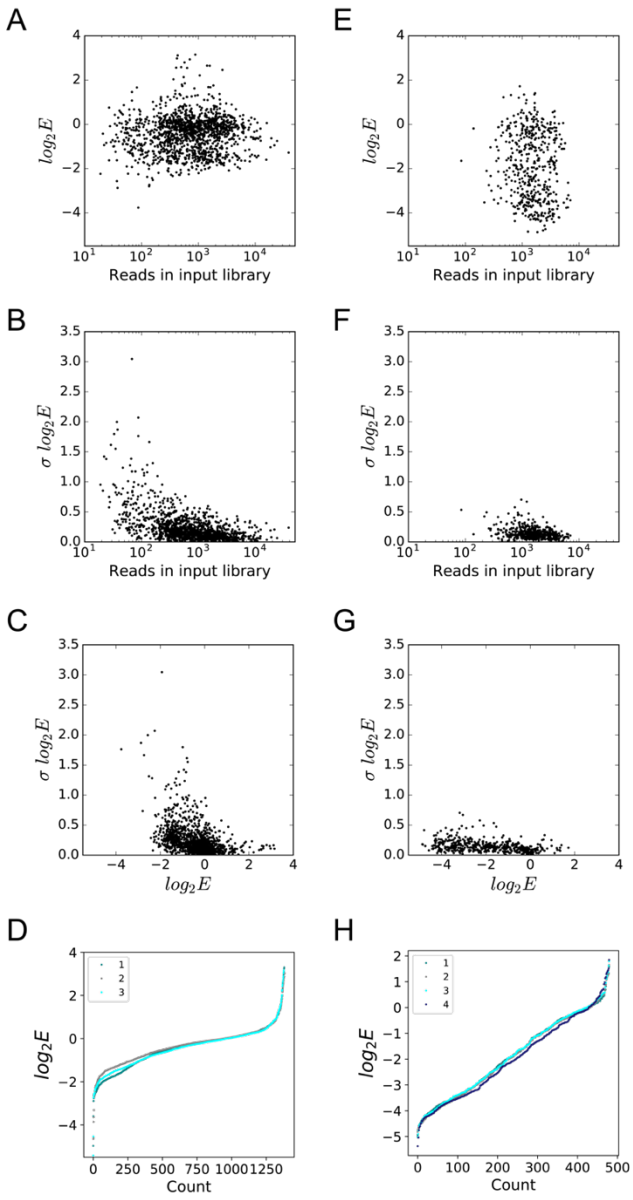


Fig. S15. Error analysis of enrichment scores (*E*). (*A*) $\log_2 E$ for PUMA mutants do not correlate with the reads in the input library. (*B*) Standard deviation for the enrichment scores ($\sigma\log_2 E$) for three replicates is larger for those PUMA mutants with fewer reads in the input library. (*C*) $\sigma\log_2 E$ is also larger for highly destabilizing mutants (lower number of reads in the output, after binding DNA library). (*D*) Count of reads with at least $\log_2 E$, for repeats of binding PUMA library to MCL1 and DNA recovery, showing systematic differences in $\log_2 E$. (*E*) $\log_2 E$ for CP2 mutants did not correlate with the reads from the input library. Unlike PUMA, $\sigma\log_2 E$ do not correlate with reads in the input library (*F*) or enrichment score of each mutant (*G*). This is likely due to the fact that more reads were collected per CP2 mutant (can be seen by comparing (*A*) with (*E*)). (*H*) Systematic differences in $\log_2 E$ were observed between repeats of the CP2 library binding to KDM4A and DNA recovery.

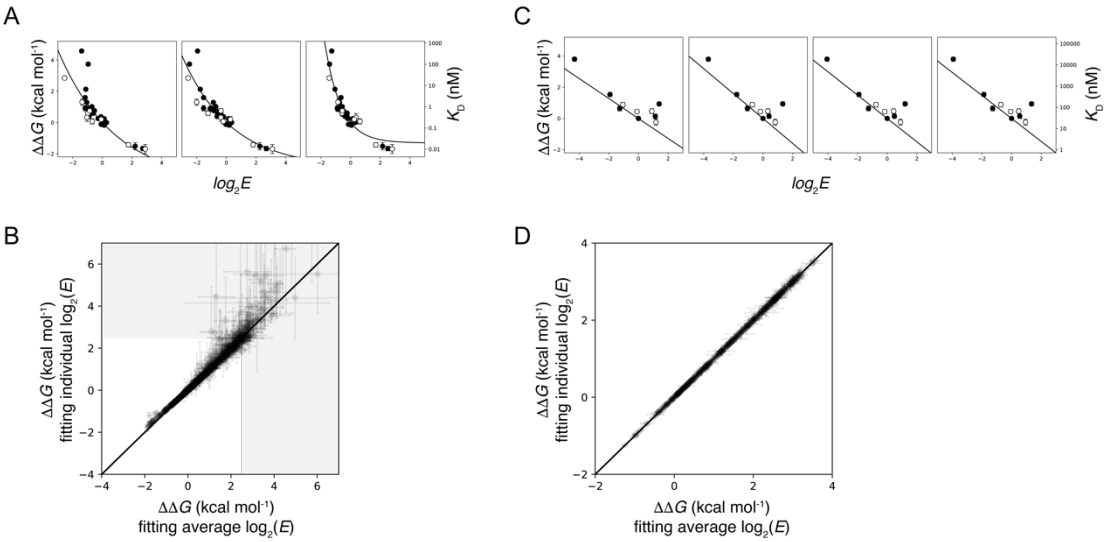


Fig. S16. Alternative calculation of ΔΔ*G* from enrichment scores (*E*). (*A*) Fitting of $\log_2 E$ for each individual repeat of binding the PUMA library to MCL1. (*B*) Average ΔΔ*G* from individual $\log_2 E$, with standard errors, produces similar results to fitting average $\log_2 E$ described in the main text. >90% of PUMA mutants fall within the white region. (*C*) Fitting of $\log_2 E$ for each individual repeat of binding the CP2 library to KDM4A. (*D*) Average ΔΔ*G* from individual $\log_2 E$, with standard errors, produces similar results to fitting average $\log_2 E$, used in the main text.

