

Target sequence-conditioned design of peptide binders using masked language modeling

In the format provided by the
authors and unedited

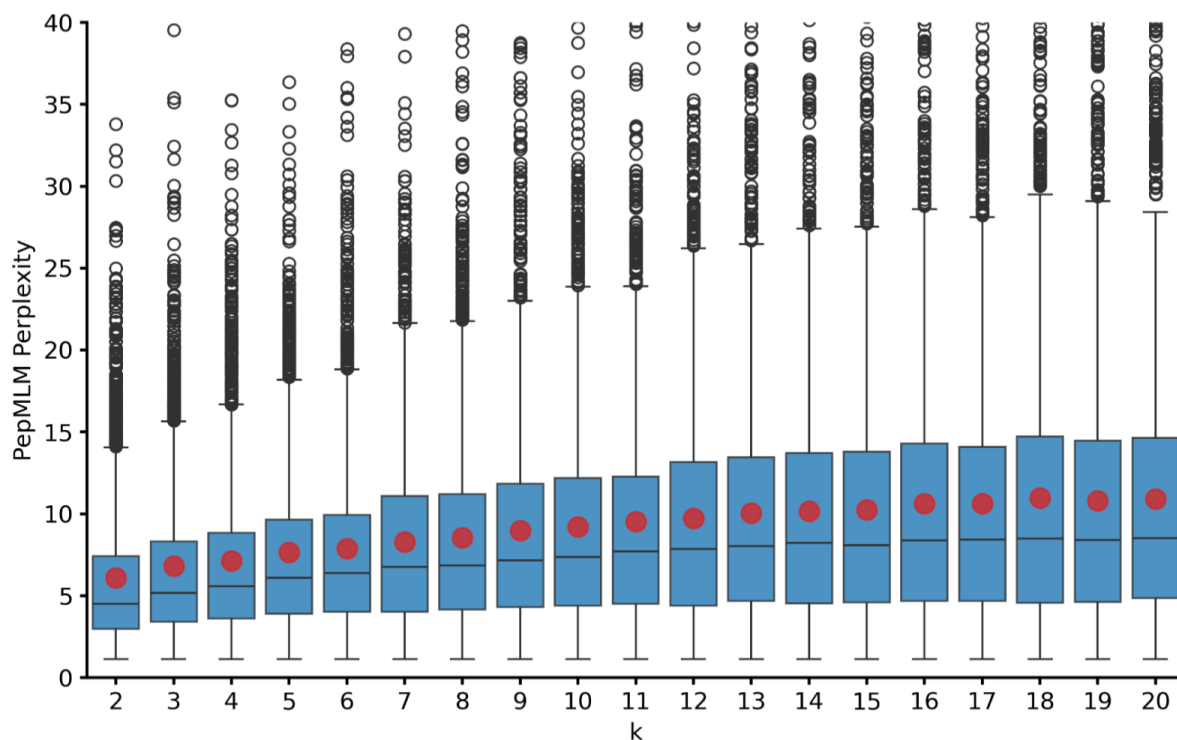
Table of Contents

Supplementary Figures

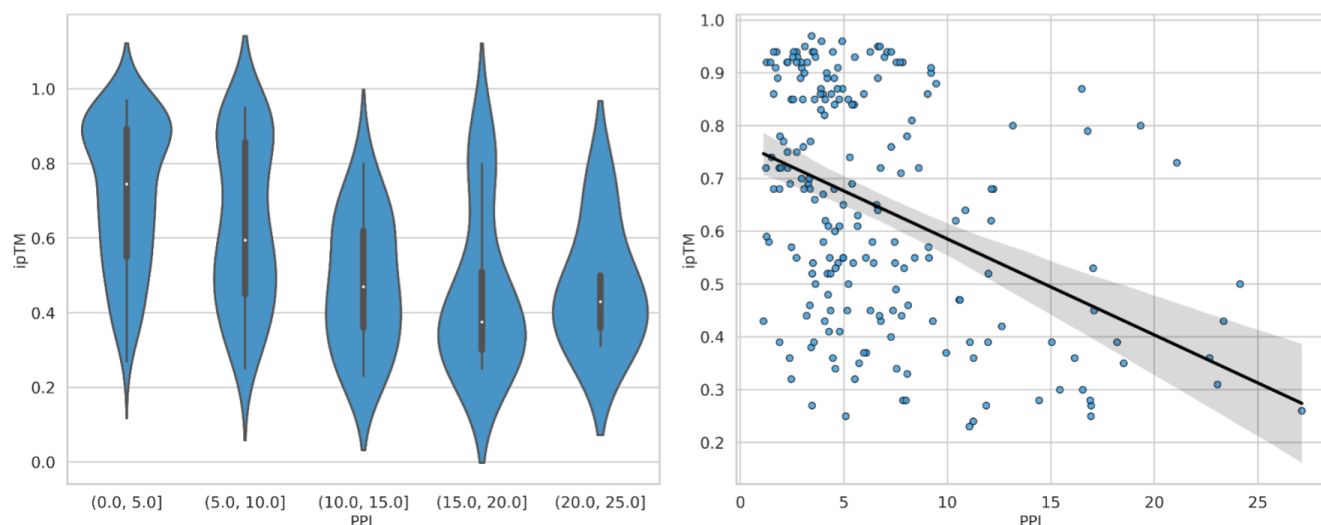
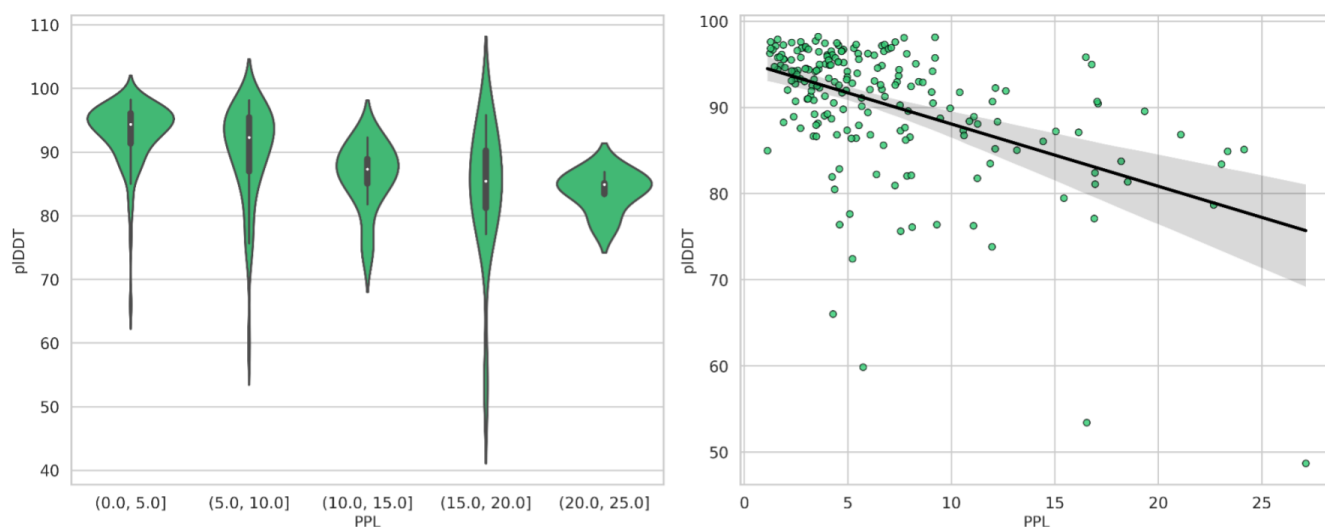
1. Top- k sampling and perplexity.
2. Association between model perplexity and co-folding metrics.
3. Benchmark RFDiffusion and PepMLM with high pLDDT filter.
4. Visualization of binder-protein complexes.
5. Amino acid specific generation matrix for all positions and contact positions.
6. Identity, ipTM, and PPL scatter plot.
7. Representative coomassie-stained sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) of anti-AMHR2 SUMO-tag peptide constructs.
8. Preliminary ELISA screen of anti-AMHR2 binders and anti-NCAM1 peptide binders.
9. PepMLM-generated uAb degraders of MESH1.
10. Evaluation of PepMLM-3B.

