PEPMLM: TARGET SEQUENCE-CONDITIONED GENERATION OF PEPTIDE BINDERS VIA MASKED LANGUAGE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Target proteins that lack accessible binding pockets and conformational stability have posed increasing challenges for drug development. Induced proximity strategies, such as PROTACs and molecular glues, have thus gained attention as pharmacological alternatives, but still require small molecule docking at binding pockets for targeted protein degradation (TPD). The computational design of protein-based binders presents unique opportunities to access "undruggable" targets, but have often relied on stable 3D structures or predictions for effective binder generation. Recently, some studies have leveraged the expressive latent spaces of protein language models (pLMs) for the prioritization of peptide binders from sequence alone. However, these methods rely on training discriminator models for ranking peptides. In this work, we introduce PepMLM, a purely target sequence-conditioned *de novo* generator of linear peptide binders. By employing a novel masking strategy that uniquely positions cognate peptide sequences at the terminus of target protein sequences, PepMLM tasks the state-of-the-art ESM-2 pLM to fully reconstruct the binder region, achieving low perplexities matching or improving upon previously-validated peptide-protein sequence pairs. After successful *in silico* benchmarking with AlphaFold-Multimer, we experimentally verify PepMLM's efficacy via fusion of model-derived peptides to E3 ubiquitin ligase domains, demonstrating endogenous degradation of target substrates in cellular models. In total, PepMLM enables the generative design of candidate binders to any target protein, without the requirement of target structure, empowering downstream programmable proteome editing applications.

1 INTRODUCTION

The development of therapeutics largely relies on the ability to design small molecule- or protein-based binders to pathogenic target proteins of interest (Chen et al., 2023). These binders can either be used as inhibitors or as functional recruiters of effector enzymes (Zhong et al., 2021). For example, proteolysis targeting chimeras (PROTACs) or molecular glues are heterobifunctional small molecules that bind and recruit endogenous E3 ubiquitin ligases for targeted protein degradation (TPD) (Békés et al., 2022; Dong et al., 2021). Still, these small molecule-based methods rely on the existence of accessible cryptic or canonical binding sites, which are not present on classically "undruggable" intracellular proteins (Gao et al., 2020; Behan et al., 2019). With the advent of deep learning-based structure prediction tools such as AlphaFold2 (Jumper et al., 2021), combined with generative modeling, algorithms such as RFDiffusion (Watson et al., 2023) and MASIF-Seed (Gainza et al., 2023) enable researchers to conduct de novo protein binder design from target structure alone. Nonetheless, much of the undruggable proteome, including dysregulated proteins such as transcription factors and fusion oncoproteins, are conformationally disordered, thus biasing design to a small subset of disease-related proteins (Behan et al., 2019; Chen et al., 2023).

Over the past few years, deep learning has revolutionized natural language processing (NLP), particularly through the implementation of the attention mechanism (Vaswani et al., 2017). This foundational advancement has transcended the boundaries of natural language analysis, finding pertinent applications in the modeling of other languages, such as proteins, which are fundamentally sequences of amino acids (Ofer et al., 2021). In recent times, several protein language models

(pLMs), trained on distinct transformer architectures, such as ProtT5 (Elnaggar et al., 2021), ProGen2 (Madani et al., 2023), ProtGPT2 (Ferruz et al., 2022), and the ESM series (Rives et al., 2021), have accurately captured critical physicochemical properties of proteins. Notably, ESM-2 (Lin et al., 2023) currently stands as the state-of-the-art model in the realm of protein sequence encoding, essentially functioning as an encoder-only model that discerns co-evolutionary patterns among protein sequences via a masked language modeling (MLM) training task (Devlin et al., 2018). These models have been extended to powerful applications, including antibody design, the creation of novel proteins, and structure prediction, offering a streamlined approach to embedding useful protein information. For peptide binder design, previous studies have used pLMs to identify and screen peptides given target protein (Palepu et al., 2022; Brixi et al., 2023). However, a purely de novo, target sequence-conditioned binder design algorithm has yet to be developed.

To achieve this goal, we introduce PepMLM, a novel **Pep**tide binder design algorithm via **M**asked **L**anguage **M**odeling, built upon the foundations of ESM-2 (Lin et al., 2023). PepMLM innovates by employing a contiguous masking strategy that uniquely positions the entire peptide binder sequence at the terminus of target protein sequences, compelling ESM-2 to reconstruct the entire binding region (Fig.1). PepMLM-derived linear peptides achieve low perplexities, matching or improving upon validated peptide-protein sequence pairs in the test dataset, outperform the state-of-the-art RFDiffusion (Watson et al., 2023) for peptide generation on structured targets *in silico*, and experimentally exhibit degradation capability of endogenous, disordered target substrates when incorporated into the ubiquibody (uAb) architecture (Chatterjee et al., 2020). Overall, by focusing on the complete reconstruction of peptide regions, PepMLM represents the first example of target-conditioned *de novo* binder design from sequence alone, thus facilitating a deeper understanding of binding dynamics and paving the way for the development of more effective, targeted binders to unstructured proteins of interest.

2 METHODS

Data Curation In the data curation phase, protein and peptide complexes were amalgamated from the PepNN and Propedia databases (Abdin et al., 2022; Martins et al., 2023). Initially, redundancy between the two datasets was eliminated, followed by the utilization of MMseqs2 to cluster the remaining protein sequences, setting a threshold of 0.8 (Steinegger & Söding, 2017). When protein sequences were identified within the same cluster and exhibited identical binder sequences, a single sequence was retained. This was followed by a manual filtering process, wherein protein sequences were sorted and those exhibiting high similarity were removed to further mitigate homology issues. Consequently, a dataset comprising 10,203 entries was amassed, from which 10,000 were randomly allocated for training and 203 for testing. The maximum lengths for the binder and protein sequences were established at 50 and 500, respectively.