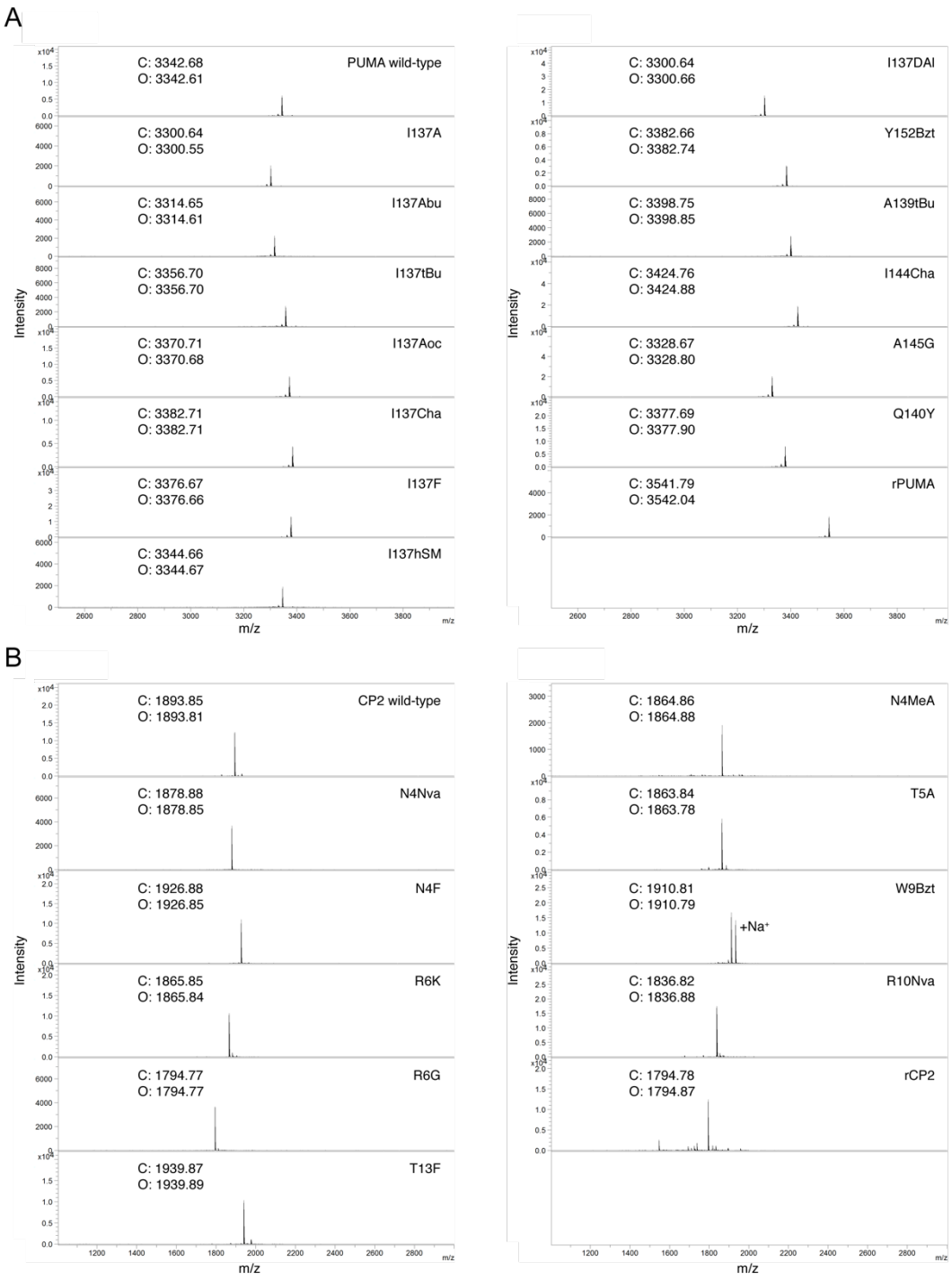


Fig. S17. Identity of synthesized peptides. MALDI-TOF-MS of SPPS, HPLC purified (*A*) PUMA (27 amino acid long) wild-type and mutant peptides, (*B*) CP2 wild-type and mutant peptides. C: calculated mass+H^+ O: observed mass.