Supplementary Tables

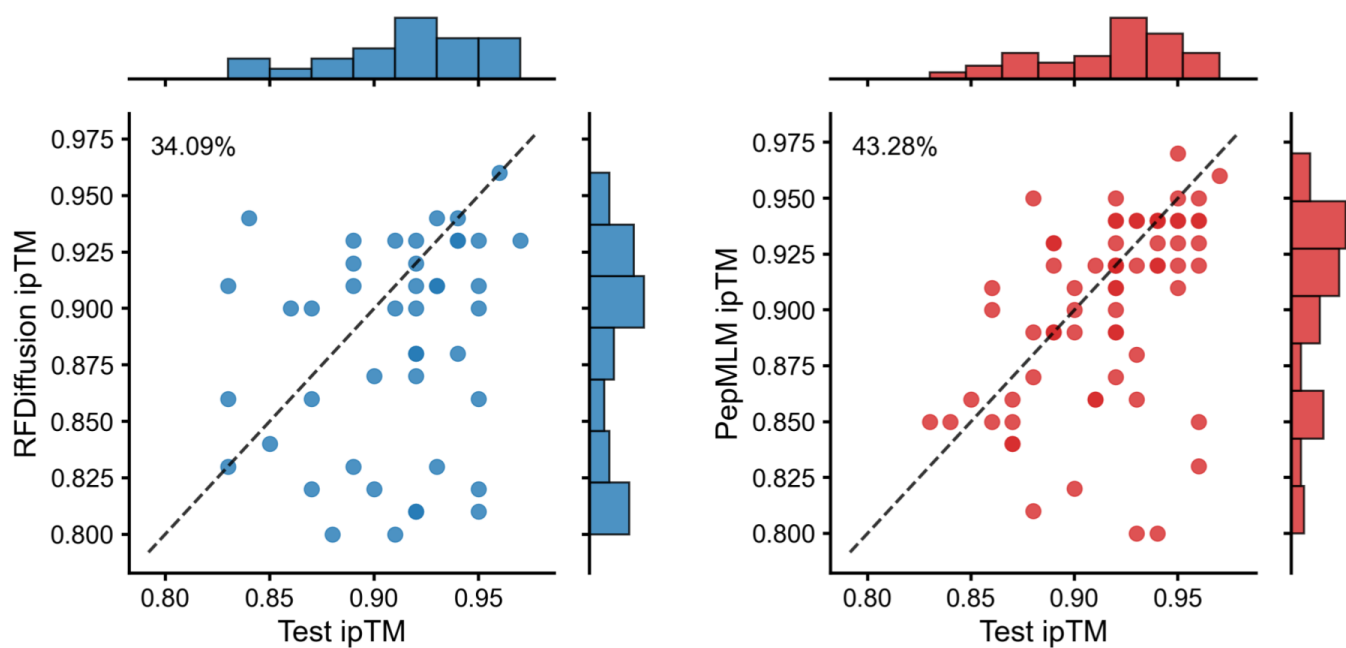
1. Settings and hyperparameters used to train PepMLM-650M.
2. Outlier analysis of protein-peptide complexes.
3. Sequence information and folding metrics for complexes in Supplementary Figure 4.
4. Peptide sequences and PPL scores for experimental testing.
5. RFDiffusion parameter setting used for generation of peptides for Figure 1.



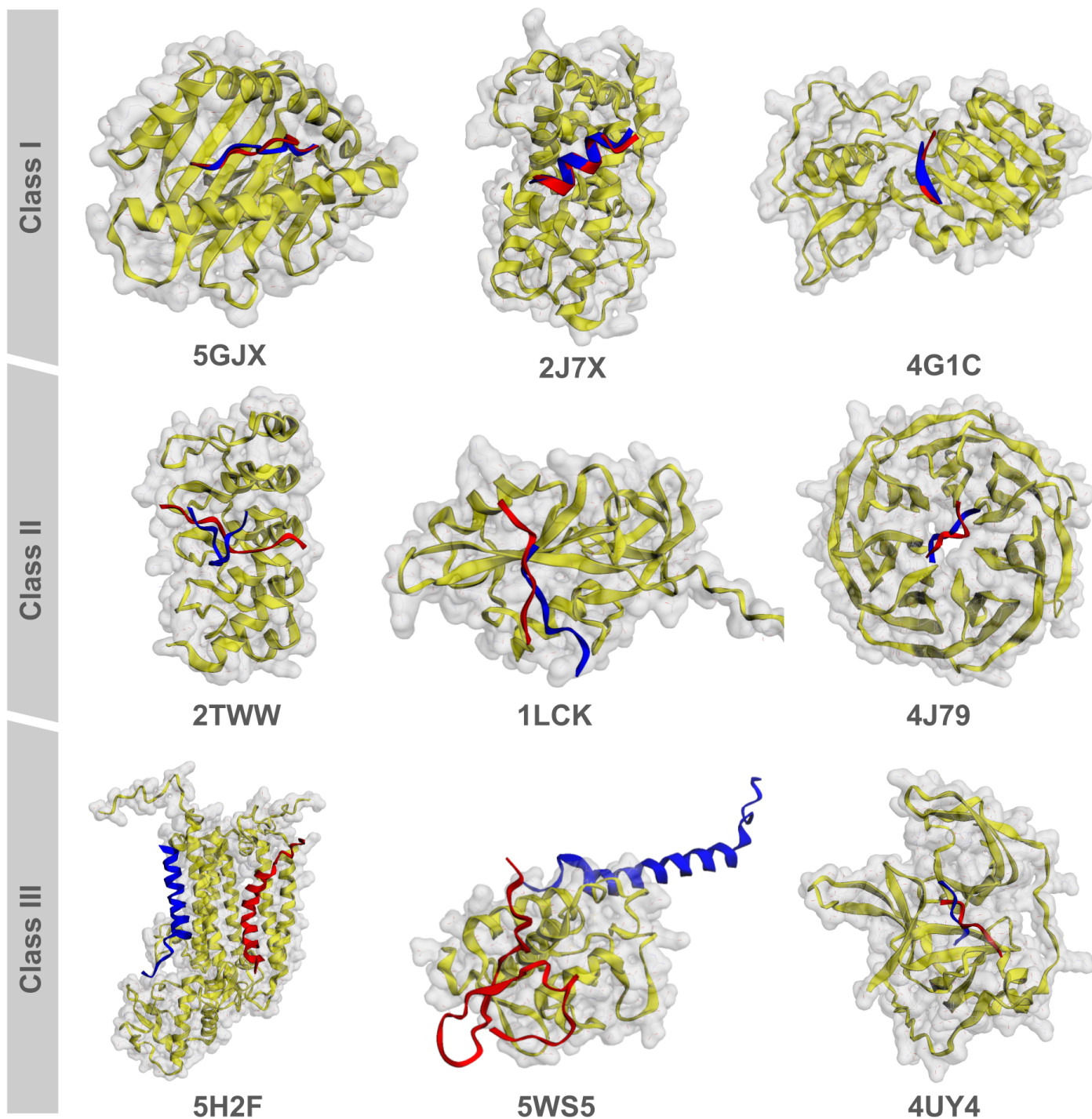
Supplementary Figure 1. Top- k sampling and perplexity. This figure illustrates the correlation between selected k values and the resulting Perplexity when generating binder sequences for target proteins, where each k value corresponds to the generation of 100 binders with lengths equating to the reference binder. The red point refers to the average PPL score. As k escalates, there is a corresponding increase in Perplexity, indicating reduced model confidence. At $k=3$, the model maintains a lower Perplexity, signifying higher reliability in its predictions. The choice of $k=3$ is supported not only by this enhanced assurance but also by considerations of diminishing returns; beyond this point, the gain in diversity is outweighed by the increase in complexity and the potential for atypical predictions. Moreover, practical considerations favor $k=3$, as it offers computational efficiency without compromising the diversity necessary for effective binder design. Consequently, $k=3$ is endorsed as the optimal value to achieve a balance between model confidence, diversity of binder sequences, and computational pragmatism.

