Supplementary Information

Design of Peptide-Guided Protein Degraders with Structure-Agnostic Language Models

Garyk Brixi,[1,*] Tianzheng Ye,[2,*] Kalyan Palepu,[1] Lauren Hong,[1] Vivian Yudistyra,[1] Sophia Vincoff,[1] Jayani Christopher,[1] Xinning Li,[1] Suhaas Bhat,[1] Connor Monticello,[2] Natalia Lopez-Barbosa,[2] Lillian Petersen,[1] Tian-Lai Zang,[1] Tong Liu,[1] Lin Zhao,[1] Sue Zhang,[1] Manvitha Ponnapati,[1] Emma Tysinger,[1] Teodora Stan,[1] Sabrina Koseki,[1] Matthew P. DeLisa[2,3] and Pranam Chatterjee[1,4,5,†]

1. Department of Biomedical Engineering, Duke University
2. Robert F. Smith School of Chemical and Biomolecular Engineering, Cornell University
3. Cornell Institute of Biotechnology, Cornell University
4. Department of Computer Science, Duke University
5. Department of Biostatistics and Bioinformatics, Duke University

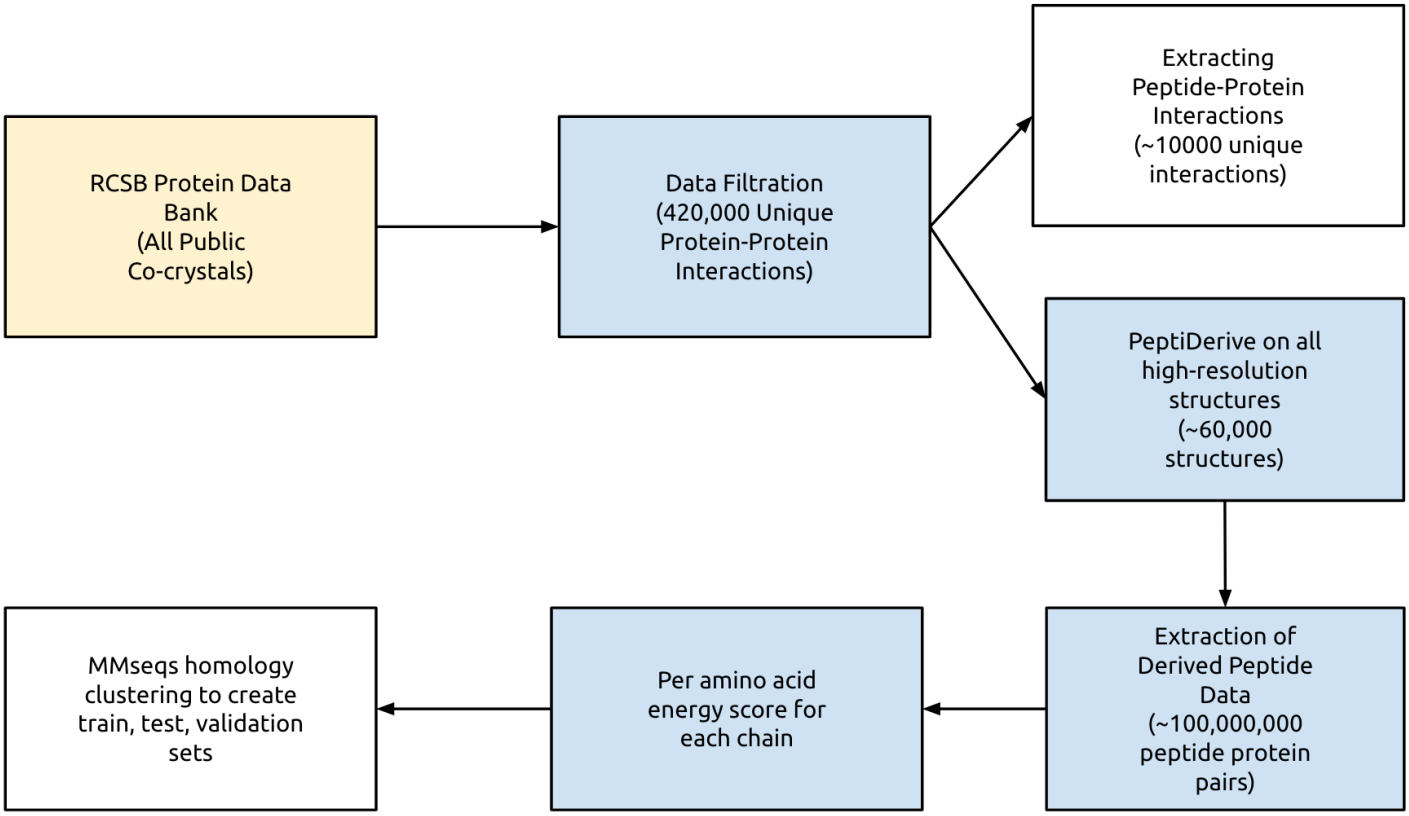
*These authors contributed equally
[†]Corresponding author: pranam.chatterjee@duke.edu

Supplementary Figures

1. PDB-Derived dataset generation for model training.
2. Benchmarking of SaLT&PepPr on isolated interacting partner structures.
3. Inference times across peptide derivation methods.
4. Per residue pLDDT scores for AlphaFold2-Multimer co-folded proteins.
5. Characterization of SaLT&PepPr peptide-guided uAbs for diverse target protein degradation.
6. Validation of β-catenin degradation via mass spectrometry.
7. Example gating strategy for flow cytometry analysis.

