DESIGN OF PEPTIDE-BASED PROTEIN DEGRADERS VIA CONTRASTIVE DEEP LEARNING

A PREPRINT

Kalyan Palepu,[1,2,*] **Manvitha Ponnapati,**[1,3,4,*] **Suhaas Bhat,**[1,2]
Emma Tysinger,[3,4] **Teodora Stan,**[3,4] **Garyk Brixi,**[1,2] **Sabrina R.T. Koseki,**[3,4]
Pranam Chatterjee[1,2,3,4,†]

[1]Department of Biomedical Engineering, Duke University,
[2]Harvard University,
[3]Center for Bits and Atoms, MIT, [4]MIT Media Lab
[*]These authors contributed equally
[†]Corresponding author: pranam.chatterjee@duke.edu

May 24, 2022

ABSTRACT

Therapeutic modalities targeting pathogenic proteins are the gold standard of treatment for multiple disease indications. Unfortunately, a significant portion of these proteins are considered "undruggable" by standard small molecule-based approaches, largely due to their disordered nature and instability. Designing functional peptides to undruggable targets, either as standalone binders or fusions to effector domains, thus presents a unique opportunity for therapeutic intervention. In this work, we adapt recent models for contrastive language-image pre-training (CLIP) to devise a unified, sequence-based framework to design target-specific peptides. Furthermore, by leveraging known experimental binding proteins as scaffolds, we create a streamlined inference pipeline, termed **Cut**&**CLIP**, that efficiently selects peptides for downstream screening. Finally, we experimentally fuse candidate peptides to E3 ubiquitin ligase domains and demonstrate robust intracellular degradation of pathogenic protein targets in human cells, motivating further development of our technology for future clinical translation.

Introduction

It has been estimated that while nearly 15% of human proteins are disease-associated, only 10% of such proteins interact with currently-approved small molecule drugs. Even more strikingly, of the 650,000 protein-protein interactions (PPIs) in the proteome, only 2% are considered "druggable" by pharmocological means [Shin et al., 2020]. Aside from small molecule-based approaches, monoclonal antibodies have found significant success in the clinic as biologics, but while they are highly selective and can bind antigens with high specificity, they are limited to extracellular PPIs and cannot naturally permeate the cell membrane [Slastnikova et al., 2018].

Peptides have been widely recognized as a more selective, effective, and safe method for targeting pathogenic proteins, due to their sequence-specific binding to regions of partner molecules [Padhi et al., 2014, Buchwald et al., 2014]. They have further demonstrated targeting of both extracellular and intracellular proteins, due to their small size and enhanced permeability, with or without conjugation to cell penetrating peptide (CPP) sequences [Lindgren et al., 2000, Lozano et al., 2017, Adhikari et al., 2018]. Beyond standalone peptide binders, our group has recently fused computationally-designed peptides to effector domains, such as E3 ubiquitin ligases, to enable binding and selective intracellular degradation of pathogenic targets of interest [Chatterjee et al., 2020]. Extending this "ubiquibody" (uAb) strategy to undruggable targets, including numerous oncogenic and viral proteins, represents a promising new therapeutic approach.

Current approaches for peptide engineering have relied on high-throughput screening and structure-based rational design, with the goal of redirecting to alternate targets, extending half-life *in vivo*, improving solubility, or preventing aggregation [Fosgerau and Hoffmann, 2015]. Experimental methods, such as large phage display libraries and quantitative binding assays, while effective at selecting strong candidate sequences, are laborious and expensive [Wu et al., 2016, Kong et al., 2020, Carle et al., 2021]. Structure-based methods for peptide design consist of interface predictors and peptide-protein docking softwares [Raveh et al., 2011, Sedan et al., 2016, Tsaban et al., 2022]. These approaches, however, rely heavily on the existence of co-crystal complexes consisting of the target protein, thus excluding disordered or unstable proteins, such as transcription factors, which have significant disease implications and are difficult to solve via experimental or computational protein structure determination methods [Peterson et al., 2017, Das et al., 2018, Jumper et al., 2021]. Therefore, there is a need for the development of a sequence-based peptide generation platform, so as to rapidly and programmbly design peptides to any target protein, especially those for which minimal structural information exists.

The sequential structure of proteins, along with their hierarchical semantics, makes them a natural target for language modeling. Recently, language models have been pre-trained on over 200 million natural protein sequences to generate latent embeddings that grasp relevant physicochemical, functional, and most notably, tertiary structural information [Rives et al., 2021, Elnaggar et al., 2020, Vig et al., 2020, Rao et al., 2020]. Even more interestingly, generative protein transformers have produced novel protein sequences with validated functional capability [Madani et al., 2021]. Additionally, by augmenting input sequences with their evolutionarily-related counterparts, in the form of multiple sequence alignments (MSAs), the predictive power of protein language models can be further strengthened, as demonstrated by state-of-the-art contact prediction results [Rao et al., 2021].

In this study, we combine pre-trained protein language embeddings with novel contrastive learning architectures for the specific task of designing peptide sequences that bind target proteins and induce their degradation when fused to E3 ubiquitin ligase domains. By jointly training protein and peptide encoders to capture similarities between known peptide-protein pairs, our model accurately evaluates peptide inputs as potential binders for embedded target proteins. To further downselect initial peptide candidate lists for queried targets, we use predicted or experimentally-validated binding proteins as scaffolds for splicing, thus creating an integrated inference pipeline known as Cut&CLIP. We demonstrate that Cut&CLIP reliably and efficiently generates peptides which, when experimentally integrated within a uAb construct, induce robust degradation of pathogenic proteins in human cells.