A**B**

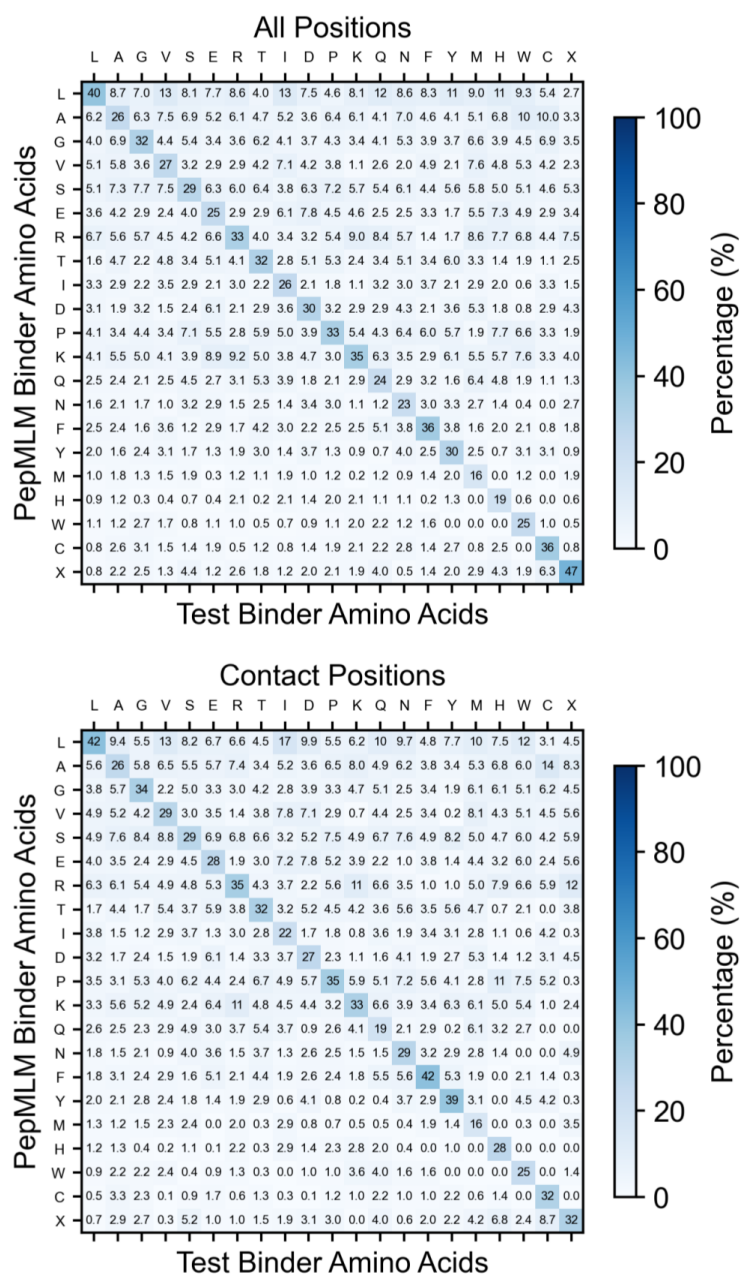
Supplementary Figure 2. Association between model perplexity and co-folding metrics. (A) Relationship between ipTM and perplexity (PPL). The initial segment of Figure A presents a violin plot, categorizing perplexity in 5-unit intervals. The subsequent segment delineates the raw data points, accompanied by a regression analysis, indicating a negative correlation (Pearson correlation coefficient -0.414, $p < 0.001$). The shaded area represents a 95% confidence interval. (B) Negative correlation between PPL and pLDDT, identified by Pearson correlation coefficient of -0.490 ($p < 0.001$). The violin plot underscores a marked decrement in specific folding metrics, most pronounced in ipTM, commensurate with elevated perplexity levels.