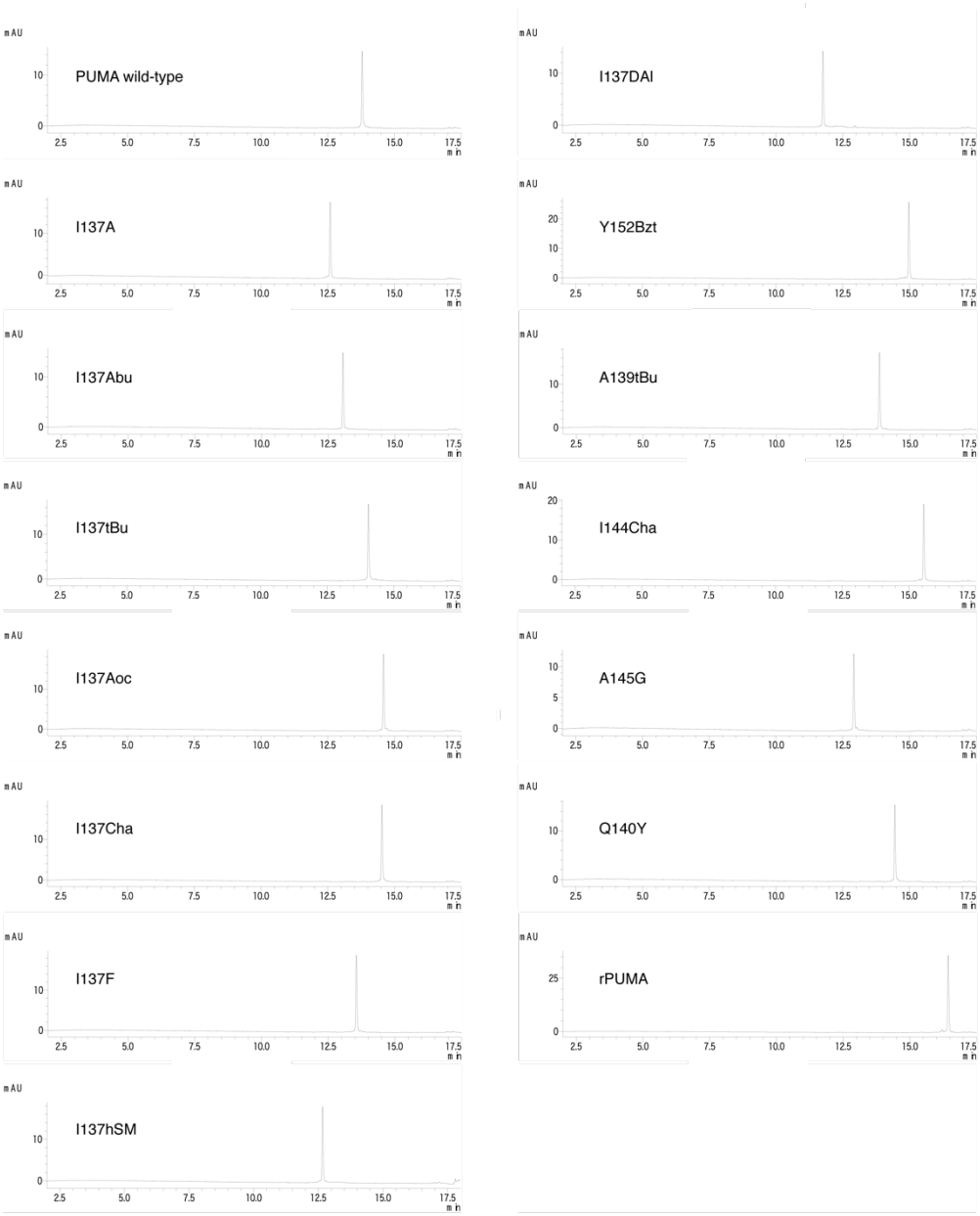


Fig. S18. Purity of synthesized PUMA peptides. UHPLC traces of SPPS made PUMA peptides.

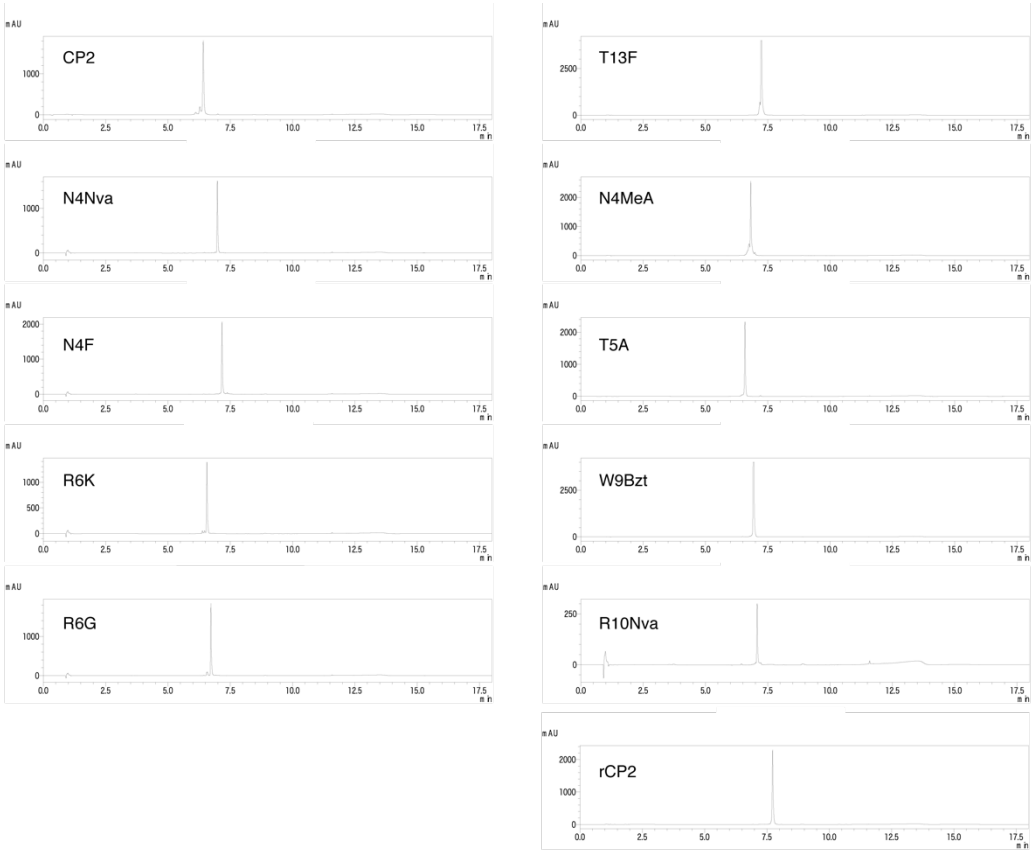


Fig. S19. Purity of synthesized CP2 peptides. UHPLC traces of SPPS made CP2 peptides.

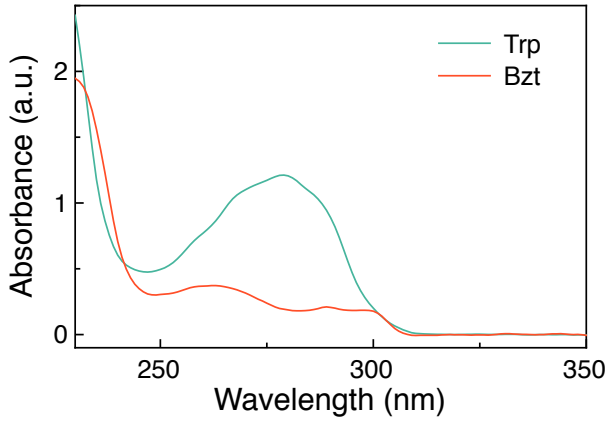


Fig. S20. Absorbance spectra for Trp and Bzt containing peptides. Spectra taken with the same mass per volume of Ac-LNAQ(Bzt)AGSS-NH_2 and Ac-LNAQ(Trp)AGSS-NH_2.

Table S1. BIACORE SPR K_D and $\Delta\Delta G$ values for PUMA mutants binding MCL1 for the sample of PUMA mutants made by SPPS. *indicated mutants combined to make the redesigned PUMA, rPUMA.

	k_{on} (M^{-1} s^{-1})		k_{off} (s^{-1})		K_D (nM)		$\Delta\Delta G$ (kcal mol^{-1})	
PUMA	1.1 x 10^6	± 0.14	0.0043	± 0.0005	4.0	± 0.7	0	± 0.14
I137A	7.3 x 10^5	± 1.7	0.0138	± 0.0005	19	± 4	0.91	± 0.17
I137Abu	7.5 x 10^5	± 1.3	0.0083	± 0.0005	11	± 2	0.60	± 0.15
I137tBu	2.3 x 10^6	± 0.9	0.0163	± 0.0019	7.0	± 2.7	0.33	± 0.25
I137Aoc	1.5 x 10^6	± 0.3	0.0087	± 0.0019	5.7	± 1.6	0.20	± 0.19
I137Cha	1.9 x 10^6	± 0.3	0.0279	± 0.0019	14.3	± 2.6	0.75	± 0.15
I137F	1.5 x 10^6	± 0.4	0.0275	± 0.0022	17.7	± 4.3	0.88	± 0.18
I137hSM	7.7 x 10^5	± 2.3	0.0282	± 0.0017	37	± 11	1.30	± 0.20
I137DAl					500	± 30	2.85	± 0.10
Y152Bzt*	1.1 x 10^6	± 0.1	0.0048	± 0.0010	4.5	± 11	0.07	± 0.17
A139tBu*	8.6 x 10^5	± 1.4	0.00031	± 0.00002	0.36	± 0.06	-1.43	± 0.15
A144Cha*	8.4 x 10^5	± 3.7	0.00019	± 0.00002	0.23	± 0.10	-1.69	± 0.29
A145G*	8.4 x 10^5	± 1.3	0.00020	± 0.00002	0.24	± 0.06	-1.67	± 0.15
Q140Y*	9.9 x 10^5	± 3.1	0.00030	± 0.00005	0.31	± 0.11	-1.52	± 0.23
rPUMA	6.0 x 10^5	± 1.0	< 0.00009		< 0.16		< -1.91	

