De Novo Generation and Prioritization of Target-Binding Peptide Motifs from Sequence Alone

Suhaas Bhat,[1,*] Kalyan Palepu,[1,*] Vivian Yudistyra,[1] Lauren Hong,[1]
Venkata Srikar Kavirayuni,[1] Tianlai Chen,[1] Lin Zhao,[1] Tian Wang,[1] Sophia Vincoff,[1]
Pranam Chatterjee[1,2,†]

1. Department of Biomedical Engineering, Duke University
2. Department of Computer Science, Duke University

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

Abstract

Designing binders to target undruggable proteins presents a formidable challenge in drug discovery, requiring innovative approaches to overcome the lack of putative binding sites. Recently, generative models have been trained to design binding proteins from the three-dimensional structure of a target protein alone, but thus exclude design to disordered or conformationally unstable targets. In this work, we provide a generalizable algorithmic framework to design short, target-binding peptide motifs, requiring only the amino acid sequence of the target protein. To do this, we propose a process to generate naturalistic peptide candidates through Gaussian perturbation of the peptidic latent space of the state-of-the-art ESM-2 protein language model, and subsequently screen these *de novo* linear sequences for target-selective interaction activity via a CLIP-based contrastive learning architecture. By integrating these generative and discriminative steps, we create a **Pep**tide **Pr**ioritization via **CLIP** (**PepPrCLIP**) pipeline and validate highly-ranked, target-specific peptide motifs experimentally via fusion to E3 ubiquitin ligase domains, demonstrating functionally potent degradation of conventionally undruggable targets *in vitro*. Overall, our design strategy provides a modular toolkit for designing short binding motifs to any target protein without the reliance on stable and ordered tertiary structure, enabling generation of programmable modulators to undruggable and disordered proteins such as transcription factors and fusion oncoproteins.

Introduction

Over 80% of pathogenic proteins are considered “undruggable” by standard small molecule inhibitors, as they often lack putative stable binding pockets.[1,2] Targeted protein degradation (TPD) has emerged as a promising alternative to treat disease, but largely rely on small molecule warheads to bind and recruit endogenous E3 ubiquitin ligases to targets, thus still requiring the presence of accessible surface binding sites.[3] Protein-based modalities, either as standalone binders or recruiters of effector domains, thus represent attractive platforms to target these proteins, as they do not require surface pockets for binding. However, protein binder design is challenging: one would either need to conduct laborious experimental screening via strategies such as yeast or phage display, or deploy structure-based computational design methods.[4–7] These methods, such as the current state-of-the-art RFDiffusion generative model, require accurate three-dimensional (3D) structures of target proteins for binder design, precluding design to unstructured and intrinsically-disordered proteins, which make up a large fraction of the undruggable proteome.[6–9]

Recently, protein language models (pLMs) have been pre-trained on millions of natural protein sequences to generate latent embeddings that grasp relevant physicochemical, functional, and most notably, tertiary structural information.[10–13] Even more interestingly, early results suggest that sequence-based protein transformers can produce novel protein sequences with functional capability.[14,15] Motivated by these findings, here, we develop a pipeline to design *de novo* target-binding peptide motifs, only requiring the amino acid sequence of the target protein. Inspired by the groundbreaking work of OpenAI to generate photorealistic images from natural language descriptions via contrastive language-image pretraining (CLIP),[16] we first leverage the state-of-the-art ESM-2 pLM to generate distinct, but naturalistic peptide candidates by applying Gaussian noise to the embeddings of bona-fide peptide binder sequences.[10] We then screen these candidates for target-selective interaction activity through a CLIP-based contrastive latent space, trained to jointly encode cognate peptide-protein pairs. Our resulting **Pep**tide **Pr**ioritization via **CLIP (PepPrCLIP)** pipeline integrates our generative-discriminative framework to prioritize *de novo* candidates with binding affinity to an input target sequence. We extensively validate that PepPrCLIP-prioritized “guide” peptides, when experimentally fused E3 ubiquitin ligase domains, exhibit robust degradation of diverse undruggable targets, including transcription factors, membrane proteins, and fusion oncoproteins proteins *in vitro*, leading to functional phenotypic outcomes in cancer cell models. Our integrative methodology thus provides a modular, CRISPR-like system for programmable proteome editing applications.