Conditional Peptide Modeling Peptide binders are modeled in a distinctive manner, wherein the peptides are modeled conditionally based on the full protein sequence. Let $p = (p_1, p_2, p_3, \ldots, p_n)$ represent the target protein sequence of length n and $b = (b_1, b_2, b_3, \ldots, b_m)$ denote the binder of length m. The protein and peptide sequences are concatenated, incorporating special tokens of start, end, and padding. Mask language modeling transforms this into a conditional modeling problem, where the objective is to reconstruct b given p, as the entire b region is masked during both training and generation phases. The entire model is updated with Cross Entropy loss, which can be represented as: $\mathcal{L} = -\sum_{i=1}^{m} b_i \log(\hat{b}_i)$ Through this methodology, the discrepancy between the generated binders and the ground truth is minimized, facilitating the learning of the conditional probability, $\prod_{i=1}^{m} P(b_i|p)$.

PepMLM The pre-trained protein language model, ESM-2 (Lin et al., 2023), was utilized to facilitate full parameter fine-tuning. ESM-2, a transformer-based model, is adept at discerning co-evolutionary patterns across protein sequences. The concatenated protein and peptide sequences were tokenized at the amino acid level and input into the model. Deviating from the original training strategy of ESM, the entire binder sequence was exclusively masked, compelling the model to learn the relationship between the peptide binder and the protein (Fig.1). The ESM-2-650M and ESM-2-3B models were both trained for PepMLM. During the generation phase, the target protein sequence, along with a designated number of mask tokens (at the end), was input into the model.

Subsequently, the model greedily decodes logits at each masked position to identify peptide binders. To infuse greater diversity into the generation process, top k sampling was implemented, wherein the model randomly selects the top k highest probability logits at each masked position (Fig.1).

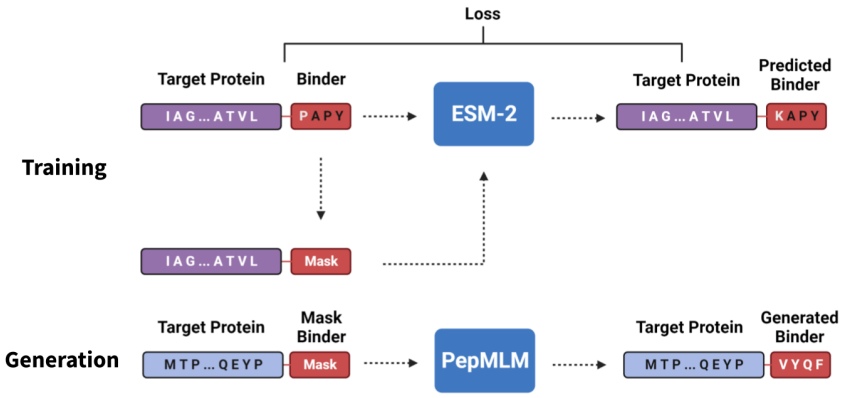

Figure 1: PepMLM Overview

Pseudo-Perplexity of PepMLM The pseudo-perplexity (Salazar et al., 2019) of ESM-2 was adapted to focus specifically on the evaluation of peptide binder generation. Notably, the perplexity calculation is confined to the binder region, or, in other words, the masked regions. Mathematically, the pseudo-perplexity is defined as:

$$\text{PseudoPerplexity}(b) = \exp\left\{-\frac{1}{m}\sum_{i=1}^{m}\log p\left(b_i|b_{j\neq i}, p\right)\right\}$$

In this equation, b represents the binder sequence and m is the length of the binder sequence. This modification ensures a more focused evaluation of the generated peptide binders, aligning with the conditional modeling approach adopted in this study.

Benchmarking To assess the efficacy of the generated peptide binders, two benchmarking studies were conducted: one on the test set and another on selected critical proteins. In the test set benchmarking, top-k sampling (k = 3) was employed to generate a single peptide binder for each target protein. Additionally, the original ESM-2 model was utilized to generate peptides, and random peptides of equivalent length were created. For ESM-2 generation, specifically, mask tokens of the same length were added at the end of target protein sequences for analogous model prediction and decoding as for PepMLM. The perplexity of the PepMLM was compared across four groups. PepMLM-generated binders and test binders were folded using the AlphaFold2 ColabFold version 1.5.2 (Jumper et al., 2021; Mirdita et al., 2022), in conjunction with the protein sequences. Folding metrics including pLDDT and ipTM were gathered, which were utilized to correlate perplexity findings. For each test target protein, the ipTM scores of the test and generated binders were compared to determine the overall hit rate. Notice, as top-k sampling generates with randomness, the hit rate might vary or increase with different runs or k options. For the proteins identified as critical, the model produced eight binders, each of a length of 15 residues, using top-k sampling (k = 3). These binders were synthesized for specific target proteins to facilitate subsequent experimental evaluations (see C).

In parallel to the PepMLM approach, RFDiffusion (Watson et al., 2023) was employed to design peptide binders for both cases. For the given test set, RFDiffusion was tasked with generating one peptide binder per target protein, matching the length specified by the ground truth binders. The predicted structures were then converted into sequences using ProteinMPNN (Dauparas et al., 2022) with initial guess and number of cycles of 3. For the selected critical proteins, RFDiffusion and ProteinMPNN generated 8 candidate binders, each comprising 15 residues, under identical parameter settings as testset generation.

3 Results and Discussion

Pseudo Perplexity For a majority of the test set, known binders exhibited a reasonable perplexity range, with only a few outliers (those with a perplexity ≥ 40), validating the PepMLM's effective ability to model them accurately (Fig.A1 and Table A1). A comparative analysis revealed that the binders generated by PepMLM exhibited lower perplexity values, suggesting a higher likelihood of them making stable binding interactions with the target. Moreover, our distribution analysis revealed that PepMLM closely mirrors the distribution peak of real binders, a deviation from the distribution shifts observed with the original ESM-2 model alone and with randomly generated binders (Fig.A1).