Supplementary Tables

1. SaLT&PepPr-derived peptide sequences and scores.
2. AlphaFold2-Multimer+PeptiDerive-derived peptide sequences.

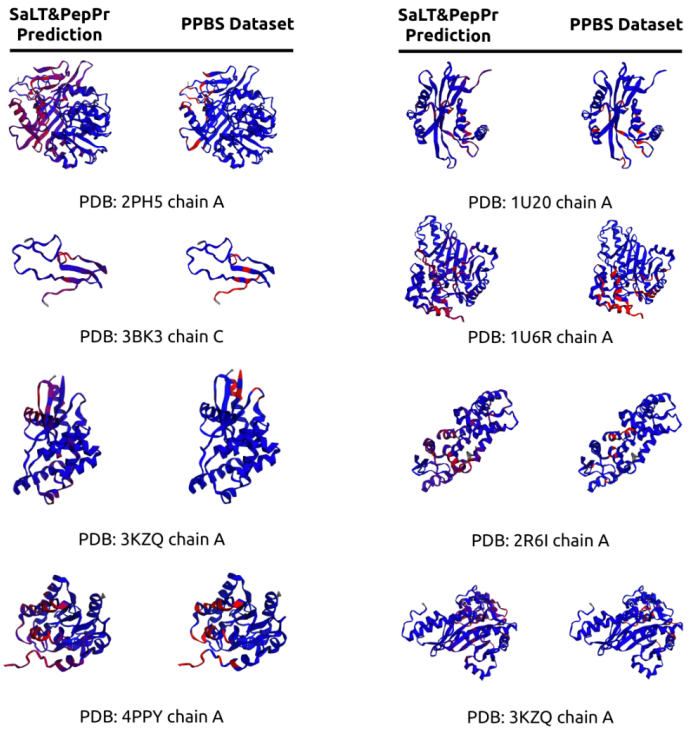


Supplementary Figure 1. PDB-derived dataset generation for model training. The RCSB Protein Data Bank was mined for verified, high-resolution PPI structures. Every interaction of every assembly of every co-crystal in the PDB was filtered for uniqueness (a unique pair of partners or >100 $Å^2$ buried surface area for the same pair of partners), yielding 420,000 PPIs. Next, interaction structures were processed with PeptiDerive, extracting a list of derived peptide "hot sequences", and their associated Rosetta energy unit (REU) scores on a per amino acid basis. Homology clustering was conducted for training and validation set derivation.

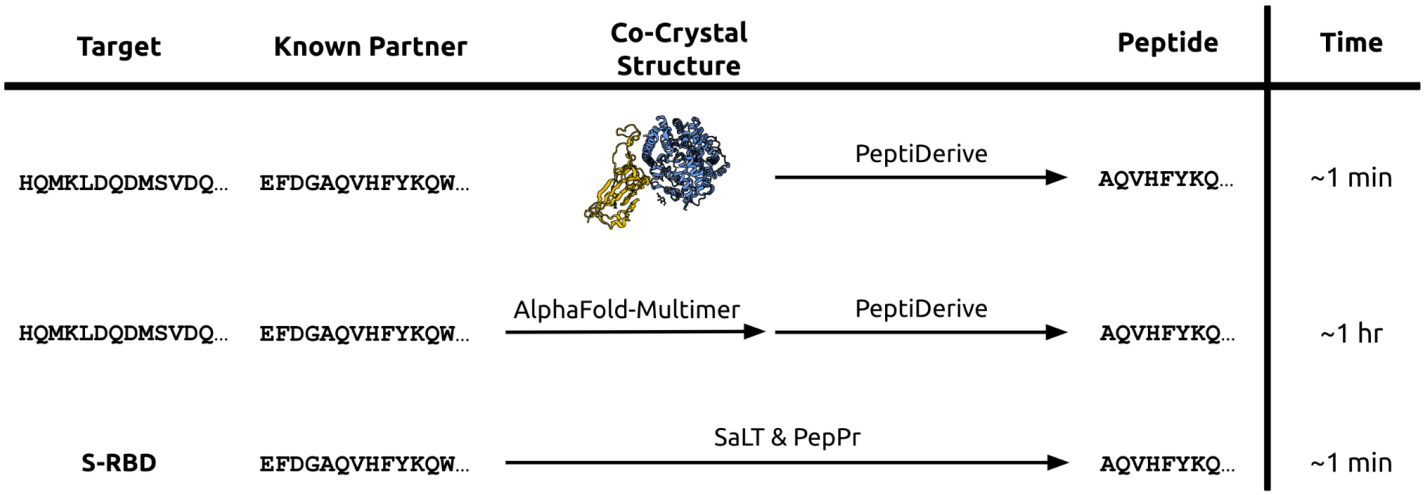
A

Algorithm	Test (70%)	Test (homology)	Test (topology)	Test (none)	Test (all)
SaLT&PepPr	0.710	0.64	0.55	0.458	0.58
ScanNet	0.732	0.712	0.735	0.605	0.694
Structural homology baseline	0.828	0.696	0.535	0.387	0.613
Handcrafted features baseline	0.596	0.567	0.568	0.432	0.537
MaSIF-site	N/A	N/A	N/A	N/A	0.533