Results

Dataset Curation and Augmentation

Previously, we demonstrated the utility of using scaffold proteins to derive functional peptides for uAb generation by executing the PeptiDerive protocol on co-crystals containing the target protein, thus identifying the linear polypeptide segments suggested to contribute most to binding energy [Chatterjee et al., 2020, Sedan et al., 2016]. To create a comprehensive dataset of these computationally-derived presumptive peptides, we applied PeptiDerive to complexes in the Database of Interacting Protein Structures (DIPS) [Sedan et al., 2016, Townshend et al., 2018]. We ran PeptiDerive on every co-crystal in DIPS with a resolution of ≤ 2 Å, and selected the top 20-mer peptides for each to include in the dataset, generating a total of 28,517 peptide-receptor pairs. Finally, we appended the Propedia dataset, an experimentally-derived database that includes an additional 19,814 peptide-receptor complexes from the Protein Data Bank (PDB) [Martins et al., 2021]. Together, our database represents one of the most comprehensive collections of peptide-protein pairs, and can thus serve as a standardized training set for future interface modeling.

For training, we specifically clustered the protein receptor sequences at 50% sequence identity using MM-Seq2, yielding 7434 clusters, and split the clusters into train, validation, and test splits according at a 0.7/0.15/0.15 ratio, respectively [Steinegger and Söding, 2018]. All sequences from the selected clusters were used in the train and validation splits, but only a single representative sequence for each cluster was employed for the test split in order to ensure a reasonable balance of sequence diversity.

Model Architecture and Training

One of the core problems in computer vision and natural language processing (NLP) is zero-shot classification. In the last year, OpenAI has introduced a new architecture, termed CLIP (Contrastive Language-Image Pre-Training), which utilizes zero-shot transfer and multimodal learning to associate visual concepts in images and link them with

their captions [Radford et al., 2021]. Thus, we hypothesized that just as CLIP connects images to their corresponding captions using jointly-trained image and caption encoders, we can leverage CLIP-based architecture to map target proteins to their corresponding peptides using jointly trained receptor and peptide encoders. To do this, we trained encoders such that the cosine similarity between receptor 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 receptor-peptide pairs which do bind to each other, and is near -1 for receptor-peptide pairs which do not bind to each other. As input, the receptor encoder uses an MSA, while the peptide encoder simply uses the peptide sequence.

Next, the receptor and peptide encoders were trained on batches of n pairs of receptors and peptides which are known to interact. Receptor MSAs 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 receptor and peptide pairs is computed in a matrix $\boldsymbol{K}$, defined such that

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

We then interpreted these cosine similarities as softmax logits. Logits were scaled by a learned temperature parameter t, which controls the model's degree of uncertainty in output probabilities [Hinton et al., 2015]. 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 receptor 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 $n^2 - n$ receptor-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 model was then trained on the average of these two losses. The entire training process is illustrated in Figure 1.

Specifically, receptor MSAs and peptide sequences were first input into the ESM pre-trained transformer protein language models introduced previously by Facebook [Rives et al., 2021, Rao et al., 2020]. These pre-trained models were trained on millions of diverse amino acid sequences, allowing our encoders to extract feature-rich embeddings, which are robust to sequence diversity while being trained on a relatively small dataset. We employed the ESM-MSA-1b model for the receptor MSAs, and ESM-1b for the peptide sequences, which does not require MSA inputs (Figure 1). We then trained the receptor and peptide encoders taking these ESM embeddings as input.

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 embedding (where l is the input sequence length and e_i is the dimension of the ESM embedding), we applied h_1 feedforward layers 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. Then, we then averaged the embedding over the length dimension, producing an embedding vector of length e_o. Finally, we applied h_2 feedforward layers with ReLU activation on the embedding vector to get 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.