Results

A CLIP discriminator accurately captures peptide-protein binding

Recently, OpenAI released CLIP, a model that is trained to associate images and their textual descriptions by learning to rank the correct matching pairs higher than incorrect ones via contrastive learning.[16,17] DALL-E 2, a denoising diffusion model, was then trained using the CLIP latent space to generate photorealistic images conditioned on an associated text prompt.[17] We thus hypothesized that just as CLIP connects images to their corresponding captions using jointly-trained image and caption encoders, we can leverage a CLIP-based architecture to map target proteins to their corresponding binding peptides using jointly-trained peptide and target protein encoders. Specifically, we task our model to predict the cosine similarity between all n^2 target and peptide pairs in a matrix, and subsequently calculate the average of the cross-entropy losses on the rows and columns of this matrix during training. We further employ Meta AI’s state-of-the-art ESM-2 model weights to generate feature-rich embeddings for cognate targets and peptides as inputs to the model (Figure 1A).[10]

In order to construct the dataset for training this model, we first considered every interaction structure in the RCSB PDB, and specifically filtered to create appropriate training and testing datasets. To train on a wide distribution of peptide-protein pairs, we generate a "noisy" dataset, in which we consider both strong and weaker interactions (co-crystal buried surface area >= 50 $Å^2$), and consider interactions between all short proteins (< 50 AA) and another larger protein. For stringent evaluation of our model, we create a "strict" peptide-protein test dataset, in which we only include strong interactions (co-crystals with a buried surface area >= 400 $Å^2$), with peptides of length <= 25 AA, and target proteins of length >= 30 AA. These parameters are equivalent to the datasets for the recent PepNN model, a deep attention model which accurately predicts

peptide binding sites on target proteins.[18] Overall, the datasets are clustered at 30% sequence identity, using MMSeqs2, ensuring significant structural diversity between the train and test sets.[19]

For evaluation, we use three key metrics: binary accuracy, Top-1 accuracy, Top 10% accuracy. Binary accuracy is the accuracy of the model in predicting the correct binding pairs when given 2 protein-peptide pairs. Top-1 accuracy is the likelihood that for a given protein, the model has chosen the correct binding peptide (i.e. in a batch of 64), and Top 10% accuracy is the likelihood that this peptide is in the top 10% of peptides when ranked for binding affinity to this protein. Our results demonstrate that training the CLIP model on the noisy dataset — a more difficult binding prediction task — allows for robust performance on the stricter held-out test set, exhibiting an over 93% binary accuracy, as well as a Top 10% accuracy of 0.89. In contrast, training and testing on the strict dataset alone yields a poorer 87% binary accuracy and Top 10% accuracy of 0.63, justifying our transfer learning approach (Figure 1B). As an additional method to validate the optimized model's ability to discriminate between peptide-protein pairs that bind and do not bind, we compare predicted CLIP scores on held-out true interacting pairs from our dataset and mismatched pairs that do not interact. We find that the distribution of CLIP scores for interacting peptide-protein pairs is heavily skewed towards +1, while mismatched non-interacting pairs have CLIP scores distributed with a peak around -0.1 (Figure 1C). Overall, these results support the usage of our final CLIP model for target-specific peptide prioritization.

Gaussian perturbation generates naturalistic peptides for CLIP prioritization

There exist approximately 10^{26} possible 20-mer peptides — far too large of a space to exhaustively screen either *in silico* or *in vitro* for a given target. Furthermore, a small fraction of these possible peptides are likely to be biologically viable. Here, we propose a method to generate biologically naturalistic peptides. We begin with an ensemble of bona-fide, interacting peptides, which appear in co-crystals in the RCSB PDB. Next, we embed the sequences of these peptides using ESM-2, whose latent space contains an understanding of naturalistic protein sequences. For each residue in the source peptide, noise is sampled from a standard Gaussian distribution and multiplied by a scaling factor k to create a perturbation, which we then add to the embedding for its respective residue. By analyzing the dependence of Hamming distance between generated peptides and their source peptides, we select values of k in the range 5 to 22 to maintain naturalistic peptide sequences (Figure 1D). By varying the amount of noise and the number of real peptides sampled, we generate an arbitrarily large set of biologically viable *de novo* peptides. Finally, we decode the perturbed embeddings with ESM-2, acquiring new peptide sequences for CLIP screening, thus completing the end-to-end PepPrCLIP pipeline (Figure 1A).

Prior to experimental testing, we computationally benchmarked the full PepPrCLIP pipeline via AlphaFold-Multimer, a massively improved method for peptide-protein docking as compared to previous docking strategies.[20,21] Using ColabFold,[22] we co-folded the top eight generated peptides, ranked based on their CLIP score, with four highly-structured target protein chains (5YXD-A, 2ZFX-A, 4G1Z-A, 4IWF-A) from our evaluation dataset, each of which possess a PDB complex with an experimentally-validated peptide (5YXD-B, 2ZFX-C, 4G1Z-C, 4IWF-C). As comprehensive negative controls, we also co-folded these four target chains each with four *de novo* peptides that have negative CLIP scores, three *de novo* peptides that were created via our generator but not screened through the CLIP module, as well as a peptide that was purely randomly generated. All peptide sequences for benchmarking can be found in Supplementary Table 1. We scored the likelihood of complex formation via AlphaFold-Multimer's predicted interface TM-score (ipTM), which evaluates the confidence of the predicted peptide-protein interface in an assessed model.[20,21] Our benchmarking results demonstrate that the full PepPrCLIP pipeline reliably identifies peptides with ipTM scores comparable to or exceeding that of the experimentally-validated peptide for each target, while negative control peptides demonstrate lowered ipTM scores on average, as expected (Figure 1E and Supplementary Figure 1). Overall, these controlled benchmarking results motivated us to experimentally test our algorithmic pipeline in a therapeutically-relevant context.