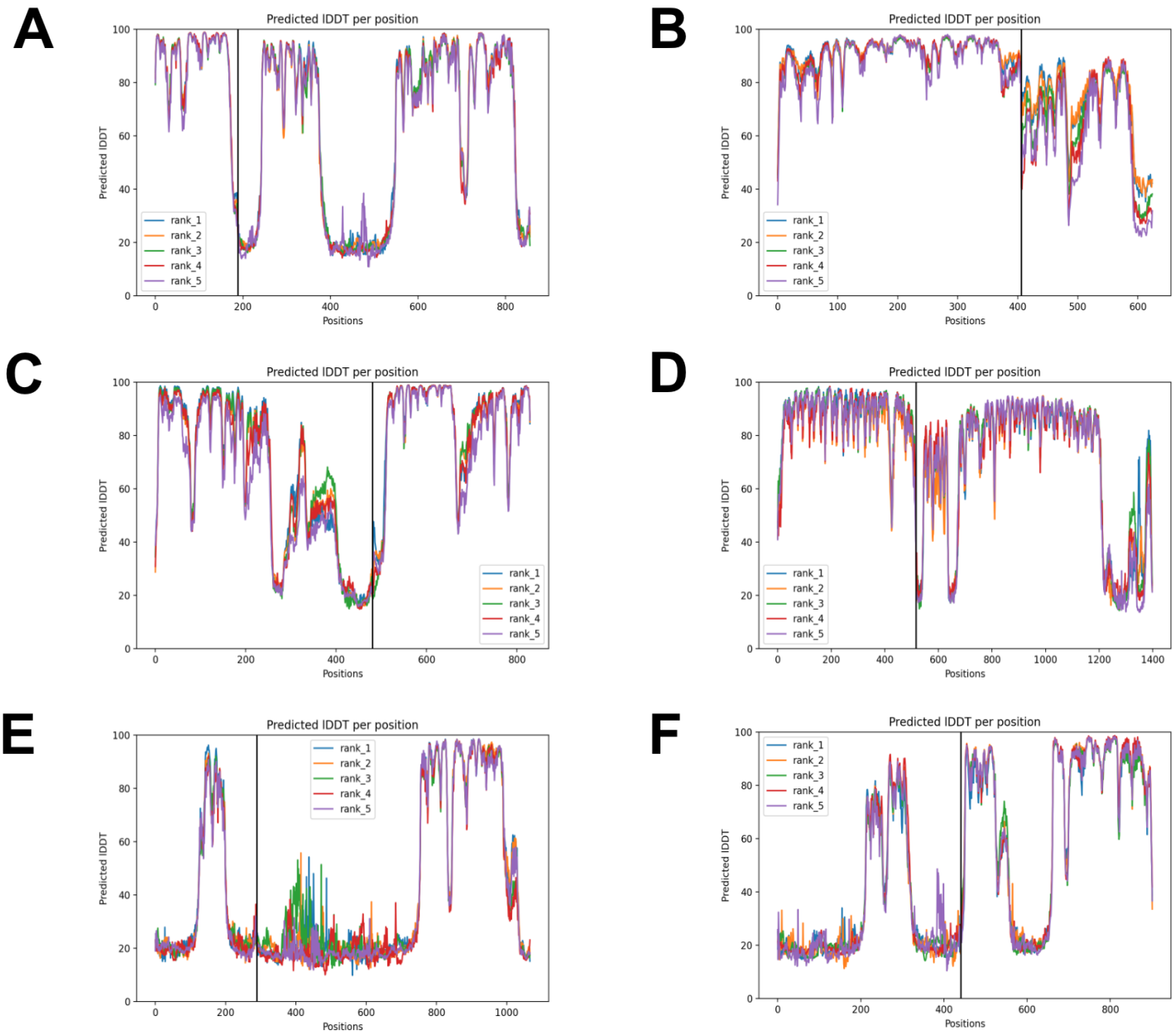
B



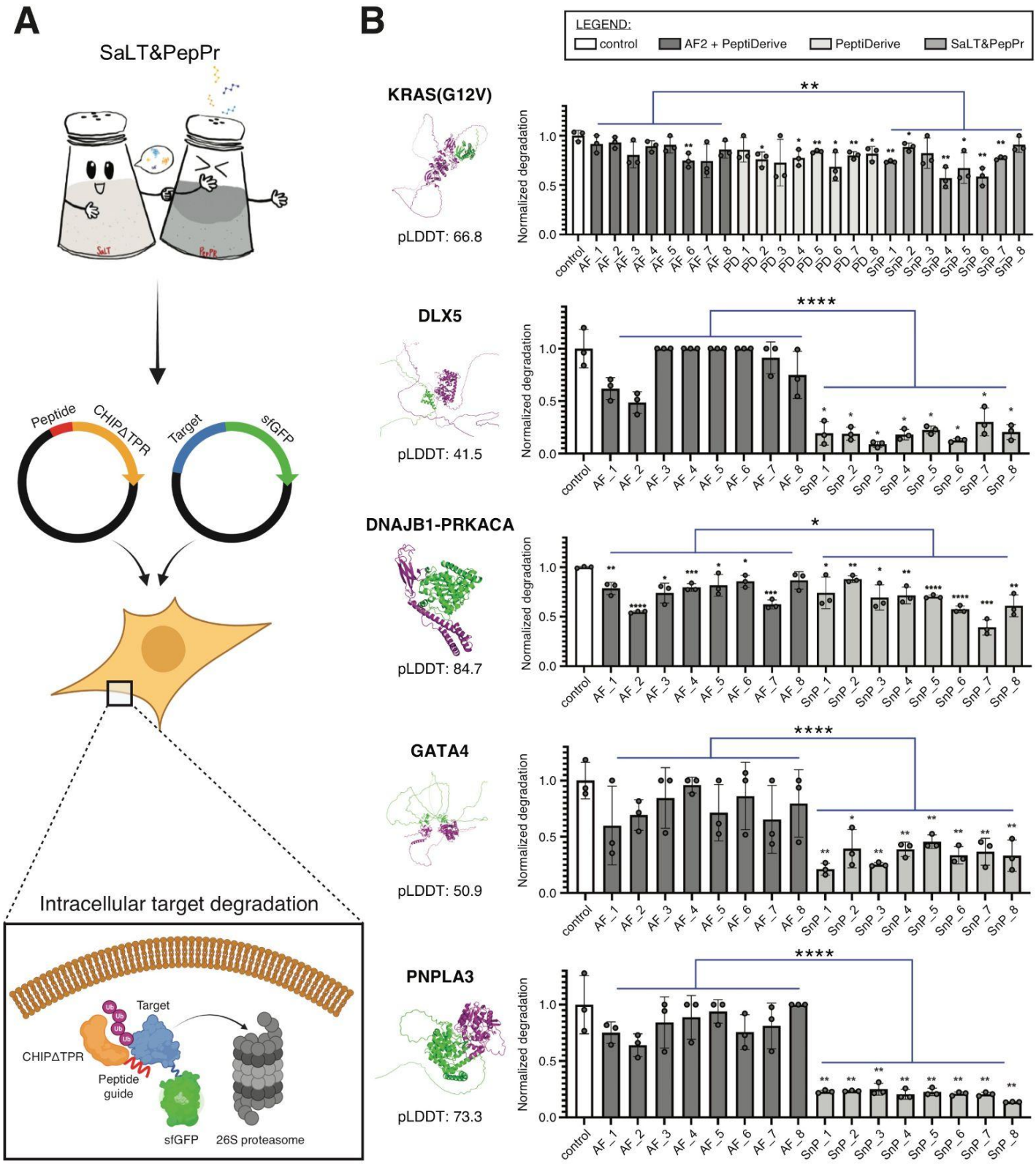
Supplementary Figure 2. Benchmarking of SaLT&PepPr on isolated interacting partner structures. A) Benchmarking of SaLT&PepPr trained on PPBS and then tested on different test splits. PPBS dataset, test splits, as well as model test scores are obtained from Tubiana, et al.[10] The PPBS dataset splits represent: 70% (at least 70% sequence homology with one training example), homology (at most 70% homology with a train set example, although at least one train set belongs to the same protein superfamily), topology (at least one train set has a similar protein topology with none in the similar protein family) and None (none of the above groups). The structural homology baseline uses template protein chains with known binding sites, a local pairwise structural comparison, and an alignment weighting