As a relevant metric for model assessment in the context of screening, we calculated the top-k accuracy, which is the probability that the correct peptide is in the top k when provided a fixed batch of 250 candidate peptides, a
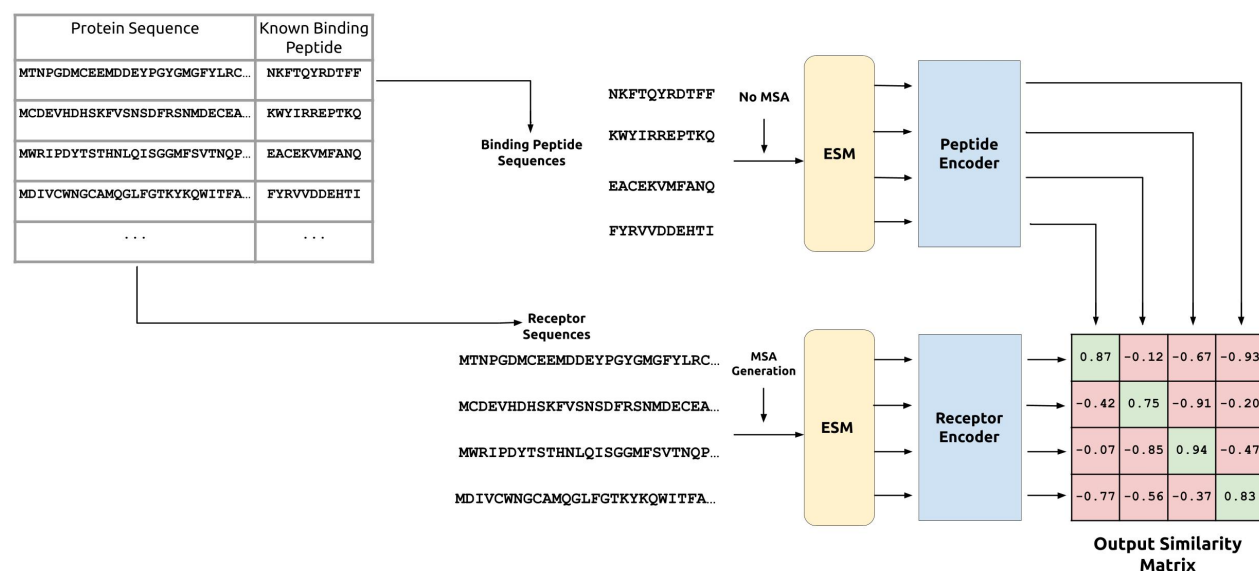
Figure 1: **CLIP training process for peptide-protein pairs.** Peptide and receptor encoders are jointly trained on ESM embeddings to predict high cosine similarities between known peptide-receptor embedding pairs and low cosine similarities for all other pairs.
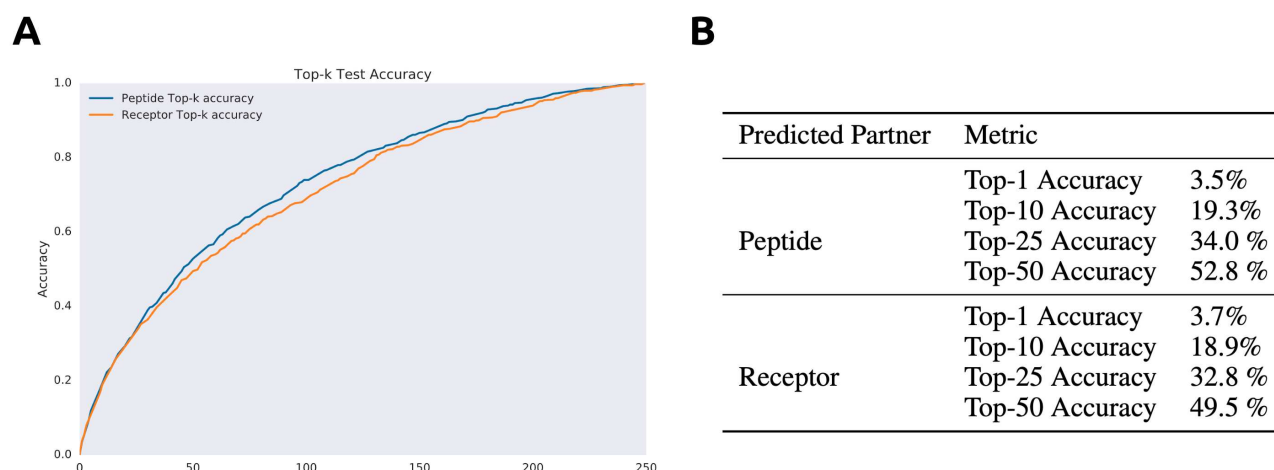

Figure 2: **Results of model validation and testing.** A) Top-k accuracy of predicting the correct binding partner out of a batch of 250. B) Selected test results. Accuracies are calculated via selection of the known binding partner out of a batch of 250 to a queried target. All results from an ensemble of the top four trained models.

suitable threshold for genetic screening. To calculate this metric, during prediction, we provide the model with a single protein target receptor and 250 peptides from our training set, one of which is a known binder.

Post-training, our final models demonstrate accurate ranking of known targeting peptides for a given target and vice versa, achieving 50% probability of identifying a correct candidate in the ranked top 50 out of 250, for example. These results motivate not only model application for tractable peptide screening assays, but also its utilization to evaluate peptide specificity to a desired target, in comparison to off-target receptor proteins (Figure 2).
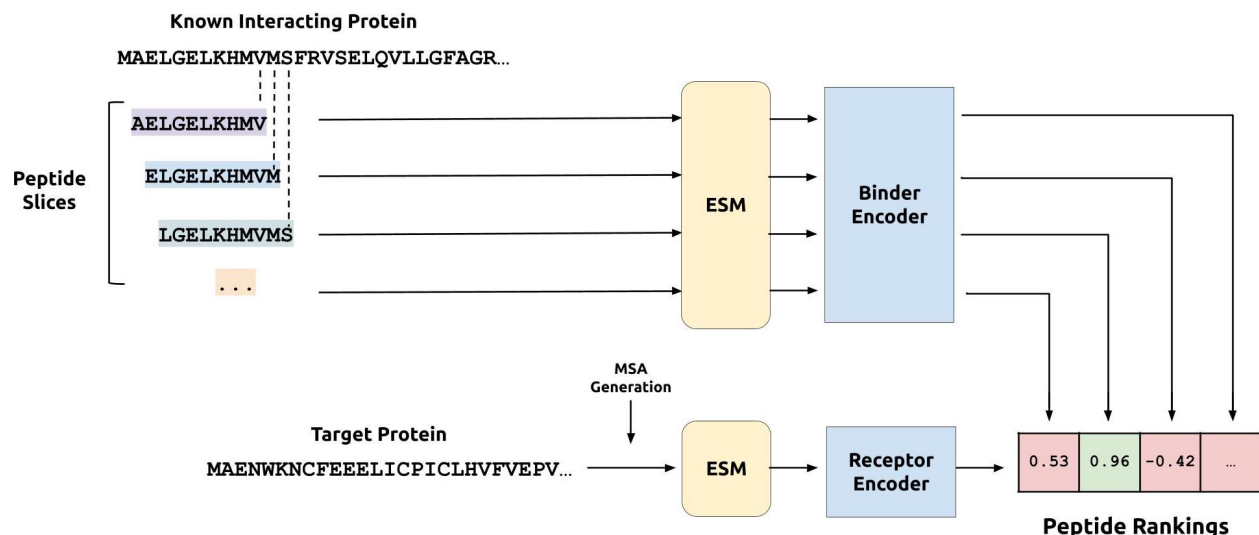
4