Folding Metrics Next, to benchmark PepMLM's generation quality, we co-folded the test and generated binders with their respective target proteins utilizing AlphaFold-Multimer, which has been proven effective at predicting peptide-protein complexes (Evans et al., 2021; Johansson-Åkhe & Wallner, 2022). The pLDDT and ipTM scores, verified metrics within AlphaFold2 (Jumper et al., 2021), function as critical indicators of the structural integrity and the potential interface binding affinity of peptide-protein complex, respectively, providing a quantitative assessment of our generation. The extracted ipTM and pLDDT values from our benchmarking indicated a statistically significant negative correlation ($p<0.01$) with PepMLM perplexity, affirming the model's reliability at prioritizing binders with stable binding capacity to the target (Fig.A2). Subsequent analysis involved sorting the test set based on their ipTM values and contrasting these with the ipTM values of the associated PepMLM-generated binders. Our analysis yielded a hit rate exceeding 38% (Fig.2A). When applying the same evaluation process to RFDiffusion for binder design on the test set, the hit rate was below 30% (Fig.2B), suggesting PepMLM's comparative advantage in designing peptide binders, potentially reducing the need for extensive downstream experimental screening. Here, We also present two top-generated binders, exhibiting high ipTM scores, with their respective target proteins, and overlayed their positions with that of PDB-validated test binders to those targets (Fig.2C). We observe high alignment between the generated and test peptides, highlighting the model's proficiency in capturing the inherent conditional distributions associated with peptide-protein binding.

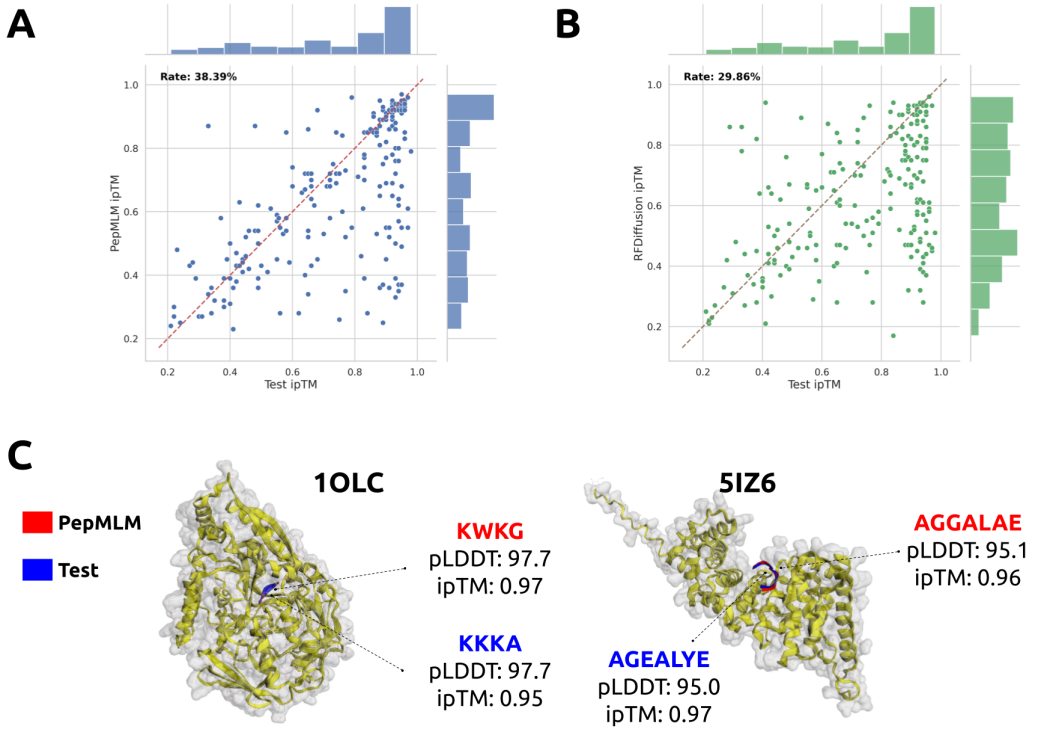

Figure 2: A) ipTM comparison of PepMLM and test binders. B) ipTM comparison of RFDiffusion and test binders. C) Binding Visualization of top-2 PepMLM binders (red) with corresponding test binders (blue).

Binding Analysis When evaluating generated peptide binders with ipTM scores surpassing those of the test binders, we classified them into three distinct groups based on ipTM score thresholds: Class I (both test and generated binders with ipTM ≥ 0.7), Class II (generated binders with ipTM ≥ 0.7, but test binders with ipTM ≤ 0.7), and Class III (both generated and test binders with ipTM ≤ 0.7). For each class, three representative complexes were chosen for joint visualization with the test binder ((Fig.A3) and Supplementary Table 3). Observations from Classes I and II indicate that despite the generated binders possessing distinct sequence compositions compared to the test binders, they tend to target the same binding pocket and exhibit similar structural conformations. This pattern suggests that our language model-based design approach successfully captures structural information of peptide-protein binding. Conversely, in Class III, characterized by lower ipTM values, we noted distinct binding modes between generated and test binders. The generated binders appeared to occupy more optimal binding positions according to AlphaFold-Multimer predictions. However, even with the high pLDDT values from AlphaFold, it remains challenging to definitively ascertain whether our binders exhibit unique binding modes or if these observations are attributable to limitations in AlphaFold-Multimer modeling.

To this point, we had utilized the lightweight ESM-2-650M model, enabling flexible fine-tuning and inference. To assess the performance of larger models, we additionally fine-tuned ESM-2-3B16 for peptide generation (PepMLM-3B) and evaluated it using the same methodology as employed for the ESM-2-650M version of PepMLM (PepMLM-650M). However, as illustrated in Appendix Figure A4, we did not observe a substantial improvement in either perplexity or hit rate for PepMLM-3B (36.02%). Considering the associated resource and inference costs, we provide our PepMLM-650M model as an accessible resource for effective linear peptide generation.