scheme. The handcrafted features baseline includes 58 features based on the structural, atomic, and sequence information, with an XGBoost algorithm. Note that MaSIF-site was not retrained for the per-residue task. Additional details and implementation of baseline models and test data splits can be found in Tubiana, et al.[10] B) Comparison between the predicted SaLT&PepPr scores and experimentally-annotated PPBS binding sites on different protein structures in the PPBS dataset. Red indicates high binding probability amino acids, with blue as low binding probability, normalized for each protein chain.



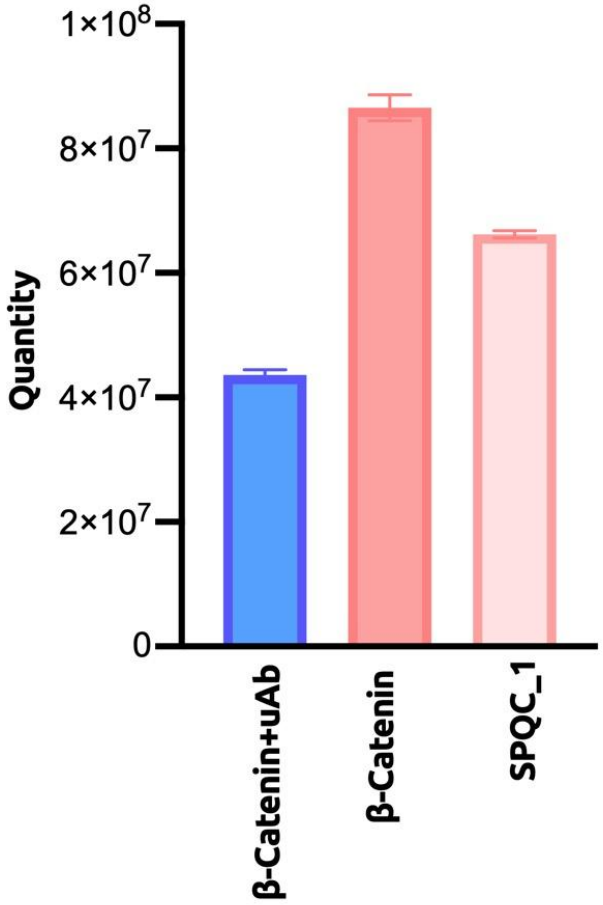
Supplementary Figure 3. Inference times across peptide derivation methods. Observed average times for inference using indicated methods are shown under “Time”. A standard machine with 2 CPU cores, 8 GB of RAM, and no GPU was used for inference.