Figure 3: **The Cut&CLIP inference protocol.** A known interacting protein which is validated to interact with the target protein is cut up into peptide-size slices, enabling downstream ranking via the trained CLIP model.

Cut&CLIP Inference

Inspired by our previous results leveraging co-crystals of target proteins and known binders to design effective peptide-based degraders, we decided to employ our CLIP model to predict binding peptides using experimentally-validated interacting proteins for a queried target [Chatterjee et al., 2020]. As opposed to our previous work, the current inference pipeline simply requires the sequence of potential binders from established PPI databases or experimental screening results, enabling us to be more flexible in identifying starting scaffolds [Szklarczyk et al., 2020, Johnson et al., 2021]. We then compute the CLIP peptide embedding for all k-mers of the interacting protein (where k is the desired size of our peptide), and rank them by their cosine similarities with the CLIP receptor embedding of the target protein. Our final Cut&CLIP pipeline is illustrated in Figure 3.

While this strategy may not be sufficient to identify stand-alone peptide binders with high target affinity, our motivation derives from the well-understood catalytic nature of E3 ubiquitin ligase activity, where selective target binding is sufficient to induce degradation [Békés et al., 2022, Buetow and Huang, 2016, Portnoff et al., 2014].

Target Protein Degradation with Cut&CLIP-Derived Peptides

Numerous previous works have attempted to redirect E3 ubiquitin ligases by replacing their natural protein binding domains with those targeting specific proteins [Gosink and Vierstra, 1995, Zhou et al., 2000, Su et al., 2003]. Recently, based on the seminal work of Portnoff, et al., our group 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 peptides to generate minimal and programmable uAb architectures [Portnoff et al., 2014, Chatterjee et al., 2020]. To evaluate Cut&CLIP's utility as compared to a slower structure-based method using AlphaFold [Jumper et al., 2021], we selected three target proteins for experimental characterization: the spike receptor binding domain (RBD) of SARS-CoV-2, the TRIM8 E3 ubiquitin ligase, and the KRAS oncoprotein.

Previously, we demonstrated robust degradation of RBD using peptide-based uAbs, and with stable co-crystal structures of RBD and the human ACE2 receptor, it represents a very tractable target for standard structure-based peptide generation [Chatterjee et al., 2020, Lan et al., 2020]. TRIM8 regulates EWS-FLI1 protein degradation in Ewing sarcoma and its depletion results in EWS/FLI1-mediated oncogene overdose, driving DNA damage and apoptosis of tumor cells [Seong et al., 2021]. Thus, as an E3 ubiquitin ligase itself, TRIM8 presents a unique target for therapeutic degradation. Finally, KRAS is the most frequently mutated oncoprotein, occurring in over 25% of all cancer patients. Due to its smooth and shallow surface, it is considered largely undruggable by standard small molecules, and its structure is evasive due to its conformational disorder as a transcription factor protein [Huang et al., 2021].