PepPrCLIP-generated uAbs induce robust degradation of pathogenic target proteins

Recently, based on the seminal work of Portnoff *et al.*,[23] we reprogrammed the specificity of a modular human E3 ubiquitin ligase called CHIP (carboxyl-terminus of Hsc70-interacting protein) by replacing its natural

substrate-binding domain, TPR, with designer “guide” peptide motifs to generate minimal, genetically-encodable “ubiquibody” (uAb) architectures for target protein degradation.[24] In multiple studies, we have demonstrated the suitability of the uAb architecture to incorporate linear guide peptides designed via both structure- and sequence-based algorithms.[24–26] We thus employ uAbs to evaluate the binding and degradation capacities of PepPrCLIP-designed peptides on four disease-related substrates, including a transcription factor (ETV6), a membrane protein (AMHR2), and two fusion oncoproteins (EWS-FLI1, PAX3-FOXO1).

For each target, we ranked generated *de novo* candidates based on their CLIP score, and the top set were selected for testing, as compared to a non-targeting normalizing control and a random peptide negative control (Supplementary Table 2). We experimentally cloned DNA plasmids expressing these peptides for each target, directly fused to CHIPΔTPR via a short glycine-serine linker, and subsequently transfected these vectors into human cells endogenously expressing target proteins tagged with an mCherry fluorescent reporter to quantify degradation efficiency (Figure 2A). Our results demonstrate that, within the top candidates per target, PepPrCLIP effectively prioritizes “guide” peptides for the design of potent degraders to these disordered substrates, with at least one and oftentimes, multiple uAbs demonstrating statistically-significant degradation (Figure 2B). On the other hand, negative control “random” degraders exhibit negligible degradation activity, emphasizing the target-specific binding capacity of PepPrCLIP-designed peptides.

TRIM8 is an E3 ubiquitin ligase itself that maintains the tumorigenic state of Ewing sarcoma by regulating levels of the EWS-FLI1 fusion oncoprotein.[27] Specifically, it has been shown that CRISPR-mediated knockout of TRIM8 causes Ewing sarcoma cells to overdose on EWS-FLI1, inducing higher levels of apoptosis.[27] As neither a PROTAC nor molecular glue has been developed to degrade TRIM8, we co-transfected A673 Ewing sarcoma cells with our eight TRIM8-targeting uAbs, all of which also demonstrate potent TRIM8 degradation (Figure 2C), alongside a GFP-based fluorogenic caspase reporter of apoptosis.[28] Remarkably, we show that all eight PepPrCLIP-generated uAbs induce significant upregulation of apoptosis in Ewing sarcoma cells, motivating further preclinical development and demonstrating a potential therapeutic application of PepPrCLIP (Figure 2C).

Discussion

Here, we describe a purely sequence-based approach for the *de novo* design of binding peptide motifs, which we use to programmably “guide” E3 ubiquitin ligases to diverse target proteins for degradation in a CRISPR-like manner. Of note, current structure-based *de novo* binder design models, including RFDiffusion and MaSIF-seed, are not specifically trained to design peptide binders,[8,9] further highlighting the novelty of PepPrCLIP. For future work, we plan to expand our *in silico* screening strategies to optimize other key properties, including selectivity and affinity to the target protein. Overall, the success of this approach motivates the future development of methods utilizing a candidate generator and contrastive discriminator screening models for molecular design. For example, we are currently exploring alternative generative processes, such as techniques leveraging databases of antibody CDRs or human protein motifs, and autoregressive models such as GPT,[29] as well as improved contrastive learning methodologies, such as the Contrastive Leave One Out Boost (CLOOB) architecture.[30] In total, our two-step method represents one of the first *de novo* protein binder design modules from sequence alone, without the requirement of 3D structural information. Our easy-to-use model, coupled with our unique protein-targeting architectures, thus establish a more modular, CRISPR-like alternative to PROTACs and molecular glues for TPD, and open the door for programmable proteome editing and screening applications.

Methods

Dataset Generation

The datasets for this paper were generated by mining the RCSB PDB (https://www.rcsb.org/) for verified, high-resolution protein-protein and peptide-protein interaction structures. Two datasets were curated: 1) A “noisy” dataset which filters the PDB-derived data for weaker interactions (cocrystal buried surface area >= 50 $Å^2$), and considers interactions between all proteins (<50 AA) paired with a larger protein (>50 AA), 2) A "strict" peptide-protein dataset, which filters for strong interactions (co-crystals with a buried surface area >= 400 $Å^2$), with peptides of length <= 25 AA, and target proteins of length >= 30 AA. Protein sequences were

then clustered using MMseqs2 connected component clustering at 0.3 minimum sequence identity to group homologous proteins.[19] Additional parameters used for data filtration can be found in Abdin, et al.[18]

CLIP Architecture and Training

The CLIP architecture was adapted from Radford, et al.[16] Encoders were trained such that the cosine similarity between target protein embeddings and peptide embeddings, defined as

$$k(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|} \tag{1}$$

is near 1 for target-peptide pairs which do bind to each other, and is near -1 for target-peptide pairs which do not bind to each other.