Table S2. BIACORE SPR K_D and $\Delta\Delta G$ values for CP2 mutants binding KDM4A for the sample of CP2 mutants made by SPPS. *indicated mutants combined to make the redesigned CP2, rCP2.

	K_d (nM)		$\Delta\Delta G$ (kcal mol^{-1})	
CP2	6.6	± 1.1	0	
N4Nva	29	± 5	0.88	± 0.15
N4F	88	± 15	1.53	± 0.14
R6K	19	± 4	0.64	± 0.16
R6G	4100	± 600	3.79	± 0.14
T13F	32	± 4	0.93	± 0.13
N4MeA*	13.9	± 1.5	0.44	± 0.12
T5A*	8.4	± 1.6	0.14	± 0.15
W9Bzt*	15	± 3	0.47	± 0.16
R10Nva*	4.4	± 1.2	-0.24	± 0.19
rCP2	7.0	± 0.6	0.03	± 0.11

Table S3. Non-proteinogenic amino acids, their ester activation for flexizyme, appropriate flexizyme and aminoacylation times. All aminoacylation reactions carried out at 0 °C.

Three letter code	Amino acid	Activation ester	Flexizyme	Aminoacylation time (hours)
Abu	$_{L}$-2-aminobutyric acid	DBE	dFx	6
Nva	$_{L}$-norvaline	DBE	dFx	2
Nle	$_{L}$-norisoleucine	DBE	dFx	2
Ahp	$_{L}$-2-aminoheptanoic acid	CME	eFx	10
Aoc	$_{L}$-2-aminooctanoic acid	CME	eFx	6
TMe	*O*-methyl-$_{L}$-threonine	DBE	dFx	6
hSM	*O*-methyl-$_{L}$-homoserine	DBE	dFx	6
tBu	$_{L}$-γ-methyl-leucine	DBE	dFx	6
Cpa	cyclopentyl-$_{L}$-alanine	DBE	dFx	6
Cha	cyclohexyl-$_{L}$-alanine	DBE	dFx	6
Aib	2-aminoisobutyric acid	DBE	dFx	2
MeG	*N*-methyl-$_{L}$-glycine	DBE	dFx	2
MeA	*N*-methyl-$_{L}$-alanine	DBE	dFx	2
MeB	*N*-methyl-$_{L}$-2-aminobutyric acid	DBE	dFx	6
MeF	*N*-methy-$_{L}$-phenylalanine	CME	eFx	6
2Th	(2-thienyl)-$_{L}$-alanine	CME	eFx	10
3Th	(3-thienyl)-$_{L}$-alanine	CME	eFx	10
YMe	O-methyl-$_{L}$-tyrosine	CME	eFx	2
2Np	(2-Naphthyl)-$_{L}$-alanine	CME	eFx	2
Bzt	(3-Benzothienyl)-$_{L}$-alanine	CME	eFx	6
DAl	$_{D}$-alanine	DBE	dFx	2
	N-chloroacetyl-$_{D}$-tyrosine	CME	eFx	2
	N-acetyl-$_{L}$-phenylalanine	CME	eFx	2

Supporting Methods

Preparation of activated amino acids, flexizymes, tRNAs, and reconstituted *in vitro* translation system. Non-proteinogenic amino acids activated as cyanomethylesters (CME) or dinitrobenzylesters (DBE) esters (Table. S3) were prepared as described recently. Flexizymes eFx and dFx were prepared as described elsewhere (1). $tRNA^{fMet}_{CAU}$ (2) and $tRNA^{EnGlu}_{CAU}$ (3) were prepared as described previously. The custom *in vitro* translation mixture lacking methionine and lacking the formyl donor 10-formyl-5,6,7,8-tetrahydrofolic acid was prepared as described previously (4).

Fidelity of non-proteinogenic amino acids incorporation during translation detected using MALDI-TOF-MS. $tRNA^{fMet}_{CAU}$ aminoacylated with N-acetyl L-phenylalanine (final concentration 50 μM), $tRNA^{EnGlu}_{CAU}$ aminoacylated with one of the 21 non-proteinogenic amino acids (final concentration 50 μM) and DNA template YK5_ATG_DNA were added to the above *in vitro* translation system and incubated for 30 min at 37 °C. The peptide products were isolated using C18 C-tip (WAKO), eluted with 50 % saturated α-cyano-4-hydrocinnamic acid in 80 % MeCN, 0.5 % AcOH and spotted on a MALDI plate. Peptides between 1-3 kDa were investigated using MALDI-TOF on positive reflector mode (UltraFlex, Bruker).

YK5_ATG_DNA =
GGCGTAATACGACTCACTATAGGGTTAACTTTAAGAAGGAGAAAAACATG
TACAAAAAGTACAAGATGGACTACAAGGACGACGACGACAAGTAAGCTT
CG

Biotinylated MCL1 expression and purification. An *E.coli* codon optimized synthetic gene MCL1_DNA (Eurofins genomics Japan) was made to contain an Avi BirA recognition sequence, His_6-tag, thrombin cleavage site and mouse MCL1 amino acids 152-308 (Uniprot P97287). This was cloned (In-Fusion HD, Clontech Laboratories) into an ampicillin resistant pQE80L co-expression vector that also encoded a BirA biotin ligase. This vector was transformed into Rosetta 2 (DE3) pLysS chemically competent cells, grown at 37 °C in the presence of 100 μg mL^{-1} ampicillin. When OD_{600} = 0.4-0.6, the media was supplemented with biotin to a final concentration of 40 μM, induced with 0.1 mM IPTG, and the temperature reduced to 25 °C for 18 hours before harvesting.