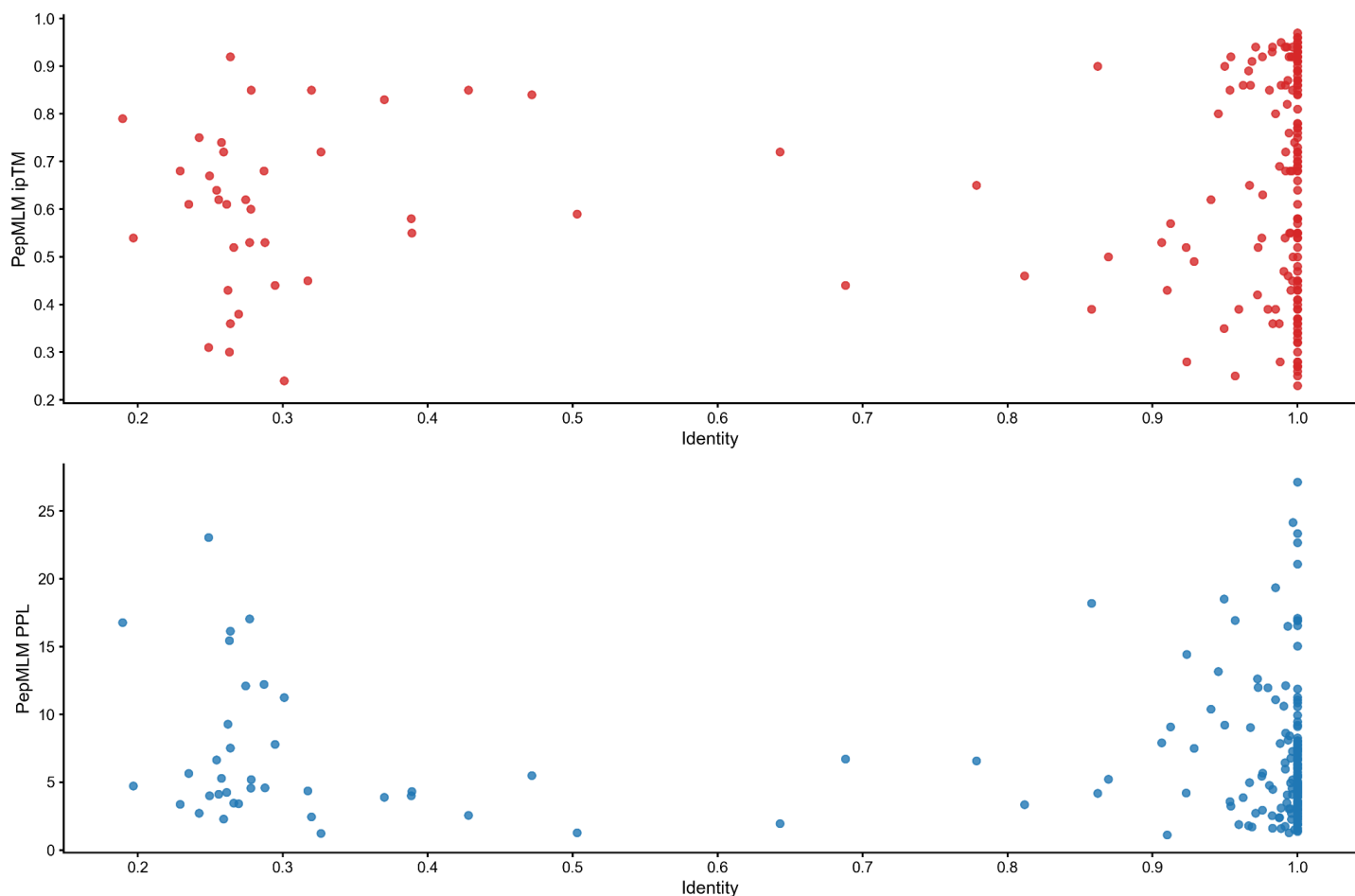
Supplementary Figure 3. Benchmark RFDiffusion and PepMLM with high pLDDT filter. Hit rate comparison between PepMLM and RFDiffusion generated binders after applying pLDDT filter (> 0.8). PepMLM achieved a 43% hit rate compared to RFDiffusion's 34%.



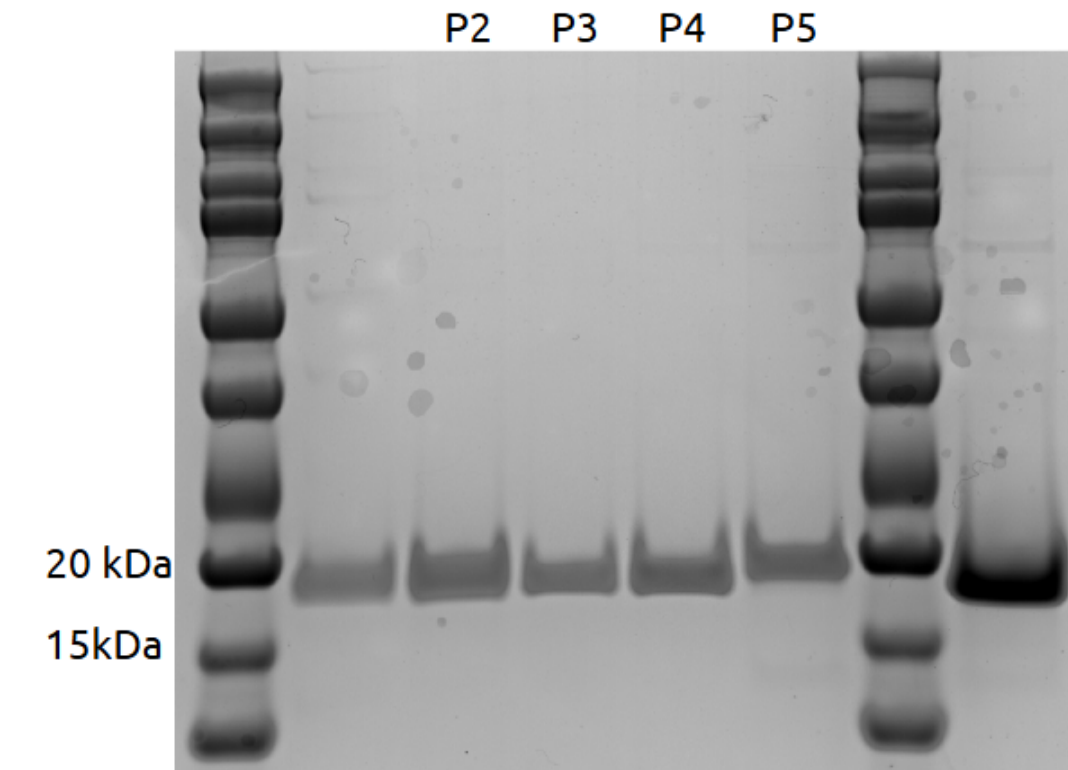
Supplementary Figure 4. Visualization of binder-protein complexes. Co-folded binder-protein complexes are categorized into three distinct classes for visualization purposes. Class I includes complexes where both the generated and test binders exhibit ipTM scores ≥ 0.7 , Class II encompasses those with generated binders having ipTM scores ≥ 0.7 and test binders with ipTM scores < 0.7 , and Class III contains complexes with both generated and test binders having ipTM scores ≤ 0.7 . In these representations, the target protein is depicted in yellow, while the PepMLM-generated binders and test binders are illustrated in red and blue, respectively. This classification facilitates a detailed comparison of the structural relationships and binding patterns among the different classes of binder-protein complexes.