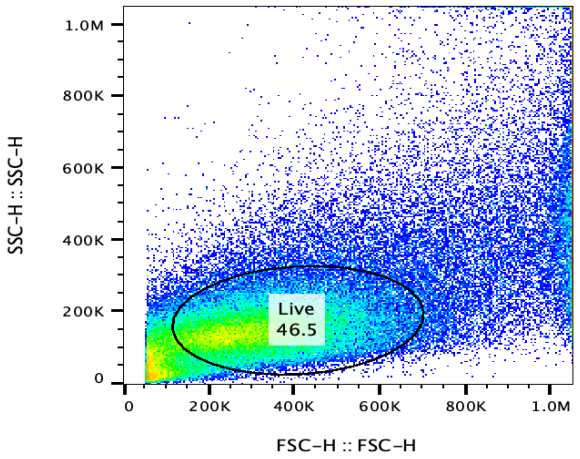
Supplementary Figure 4. Per residue pLDDT scores for AlphaFold2-Multimer co-folded proteins. Per amino acid position pLDDT scores for AlphaFold2-Multimer folds of A) KRAS(G12V) and RAF1, B) DNAJB1-PRKACA and HSPA8, C) PNPLA3 and ABHD5, D) β-Catenin and E-Cadherin, E) DLX5 and MAGED1, F) GATA4 and NR5A1.

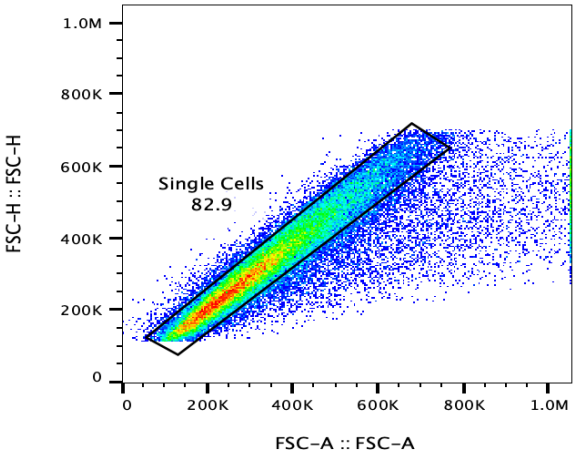

A
SaLT&PepPr
Peptide
CHIPΔTPR
Target
sfGFP
Intracellular target degradation
Target
CHIPΔTPR
Peptide guide
sfGFP
26S proteasome
B
LEGEND:
control
AF2 + PeptiDerive
PeptiDerive
SaLT&PepPr
KRAS(G12V)
pLDDT: 66.8
DLX5
pLDDT: 41.5
DNAJB1-PRKACA
pLDDT: 84.7
GATA4
pLDDT: 50.9
PNPLA3
pLDDT: 73.3
Normalized degradation
control
AF_1
AF_2
AF_3
AF_4
AF_5
AF_6
AF_7
AF_8
PD_1
PD_2
PD_3
PD_4
PD_5
PD_6
PD_7
PD_8
SnP_1
SnP_2
SnP_3
SnP_4
SnP_5
SnP_6
SnP_7
SnP_8

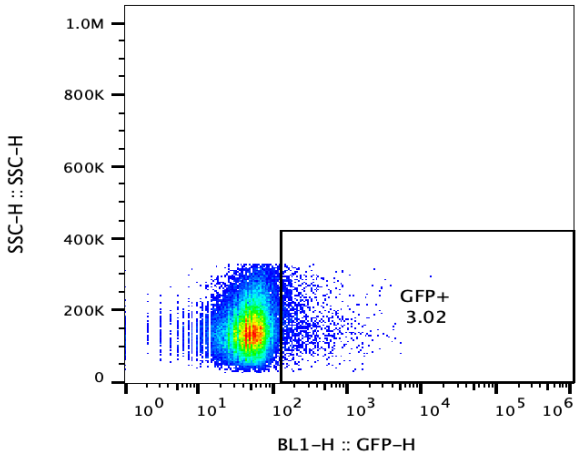
Supplementary Figure 5. Characterization of SaLT&PepPr peptide-guided uAbs for diverse target protein degradation. (A) Schematic of uAb degradation system. Briefly, SaLT&PepPr-derived peptides are fused to the N-terminus of CHIPΔTPR in mammalian expression vectors. The resulting vectors are transiently co-transfected in cell lines along with a plasmid encoding the target protein fused to superfolder GFP (sfGFP) and uAb-mediated silencing of the target is quantified by measuring cell fluorescence. (B) Analyses of AlphaFold2-Multimer co-complexes (magenta = partner, green = target) and uAb-mediated sfGFP degradation via flow cytometry were conducted for the following targets (from top to bottom): KRAS (G12V) with interaction partner, RAF1 (PeptiDerive was conducted on PDB 6XHA); DLX5 with interaction partner, MAGED1; DNAJB1-PRKACA with interaction partner, HSPA8; GATA4 with interaction partner, NR5A1; and PNPLA3 with interaction partner, ABHD5. AlphaFold2 pLDDT scores (1-100) represent average model confidence across the entire co-complex sequence, with <70 being low confidence, 70-90 medium confidence, and >90 as high accuracy predictions. All experimental samples were performed in independent transfection triplicates (n = 3) and gated on sfGFP fluorescence relative to the non-transfected control. Normalized degradation was calculated by dividing the percentage of sfGFP+ cells in a sample by that of the control for each target. For individual samples, statistical significance was determined by two-tailed Student's t test. For experimental comparison between groups (blue), a one-sided Mann-Whitney U Test was used to determine statistical significance for SaLT&PepPr (SnP) vs. the AlphaFold2-Multimer + PeptiDerive (AF) method or the PeptiDerive (PD) method. Calculated p values are represented as follows: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$; ns, not significant.