4 Conclusion

In this paper, we introduce PepMLM, the first *de novo* binder design algorithm directly conditioned on the target sequence of a protein. The model works simply by concatenating peptides with proteins and masking them out. Despite its simplicity, it achieves satisfactory performance for peptide design both *in silico* and in human cells. We envision that further improvements can be made to PepMLM, for example incorporating diverse sampling algorithms. To enable PepMLM as a universal tool for peptide binder design, we can retrain with modification-aware and variant-aware pLM embeddings to enable specificity to post-translational isoforms over wild-type protein states. We also plan to integrate PepMLM generation with high-throughput lentiviral screening to both evaluate its hit rate experimentally and input experimental data back into the algorithm, creating an active learning-based optimization loop. As a note, we have not validated PepMLM's ability to generate high affinity, standalone peptide binders, those that can be further stabilized via cyclization or stapling, though this may prove possible via the current algorithm (Vinogradov et al., 2019; Moiola et al., 2019).Nonetheless, we envision that through additional development, our accessible peptide generator, coupled with variants of our uAb architecture (Chatterjee et al., 2020), will enable a protein editing system to bind and modulate any target protein, whether structured or not.

References

Osama Abdin, Satra Nim, Han Wen, and Philip M Kim. Pepnn: a deep attention model for the identification of peptide binding sites. *Communications biology*, 5(1):503, 2022.

Fiona M Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, et al. Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature*, 568(7753):511–516, 2019.

Miklós Békés, David R Langley, and Craig M Crews. Protac targeted protein degraders: the past is prologue. *Nature Reviews Drug Discovery*, 21(3):181–200, 2022.

Garyk Brixi, Tianzheng Ye, Lauren Hong, Tian Wang, Connor Monticello, Natalia Lopez-Barbosa, Sophia Vincoff, Vivian Yudistyra, Lin Zhao, Elena Haarer, et al. Salt&peppr is an interface-predicting language model for designing peptide-guided protein degraders. *Communications Biology*, 6(1):1081, 2023.

Pranam Chatterjee, Manvitha Ponnapati, Christian Kramme, Alexandru M Plesa, George M Church, and Joseph M Jacobson. Targeted intracellular degradation of sars-cov-2 via computationally optimized peptide fusions. *Communications Biology*, 3(1):715, 2020.

Tianlai Chen, Lauren Hong, Vivian Yudistyra, Sophia Vincoff, and Pranam Chatterjee. Generative design of therapeutics that bind and modulate protein states. *Current Opinion in Biomedical Engineering*, pp. 100496, 2023.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Mei Ding, Theodorus H Van der Kwast, Ravi N Vellanki, Warren D Foltz, Trevor D McKee, Nahum Sonenberg, Pier P Pandolfi, Marianne Koritzinsky, and Bradly G Wouters. The mtor targets 4e-bp1/2 restrain tumor growth and promote hypoxia tolerance in pten-driven prostate cancer. *Molecular Cancer Research*, 16(4):682–695, 2018.

Guoqiang Dong, Yu Ding, Shipeng He, and Chunquan Sheng. Molecular glues for targeted protein degradation: from serendipity to rational discovery. *Journal of medicinal chemistry*, 64(15): 10606–10620, 2021.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, et al. De novo design of protein interactions with learned surface fingerprints. *Nature*, pp. 1–9, 2023.

Hongying Gao, Xiuyun Sun, and Yu Rao. Protac technology: opportunities and challenges. *ACS medicinal chemistry letters*, 11(3):237–240, 2020.

Isak Johansson-Åkhe and Björn Wallner. Improving peptide-protein docking with alphafold-multimer using forced sampling. *Frontiers in Bioinformatics*, 2:85, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.

Pedro Martins, Diego Mariano, Frederico Chaves Carvalho, Luana Luiza Bastos, Lucas Moraes, Vivian Paixão, and Raquel Cardoso de Melo-Minardi. Propedia v2. 3: A novel representation approach for the peptide-protein interaction database using graph-based structural signatures. *Frontiers in Bioinformatics*, 3:1103103, 2023.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Mattia Moiola, Misal G Memeo, and Paolo Quadrelli. Stapled peptides—a useful improvement for peptide-based drugs. *Molecules*, 24(20):3654, 2019.

Dan Ofer, Nadav Brandes, and Michal Linial. The language of proteins: Nlp, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758, 2021.

Kalyan Palepu, Manvitha Ponnapati, Suhaas Bhat, Emma Tysinger, Teodora Stan, Garyk Brixi, Sabrina RT Koseki, and Pranam Chatterjee. Design of peptide-based protein degraders via contrastive deep learning. *bioRxiv*, pp. 2022–05, 2022.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. *arXiv preprint arXiv:1910.14659*, 2019.

Bo Kyung A Seong, Neekesh V Dharia, Shan Lin, Katherine A Donovan, Shasha Chong, Amanda Robichaud, Amy Conway, Amanda Hamze, Linda Ross, Gabriela Alexe, et al. Trim8 modulates the ews/fli oncoprotein to promote survival in ewing sarcoma. *Cancer Cell*, 39(9):1262–1278, 2021.

Shuang Shang, Fang Hua, and Zhuo-Wei Hu. The regulation of β-catenin activity and function in cancer: therapeutic opportunities. *Oncotarget*, 8(20):33972, 2017.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alexander A Vinogradov, Yizhen Yin, and Hiroaki Suga. Macrocyclic peptides as drug candidates: recent progress and remaining challenges. *Journal of the American Chemical Society*, 141(10): 4167–4181, 2019.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Zibo Zhao and Ali Shilatifard. Epigenetic modifications of histones in cancer. *Genome biology*, 20: 1–16, 2019.