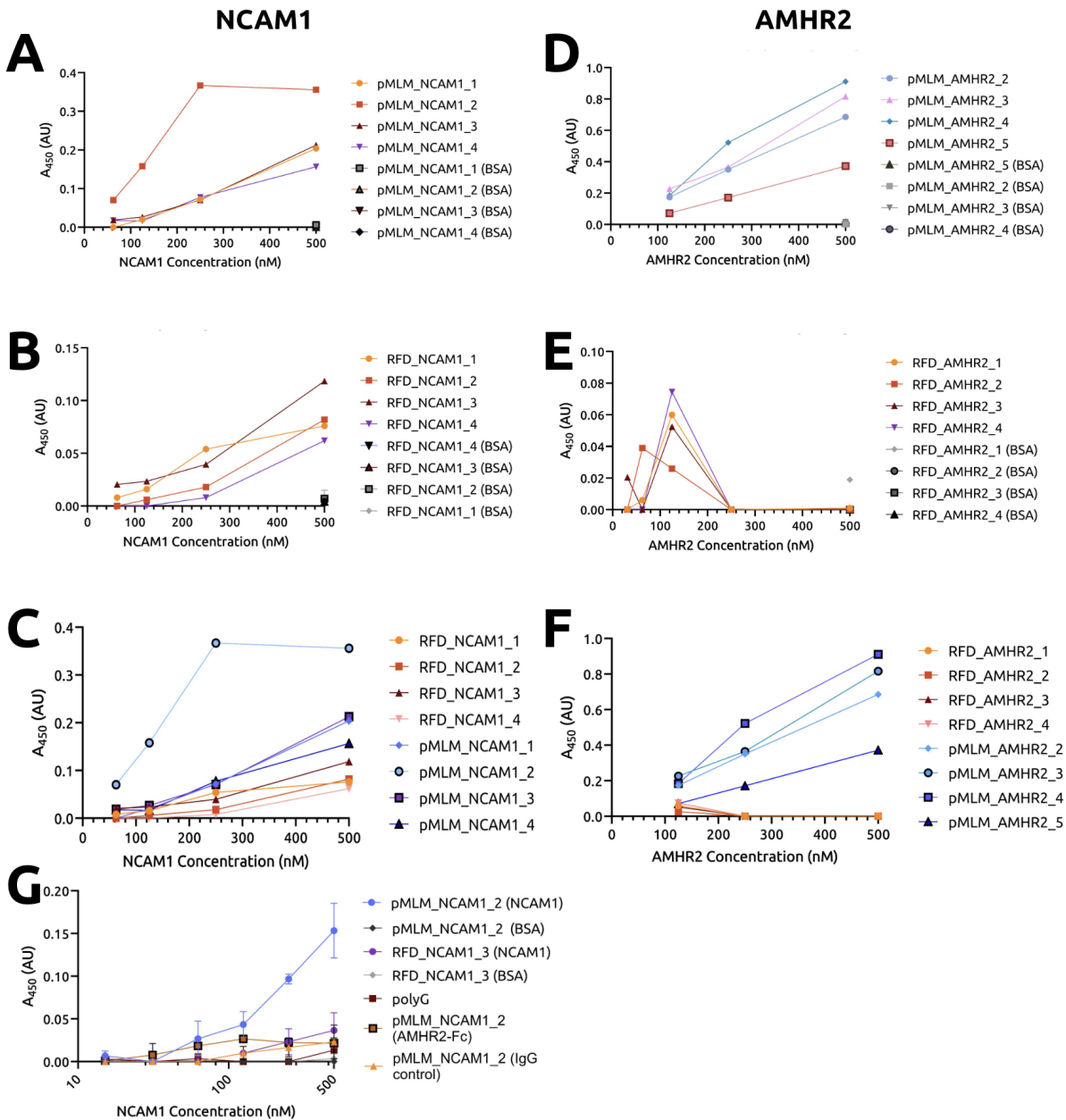
Supplementary Figure 5. Amino acid specific generation matrix for all positions and contact positions. (Top) Generation percentage across all positions for each AA, with rows representing amino acids in test binders and columns showing corresponding amino acids in generated binders (n=100 generations per test binder). (Bottom) Generation percentage specifically for contact positions, defined using an 8Å distance threshold. Color intensity indicates substitution frequency, with diagonal elements representing conservation of original amino acids. Position-specific changes of 69.2% and 68.4% were observed for all positions and contact positions, respectively.



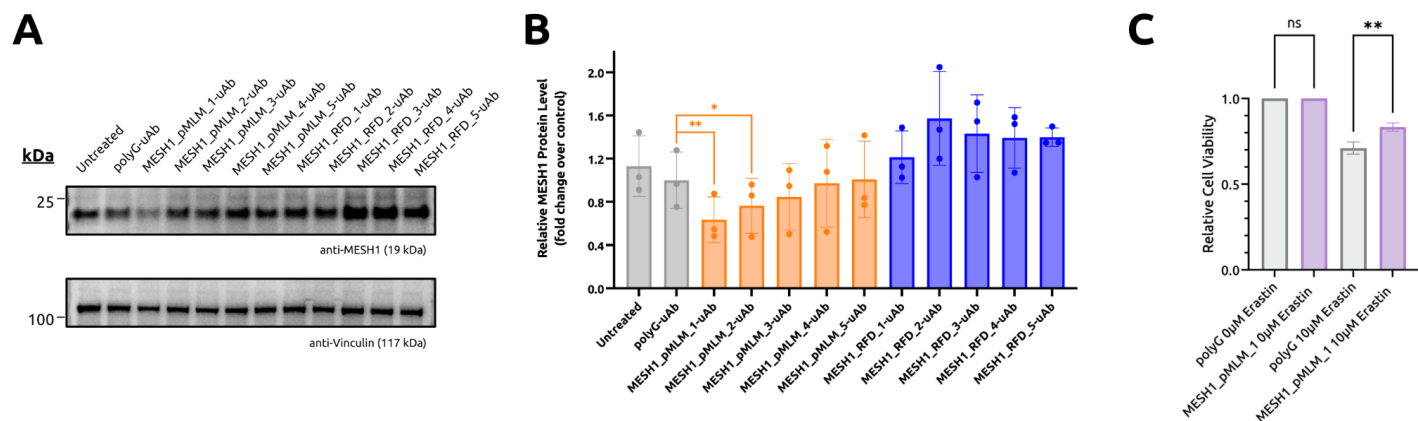
Supplementary Figure 6. Identity, ipTM, and PPL scatter plot. (Top) Relationship between maximum sequence identity of test targets against training set and AlphaFold-predicted ipTM scores of generated binders. The x-axis shows the maximum sequence identity computed using biotite Alignment, while the y-axis represents the predicted ipTM scores for PepMLM-generated binders. (Bottom) Relationship between maximum sequence identity and PPL scores. The x-axis maintains the same sequence identity metric, while the y-axis shows the PPL scores for generated binders. Scatter points represent individual target-binder pairs.