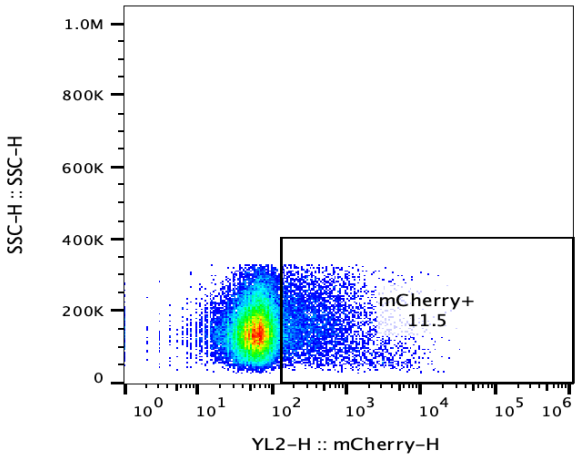


Supplementary Figure 6. Validation of β-catenin degradation via mass spectrometry. The abundances of β-Catenin (CTNNB1) were analyzed in triplicates with 1D-LCMS/MS in the presence and absence of the uAb and the SPQC pool. The average abundance of CTNNB1 was calculated for the triplicates and plotted using GraphPad Prism.









Supplementary Figure 7. Example gating strategy for flow cytometry. ~10,000 gated events for data analysis based on default FSC/SSC parameters for HEK293T cells. The GFP+ and mCherry+ gates were established both by a GFP- negative control and an mCherry- control. All analysis was conducted in FlowJo.

Supplementary Table 1. SaLT&PepPr-derived peptide sequences and scores. Eight peptides were built per target protein based on specific binders using SaLT&PepPr (SnP) with provided scores indicating the cumulative binding site probability of the peptide. These were cloned into uAb constructs for subsequent analysis via flow cytometry.

Target	Binder	Peptide Name	Sequence	SnP Score
DNAJB1-PRKACA	HSPA8	DJ1PKA_SnP_1	TERLIGDAAKNQVAMNPT	0.39120978
DNAJB1-PRKACA	HSPA8	DJ1PKA_SnP_2	GDAAKNQVAMNPTNTVF	0.4198106
DNAJB1-PRKACA	HSPA8	DJ1PKA_SnP_3	NKRAVRRLRTACERAKRT	0.3084646
DNAJB1-PRKACA	HSPA8	DJ1PKA_SnP_4	QASIEIDSLYEGIDFYTS	0.42268646
DNAJB1-PRKACA	HSPA8	DJ1PKA_SnP_5	KRNTTIPTKQTQTFTTYS	0.4849106
DNAJB1-PRKACA	HSPA8	DJ1PKA_SnP_6	LLLLDVTPLSLGIETAGG	0.5474601
DNAJB1-PRKACA	HSPA8	DJ1PKA_SnP_7	IEIDSLYEGID	0.44183227
DNAJB1-PRKACA	HSPA8	DJ1PKA_SnP_8	TFTTYSDNQPGVLIQ	0.3081186
PNPLA3	ABHD5	PNPLA3_SnP_1	MAAEEEEVDSADTG	0.502174
PNPLA3	ABHD5	PNPLA3_SnP_2	YSSMFEDDTVTEYIYHCN	0.56430477
PNPLA3	ABHD5	PNPLA3_SnP_3	SMFEDDTVTEY	0.6153683
PNPLA3	ABHD5	PNPLA3_SnP_4	YSSMFEDDTVTEY	0.6076508
PNPLA3	ABHD5	PNPLA3_SnP_5	FEDDTVTEYIYHCNVQTP	0.5483647
PNPLA3	ABHD5	PNPLA3_SnP_6	CNVQTPSGETAFKNMTIP	0.38931718
PNPLA3	ABHD5	PNPLA3_SnP_7	LAGLRIAGPFGLSLVQRL	0.5249392
PNPLA3	ABHD5	PNPLA3_SnP_8	GAALTPFNPLAGLRIAGP	0.4830475
KRAS (G12V)	RAF1	KRAS_SnP_1	LDHVPLTTHNFARKTFLK	0.7078685
KRAS (G12V)	RAF1	KRAS_SnP_2	AFCDICQKFLLNGFRCQ	0.73019314
KRAS (G12V)	RAF1	KRAS_SnP_3	AFCDICQKFLLNGFRCQT	0.72925234
KRAS (G12V)	RAF1	KRAS_SnP_4	CQKFLLNGFRCQT	0.7296444
KRAS (G12V)	RAF1	KRAS_SnP_5	DWSNIRQLLLFPN	0.56445897
KRAS (G12V)	RAF1	KRAS_SnP_6	QGMDYLHAKNI	0.23349328
KRAS (G12V)	RAF1	KRAS_SnP_7	HINNRDQIIFMVGRGY	0.48506093
KRAS (G12V)	RAF1	KRAS_SnP_8	FARKTFLKLAFCDICQ	0.7431202