MCL1_DNA =

ATGTCCGGCCTGAACGACATCTTCGAGGCTCAGAAAATCGAATGGCACGA
AGGTCACCATCATCACCACCACCTTGTGCCCCGCGGATCGGAGGATGACC
TGTACCGCCAATCCTTAGAAATCATTAGCCGTTACTTGCGCGAGCAGGCT
ACCGGCTCGAAAGATAGTAAACCCCTGGGTGAGGCTGGCGCGGCGGGGC
GTCGTGCACTGGAGACACTTCGCCGTGTAGGCGACGGAGTCCAACGTAAT
CATGAAACGGCGTTCCAAGGGATGTTGCGTAAATTGGACATTAAGAATGA
GGGGGACGTTAAATCGTTTTCACGTGTCATGGTCCACGTTTTCAAGGATGG
CGTCACAAATTGGGGGCGTATCGTAACACTGATCTCTTTCGGAGCGTTCGT
CGCAAAGCATTTAAAATCCGTGAACCAGGAATCGTTTATTGAACCATTAG
CAGAGACAATTACGGATGTGCTTGTTCGCACCAAGCGCGACTGGTTAGTA
AAGCAGCGTGGCTGGGACGGGTTCGTTGAGTTCTTTCACGTCCAAGATTT
AGAAGGGGGTTAAGGTACCCAATAAGCTTACAAT

Cell pellets were sonicated in the presence of DNase, resuspended in cold 20 mM Tris pH 7.4 150 mM NaCl buffer, and cell debris removed by centrifugation. His tagged MCL1 was isolated with Ni-NTA resin (Qiagen), washed with 20 mM Tris pH 7.4, 500 mM NaCl, 40 mM imidazole, and MCL1 eluted with 20 mM Tris pH 7.4, 500 mM NaCl, 250 mM imidazole. Elutant was 0.22 μm filtered and purified using ÄKTA FPLC (GE Healthcare) HiLoad 16/600 Superdex 200 pg in 50 mM sodium phosphate pH 7.0. Fractions with >95% pure MCL1 (assessed using SDS-PAGE) were combined, 0.22 μm filtered, and were concentrated (Amicron 10K cut-off). Concentration was determined using UV absorbance at A280 with an extinction coefficient of 24980 M^{-1} cm^{-1} (ExPASy ProtParam), flash frozen using $N_{2(l)}$ and stored at -80 °C.

Biotinylated KDM4A. Biotinylated KDM4A was a gift from Akane Kawamura (University of Oxford). Expression described previously (5).

Assembly of saturation mutagenesis DNA/mRNA libraries. Wild-type DNA sequences were designed to contain (5' to 3') a T7 polymerase binding site, ribosome binding site, peptide encoding region, 'TAG' stop codon and finally, a sequence complementary to a DNA splint used for puromycin ligation. The peptide region encodes the CP2/PUMA peptide followed by a $(Gly\text{-}Ser)_2$ linker, HA-tag and $(Gly\text{-}Ser)_3$ linker. CP2 peptide sequence has been described previously (5). A 35 amino acid segment of PUMA, amino acids 127-161 (Uniprot Q99ML1) of the full-length

protein, was chosen as it encompasses the entire MCL1 binding site. The chosen PUMA sequence also includes the artificial mutation M144A to match that used in previous studies (6, 7).

CP2_wild-type_DNA =
TAATACGACTCACTATAGGGTTAACTTTAAGAAGGAGATATACATATGGT
GTATAATACCCGTAGTGGGTGGCGTTGGTACACTTGCGGCAGCGGCAGCT
ACCCATACGACGTGCCCGACTATGCAGGTTCTGGTTCTGGTTCTTAGGACG
GGGGGCGGAAA

CP2_wild-type_prot = (ClAc D-tyrosine)VYNTRSGWRWYTCGSGSYPYDVPDYAGSGSGS

PUMA_wild-type_DNA =
TAATACGACTCACTATAGGGTTAACTTTAAGAAGGAGATATACATATGCG
TGTAGAGGAGGAGGAGTGGGCACGCGAGATCGGAGCCCAATTACGTCGT
GCGGCTGACGACCTTAACGCACAATACGAGCGTCGTCGTCAGGAAGAACA
ACACGGCAGCGGCAGCTACCCATACGACGTGCCCGACTATGCAGGTTCTG
GTTCTGGTTCTTAGGACGGGGGGCGGAAA

PUMA_wild-type_prot =
MRVEEEEWAREIGAQLRRAADDLNAQYERRRQEEQHGSGSYPYDVPDYAGS
GSGS

Based on these ‘wild-type’ sequences, site saturation mRNA libraries were constructed so that, after translation, every single amino acid change to the wild-type sequence was sampled i.e. one codon change per molecule of mRNA.

As the CP2 peptide is short, a site saturation DNA library could be easily assembled using primer extension and PCR. Multiple primers were used, each containing a single ‘NNK’ degenerate codon at different positions in the CP2 encoding region. The ‘ATG’ initiation codon and the codon for cysteine (required for cyclization) were not modified. Primers were annealed and extended using Phusion polymerase (NEB), PCR amplified, purified using phenol:chloroform:isoamyl alcohol (PCI) extraction and ethanol precipitated. To ensure an equal mixture of proteinogenic to non-proteinogenic amino acids in the final peptide library, and additional DNA library was prepared as above using an ‘ATG’ codon in place of the ‘NNK’ codon in each of the primers. This ‘ATG’ library was mixed with the ‘NNK’ library in equal proportions.

The final CP2 mutant DNA library was added to an *in vitro* transcription reaction using in-house T7 polymerase, treated with DNase, purified using PCI extraction and the resulting mRNA library ethanol precipitated, resuspended in H_2O and stored at -80 °C.

The PUMA wild-type DNA construct is significantly longer than the CP2 construct, and a mutant DNA library would be laborious to assemble using primer annealing and PCR. Instead, the protocol of Olson et.al (8) was followed to generate a site saturation mutagenesis library with 34 NNK codons. To ensure an equal mixture of proteinogenic to non-proteinogenic amino acids in the final peptide library, it was desirable to enrich the PUMA mRNA library with 'AUG' codons. Instead of reassembling the PUMA library using 'ATG' containing primers, a single round of RaPID selection (4) was preformed, with the genetic code reprogrammed so that 'ATG' would code for biotinylated lysine. Any 'AUG' containing mRNA would then be ligated to biotin containing peptide, and could be isolated using M280 Streptavidin Dynabeads. Recovered mRNA were reverse transcribed to DNA, PCR amplified, used as a template for *in vitro* transcription, and mutant 'AUG' enriched PUMA mRNA library recovered. The 'AUG' enriched library and initial 'NNK' mRNA libraries were mixed in equal proportions.