We searched existing PPI databases to identify putative interacting partners of our three targets: ACE2 for
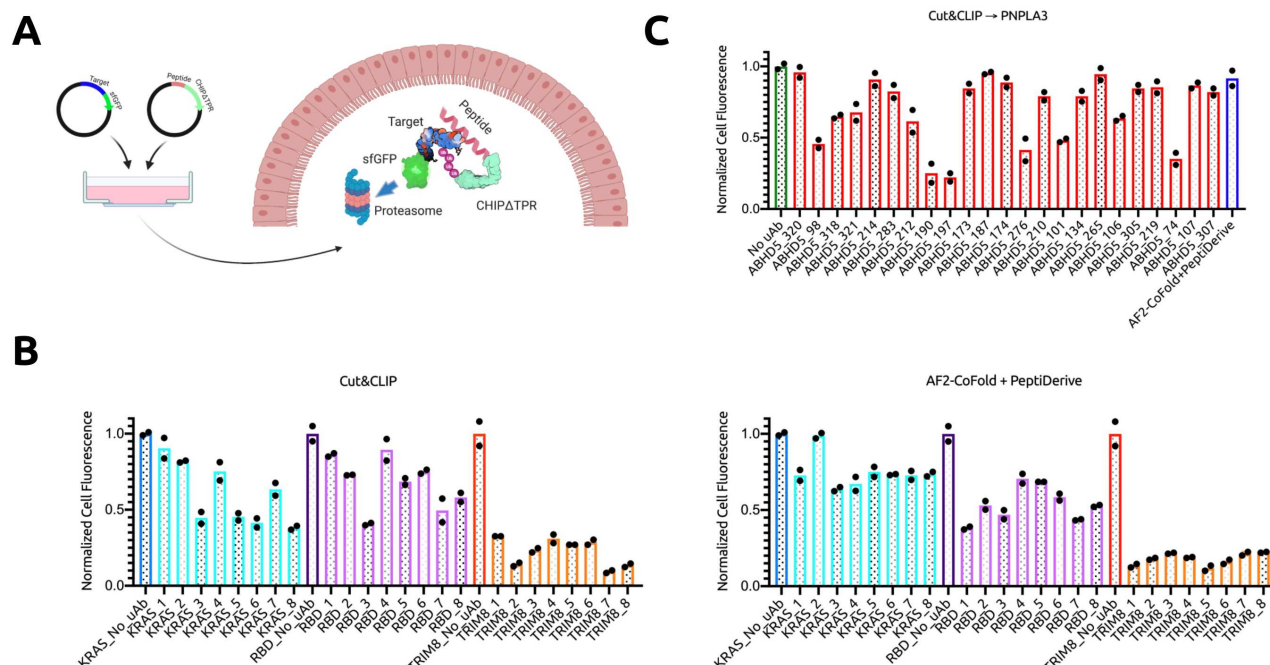
Figure 4: **Experimental validation of Cut&CLIP.** A) Architecture and mechanism of uAb degradation system. CHIPΔTPR is fused to the C-terminus of targeting peptides, and can thus tag target-sfGFP complexes for ubiquitin-mediated degradation in the proteasome, post-plasmid transfection. B) Analysis of KRAS-sfGFP, RBD-sfGFP, and TRIM8-sfGFP degradation via flow cytometry. All samples were performed in independent transfection duplicates (n=2) and gated on GFP+ fluorescence. Normalized cell fluorescence was calculated by dividing the %GFP+ of samples to that of their respective "No uAb" control. C) Analysis of PNPLA3-sfGFP degradation via flow cytometry. All samples were performed in independent transfection duplicates (n=2) and gated on GFP+ fluorescence. Normalized cell fluorescence was calculated by dividing the %GFP+ of samples to that of the "No uAb" control. The final peptide was derived from the AF2-CoFold+PeptiDerive strategy on PNPLA3-ABDH5.

RBD, PIAS3 for TRIM8, and RAF1 for KRAS [Szklarczyk et al., 2020]. We input these pairs into both our Cut&CLIP pipeline, as well as a co-folding pipeline that adapts the AlphaFold-Multimer complex prediction algorithm followed by PeptiDerive (AF2-CoFold+PeptiDerive) [Evans et al., 2021, Sedan et al., 2016]. After candidate peptide derivation, we experimentally cloned plasmids expressing eight peptides of variable lengths (≤18 amino acids) for each target, directly fused to the CHIPΔTPR uAb domain via a short glycine-serine linker (GSGSG). We subsequently co-transfected these vectors into human HEK293T cells alongside plasmids expressing the target protein fused to superfolder green fluorescent protein (sfGFP) and analyzed the reduction of GFP+ signal (and thus target degradation) via flow cytometry (Figure 4A). Our results demonstrate that select Cut&CLIP-derived peptides induce robust target degradation for all three targets, even on the "undruggable" KRAS oncoprotein. In comparison, the structure-based strategy, while successfully identifying degraders to RBD and TRIM8, failed to produce constructs that induce $> 50\%$ degradation of KRAS (Figure 4B).

Finally, as uAbs are genetically-encoded constructs, their therapeutic application is limited by the need for *in vivo* delivery vehicles, most of which home to the liver, such as lipid nanoparticles (LNPs) [Hou et al., 2021]. Thus, to extended Cut&CLIP's utility to a viable therapeutic target, we designed peptides to PNPLA3, a known driver of fatty liver disease, by employing its direct interacting protein, ABHD5 [Yang et al., 2019]. Post transfection and flow cytometry, we show that Cut&CLIP identifies potent ABHD5-derived peptides that enable over 80% degradation of PNPLA3, thus motivating potential clinical translation of our technology (Figure 4C).

Discussion

Eradicating disease is one of the greatest challenges for the future of human health, and protein-targeting therapeutics have served as potent solutions to this problem. As an example, targeted protein degradation with proteolysis targeting chimeras (PROTACs) and molecular glues utilize small molecules to bind to intracellular proteins transiently and direct

their proteolysis by recruiting endogenous E3 ubiquitin ligases [Békés et al., 2022]. More recently, the development of the uAb technology has provided a modular, genetically-encoded alternative to achieve selective degradation of potentially "undruggable" proteins [Portnoff et al., 2014, Chatterjee et al., 2020].

In this work, we exploit recent advancements in contrastive deep learning to design peptides to specified target proteins. Our final models accurately retrieve peptides for known protein-peptide pairs, and more importantly, prioritize candidates that demonstrate effective intracellular target degradation when integrated into the uAb architecture. Our final Cut&CLIP model employs natural binding partners as scaffolds for peptide generation, thus representing a streamlined, efficient, sequence-based pipeline to generate degraders to diverse proteins in the proteome.