Lei Zhong, Yueshan Li, Liang Xiong, Wenjing Wang, Ming Wu, Ting Yuan, Wei Yang, Chenyu Tian, Zhuang Miao, Tianqi Wang, et al. Small molecules in targeted cancer therapy: Advances, challenges, and future perspectives. *Signal transduction and targeted therapy*, 6(1):201, 2021.

Stefan K Zöllner, James F Amatruda, Sebastian Bauer, Stéphane Collaud, Enrique de Álava, Steven G DuBois, Jendrik Hardes, Wolfgang Hartmann, Heinrich Kovar, Markus Metzler, et al. Ewing sarcoma—diagnosis, treatment, clinical challenges and future perspectives. *Journal of clinical medicine*, 10(8):1685, 2021.

A SUPPLEMENTARY FIGURES

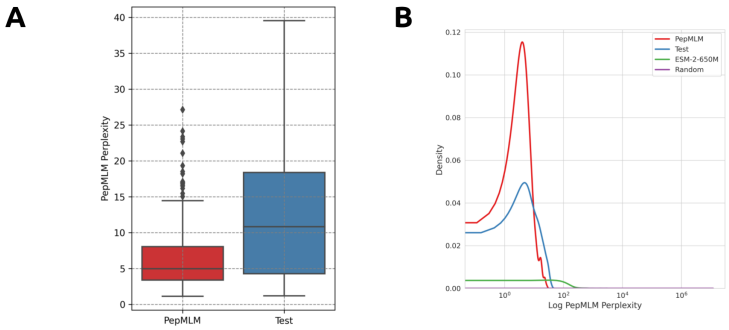

Figure A1: (A) Perplexity distribution comparison. The perplexity values were calculated for test and generated peptides, encompassing the target proteins in the test set. (B) The density distribution visualization of the log perplexity values for target-peptide pairs, encompassing test peptides, PepMLM-650M-generated peptides, ESM-2-650M-generated peptides, and random peptides.

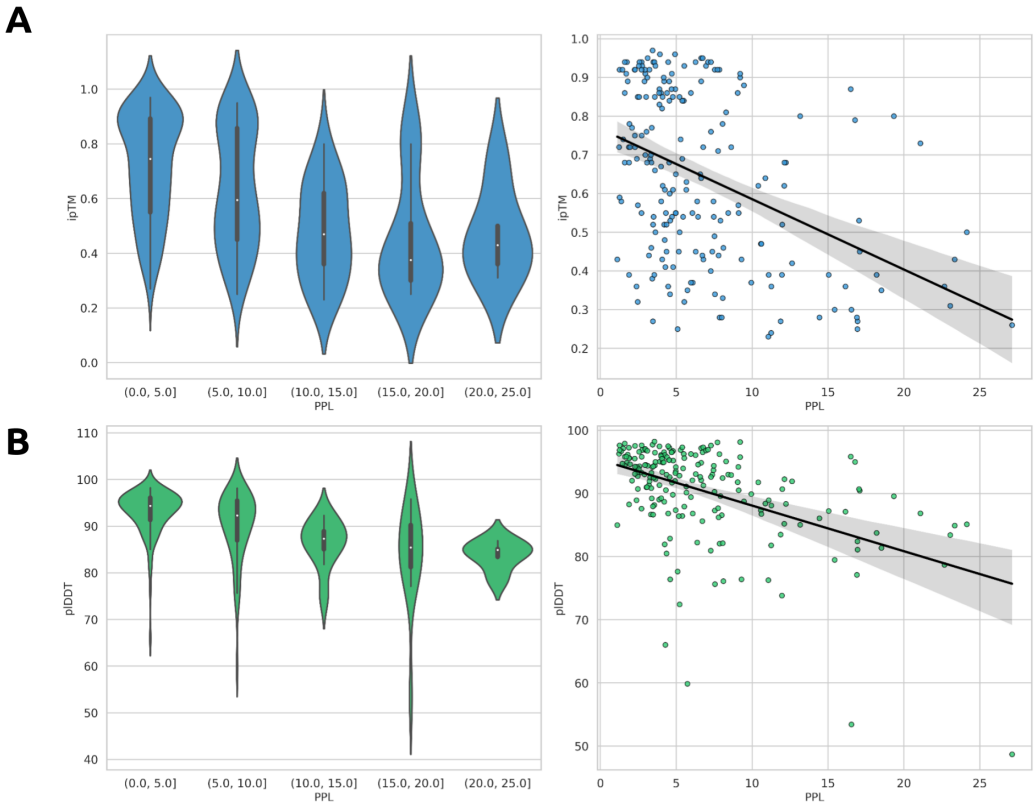

Figure A2: Association between model perplexity and co-folding metrics. (A) Relationship between ipTM and perplexity (PPL). The initial segment of Figure A presents a violin plot, categorizing perplexity in 5-unit intervals. The subsequent segment delineates the raw data points, accompanied by a regression 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.