Each PUMA and CP2 mRNA library was ligated to a puromycin-PEG-DNA splint using T4 RNA ligase according to the standard RaPID system protocol (4).

***In vitro* translation with genetic code reprogramming.** Each mRNA-puromycin library was added to 21 separate, small-scale (3.75 μL) *in vitro* translation systems, each with a different reprogrammed genetic code. In each genetic code one of the 21 non-proteinogenic amino acids (Fig. 2*A*) was assigned to elongation 'AUG' codon and, simultaneously, a non-proteinogenic amino acid (N-chloroacetyl D-tyrosine for CP2, N-acetyl L-phenylalanine for PUMA) was assigned to the initiation 'AUG' codon. For the PUMA experiment, an additional translation reaction was set up without genetic code reprogramming i.e. with elongation 'AUG' assigned to methionine and, initiation to formyl methionine. Each translation system was incubated for 30 mins at 37 °C, then 25 °C for 12 mins to promote ligation. 16.7 mM EDTA pH 8.0 was added before the reaction was incubated for 30 mins at 37 °C to dissociate the mRNA-peptide fusions from the ribosome and, for CP2, allow cyclization to reach completion.

Barcoding. Each reaction was reverse transcribed (M-MLV RTase, RNase H minus, Promega), initiated using a DNA primer, to form cDNA-mRNA-peptide fusions. Each DNA primer contained a barcode unique to the reprogrammed genetic code used. This barcode was later sequenced, along with the peptide coding region, to identify the genetic code was used to generate each peptide.

HA-tag purification. All translated mutant and wild-type cDNA-peptides were combined, diluted with buffered acetylated-BSA 0.1% and incubated with anti-HA magnetic beads (Pierce, Thermo Fisher) at 4°C for 1 hour. Beads were washed three times with ice-cold buffer to remove the translation system, unligated mRNA and incompletely translated peptides. The C-terminal HA-tag ensured that only fully translated peptide products were retained, and these were eluted using a solution containing HA-peptide (2 mg mL-1, Pierce, Thermo Fisher) in a 0.1 % acetylated-BSA containing buffer appropriate to the partner protein and incubated at 37°C for 30 mins. The buffer used for the PUMA experiments was 50 mM sodium phosphate, pH 7.0, 2 mM DTT, 0.05% Tween20. The equivalent buffer in the CP2 experiments contained 20 mM Tris pH 7.6, 150 mM, 0.05% Tween20, 100 μM 2-oxoglutatate, 100 μM $NiCl_2$. HA-tag purified libraries of DNA-mRNA-peptides make up the 'input' library in the following analysis.

Binding to partner protein. Each input library was incubated with partner protein (MCL1 for the PUMA library, KDM4A for the CP2 library) immobilized on magnetic beads. The amount of beads was chosen so that, for the reaction volume, the partner protein would be present at a concentration of 200 nM. Based on the kinetics of wild-type PUMA and CP2 binding their partner proteins (5, 6), 3 hours was deemed sufficient for the binding reactions to have reached equilibrium. After 3 hours at 25 °C, the beads were collected at the bottom of the tube using a magnet, and the supernatant discarded. One reaction volume of buffer was added to the beads, which were resuspended, and the reaction incubated for a further 3 hours. The supernatant was discarded and an additional volume of buffer added, and incubated for a final 3 hours. The supernatant was removed and the beads resuspended in 0.1% Triton containing 100 mM biotin, heated to 95 °C for 5 mins, to denature streptavidin and the mRNA/DNA duplex, and the 'output' DNA library recovered. Binding experiments were performed in triplicate.

Illumina sequencing and enrichment score (*E*) calculation. The concentration of recovered input and output DNA was assessed by real-time PCR (lightCycler 96, Roche) using the primers T7g10M_F46 and R19_JMR_qPCR_barcode.

T7g10M_F46 = TAATACGACTCACTATAGGGTTAACTTTAAGAAGGAGATATACATATG

R19_JMR_qPCR_barcode = ACGTGTGCTCTTCCGATCT

Using the DNA concentrations, PCR reactions were designed that could be stopped before the non-exponential phase was reached. The DNA libraries were PCR amplified using high fidelity Phusion polymerase (NEB), and primers containing the indexes required for Illumina sequencing. First PCR reaction using primers F60_JMR_Rd1T7g10M and R32_JMR_an13Rd2_barcode. Second PCR reaction using primers P5XXXXRd1F57 and Rd2XXXXP7_R52, where XXXX refers to indexes unique for each sample.

F60_JMR_Rd1T7g10M = CACTCTTTCCCTACACGACGCTCTTCCGATCTGGGTTAACTTTAAGAAGGAGATATACAT

R32_JMR_an13Rd2_barcode = GACTGGAGTTCAGACGTGTGCTCTTCCGATCT

P5XXXXRd1_F57 = AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCTTTCCCTACACGAC

Rd2XXXXP7_R52 = CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTG

PCR reactions were column purified (Macherey-Nagel), and concentration determined using Qubit dsDNA BR kit (Thermo Fisher). CP2 input and output DNA libraries were run on the MiSeq (Illumina) platform, using single 151 cycle read mode and v3 chip. PUMA libraries were run on paired-end 2x151 cycle mode and v2 chip. Using a custom python script based on the ENRICH pipeline (9, 10), where reads were filtered for quality: any reads that did not have >30 Phred+33 score across the entire peptide coding region were discarded. Each DNA sequence was converted to peptide sequence, where elongation ‘ATG’ codons were encountered, the genetic-

code-specifying barcode was checked to assign the appropriate non-proteinogenic amino acid (or methionine). Mutations in these peptide regions were tallied. A key assumption of the deep mutational scanning method is that the population of a particular mutant in the output library will be a function of the 1) strength of binding to the partner protein (K_D) and 2) the initial population in the input library. By sequencing both the input and output libraries, the latter can be corrected for, and an enrichment score (E) can be calculated for each peptide that reports on the strength of binding. Equations [1-3] were used to calculate an enrichment score E for each peptide, $E = 0$ means identical enrichment to the wild-type peptide.

$$F_i = \frac{\text{reads}_i}{\sum \text{reads}} \quad [1]$$

Fraction of reads F_i. Number of reads for peptide i (wild-type or mutant) relative to the total number of reads (for wild-type and all mutants).