β-catenin	CADH1	β-cat_SnP_1	DYEGSGSEAASLSSLNSSESDKDQ	0.509226
β-catenin	CADH1	β-cat_SnP_2	YDSLLVFDYE	0.39451852
β-catenin	CADH1	β-cat_SnP_3	AADTDPTAPPYDSLLVFDYE	0.36254913
β-catenin	CADH1	β-cat_SnP_4	PTAPPYDSLLVFDYE	0.38101247
β-catenin	CADH1	β-cat_SnP_5	YDSLLVFDYEG	0.40408888
β-catenin	CADH1	β-cat_SnP_6	TAPPYDSLLVFDYE	0.38535473
β-catenin	CADH1	β-cat_SnP_7	DPTAPPYDSLLVFDYEGS	0.39557245
β-catenin	CADH1	β-cat_SnP_8	PTAPPYDSLLVFDYEG	0.38843626
DLX5	MAGED1	DLX5_SnP_1	ADEALDALDAAA	0.7373449
DLX5	MAGED1	DLX5_SnP_2	LAGILGTTKDTP	0.70572895
DLX5	MAGED1	DLX5_SnP_3	EHLYILISTPES	0.69121355
DLX5	MAGED1	DLX5_SnP_4	ISTPESLAGILG	0.6980049
DLX5	MAGED1	DLX5_SnP_5	KEEHLYILISTP	0.6835796
DLX5	MAGED1	DLX5_SnP_6	FGIQLKEIDKEE	0.61468536
DLX5	MAGED1	DLX5_SnP_7	QLKEIDKEEHLY	0.63843673
DLX5	MAGED1	DLX5_SnP_8	GAAQPRDVALLQ	0.61574763
GATA4	NR5A1	GATA4_SnP_1	CGDKFQQLLLCL	0.6886584
GATA4	NR5A1	GATA4_SnP_2	CHYPHCGDKFQQ	0.67182046
GATA4	NR5A1	GATA4_SnP_3	DYTLCHYPHCGD	0.664633
GATA4	NR5A1	GATA4_SnP_4	ANAALLDYTLCH	0.67642945
GATA4	NR5A1	GATA4_SnP_5	ALLDYTLCHYPH	0.66024536
GATA4	NR5A1	GATA4_SnP_6	PPPPPAPDYVLP	0.41717997
GATA4	NR5A1	GATA4_SnP_7	LPPSLHGPEPKG	0.411788
GATA4	NR5A1	GATA4_SnP_8	YVLPPSLHGPEP	0.41301522

Supplementary Table 2. AlphaFold2-Multimer+PeptiDerive-derived peptide sequences. Eight peptides were built per target protein based on specific binders using AlphaFold2-Multimer + PeptiDerive (AF2 + Peptiderive) and PeptiDerive only. These were cloned into uAb constructs for subsequent analysis via flow cytometry.

Target	Binder	Method	Peptide Name	Sequence
DNAJB1-PRKACA	HSPA8	AF2+PeptiDerive	DJ1PKA_AF2_1	AAILSGDKSENVQDLLLL
DNAJB1-PRKACA	HSPA8	AF2+PeptiDerive	DJ1PKA_AF2_2	AILSGDKSENVQDLLLLD
DNAJB1-PRKACA	HSPA8	AF2+PeptiDerive	DJ1PKA_AF2_3	ILSGDKSENVQDLLLLDV
DNAJB1-PRKACA	HSPA8	AF2+PeptiDerive	DJ1PKA_AF2_4	LSGDKSENVQDLLLLDVT
DNAJB1-PRKACA	HSPA8	AF2+PeptiDerive	DJ1PKA_AF2_5	QAAILSGDKSENVQDLLL
DNAJB1-PRKACA	HSPA8	AF2+PeptiDerive	DJ1PKA_AF2_6	VQAAILSGDKSENVQDLL
DNAJB1-PRKACA	HSPA8	AF2+PeptiDerive	DJ1PKA_AF2_7	PRGVPQIEVTFDIDANGI
DNAJB1-PRKACA	HSPA8	AF2+PeptiDerive	DJ1PKA_AF2_8	LLLDVTPLSLGIETAGGV
PNPLA3	ABHD5	AF2+PeptiDerive	PNPLA3_AF_1	LADQDRPIPVWI
PNPLA3	ABHD5	AF2+PeptiDerive	PNPLA3_AF_2	QDRPIPVWIRAL
PNPLA3	ABHD5	AF2+PeptiDerive	PNPLA3_AF_3	DRPIPVWIRALG
PNPLA3	ABHD5	AF2+PeptiDerive	PNPLA3_AF_4	ADQDRPIPVWIR
PNPLA3	ABHD5	AF2+PeptiDerive	PNPLA3_AF_5	DQDRPIPVWIRA
PNPLA3	ABHD5	AF2+PeptiDerive	PNPLA3_AF_6	PIPVWIRALGAA
PNPLA3	ABHD5	AF2+PeptiDerive	PNPLA3_AF_7	RPIPVWIRALGA
PNPLA3	ABHD5	AF2+PeptiDerive	PNPLA3_AF_8	IPVWIRALGAAL
KRAS	6XHA	PeptiDerive	KRAS_PD_1	SNTIRVFLPNKQRTVVNV
KRAS	6XHA	PeptiDerive	KRAS_PD_2	NTIRVFLPNKQRTVVNVR
KRAS	6XHA	PeptiDerive	KRAS_PD_3	TIRVFLPNKQRTVVNVRN
KRAS	6XHA	PeptiDerive	KRAS_PD_4	IRVFLPNKQRTVVNVRNG
KRAS	6XHA	PeptiDerive	KRAS_PD_5	RVFLPNKQRTVVNVRNGM
KRAS	6XHA	PeptiDerive	KRAS_PD_6	VFLPNKQRTVVNVRNGMS
KRAS	6XHA	PeptiDerive	KRAS_PD_7	GFRCQTCGYKFHEHCSTK
KRAS	6XHA	PeptiDerive	KRAS_PD_8	LNGFRCQTCGYKFHEHCS
KRAS (G12V)	RAF1	AF2+PeptiDerive	KRAS_AF2_1	NTIRVFLPNKQRTVVNVR