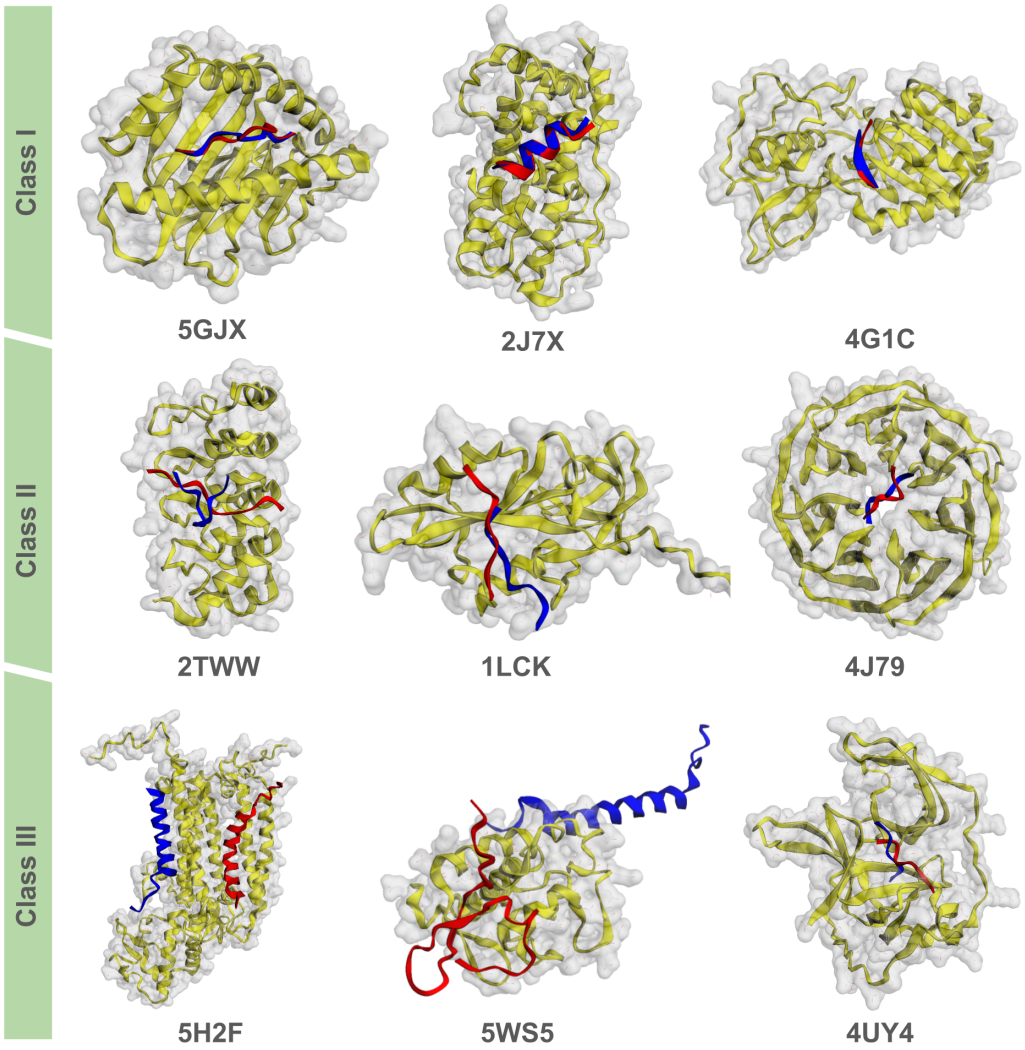

Figure A3: Visualization of Binder-Protein Complexes. Co-folded binder-protein complexes are categorized into three distinct classes for visualization purposes. Class I includes complexes where both the generated and test binders exhibit ipTM scores ≥ 0.7, Class II encompasses those with generated binders 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.

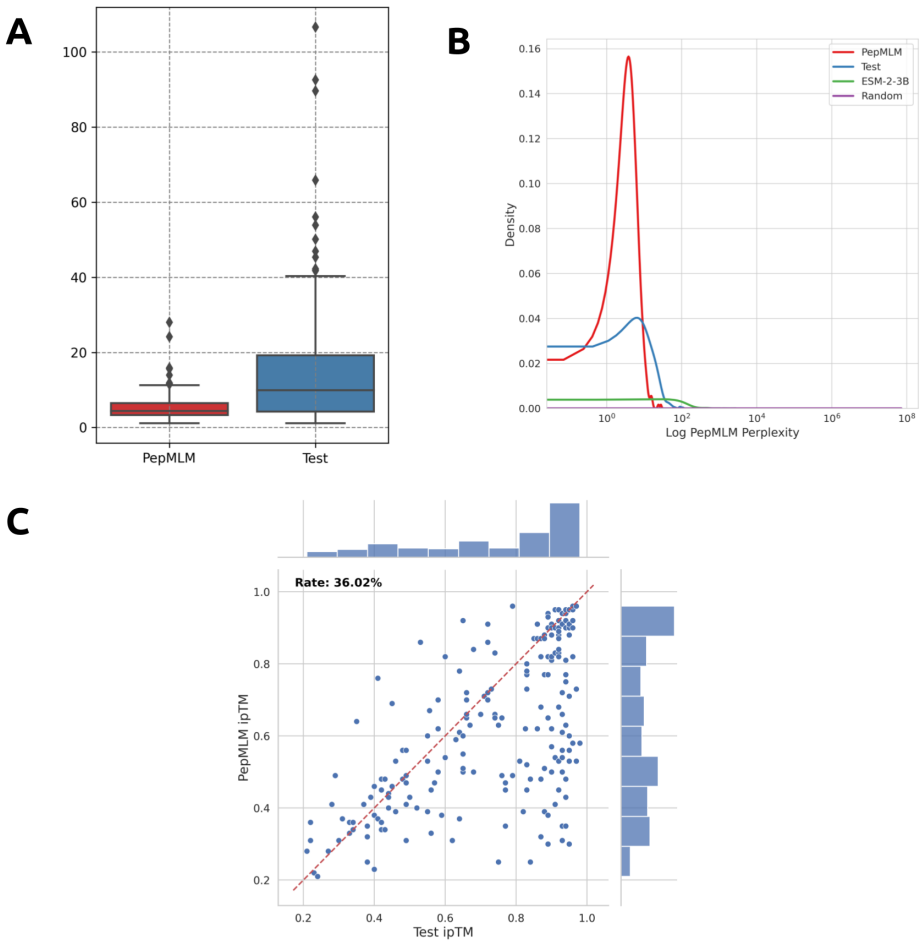

Figure A4: Evaluation of PepMLM-3B. (A) Perplexity distribution comparison. The perplexity values were calculated for test and generated peptides, encompassing the target proteins in the test set. (B) The density distribution visualization of the log perplexity values for target-peptide pairs, encompassing test peptides, PepMLM-3B-generated peptides, ESM-2-3B-generated peptides, and random peptides. (C) 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.