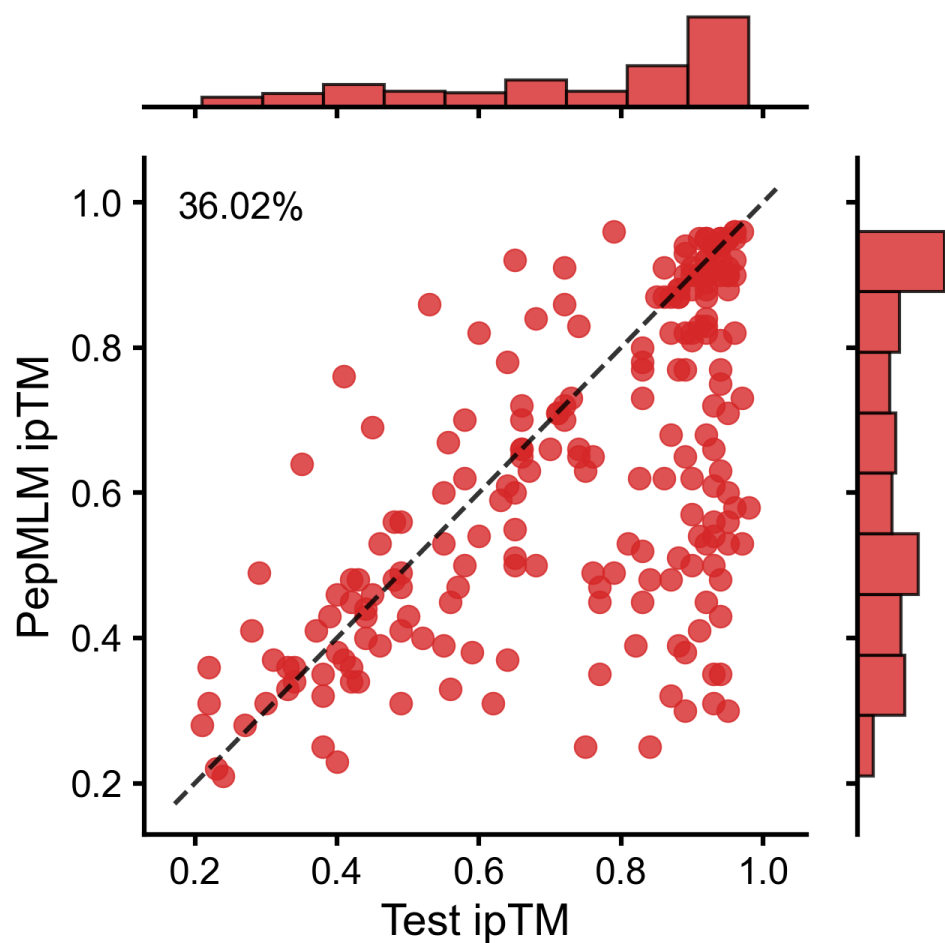
Supplementary Figure 7. Representative coomassie-stained sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) of anti-AMHR2 SUMO-tag peptide constructs. While constructs have a predicted molecular weight of 14.2 kDa based on amino acid sequence, the addition of a SUMO-tag fusion results in a higher apparent molecular weight of around 20 kDa.



Supplementary Figure 8. Preliminary ELISA screen of anti-AMHR2 binders and anti-NCAM1 peptide binders. Four NCAM1-targeting peptides (N were generated for both RFDiffusion and PepMLM as C-terminal SUMO-tag fusions and screened for potential binding to NCAM1-Fc and AMHR2-Fc by ELISA (Supplementary Figure 8). Limited dilutions of each antigen were tested and 500 nM BSA was used as a control antigen. (A) Binding analysis of four potential PepMLM-generated anti-NCAM1 peptides (B) Binding analysis of four potential RFDiffusion-generated anti-NCAM1 peptides (C) PepMLM and RFDiffusion screens from A and B plotted on the same axes. (D) Binding analysis of four potential PepMLM-generated anti-AMHR2 peptides (E) Binding analysis of four potential RFDiffusion-generated anti-AMHR2 peptides. (F) PepMLM and RFDiffusion screens from D and E plotted on the same axes. (G) Binding analysis of top candidate peptide against NCAM1 for each algorithm fused to C-terminal SUMO-tag as measured by ELISA. PepMLM-designed peptides were immobilized onto 96-well plates, incubated with serial dilutions of either NCAM1-Fc or BSA, and subsequently detected with an anti-IgG-HRP conjugate. An anti-NCAM1 antibody and AMHR2-Fc was used as IgG control. For preliminary screens (A-F). Data are single point measurements (besides blanks) and each absorbance signal is subtracted by the average of its associated blank values (in duplicate). Axes for subplots A-F are linear, the axis of subplot G is log(concentration). No outliers were removed.



Supplementary Figure 9. PepMLM-generated uAb degraders of MESH1. A) Western blot analysis of HeLa cell lysate post-transfection of uAb vectors. polyG-uAb was used as a transfection control and Vinculin was used as the loading control. B) Densitometry analysis of accompanying Western blots, normalized to polyG and Vinculin controls. Statistical significance is determined by a paired one-tailed Student's t test. Calculated p values are represented as follows: **p < 0.01; ***p < 0.001). C) Inhibition of ferroptosis-related cell death upon transfection of anti-MESH1 uAbs into HEK293T cells in the presence of erastin.



Supplementary Figure 10. Evaluation of PepMLM-3B. *In silico* hit-rate assessment. Utilizing AlphaFold-Multimer, the ipTM scores were computed for both the generated and test peptides in conjunction with the target protein sequence. The entries are organized in accordance with the ipTM scores attributed to the test set peptides. The hit rate is characterized by the generated peptides exhibiting ipTM scores $\geq$ those of the test set peptides.

Supplementary Table 1. Settings and hyperparameters used to train PepMLM-650M.

Name	Setting
Learning Rate	0.000798
Batch Size	16
Gradient Accumulation	2
Warm up steps	501
Number of training epochs	5
Optimizer	AdamW (default)
Trainer	Default HuggingFace Settings

Supplementary Table 2. Outlier analysis of protein-peptide complexes. This table displays 12 protein-peptide complexes with pseudo-perplexity (PPL) values exceeding 40. Included are evaluation metrics for both the test complexes and the PepMLM generation results, as well as the binder sequences. ipTM scores for test and generated complexes are highlighted in different colors for comparison. Notably, even though these outliers exhibit high PPL values indicative of accurate modeling by PepMLM, the model remains proficient in generating binders that perform equivalently well *in silico* as per AlphaFold-Multimer ipTM score.