In the past few years, protein structure prediction has experienced a wave of excitement with the advent of AlphaFold2 [Jumper et al., 2021]. With these prediction methods in hand, the protein design community now possesses new tools to generate custom proteins with enhanced or novel functionality [Anishchenko et al., 2021, Cao et al., 2022]. In this work, we provide a use-case for which methods like AlphaFold2 may be inferior to pure sequence-based models like Cut&CLIP. Though the AF2-CoFold+PeptiDerive pipeline described in this study managed to produce viable degraders, it struggled to predict large and disordered protein complexes, highlighting its main drawback: efficiency. To generate TRIM8-targeting peptides from PIAS3, the AF2-CoFold+PeptiDerive pipeline required 3 hours, 17 minutes, and 50 seconds on a powerful AWS p3.2xlarge instance with 8 CPU cores and a Nvidia V100 GPU, resources to which many researchers do not have access. Cut&CLIP, on the other hand, only required 15 minutes and 58 seconds for the equivalent design task (including receptor MSA generation) on a standard machine with 2 CPU cores, 8 GB of RAM, and no GPU. Additionally, while both models produced highly effective peptides for TRIM8 and RBD, only Cut&CLIP produced effective degraders for one of the most challenging cancer targets, KRAS.

To further contextualize the power of contrastive sequence-based models for protein design and screening, the model results shown here are based upon the strong assumption that within a batch of 250 peptides, only one is a viable binder. In most applications, especially when using a known interacting partner as a scaffold for peptide generation, there are likely multiple candidates that bind to the queried target. Our experimental results support this observation, as potent degraders were identified by only testing 8 candidates for KRAS, RBD, and TRIM8.

Furthermore, just as OpenAI's CLIP joint image-text embedding space enables generative models like DALL-E 2 to generate images conditioned on a caption embedding, our joint receptor-peptide embedding space naturally indicates opportunities for *de novo* generation of peptides [Ramesh et al., 2022]. Once we have a target protein embedding, the optimal peptide embedding is known, and all that needs to be done is to decode the peptide embedding back into a sequence. This could be especially critical for generating degraders to targets where minimal experimental evidence exists of potential interacting partners. Considering this versatility, the CLIP architecture thus serves as an ideal framework for target-specified peptide design, as compared to more standard models.

Overall, this work represents an initial application of sequence-based language models to therapeutically-relevant protein design. Future iterations of Cut&CLIP, for example, will incorporate K_d values for high-affinity peptide design and predict the off-targeting propensity of generated sequences. Most importantly, by integrating Cut&CLIP and uAb technology with effective delivery vehicles, such as adeno-associated vectors (AAVs) or LNPs, the peptide-guided protein degradation platform presented here may eventually serve as a potent therapeutic strategy to address a host of diseases deemed untreatable by standard small molecule-based means.

Methods

Resources and Model Evaluation

Model architecture, loss functions, and inference methodologies are described in the Results section. Models were trained on a single Nvidia V100 GPU with 32 GB VRAM, as well as 10 CPU cores with a total of 90 GB of RAM. Top-k accuracy was calculated to be the probability that the correct peptide is in the top k when provided a fixed batch of 250 candidate peptides. Peptide inference was conducted on a machine with 2 CPU cores and 8 GB of RAM.

Generation of Plasmids

pcDNA3-SARS-CoV-2-S-RBD-sfGFP (Addgene #141184) and pcDNA3-R4-uAb (Addgene #101800) were obtained as gifts from Erik Procko and Matthew DeLisa, respectively. Target coding sequences (CDS) were synthesized as gBlocks from Integrated DNA Technologies (IDT). Sequences was amplified with overhangs for Gibson Assembly-mediated insertion into the pcDNA3-SARS-CoV-2-S-RBD-sfGFP backbone linearized by digestion with NheI and BamHI. An

Esp3I restriction site was introduced immediately upstream of the CHIPΔTPR CDS and GSGSG linker via the KLD Enzyme Mix (NEB) following PCR amplification with mutagenic primers (Genewiz). For peptide CDS assembly, 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.

Cell Culture and Flow Cytometry

HEK293T 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). Target-sfGFP (50 ng) and peptide-CHIPΔTPR (50 ng) plasmids were transfected into cells as duplicates ($2x10^4$/well in a 96-well plate) with Lipofectamine 3000 (Invitrogen) in Opti-MEM (Gibco). After 3 days post transfection, cells were harvested and analyzed on a FACSCelesta for GFP fluorescence (488-nm laser excitation, 530/30 filter for detection). Cells expressing GFP were gated as compared to a GFP- control, and normalized cell fluorescence was calculated to the "No uAb" control.

Statistics and Reproducibility

All samples were performed in independent transfection duplicates (n=2), and normalized cell fluorescence values were averaged.

Declarations

Acknowledgements

Author Contributions

P.C. conceived, designed, directed, and supervised the study. K.P. and M.P. curated peptide-protein data. K.P. conceived, designed, and implemented the Cut&CLIP workflow, with assistance from S.B. and G.B. M.P. and S.B. conceived, designed, and implemented the AF2-CoFold+PeptiDerive workflow. E.T., T.S., and S.R.T.K. built constructs, conducted transfections, carried out degradation experiments, and performed data analyses. P.C. and K.P. wrote the paper, with input from all authors.

Data and Materials Availability

All data needed to evaluate the conclusions in the paper are present in the paper and tables. All peptide sequences, source data, and results can be found at: `https://tinyurl.com/cutnclip`. Cut&CLIP code will be made available on GitHub upon publication. uAb expression plasmids will be deposited to Addgene.