Next, the receptor and peptide encoders were trained on batches of n pairs of receptors and peptides which are known to interact from the datasets described above. Target and peptide sequences were encoded by their respective encoders, producing receptor embeddings $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$, and peptide embeddings $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$. The cosine similarity between all n^2 target and peptide pairs is computed in a matrix $\boldsymbol{K}$, defined such that:

$$K_{ij} = k(\boldsymbol{r_i}, \boldsymbol{p_j}) \tag{2}$$

Cosine similarities are interpreted as softmax logits. We define two cross-entropy losses, one on the matrix rows and one the matrix columns:

$$\mathcal{L}_r = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp(K_{ii})}{\sum_{j=1}^{n} \exp(K_{ij})} \right) \tag{3}$$

$$\mathcal{L}_p = -\frac{1}{n} \sum_{j=1}^{n} \log \left(\frac{\exp(K_{jj})}{\sum_{i=1}^{n} \exp(K_{ij})} \right) \tag{4}$$

$\mathcal{L}_r$ represents the loss on the model's ability to predict the correct target given a single peptide, while $\mathcal{L}_p$ represents the loss on the model's ability to predict the correct peptide given a single receptor. By using these cross-entropy losses, we implicitly assumed that the $n^2 - n$ target-peptide pairs in the batch which are not known interactions do not bind at all. While this may not be a completely accurate assumption, it is approximately true.

The receptor encoder and peptide encoder have identical architectures, though they differ in hyperparameters such as the number of layers. Starting with an input $l \times e_i$ ESM-2 embedding (where l is the input sequence length and e_i is the dimension of the ESM-2 embedding), h_1 feedforward layers were applied with ReLU activation on each amino acid embedding separately, producing a $l \times e_o$ embedding, where e_o is the output embedding dimension produced by our encoder. The embedding over the length dimension was subsequently averaged, producing an embedding vector of length e_o. Finally, h_2 feedforward layers were applied with ReLU activation on the embedding vector to obtain the output embedding. Notably, since the first set of hidden layers operate on single amino acids in isolation, and the second set of hidden layers operate on the embedding averaged over the length dimension, the encoder has minimal dependence on the sequence length. This is particularly important for helping the peptide encoder generalize to a broad range of peptide lengths.

Models were trained and evaluated in Google CoLab Pro+ via a provided V100 GPU in a "High RAM" setting.

Generation of peptide candidate sequences

Peptides are generated from an ensemble of bona-fide interacting peptides, which are embedded within ESM-2 pLM (https://huggingface.co/docs/transformers/model_doc/esm) via HuggingFace in PyTorch. For a given peptide embedding, variance (σ^2) is computed over all dimensions of the embedding. For each residue in the source peptide, noise is sampled from the standard Normal distribution $N(0,1)$ and multiplied by a scaling factor given by $\sigma^2 \times k$ to create a perturbation p_{r_i}. This perturbation is added to the embedding for its respective residue. The above scaling factor k can be varied to produce peptides that are similar to the source interacting peptides. For generation of peptides for this study, k values in the range 5 to 22 were selected. All peptide sequences evaluated in this study can be found in Supplementary Tables 1 and 2.

AlphaFold-Multimer benchmarking

Generated peptides from PepPrCLIP were ranked based on their CLIP score for four highly-structured target protein chains (5YXD-A, 2ZFX-A, 4G1Z-A, 4IWF-A). The top eight were selected, alongside the four lowest ranked peptides, three randomly-selected generated peptides that were not screened through the CLIP module, as well as a peptide that was purely randomly generated via a random sequence generator. The four experimentally-validated peptide chain sequences for each target (5YXD-B, 2ZFX-C, 4G1Z-C, 4IWF-C) were also selected. AlphaFold-Multimer_v3 was used in ColabFold with one relaxation step, according to provided instructions: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb.[22] The predicted interface TM-score (ipTM) was extracted from the post-run output files and plotted via Matplotlib. 3D images shaded with pLDDT scores were extracted from the ColabFold interface.[22]

Generation of plasmids

All uAb plasmids were generated from the standard pcDNA3 vector, harboring a cytomegalovirus (CMV) promoter and a C-terminal P2A-eGFP cassette. 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, oligos for candidate peptides 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

HEK293T and Ewing sarcoma A-673 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 screening, pcDNA-uAb (500 ng) plasmids were transfected into cells as triplicates (3×10^5/well in a 24-well plate) with Lipofectamine 2000 (Invitrogen) in Opti-MEM (Gibco). After 3 days post transfection, cells were harvested and analyzed on a Attune NxT Flow Cytometer (ThermoFisher) for GFP fluorescence (488-nm laser excitation, 530/30 filter for detection) and mCherry fluorescence (561-nm laser excitation, 620/15 filter for detection). 10,000 events were gated for data analysis based on default FSC/SSC parameters for the analyzed cells. Cells expressing eGFP were gated, and normalized mCherry cell fluorescence was calculated to a sample transfected with a non-targeting uAb using the FlowJo software (https://flowjo.com/). Representative flow cytometry gating strategies are indicated in Supplementary Figure 2.