KRAS (G12V)	RAF1	AF2+PeptiDerive	KRAS_AF2_2	TIRVFLPNKQRTVVNVRN
KRAS (G12V)	RAF1	AF2+PeptiDerive	KRAS_AF2_3	SNTIRVFLPNKQRTVVNV
KRAS (G12V)	RAF1	AF2+PeptiDerive	KRAS_AF2_4	IRVFLPNKQRTVVNVRNG
KRAS (G12V)	RAF1	AF2+PeptiDerive	KRAS_AF2_5	RVFLPNKQRTVVNVRNGM
KRAS (G12V)	RAF1	AF2+PeptiDerive	KRAS_AF2_6	VFLPNKQRTVVNVRNGMS
KRAS (G12V)	RAF1	AF2+PeptiDerive	KRAS_AF2_7	FRCQTCGYKFHEHCSTKV
KRAS (G12V)	RAF1	AF2+PeptiDerive	KRAS_AF2_8	GFRCQTCGYKFHEHCSTK
β-catenin	CADH1	AF2+PeptiDerive	β-cat_AF_1	EWGNRFKKLADM
β-catenin	CADH1	AF2+PeptiDerive	β-cat_AF_2	LNEWGNRFKKLA
β-catenin	CADH1	AF2+PeptiDerive	β-cat_AF_3	NEWGNRFKKLAD
β-catenin	CADH1	AF2+PeptiDerive	β-cat_AF_4	YLNEWGNRFKKL
β-catenin	CADH1	AF2+PeptiDerive	β-cat_AF_5	WGNRFKKLADMY
β-catenin	CADH1	AF2+PeptiDerive	β-cat_AF_6	RFKKLADMYGGG
β-catenin	CADH1	AF2+PeptiDerive	β-cat_AF_7	NRFKKLADMYGG
β-catenin	CADH1	AF2+PeptiDerive	β-cat_AF_8	GNRFKKLADMYG
DLX5	MAGED1	AF2+PeptiDerive	DLX5_AF_1	LLTWDEEGDFGD
DLX5	MAGED1	AF2+PeptiDerive	DLX5_AF_2	ELLTWDEEGDFG
DLX5	MAGED1	AF2+PeptiDerive	DLX5_AF_3	WARYHQNARSRF
DLX5	MAGED1	AF2+PeptiDerive	DLX5_AF_4	FWARYHQNARSR
DLX5	MAGED1	AF2+PeptiDerive	DLX5_AF_5	TFWARYHQNARS
DLX5	MAGED1	AF2+PeptiDerive	DLX5_AF_6	FTFWARYHQNAR
DLX5	MAGED1	AF2+PeptiDerive	DLX5_AF_7	PFTFWARYHQNA
DLX5	MAGED1	AF2+PeptiDerive	DLX5_AF_8	IPFTFWARYHQN
GATA4	NR5A1	AF2+PeptiDerive	GATA4_AF_1	LQLLALQLDRQE
GATA4	NR5A1	AF2+PeptiDerive	GATA4_AF_2	LLALQLDRQEFV
GATA4	NR5A1	AF2+PeptiDerive	GATA4_AF_3	QLLALQLDRQEF
GATA4	NR5A1	AF2+PeptiDerive	GATA4_AF_4	QELVLQLLALQL
GATA4	NR5A1	AF2+PeptiDerive	GATA4_AF_5	ELVLQLLALQLD
GATA4	NR5A1	AF2+PeptiDerive	GATA4_AF_6	LVLQLLALQLDR

GATA4	NR5A1	AF2+PeptiDerive	GATA4_AF_7	VLQLLALQLDRQ
GATA4	NR5A1	AF2+PeptiDerive	GATA4_AF_8	LALQLDRQEFVC