	Test				PepMLM Generation			
PDB ID	Binder	PPL	ipTM	pLDDT	Binder	PPL	ipTM	pLDDT
5B5V	FLFGSRSS	42.8	0.45	88.9	YHYVMRYA	4.2	0.52	88.7
4G1C	AVXCAX	82.5	0.86	97.0	TAKXST	3.0	0.91	96.9
2L1C	RAKWDTANNPLXKE ATSTFTNITXRGT	45.4	0.49	61.9	HIAEEPHFFESMQN NYEKPTTYKFQQK	12.0	0.39	73.8
6GHJ	FAQ	209.7	0.71	93.4	MXL	3.4	0.68	93.3
6AMU	MMWDRGLGMM	59.2	0.34	87.9	YQALIGGFNA	14.4	0.28	86.1
5WMR	QIKVRVDMV	76.3	0.91	92.3	LRFWRARTL	9.0	0.86	91.8
5NJC	VLEDRI	63.0	0.83	97.5	AAAAAA	1.5	0.74	97.1
5FML	LSNDISQGIKRQRM TVESM	42.8	0.91	94.0	AAMTKLALAAKTRA QLFKK	16.1	0.36	87.1
6DQU	GIINTL	65.8	0.87	97.7	YLGANG	5.4	0.84	97.3
2IWB	GHMS	194.0	0.64	96.1	XPPX	4.0	0.67	95.9
4MLI	AHIVMVDAYKPT	62.6	0.87	97.5	GPTPVQVLKRRG	17.0	0.53	90.7
5DHM	RSIEISIRVDDFTKTG ETVRY	64.2	0.93	94.4	AQSPEITADVVTST DEFTTT	19.3	0.8	89.6

Supplementary Table 3. Sequence information and folding metrics for complexes in Supplementary Figure 4.

PDB ID	Generated Binder	pIDDT	ipTM	Test Binder	pIDDT	ipTM	Target Identity	Binder Identity
5GJX	RLLEWMIYI	96.3	0.92	RLIQNSITI	96.0	0.92	0.99	0.44
2J7X	HHLLLHLLTQD	91.9	0.92	IQSLINLLADN	91.9	0.91	0.99	0.28
4G1C	TAKXST	96.9	0.91	AVXCAX	97.0	0.86	0.99	0.15
3TWW	RREPPGGAFRX	97.4	0.87	RQSPDGQSFRX	92.7	0.48	0.99	0.55
1LCK	PPXEEIPP	87.3	0.92	EGQQPQPA	86.1	0.68	0.60	0.13
4J79	AARHLD	97.3	0.72	EKVHVQ	97.2	0.64	0.99	0.17
5H2F	XETNTLVRYVVA HFVLLVSVILIREA PRIESSKXX	84.9	0.43	XETITYVFIFACIAL FFFAIFFREPPRIT XXXXX	86.7	0.27	0.93	0.4
5WS5	SSEEGRPILWIAT TTGGGGVIVLFL FYAYYGSLSLXX X	77.7	0.24	MSEGGRIPLWIVA TVAGMGVIVIVGLF FYGAYAGLGSSLX X	86.4	0.24	0.99	0.48
4UY4	ARTKQT	90.1	0.63	ARTXQT	89.0	0.43	0.97	0.83

Supplementary Table 4. Peptide sequences and PPL scores for experimental testing. The accession IDs for NCAM1, MSH3, and HTT are from UniProt. Viral target sequences are derived from NCBI/GenBank accession IDs.

Target	Accession ID	uAb Name	Peptide Sequence	Derivation Algorithm	PPL
Figure 2					
NCAM1	P13591	NCAM1_pMLM_1	GKLPLPSLPCK	PepMLM	5.343684945
NCAM1	P13591	NCAM1_pMLM_2	GLGPSPVLPRC	PepMLM	5.299030403
NCAM1	P13591	NCAM1_pMLM_3	GLGPLPVLPCCK	PepMLM	6.329989946
NCAM1	P13591	NCAM1_pMLM_4	HSLGQPLSPIC	PepMLM	4.753515747
NCAM1	P13591	NCAM1_RFD_1	SLPIENIYIEA	RFDiffusion	N/A
NCAM1	P13591	NCAM1_RFD_2	MKPIEVVYEKA	RFDiffusion	N/A
NCAM1	P13591	NCAM1_RFD_3	ELPEQVIYIEA	RFDiffusion	N/A
NCAM1	P13591	NCAM1_RFD_4	EKPIEVIYEKA	RFDiffusion	N/A
AMHR2	Q16671	pMLM_AMHR2_2	LTRYTSLAAQGC	PepMLM	13.09
AMHR2	Q16671	pMLM_AMHR2_3	DTTLYSGFAQYG	PepMLM	13.48
AMHR2	Q16671	pMLM_AMHR2_4	DESLRSFLAHYC	PepMLM	15.31
AMHR2	Q16671	pMLM_AMHR2_5	LETYRSGLAQYC	PepMLM	16.09
AMHR2	Q16671	RFD_NCAM_1	APAVTGSILADL	RFDiffusion	N/A
AMHR2	Q16671	RFD_NCAM_2	SAAVTGSILADL	RFDiffusion	N/A
AMHR2	Q16671	RFD_NCAM_3	RSIRLSYPVELP	RFDiffusion	N/A
AMHR2	Q16671	RFD_NCAM_4	RSVRLVHPVELP	RFDiffusion	N/A
Figure 3					
MSH3	P20585	MSH3_pMLM_1	SRREQLARILEGAFLA SK	PepMLM	7.65
MSH3	P20585	MSH3_pMLM_2	SRLEESASAMEASAA QAS	PepMLM	9.12
MSH3	P20585	MSH3_pMLM_3	SRLKQAKSIMGGSLLL AE	PepMLM	9.61
MSH3	P20585	MSH3_pMLM_4	NRLVEALASLEFSAQL SE	PepMLM	9.91
MSH3	P20585	MSH3_pMLM_5	SRNKELKSILEFSLAQ QK	PepMLM	11
MSH3	P20585	MSH3_pMLM_6	SRLKQLASALDGSFLQ AS	PepMLM	11.84

MSH3	P20585	MSH3_pMLM_7	SLRKELASAMEFAAAQ SK	PepMLM	12.24
MSH3	P20585	MSH3_pMLM_8	SLNEQAASILEAFFAQ SS	PepMLM	13.47
MSH3	P20585	MSH3_pMLM_9	SYNVELASISEASLAAA K	PepMLM	13.69
MSH3	P20585	MSH3_pMLM_10	SLNEQLASIMGGSACL AE	PepMLM	14.11
MSH3	P20585	MSH3_pMLM_11	SRRVELLSILEFALAAA S	PepMLM	14.45
HTT (Q43)	P42858	Q43_pMLM_1	SAAPQLLGSGLAGLK	PepMLM	Q43: 5.302052023 Q17: 21.74186374
HTT (Q43)	P42858	Q43_pMLM_2	TAPQLSQASGLAGGK	PepMLM	Q43: 6.223211251 Q17: 22.05928384
HTT (Q43)	P42858	Q43_pMLM_3	LAPQLLLLGLGGLAK	PepMLM	Q43: 6.122010055 Q17: 21.22001317
HTT (Q43)	P42858	Q43_pMLM_4	SAPPQLAAAGGLLLA	PepMLM	Q43: 5.42110285 Q17: 19.667137
HTT (Q43)	P42858	Q43_pMLM_5	SPPPQAAAGAALGAK	PepMLM	Q43: 6.196895246 Q17: 20.36823003
Figure 4					
Hendravirus P	MN062017.1	HeV_1	RLPVYLSLQG	PepMLM	3.635173
Hendravirus P	MN062017.1	HeV_2	HSPVHLSLLG	PepMLM	4.296802
Hendravirus P	MN062017.1	HeV_3	SRSVLHSLQGR	PepMLM	4.974103
Hendravirus P	MN062017.1	HeV_4	HSSVLQSLFGG	PepMLM	5.449664
Hendravirus P	MN062017.1	HeV_5	SESLYLSLFGK	PepMLM	6.702855
Hendravirus P	MN062017.1	HeV_6	SMSRRRQLAKKLLLLA IKS	PepMLM	6.707319
Hendravirus P	MN062017.1	HeV_7	RQSLRQQLLLDLGR	PepMLM	6.74773
Hendravirus P	MN062017.1	HeV_8	SSLVYLSLGA	PepMLM	6.824329
Hendravirus P	MN062017.1	HeV_9	HLSLPHSLQKR	PepMLM	6.850688
Hendravirus P	MN062017.1	HeV_10	SMSVEKSLSKKLGGKL IKS	PepMLM	6.855796
Hendravirus P	MN062017.1	HeV_11	RSLVKKQLLLKLLG	PepMLM	6.950253
Hendravirus P	MN062017.1	HeV_12	MQSVKLKLLKGLLR	PepMLM	7.039044
Hendravirus P	MN062017.1	HeV_13	HSSVLQQLFGE	PepMLM	7.046471
Hendravirus P	MN062017.1	HeV_14	HHSLLQSLQGT	PepMLM	7.287729