Lentiviral production

For target-reporter packaging, HEK293T cells were seeded in a 6-well plate and transfected at ~50% confluency. For each well, 0.5 µg pMD2.G (Addgene #12259), 1.5 µg psPAX2 (Addgene #12260) and 0.5 µg of the target-mCherry reporter transfer vector were transfected with Lipofectamine 3000 (Invitrogen) according to the manufacturer's protocol. The medium was exchanged 8 hours post transfection, and the viral supernatant was harvested at 48 and 72 hours post-transfection. The viral supernatant was concentrated to 100x in 1x DPBS using Lenti-X Concentrator (Clontech, 631232) according to the manufacturer's instruction, and stored at -80°C for further use.

Target-mCherry reporter monoclonal cell line generation

For target-reporter cell line generation, $1x10^5$ HEK293T cells were mixed with 20 µL of the concentrated virus in a 6-well plate. Media was changed 24 hours post transduction. Antibiotic selection was started 36 hours post transduction by adding 2 µg/mL puromycin (Sigma, P8833). Cells were harvested for sorting at 5 days post antibiotic selection, and a single cell of mCherry positive was plated in a 96 well plate. Genomic PCR was performed after cell growth to validate the genotype of the monoclonal cell line.

Functional Assays

For the apoptosis assay, $3x10^5$ A673 cells/well were seeded on a 24-well plate 20-24 hours prior to transfection. On the day of transfection, each well received the following plasmids: ZipGFP-Casp3 plasmid (Addgene plasmid #81241) and pcDNA3-PpC_TRIM8_#. A total of 500 ng of plasmid DNA in a ratio of ZipGFP-Casp3:pcDNA3-PpC_TRIM8_# = 1:1 was mixed with Lipofectamine 2000 reagent in serum-free Opti-MEM medium and added dropwise to each well after incubation at room temperature for 20 min. After 60 h of incubation, cells were harvested and analyzed similarly as mentioned for uAb screening. Cells expressing mCherry were gated, and normalized EGFP cell fluorescence was calculated as compared to a sample transfected with a non-targeting uAb, using the FlowJo software (https://flowjo.com/).

Statistical analysis and reproducibility

To ensure robust reproducibility of all results, experiments were performed with at least three biological replicates and at least three technical measurements. 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), which gave sufficient statistics for the effect sizes of interest. All data were reported as average values with error bars representing standard deviation (SD). For individual samples, statistical significance was determined by unpaired *t* test with Welch's correction (*$p < 0.05$, **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$). All graphs were generated using Prism 9 for MacOS version 9.2.0. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Supplementary Figures

Supplementary Figure 1. AlphaFold-Multimer benchmarking of PepPrCLIP.

Supplementary Figure 2. Example gating strategy for flow cytometry analysis.

Supplementary Table 1. Peptide sequences and CLIP scores for AlphaFold-Multimer benchmarking.

Supplementary Table 2. Peptide sequences and CLIP scores for experimental testing.

Author Contributions

P.C. conceived, designed, and directed the study. S.B. and K.P. designed the PepPrCLIP architecture, curated peptide-protein data, trained the CLIP module, and evaluated trained models. S.B. developed the peptide generation method. K.P. developed the CLIP module, with assistance from S.B. V.S.K. performed *in silico* benchmarking with assistance from T.C., K.P., S.B, and S.V. V.Y., L.H, and T.W. conducted *in vitro* assays for uAb screening. L.Z. generated reporter cell lines for uAb screening.

Data and Materials Availability

All data needed to evaluate the conclusions in the paper are present in the paper and supplementary tables and figures. All training data, code, and Colab notebooks to run PepPrCLIP can be found at: https://github.com/programmablebio/pepprclip. All raw and processed data, including peptide sequences, have been deposited to the Zenodo repository: https://doi.org/10.5281/zenodo.8083954.

Competing Interests

P.C., K.P, and S.B. are listed as inventors for U.S. Provisional Application No. 63/344,820, entitled: "Contrastive Learning for Peptide Based Degrader Design and Uses Thereof." P.C. is listed as an inventor for U.S. Provisional Application No. 63/032,513, entitled: "Minimal Peptide Fusions for Targeted Intracellular Degradation." P.C. is a co-founder of UbiquiTx, Inc. S.B. and K.P. are consultants for UbiquiTx, Inc.

Acknowledgements

We thank the Duke Compute Cluster for providing database resources that have contributed to the research reported within this manuscript. We further thank Dr. Matthew DeLisa and Dr. Chris Counter for providing cell lines for functional assays.