B SUPPLEMENTARY TABLES

Table A1 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.

Table A1: Outlier analysis of protein-peptide complexes.

	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	RAKWDT ANNPLXKE ATSTFTNITX RGT	45.4	0.49	61.9	HIAEEP HFFESMQ NNYEKPT TYKFQQK	12.0	0.39	73.8
6GHJ	FAQ	209.7	0.71	93.4	MXL	3.4	0.68	93.3
6AMU	MMWDR GLGMM	59.2	0.34	87.9	YQALI GGFNA	14.4	0.28	86.1
5WMR	QIKV RVDMV	76.3	0.91	92.3	LRFW RARTL	9.0	0.86	91.8
5NJC	VLEDRI	63.0	0.83	97.5	AAAAAA	1.5	0.74	97.1
5FML	LSNDI SQGIK RQRMT VESM	42.8	0.91	94.0	AAMTKLAL AAKTRAQ LFKK	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	AHIVM VDAYKPT	62.6	0.87	97.5	GPTPVQ VLKRRG	17.0	0.53	90.7
5DHM	RSIEISIR VDDFTKT GETVRY	64.2	0.93	94.4	AQSPEIITAD VVVTSD EFTTT	19.3	0.8	89.6

Table A2: Sequence information and folding metrics for complexes in supplementary Figure A3

PDB ID	**Generated Binder**	**plDDT**	**ipTM**	**Test Binder**	**plDDT**	**ipTM**
5GJX	RLLEWMIYI	96.3	0.92	RLIQNSITI	96.0	0.92
2J7X	HHLLLHLLTQD	91.9	0.92	IQSLINLLADN	91.9	0.91
4G1C	TAKXST	96.9	0.91	AVXCAX	97.0	0.86
3TWW	RREPPGGAFRX	97.4	0.87	RQSPDGQSFRX	92.7	0.48
1LCK	PPXEEIPP	87.3	0.92	EGQQPQPA	86.1	0.68
4J79	AARHLD	97.3	0.72	EKVHVQ	97.2	0.64
5H2F	XETNTLVRYV VAHFVLLVSVIL IREAPRIESSKXX	84.9	0.43	XETITYVF IFACIIALFFFA IFFREPPRITXX XXX	86.7	0.27
5WS5	SSEEGRPIL WIATTTGGGGV IIIVLFLFYAYYGSL SXLXXX	77.7	0.24	MSEGGRIPL WIVATVAGM GVIVIVGLFF YGAYAGLGSSL XX	86.4	0.24
4UY4	ARTKQT	90.1	0.63	ARTXQT	89.0	0.43

C Experimental Validation

To corroborate our in silico results on more unique targets, we first sought to test PepMLM-generated binders in our uAb architecture to degrade pathogenic proteins in a cellular model of Ewing sarcoma, a pediatric bone malignancy with no approved targeted therapies (Zöllner et al., 2021). As our targets, we chose two cancer-related proteins, 4E-BP2 and β-catenin, as well as the more structured

histone H3 protein, a core epigenetic nuclear protein that comprises chromatin (Ding et al., 2018; Shang et al., 2017; Zhao & Shilatifard, 2019). To design peptides, we first employed greedy decoding to determine the optimal binder length that yielded the lowest perplexity, followed by the generation of binders for each target sequence using top k sampling, where k was fixed at 3 as previously described. After cloning these peptides into our uAb backbone and transfecting into A673 Ewing sarcoma cells, we conducted Western blotting on whole-cell protein extracts with target-selective primary antibodies (Fig.A5A). Our results demonstrate that select PepMLM-generated "guide" peptides induce binding and subsequent degradation of endogenous targets when fused to E3 ubiquitin ligase domains, demonstrating reduced protein levels relative to that of the non-targeting control uAb (Fig.A5B). Next, we sought to compare PepMLM-derived degraders with that of RFDiffusion on TRIM8, a known regulator of the fusion oncoprotein, EWS-FLI1, that drives Ewing sarcoma (Seong et al., 2021). For both models, we generated 15 amino acid binders provided the target sequence (to PepMLM) or target structure (to RFDiffusion) of TRIM8. After integrating these peptides into uAb-expressing plasmids, we evaluated TRIM8 protein degradation via Western blotting and observed comparable performance between the two models (Fig.A5C). In total, our results motivate further design and testing of effective PepMLM-derived degraders to diverse pathogenic targets, whether they are conformationally stable or not.

Generation of plasmids All uAb plasmids were generated from the standard pcDNA3 vector, harboring a cytomegalovirus (CMV) promoter and a C-terminal IRES-mCherry cassette as a transfection control. An Esp3I restriction site was introduced immediately upstream of the CHIPΔTPR CDS and flexible GSGSG linker via the KLD Enzyme Mix (NEB) following PCR amplification with mutagenic primers (Genewiz). For uAb assembly, peptide sequences were human codon-optimized for complementary oligo generation (Genewiz). Oligos were annealed and ligated via T4 DNA Ligase into the Esp3I-digested uAb backbone. Assembled constructs were transformed into 50 µL NEB Turbo Competent Escherichia coli cells, and plated onto LB agar supplemented with the appropriate antibiotic for subsequent sequence verification of colonies and plasmid purification (Genewiz).

Cell culture A673 cells were maintained in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 100 units/ml penicillin, 100 mg/ml streptomycin, and 10% fetal bovine serum (FBS). For uAb testing, pcDNA3-uAb (500 ng) plasmids were transfected into cells as triplicates (4x105/well in a 12-well plate) with Lipofectamine 2000 (Invitrogen) in Opti-MEM (Gibco).