Hendravirus P	MN062017.1	HeV_15	RLPLYLSLGA	PepMLM	7.372559
Hendravirus P	MN062017.1	HeV_16	HHSLLHSLLKGT	PepMLM	7.621191
Hendravirus P	MN062017.1	HeV_17	RLSVLQQLKLGG	PepMLM	7.758791
Hendravirus P	MN062017.1	HeV_18	SLSRRQQLLDLGK	PepMLM	7.797618
Hendravirus P	MN062017.1	HeV_19	SKPLYLLLGG	PepMLM	7.82979
Hendravirus P	MN062017.1	HeV_20	RLSVRKLLLLDLGK	PepMLM	7.860104
HMPV P	AAS22075.1	HMPV_1	LTLEQLQEKR	PepMLM	6.510858
HMPV P	AAS22075.1	HMPV_2	TLEEELLLKR	PepMLM	7.514123
HMPV P	AAS22075.1	HMPV_3	LTLEQLQLIR	PepMLM	7.855078
HMPV P	AAS22075.1	HMPV_4	AELLLRQQQLL	PepMLM	7.869746
HMPV P	AAS22075.1	HMPV_5	SVLTAEQLIKI	PepMLM	8.039237
HMPV P	AAS22075.1	HMPV_6	DLRRRLAEKTPELQLL LI	PepMLM	8.068756
HMPV P	AAS22075.1	HMPV_7	ALLAKKLTLEALLAL	PepMLM	8.235918
HMPV P	AAS22075.1	HMPV_8	AEEAKKLTEELLRLR	PepMLM	8.411599
HMPV P	AAS22075.1	HMPV_9	DTELAAKKLTTELLKI	PepMLM	8.626335
HMPV P	AAS22075.1	HMPV_10	TLTLQQLLKL	PepMLM	8.713968
HMPV P	AAS22075.1	HMPV_11	ADRLAALSLETRQALL KI	PepMLM	9.022841
HMPV P	AAS22075.1	HMPV_12	LLRLELSKLTAEALLI	PepMLM	9.046337
HMPV P	AAS22075.1	HMPV_13	KTLTEQQELLR	PepMLM	9.120272
HMPV P	AAS22075.1	HMPV_14	RLELTLEEQARLKL	PepMLM	9.250608
HMPV P	AAS22075.1	HMPV_15	PQRLRALELELRQQR RLL	PepMLM	9.263967
HMPV P	AAS22075.1	HMPV_16	EADALARLELLEAEAR RLI	PepMLM	9.290414
HMPV P	AAS22075.1	HMPV_17	LLELELKKLTALAAKLI	PepMLM	9.432891
HMPV P	AAS22075.1	HMPV_18	ELELALEQLLR	PepMLM	9.467909
HMPV P	AAS22075.1	HMPV_19	PTEALAKLPTREALRK R	PepMLM	9.484423
HMPV P	AAS22075.1	HMPV_20	DLERLAASVLLLEEALR KR	PepMLM	9.516725
Nipah virus P	AY029767.1	NiV_1	GKLLHHLLQGG	PepMLM	4.69
Nipah virus P	AY029767.1	NiV_2	SSKLLQLLGN	PepMLM	5.02

Nipah virus P	AY029767.1	NiV_3	HHSLLHQLGLEG	PepMLM	5.20
Nipah virus P	AY029767.1	NiV_4	HSSLLQSLLLGG	PepMLM	5.31
Nipah virus P	AY029767.1	NiV_5	MQSVRLQLLLESLLK	PepMLM	5.38
Nipah virus P	AY029767.1	NiV_6	SHSLLQSLLQQG	PepMLM	5.44
Nipah virus P	AY029767.1	NiV_7	SSSLLHQLLQQG	PepMLM	5.66
Nipah virus P	AY029767.1	NiV_8	GHSLLQSLLLQG	PepMLM	5.80
Nipah virus P	AY029767.1	NiV_9	HSSLLLSLLQEG	PepMLM	5.88
Nipah virus P	AY029767.1	NiV_10	HSLLLHQLLKKG	PepMLM	5.90
Nipah virus P	AY029767.1	NiV_11	SELLHRLLLEN	PepMLM	6.08
Nipah virus P	AY029767.1	NiV_12	HHSLLHSLLQEG	PepMLM	6.19
Nipah virus P	AY029767.1	NiV_13	HSSVLQLLLLEG	PepMLM	6.44
Nipah virus P	AY029767.1	NiV_14	GESVLQQLLQQGG	PepMLM	6.50
Nipah virus P	AY029767.1	NiV_15	SSSIPKQLRLESLLK	PepMLM	6.69
Nipah virus P	AY029767.1	NiV_16	HESLLQLLLEQG	PepMLM	6.69
Nipah virus P	AY029767.1	NiV_17	MSSVRKQLLLLLGLR	PepMLM	6.75
Nipah virus P	AY029767.1	NiV_18	SSLLLQLLLQN	PepMLM	6.79
Nipah virus P	AY029767.1	NiV_19	SSSLLQLLGQQG	PepMLM	6.84
Nipah virus P	AY029767.1	NiV_20	PSSVLLQLGLKE	PepMLM	6.93

Supplementary Table 5. RFDiffusion parameter setting used for generation of peptides for Figure 1.

Parameter	Setting
PDB	7L0J:B
Contigs	B16-124:12-12
Copies	1
Num_seqs	4
Initial guess	True
Use_multimer	False
Num_recycles	3
Rm_aa	C
Num_designs	4
Mpnn_sampling_temp	0.1