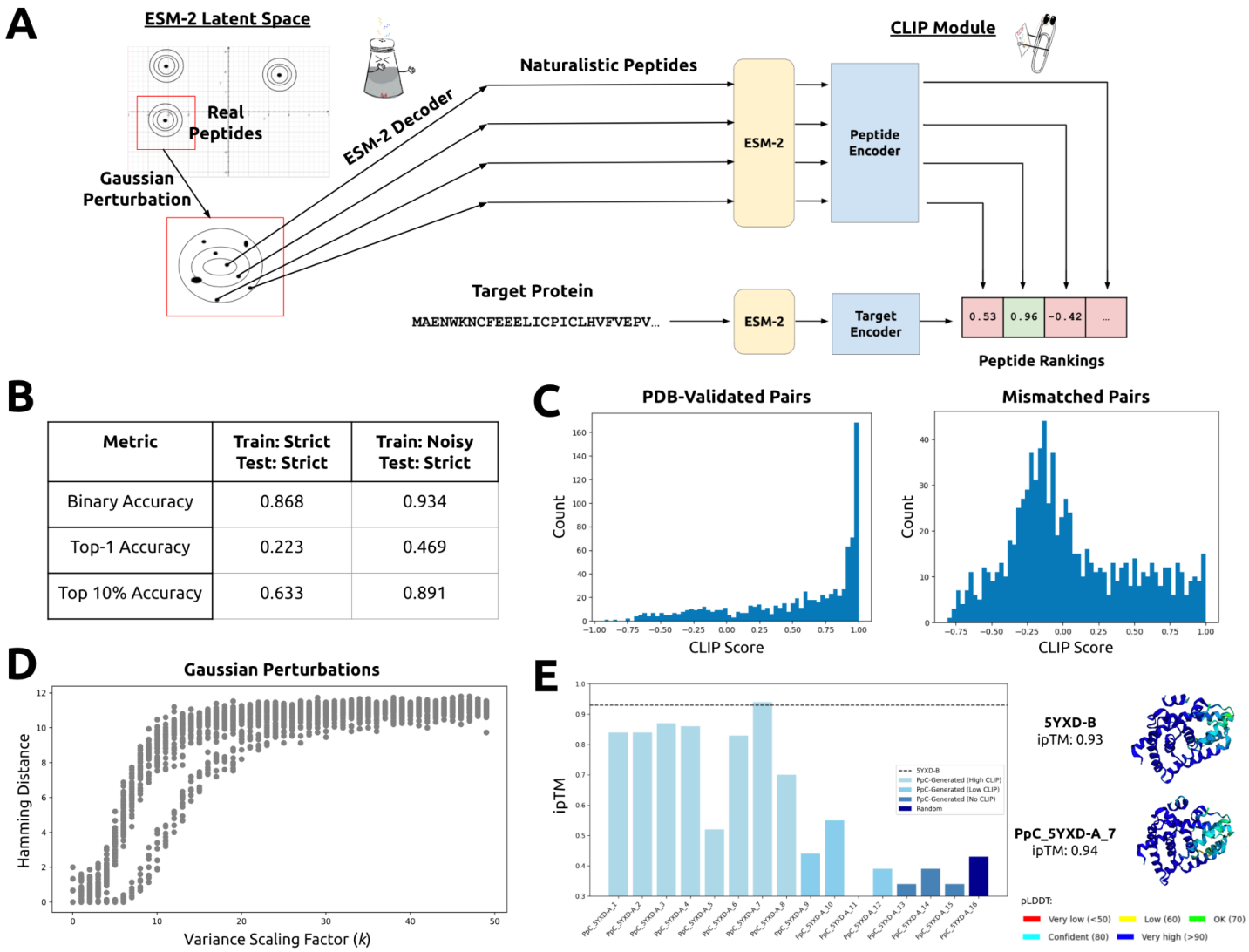
Declarations

The research was supported by The Hartwell Foundation, the Wallace H. Coulter Foundation, and institutional startup funds to the lab of P.C. from Duke University.

References

1. Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* **568**, 511–516 (2019).
2. Dang, C. V., Reddy, E. P., Shokat, K. M. & Soucek, L. Drugging the 'undruggable' cancer targets. *Nat. Rev. Cancer* **17**, 502–508 (2017).
3. Zhao, L., Zhao, J., Zhong, K., Tong, A. & Jia, D. Targeted protein degradation: mechanisms, strategies and application. *Signal Transduction and Targeted Therapy* **7**, 1–13 (2022).
4. Könning, D. & Kolmar, H. Beyond antibody engineering: directed evolution of alternative binding scaffolds and enzymes using yeast surface display. *Microb. Cell Fact.* **17**, 32 (2018).
5. Alfaleh, M. A. *et al.* Phage Display Derived Monoclonal Antibodies: From Bench to Bedside. *Front. Immunol.* **11**, 1986 (2020).
6. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).
7. Bennett, N. R. *et al.* Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 1–9 (2023).
8. Watson, J. L. *et al.* Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv* (2022) doi:10.1101/2022.12.09.519842.
9. Gainza, P. *et al.* De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176–184 (2023).
10. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
11. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
12. Elnaggar, A. *et al.* Ankh ☥: Optimized protein language model unlocks general-purpose modelling. *bioRxiv* (2023) doi:10.1101/2023.01.16.524265.
13. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
14. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* 1–8 (2023).
15. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 1–10 (2022).
16. Radford, A. *et al.* Learning transferable visual models from natural language supervision. (2021) doi:10.48550/ARXIV.2103.00020.
17. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. (2022) doi:10.48550/ARXIV.2204.06125.
18. Abdin, O., Nim, S., Wen, H. & Kim, P. M. PepNN: a deep attention model for the identification of peptide binding sites. *Communications Biology* **5**, 1–10 (2022).

19. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
20. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021) doi:10.1101/2021.10.04.463034.
21. Johansson-Åkhe, I. & Wallner, B. Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Front Bioinform* **2**, 959160 (2022).
22. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
23. Portnoff, A. D., Stephens, E. A., Varner, J. D. & DeLisa, M. P. Ubiquibodies, synthetic E3 ubiquitin ligases endowed with unnatural substrate specificity for targeted protein silencing. *J. Biol. Chem.* **289**, 7844–7855 (2014).
24. Chatterjee, P. *et al.* Targeted intracellular degradation of SARS-CoV-2 via computationally optimized peptide fusions. *Communications Biology* **3**, 1–8 (2020).
25. Palepu, K. *et al.* Design of Peptide-Based Protein Degraders via Contrastive Deep Learning. *bioRxiv* (2022) doi:10.1101/2022.05.23.493169.
26. Brixi, G. *et al.* Design of peptide-guided protein degraders with Structure-agnostic language models. *Research Square* (2023) doi:10.21203/rs.3.rs-1694183/v1.
27. Seong, B. K. A. *et al.* TRIM8 modulates the EWS/FLI oncoprotein to promote survival in Ewing sarcoma. *Cancer Cell* **39**, 1262–1278.e7 (2021).
28. To, T.-L. *et al.* Rational Design of a GFP-Based Fluorogenic Caspase Reporter for Imaging Apoptosis In Vivo. *Cell Chem. Biol.* **23**, 875–882 (2016).
29. Brown, T. B. *et al.* Language Models are Few-Shot Learners. (2020) doi:10.48550/ARXIV.2005.14165.
30. Fürst, A. *et al.* CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP. (2021) doi:10.48550/ARXIV.2110.11316.



Metric	Train: Strict Test: Strict	Train: Noisy Test: Strict
Binary Accuracy	0.868	0.934
Top-1 Accuracy	0.223	0.469
Top 10% Accuracy	0.633	0.891

Figure 1. PepPrCLIP model training, evaluation, and benchmarking. (A) Gaussian distributions centered around the ESM-2 embeddings of naturally-occurring peptides are sampled and then decoded back to amino acid sequences. A trained CLIP module, jointly encoding cognate peptide-protein embeddings, screens thousands of these peptides for specific binding activity to the target. (B) Evaluation metrics for final CLIP module. Binary accuracy is the accuracy of the model in predicting the correct binding pairs when given 2 protein-peptide pairs. Top-1 accuracy is the likelihood that for a given protein, the model has chosen the correct binding peptide, and Top 10% accuracy is the likelihood that this peptide is in the top 10% of peptides when ranked for binding affinity to this protein. Metric values are presented for the "strict" held-out test set described in the Results and Methods sections. (C) Known peptide-target sequence pairs, from the PDB-validated dataset, were scored via the trained CLIP module. Mismatched pairs, each containing a peptide and a non-matching protein, both from the validation dataset, were also scored via the CLIP module. +1 indicates a definitive binding pair while -1 indicates that the peptide and target sequences do not bind (see Methods). Scores are represented as a histogram. (D) Analysis of Hamming distance to source peptides vs. variance scaling factor (*k*) of embedding perturbation. (E) AlphaFold-Multimer benchmarking of PepPrCLIP. ipTM values were extracted post AlphaFold-Multimer co-folding with indicated peptides. Peptides are either experimentally-derived (dotted line), generated via Gaussian perturbation (PpC-Generated) or purely random. Generated peptides were selected for High CLIP scores, Low CLIP scores, or were randomly selected prior to CLIP screening (No CLIP). The 3D binding mode of the top PepPrCLIP-generated peptide and the experimentally-validated peptide are shown via ColabFold, shaded by pLDDT. Additional target results can be found in Supplementary Figure 1.