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. K.P. and S.B. are consultants for UbiquiTx, Inc.

References

[Adhikari et al., 2018] Adhikari, S., Alahmadi, T. I., Gong, Z., and Karlsson, A. J. (2018). Expression of cell-penetrating peptides fused to protein cargo. *Journal of Molecular Microbiology and Biotechnology*, 28(4):159–168.

[Anishchenko et al., 2021] Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T., and Baker, D. (2021). De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552.

[Békés et al., 2022] Békés, M., Langley, D. R., and Crews, C. M. (2022). PROTAC targeted protein degraders: the past is prologue. *Nature Reviews Drug Discovery*, 21(3):181–200.

[Buchwald et al., 2014] Buchwald, H., Dorman, R. B., Rasmus, N. F., Michalek, V. N., Landvik, N. M., and Ikramuddin, S. (2014). Effects on GLP-1, PYY, and leptin by direct stimulation of terminal ileum and cecum in humans: implications for ileal transposition. *Surgery for Obesity and Related Diseases*, 10(5):780–786.

[Buetow and Huang, 2016] Buetow, L. and Huang, D. T. (2016). Structural insights into the catalysis and regulation of e3 ubiquitin ligases. *Nature Reviews Molecular Cell Biology*, 17(10):626–642.

[Cao et al., 2022] Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J. S., Jude, K. M., Marković, I., Kadam, R. U., Verschueren, K. H. G., Verstraete, K., Walsh, S. T. R., Bennett, N., Phal, A., Yang, A., Kozodoy, L., DeWitt, M., Picton, L., Miller, L., Strauch, E.-M., DeBouver, N. D., Pires, A., Bera, A. K., Halabiya, S., Hammerson, B., Yang, W., Bernard, S., Stewart, L., Wilson, I. A., Ruohola-Baker, H., Schlessinger, J., Lee, S., Savvides, S. N., Garcia, K. C., and Baker, D. (2022). Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560.

[Carle et al., 2021] Carle, V., Kong, X.-D., Comberlato, A., Edwards, C., Díaz-Perlas, C., and Heinis, C. (2021). Generation of a 100-billion cyclic peptide phage display library having a high skeletal diversity. *Protein Engineering, Design and Selection*, 34.

[Chatterjee et al., 2020] Chatterjee, P., Ponnapati, M., Kramme, C., Plesa, A. M., Church, G. M., and Jacobson, J. M. (2020). Targeted intracellular degradation of SARS-CoV-2 via computationally optimized peptide fusions. *Communications Biology*, 3(1).

[Das et al., 2018] Das, P., Matysiak, S., and Mittal, J. (2018). Looking at the disordered proteins through the computational microscope. *ACS Central Science*, 4(5):534–542.

[Elnaggar et al., 2020] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2020). ProtTrans: Towards cracking the language of life's code through self-supervised learning. *bioRxiv*.

[Evans et al., 2021] Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Protein complex prediction with AlphaFold-multimer. *biorXiv*.

[Fosgerau and Hoffmann, 2015] Fosgerau, K. and Hoffmann, T. (2015). Peptide therapeutics: current status and future directions. *Drug Discovery Today*, 20(1):122–128.

[Gosink and Vierstra, 1995] Gosink, M. M. and Vierstra, R. D. (1995). Redirecting the specificity of ubiquitination by modifying ubiquitin-conjugating enzymes. *Proceedings of the National Academy of Sciences*, 92(20):9117–9121.

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*.

[Hou et al., 2021] Hou, X., Zaks, T., Langer, R., and Dong, Y. (2021). Lipid nanoparticles for mRNA delivery. *Nature Reviews Materials*, 6(12):1078–1094.

[Huang et al., 2021] Huang, L., Guo, Z., Wang, F., and Fu, L. (2021). KRAS mutation: from undruggable to druggable in cancer. *Signal Transduction and Targeted Therapy*, 6(1).

[Johnson et al., 2021] Johnson, K. L., Qi, Z., Yan, Z., Wen, X., Nguyen, T. C., Zaleta-Rivera, K., Chen, C.-J, Fan, X., Sriram, K., Wan, X., Chen, Z. B., and Zhong, S. (2021). Revealing protein-protein interactions at the transcriptome scale by sequencing. *Molecular Cell*, 81(19):4091–4103.e9.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

[Kong et al., 2020] Kong, X.-D., Carle, V., Díaz-Perlas, C., Butler, K., and Heinis, C. (2020). Generation of a large peptide phage display library by self-ligation of whole-plasmid PCR product. *ACS Chemical Biology*, 15(11):2907–2915.

[Lan et al., 2020] Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., and Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581(7807):215–220.

[Lindgren et al., 2000] Lindgren, M., Hällbrink, M., Prochiantz, A., and Ülo Langel (2000). Cell-penetrating peptides. *Trends in Pharmacological Sciences*, 21(3):99–103.

[Lozano et al., 2017] Lozano, T., Gorraiz, M., Lasarte-Cía, A., Ruiz, M., Rabal, O., Oyarzabal, J., Hervás-Stubbs, S., Llopiz, D., Sarobe, P., Prieto, J., Casares, N., and Lasarte, J. J. (2017). Blockage of FOXP3 transcription factor dimerization and FOXP3/AML1 interaction inhibits t regulatory cell activity: sequence optimization of a peptide inhibitor. *Oncotarget*, 8(42):71709–71724.