$$e_i = \frac{F_{i,\text{output}}}{F_{i,\text{input}}} \quad [2]$$

Raw enrichment score e_i for mutant i. Fraction of reads for mutant in the output library, $F_{i,\text{output}}$, relative to the fraction of reads in the input library, $F_{i,\text{input}}$.

$$E_i = \frac{e_i}{e_{\text{wild-type}}} \quad [3]$$

There was no correlation between E and the read count for each mutant (Fig. S15). Errors for $\log_2 E$ ($\sigma \log_2 E$) were calculated using the standard deviation between repeat binding and DNA recovery experiments. For PUMA, $\sigma \log_2 E$ did show a dependence on the input read count, with lower input reads resulting in greater error. Also, for PUMA, destabilizing mutants (low $\log_2 E$) also produced greater errors, presumably due to the low read count in the output library. Errors in $\log_2 E$ appear large relative to the absolute values of $\log_2 E$ in Fig. S15. However, the errors are standard deviations, and do not include the increase in confidence that results from repeat experiments. Second, there are systematic differences in $\log_2 E$ magnitude between repeats of the enrichment protocol (Fig. S15 *D,H*). Differences in $\log_2 E$, not the absolute values,

that are important for finding structure activity relationships and fitting for $\Delta\Delta G$ (see below).

Calculation of *E* for PUMA mutations to methionine. The translation system without genetic code reprogramming, used with the PUMA DNA/mRNA library, generated PUMA peptides initiated with formyl methionine (instead of N-acetyl L-phenylalanine in the reprogrammed genetic codes). Reads from this genetic code were used to calculate a separate $e_{\text{i,wild-type}}$ for this formyl methionine initiated 'wild-type'. This value was used in the calculation of E for all methionine mutants.

Calibration of $\Delta\Delta G$ using an empirical fit. It is possible to derive functions for $\Delta\Delta G = f(\log_2 E)$ on theoretical grounds. However, it is difficult to take into account factors such as: the fact that multiple washes are preformed, that moles of target protein might be limiting, or other (perhaps unknown) biases that affect enrichment. Here, we chose to use a simple empirical fit for $\Delta\Delta G = f(\log_2 E)$, deciding that this was the most cautious approach to calibrate the deep mutational scanning data.

$\Delta\Delta G$ calculation for each PUMA mutant. There are a number of published K_D values for PUMA mutants binding to MCL1, which can be used to calculate $\Delta\Delta G$ for each mutant.

$$\Delta G = RT \ln K_D$$

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild-type}$$

A number of extra PUMA mutants were selected, based on the deep mutational scanning data, to include proteinogenic and non-proteinogenic amino acids to cover a range of $\log_2 E$. These PUMA mutants were synthesized and binding to MCL1 assessed using SPR (see below). The resulting $\Delta\Delta G$ values were combined with the previously published $\Delta\Delta G$, and this combined set, and the corresponding $\log_2 E$ values (average $\log_2 E$ from the repeat experiments of the library binding and DNA recovery) were fit empirically to the following curve:

$$\Delta\Delta G = ae^{-log_2(E)/b} - a$$

A linear relationship might be expected between $\Delta\Delta G$ and $\log_2 E$: raw e reports on the proportion of PUMA molecules bound MCL1 relative to the total, so should be reporting on K_D. Then, $\log_2 e$ reports on $\log K_D$, and therefore $\log_2 E$ should be

proportional to $\Delta\Delta G$. However, the particular deep mutational scanning set up here was conducted with the concentration of MCL1 being lower than the total PUMA concentration (including all sequence variations, and peptides not ligated to cDNA, Fig S14). Lower K_D PUMA mutants bind in greater numbers to MCL1, but in addition to this, leave fewer MCL1 binding sites for the higher K_D PUMA mutants. This produces deviations from linearity in $\Delta\Delta G$ vs $\log_2 E$, and necessitates the curved fit used above.

$\Delta\Delta G$ calculation for each CP2 mutant. A collection of CP2 mutants were synthesized and $\Delta\Delta G$ determined by SPR (see below). These $\Delta\Delta G$ and their corresponding $\log_2 E$ were fit to a linear function.

$$\Delta\Delta G = m \times \log_2(E)$$

This linear function was deemed sufficient to empirical fit of $\Delta\Delta G$ to average $\log_2 E$, as the deep mutational scanning suggested that there were no mutants of CP2 that significantly stabilized the interaction with KDM4A. Fitting to the curved function, which was used in the case of PUMA, does not change the conclusions in the main text. The gradient of the linear fits of $log_2 E$ and $\Delta\Delta G$ were same when published IC_{50} values and BIACORE SPR K_D values are used to calculate $\Delta\Delta G$ (Fig. S7).

Leave-one-out cross validation. The above fitting procedure for PUMA and CP2 were carried out on a data set lacking one mutant, and the resulting function used to estimate the $\Delta\Delta G/K_D$ of the missing mutant. This was repeated for all mutants.

Error analysis. In the above protocol, average $\log_2 E$ were used to fit for $\Delta\Delta G$. Standard deviation in $\log_2 E$ were propagated appropriately, depending on the function $\Delta\Delta G = f(\log_2 E)$ used, to give the reported error in $\Delta\Delta G$. An alternative fitting procedure was also performed: $\log_2 E$ from each individual binding experiment repeat (Fig. S16*A*, *C*) were fit, the resulting $\Delta\Delta G$ averaged, and the standard error of these $\Delta\Delta G$ collected. This alternative analysis produced similar values of, and errors in, $\Delta\Delta G$ as the original analysis (Fig. S16*B*, *D*).

SPPS of a sample set of PUMA mutants. A 27 amino acid sequence segment of PUMA_wild-type_prot, PUMA27, was designed for solid phase peptide synthesis. This 27 amino acid peptide covers the α-helix formed by the PUMA peptide in the MCL1 bound structure (pdb 2ROC). To prevent ionizable groups at the termini, the C-terminus was amidated and the *N*-terminus acetylated.