Cell fractionation and immunoblotting On the day of harvest, cells were detached by addition of 0.05% trypsin-EDTA and cell pellets were washed twice with ice-cold 1X PBS. Cells were then lysed and subcellular fractions were isolated from lysates using a 1:100 dilution of protease inhibitor cocktail (Millipore Sigma) in Pierce RIPA buffer (ThermoFisher). Specifically, the protease inhibitor cocktail-RIPA buffer solution was added to the cell pellet, the mixture was placed at 4 oC for 30 min followed by centrifugation at 15,000 rpm for 10 min at 4 oC. The supernatant was collected immediately to a pre-chilled PCR tube, and after adding 4X Bolt™ LDS Sample Buffer (ThermoFisher) with 5% β-mercaptoethanol in a 3:1 ratio, the mixture was incubated at 95 oC for 10 min prior to immunoblotting. Immunoblotting was performed according to standard protocols. Briefly, samples were loaded at equal volumes into Bolt™ Bis-Tris Plus Mini Protein Gels (ThermoFisher) and separated by electrophoresis. iBlot™ 2 Transfer Stacks (Invitrogen) were used for membrane blot transfer, and following a 1 h room-temperature incubation in SuperBlock™ Blocking Buffer (ThermoFisher), proteins were probed with rabbit anti-4E-BP2 antibody (Cell Signaling, Cat # 2845; diluted 1:500), rabbit anti-Histone 3 antibody (Abcam, Cat # ab1791; diluted 1:500), rabbit anti-β-catenin antibody (Cell Signaling, Cat # 8480; diluted 1:500), mouse anti-TRIM8 antibody (SantaCruz Biotechnology Cat # sc-398878; diluted 1:1000), mouse anti-GAPDH antibody (Santa Cruz Biotechnology, Cat # sc-47724; diluted 1:500), or rabbit anti-Vinculin antibody (ThermoFisher, Cat # 42H89L44; diluted 1:1000) for overnight incubation at 4oC. The blots were washed three times with 1X TBST for 5 min each and then probed with a secondary antibody, goat anti-rabbit IgG (H+L), horseradish peroxidase (HRP) (ThermoFisher, Cat # 31460, diluted 1:5000) or goat anti-mouse IgG (H+L) Poly-HRP (ThermoFisher, Cat # 32230, diluted 1:2000) for 1 h at room temperature. Following three washes with 1X TBST for 5 min each, blots were detected by chemiluminescence using an iBright 1500 Imaging System (ThermoFisher). Densitometry analysis of protein bands in immunoblots was performed using ImageJ software as described here[1]. Briefly,

[1]https://imagej.nih.gov/ij/docs/examples/dot-blot/

bands in each lane were grouped as a row or a horizontal "lane" and quantified using FIJI's gel analysis function. Intensity data for the uAb bands was first normalized to band intensity of GAPDH or Vinculin in each lane then to the average band intensity for the uAb vector control cases across replicates.

Statistical analysis and reproducibility To ensure robust reproducibility of all results, experiments were performed with at least three biological replicates. Sample sizes were not predetermined based on statistical methods but were chosen according to the standards of the field (at least three independent biological replicates for each condition). All data were reported as average values with error bars representing standard deviation (SD). All graphs were generated using Prism 10 for MacOS. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

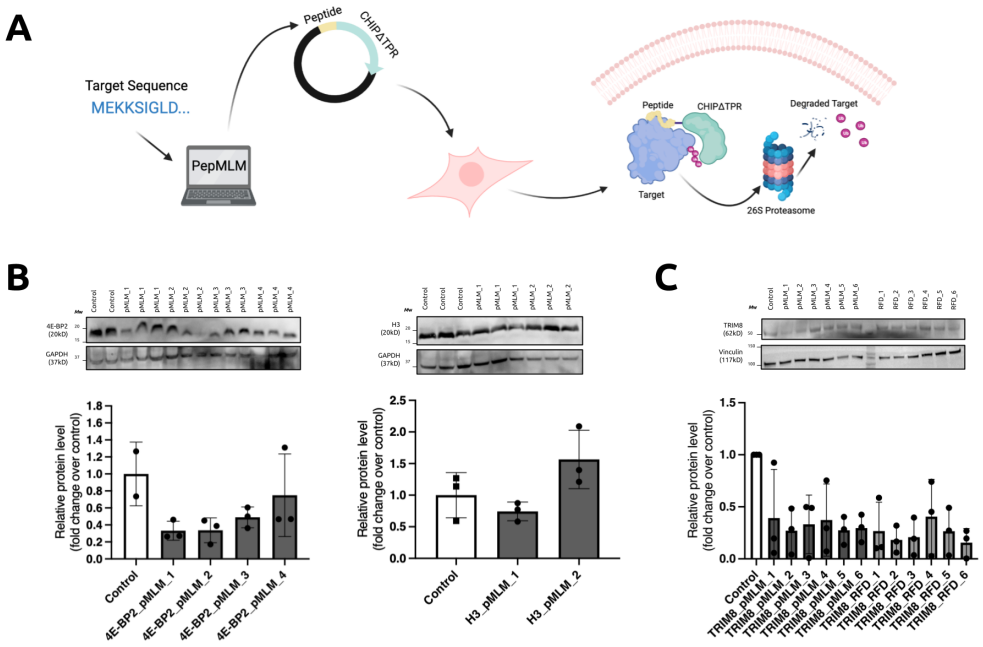

Figure A5: Experimental Validation and Results. (A) Architecture and mechanism of uAb degradation system. CHIPΔTPR is fused to the C-terminus of PepMLM-designed target-specific peptides, and can thus tag endogenous target proteins for ubiquitin-mediated degradation in the proteasome, post-plasmid transfection. (B) Degradation of endogenous targets in protein extracts of A673 Ewing sarcoma cells analyzed via immunoblotting. Blots are representative of independent transfection replicates (n = 3). Relative degradation activity was determined by densitometry analysis of target protein signal normalized to sample-specific GAPDH signal. (C) Degradation of TRIM8 analyzed via immunoblotting. Blots are representative of a single replicate of independent transfection replicates (n = 3). Relative degradation activity was determined by densitometry analysis of target protein signal normalized to sample-specific Vinculin signal.