[Madani et al., 2021] Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. (2021). Deep neural language modeling enables functional protein generation across families. *bioRxiv*.

[Martins et al., 2021] Martins, P. M., Santos, L. H., Mariano, D., Queiroz, F. C., Bastos, L. L., de S. Gomes, I., Fischer, P. H. C., Rocha, R. E. O., Silveira, S. A., de Lima, L. H. F., de Magalhães, M. T. Q., Oliveira, M. G. A., and de Melo-Minardi, R. C. (2021). Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm. *BMC Bioinformatics*, 22(1).

[Padhi et al., 2014] Padhi, A., Sengupta, M., Sengupta, S., Roehm, K. H., and Sonawane, A. (2014). Antimicrobial peptides and proteins in mycobacterial therapy: Current status and future prospects. *Tuberculosis*, 94(4):363–373.

[Peterson et al., 2017] Peterson, L. X., Roy, A., Christoffer, C., Terashi, G., and Kihara, D. (2017). Modeling disordered protein interactions from biophysical principles. *PLOS Computational Biology*, 13(4):e1005485.

[Portnoff et al., 2014] Portnoff, A. D., Stephens, E. A., Varner, J. D., and DeLisa, M. P. (2014). Ubiquibodies, synthetic e3 ubiquitin ligases endowed with unnatural substrate specificity for targeted protein silencing. *Journal of Biological Chemistry*, 289(11):7844–7855.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv*.

[Ramesh et al., 2022] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv*.

[Rao et al., 2021] Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., and Rives, A. (2021). MSA transformer. *bioRxiv*.

[Rao et al., 2020] Rao, R. M., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. (2020). Transformer protein language models are unsupervised structure learners. *biorXiv*.

[Raveh et al., 2011] Raveh, B., London, N., Zimmerman, L., and Schueler-Furman, O. (2011). Rosetta FlexPepDock ab-initio: Simultaneous folding, docking and refinement of peptides onto their receptors. *PLoS ONE*, 6(4):e18934.

[Rives et al., 2021] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.

[Sedan et al., 2016] Sedan, Y., Marcu, O., Lyskov, S., and Schueler-Furman, O. (2016). Peptiderive server: derive peptide inhibitors from protein–protein interactions. *Nucleic Acids Research*, 44(W1):W536–W541.

[Seong et al., 2021] Seong, B. K. A., Dharia, N. V., Lin, S., Donovan, K. A., Chong, S., Robichaud, A., Conway, A., Hamze, A., Ross, L., Alexe, G., Adane, B., Nabet, B., Ferguson, F. M., Stolte, B., Wang, E. J., Sun, J., Darzacq, X., Piccioni, F., Gray, N. S., Fischer, E. S., and Stegmaier, K. (2021). TRIM8 modulates the EWS/FLI oncoprotein to promote survival in ewing sarcoma. *Cancer Cell*, 39(9):1262–1278.e7.

[Shin et al., 2020] Shin, W.-H., Kumazawa, K., Imai, K., Hirokawa, T., and Kihara, D. (2020). pcurrent challenges and opportunities in designing protein–protein interaction targeted drugs/p. *Advances and Applications in Bioinformatics and Chemistry*, Volume 13:11–25.

[Slastnikova et al., 2018] Slastnikova, T. A., Ulasov, A. V., Rosenkranz, A. A., and Sobolev, A. S. (2018). Targeted intracellular delivery of antibodies: The state of the art. *Frontiers in Pharmacology*, 9.

[Steinegger and Söding, 2018] Steinegger, M. and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1).

[Su et al., 2003] Su, Y., Ishikawa, S., Kojima, M., and Liu, B. (2003). Eradication of pathogenic β-catenin by skp1/cullin/f box ubiquitination machinery. *Proceedings of the National Academy of Sciences*, 100(22):12729–12734.

[Szklarczyk et al., 2020] Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., and von Mering, C. (2020). The STRING database in 2021:

customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612.

[Townshend et al., 2018] Townshend, R. J. L., Bedi, R., Suriana, P. A., and Dror, R. O. (2018). End-to-end learning on 3d protein structure for interface prediction. *arXiv*.

[Tsaban et al., 2022] Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A., and Schueler-Furman, O. (2022). Harnessing protein folding neural networks for peptide–protein docking. *Nature Communications*, 13(1).

[Vig et al., 2020] Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. (2020). BERTology meets biology: Interpreting attention in protein language models. *biorXiv*.

[Wu et al., 2016] Wu, C.-H., Liu, I.-J., Lu, R.-M., and Wu, H.-C. (2016). Advancement and applications of peptide phage display technology in biomedical science. *Journal of Biomedical Science*, 23(1).

[Yang et al., 2019] Yang, A., Mottillo, E. P., Mladenovic-Lucas, L., Zhou, L., and Granneman, J. G. (2019). Dynamic interactions of ABHD5 with PNPLA3 regulate triacylglycerol metabolism in brown adipocytes. *Nature Metabolism*, 1(5):560–569.

[Zhou et al., 2000] Zhou, P., Bogacki, R., McReynolds, L., and Howley, P. M. (2000). Harnessing the ubiquitination machinery to target the degradation of specific cellular proteins. *Molecular Cell*, 6(3):751–756.