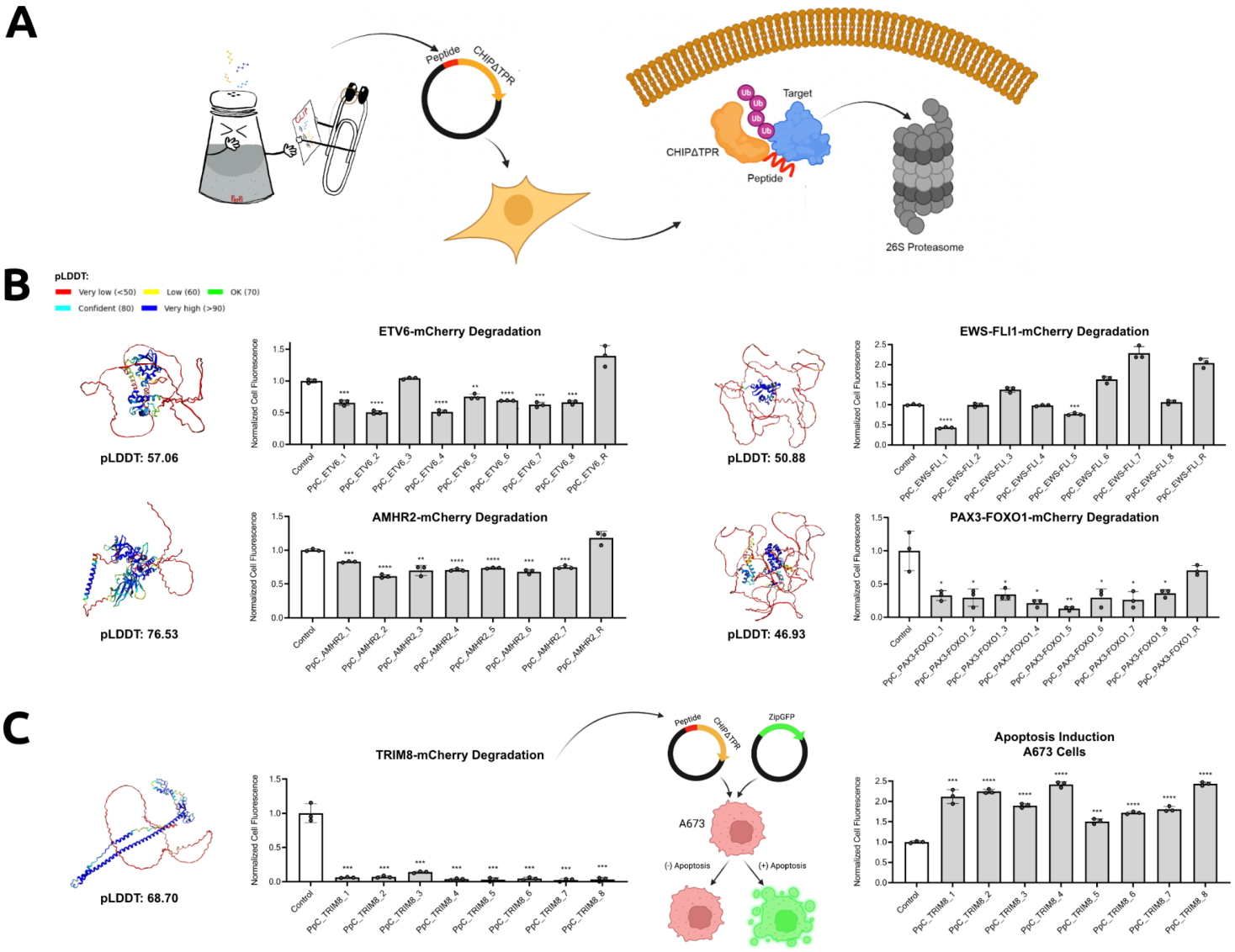


Figure 2. Characterization of PepPrCLIP-generated uAbs for target protein degradation. (A) Schematic of uAb degradation system. Briefly, PepPrCLIP-derived peptides are fused to the N-terminus of CHIPΔTPR in mammalian expression vectors. The resulting vector is transfected into mammalian cells, inducing binding, ubiquitination, and subsequent proteasomal degradation of a specified target protein. (B) Analyses of target structures via AlphaFold2 and uAb-mediated target-mCherry degradation via flow cytometry were conducted for the following targets: ETV6, AMHR2, EWS-FLI1, PAX3-FOXO1. AlphaFold2 pLDDT scores (1-100) represent average model confidence across all amino acids, 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 for reporter EGFP fluorescence. mCherry fluorescence levels were calculated relative to the control samples, which were treated with a non-targeting uAb. Normalized degradation was calculated by dividing the percentage of mCherry+ cells in a sample by that of the non-targeting control for each target. For individual samples, statistical significance was determined by a two-tailed Student's t test. Calculated p values are represented as follows: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$. (C) Degradation testing for TRIM8-mCherry was conducted as described in Figure 2B. For the apoptosis assay (shown in the schematic), A673 cells were co-transfected with equal ratios of the uAb plasmid and ZipGFP caspase reporter plasmids. Cells expressing mCherry (transfection reporter) were gated, and normalized EGFP cell fluorescence, indicating functional reporter activity, was calculated to samples transfected with a non-targeting uAb. For individual samples, statistical significance was determined by a two-tailed Student's t test. Calculated p values are represented as follows: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.