PUMA27 = EEEWAREIGAQLRRAADDLNAQYERRR

Peptides were synthesized on a 25 μmol scale using NovaPEG Rink Amide resin (Novabiochem) using standard Fmoc protection chemistry (using automated synthesizer, Biotage SYRO I). After coupling the final *N*-terminal amino acid, the terminal Fmoc group was removed. The resulting *N*-terminal α-amino group was acetylated using 10% acetic anhydride, 10% trimethylamine, 40% dichloromethane (DCM), 40% dimethylformamide (DMF), with rotation for 30 mins. After washing the resin with 1:1 DCM:DMF (three times) and then DCM (five times), the resin was dried *in vacuo*. The peptides were cleaved from the resin using a solution of trifluoroacetic acid (TFA)/1,2-ethanedithiol/triisopropyl silane/water (92.5:2.5:2.5:2.5) with rotation at room temperature for 2 hours. The cleaved solutions were concentrated *in vacuo* and the peptides precipitated using diethyl ether. The peptide pellets were dissolved in ~1:1 acetonitrile:water, 0.1% TFA, and purified by reverse-phase HPLC (Shimadzu prominence LC-20AP system with Merck Chromolith Prep column) in 0.1% aqueous TFA/acetonitrile containing 0.1% TFA gradient. The purified peptides were lyophilized and reconstituted using a small volume of DMSO. The concentration was determined using absorbance at 280 nm in neutral buffer with 1 % DMSO, using predicted extinction coefficients (ε) (11) and ε estimates for the non-proteinogenic amino acid Bzt (see below). Identity and purity of peptides confirmed using MALDI-TOF-MS and UHPLC (Fig. S17*A*, S18).

UHPLC. Ultra-high performance liquid chromatography was performed using a Nexera X2 UHPLC system (Shimadzu) fitted with a C18 reversed phase column. 0.5-5 nmol of peptide was separated using a 10 vol% to 70 vol% aqueous acetonitrile gradient supplemented with 0.1 vol% trifluoroacetic acid, and monitored by absorbance at 280 nm using a SPD-M20A (PUMA peptides) or SPD-30AM (CP2 peptides) detector.

SPPS of a sample set of CP2 mutants. CP2 peptides were synthesized as previously described (5). Briefly, after automated Fmoc assembly the *N*-terminus was chloroacetylated using a solution of 0.5 M chloroacetyl *N*-hydroxysuccinimide ester in DMF and rotation for 1 hour. After washing, TFA cleavage and ether precipitation the linear peptides were dissolved in 90% DMSO, 10% H_2O and the pH raised to > 8.0 to promote the cyclization reaction. Cyclization was followed using MALDI-TOF-MS and when complete the pH was lowered to < 3 using TFA and the cyclic

peptides purified using HPLC. Identity and purity of peptides confirmed using MALDI-TOF-MS and UHPLC (Fig. S17*B*, S19).

Bzt extinction coefficient. The non-proteinogenic amino acid Bzt (3-benzothienyl alanine) is a tryptophan mimic where the N-H of the indole has been replaced by sulfur. This aromatic non-proteinogenic amino acid was expected to contribute to the absorbance of peptides. To estimate the extinction coefficient of Bzt, two model peptides were synthesized as above: Ac-LNAQ(Bzt)AGSS-NH_2 and Ac-LNAQ(Trp)AGSS-NH_2. Weighed aliquots of HPLC purified peptides, absorbance at 280 nm, and the known ε of tryptophan (5500 M^{-1} cm^{-1}) (11)were used to estimate ε for Bzt (810 M^{-1} cm^{-1}), (Fig. S20).

PUMA MCL1 SPR. The interaction between PUMA27 (plus mutants) and MCL1 was quantified using BIACORE T100 instrument (GE healthcare) equipped with biotin CAPture kit, Series S, chip. Biotinylated MCL1 was added to a surface density of approximately 1000 R.U. 50 mM sodium phosphate buffer, 0.05% Tween20, 0.2 % DMSO was used as the running buffer for all experiments. Varying concentrations of PUMA27 peptide were flowed over the chip at 100 µL min^{-1} to examine association kinetics, followed by running buffer only to induce dissociation. The reaction was assumed to be in the pseudo-first-order regime, where the concentration of flowed peptide is constant. Association was fit to a single exponential to extract an apparent rate constant, which was used to calculate a second order rate constant. Apparent first order rate constants increased linearly with peptide concentration. Dissociation traces were fit to a single exponential decay, with a linear drift term. For the stabilizing PUMA mutants (A144Cha, A139tBu, A145Gly, Q140Tyr) a trace long enough to examine the drift could not be collected, for these traces no drift term was included. For the highly destabilizing mutant I137DAl, the association kinetics showed complex behavior, so the amplitudes of the binding were fit to estimate an apparent K_D using:

$$RU = \frac{\text{c x } RU_{max}}{K_d + c} \qquad [4]$$

Fitting BIACORE amplitudes to find K_D, where *RU* is the amplitude of the association kinetic trace (in response units) and *c* the concentration of the peptide flowed over the chip.

CP2 KDM4A SPR. The interaction between CP2 (or mutants) and KDM4A was also quantified using BIACORE, as above. Biotinylated KDM4A was added to a surface density of approximately 1000 R.U. 20 mM Tris pH 7.6, 150 mM, 0.05% Tween20, 100 μM 2-oxoglutatate, 100 μM $NiCl_2$, 0.2 % DMSO was used as the running buffer. Varying concentrations of CP2 peptide were flowed over the chip at 100 μL min^{-1} to examine association kinetics. Association was fit to a single exponential, and the amplitudes fit to determine each K_D using Equation 4.

SI References

1. Murakami H, Ohta A, Ashigai H, & Suga H (2006) A highly flexible tRNA acylation method for non-natural polypeptide synthesis. *Nat Methods* 3(5):357-359.
2. Goto Y*, et al.* (2008) Reprogramming the translation initiation for the synthesis of physiologically stable cyclic peptides. *ACS Chem Biol* 3(2):120-129.
3. Katoh T, Tajima K, & Suga H (2017) Consecutive Elongation of D-Amino Acids in Translation. *Cell Chem Biol* 24(1):46-54.
4. Yamagishi Y*, et al.* (2011) Natural product-like macrocyclic N-methyl-peptide inhibitors against a ubiquitin ligase uncovered from a ribosome-expressed de novo library. *Chem Biol* 18(12):1562-1570.
5. Kawamura A*, et al.* (2017) Highly selective inhibition of histone demethylases by de novo macrocyclic peptides. *Nat Commun* 8:14773.
6. Rogers JM, Wong CT, & Clarke J (2014) Coupled folding and binding of the disordered protein PUMA does not require particular residual structure. *J Am Chem Soc* 136(14):5197-5200.
7. Rogers JM*, et al.* (2014) Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proc Natl Acad Sci U S A* 111(43):15420-15425.
8. Olson CA, Wu NC, & Sun R (2014) A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* 24(22):2643-2651.
9. Fowler DM*, et al.* (2010) High-resolution mapping of protein sequence-function relationships. *Nat Methods* 7(9):741-746.
10. Fowler DM, Araya CL, Gerard W, & Fields S (2011) Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* 27(24):3430-3431.
11. Gill SC & von Hippel PH (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* 182(2):319-326.