muPPIt: *De Novo* Generation of Mutant-Specific Peptide Binders via Conditional Uniform Discrete Diffusion

Tong Chen,[1] **Pranam Chatterjee**[1,†]

[1]Duke University, Durham, NC

[†]Corresponding author: pranam.chatterjee@duke.edu

Abstract

The ability to selectively target disease-causing mutations in proteins, such as oncogenic mutations in cancer or pathogenic mutations in neurodegenerative diseases, is crucial for developing precise therapeutics that minimize off-target effects. Current approaches often lack the specificity required to distinguish between mutant and wildtype proteins, particularly in the absence of detailed structural information. In this work, we introduce **muPPIt**, a **mu**tant-specific **PPI t**argeting algorithm designed for the *de novo* generation of mutant-specific peptide binders based solely on mutant and wildtype sequences. At the core of muPPIt is **MutBind**, an attention-based model that differentiates between mutant and wildtype protein language model embeddings, achieving over 70% test accuracy in predicting binding probabilities. Additionally, we present **PepUDLM**, a uniform diffusion language model that generates diverse and biologically plausible peptides. By integrating MutBind's predictions into PepUDLM's sampling process, muPPIt efficiently designs peptides that specifically bind to mutant proteins. We demonstrate muPPIt's effectiveness in computationally designing mutant-specific binders for a range of targets, including disease-related protein variants. In total, muPPIt serves as a powerful tool for developing highly specific peptide therapeutics, enabling precise targeting of mutant proteins without relying on structural information or structure-dependent latent spaces.

1 Introduction

Mutant-specific targeting of protein-protein interactions (PPIs) offers a promising strategy for developing therapeutics that can precisely target disease-causing mutations without affecting the wildtype protein. Designing binders that selectively recognize mutations that drive disease can provide a new pathway for treating conditions where conventional therapies fall short. For instance, in sickle cell anemia, the E6K mutation in HBB leads to the production of abnormal hemoglobin that causes red blood cells to sickle Pauling et al. (1949); Eaton & Bunn (2017); Abraham & Tisdale (2021); binders targeting this specific mutation could prevent sickling without interfering with normal hemoglobin function. In cancers, the G12V mutation in H-Ras results in constitutive activation of signaling pathways that drive tumor growth Prior et al. (2012); Simanshu et al. (2017); mutant-specific binders could inhibit the oncogenic activity of mutant H-Ras while leaving the wildtype protein unaffected. Similarly, in ALS, the A4V mutation in SOD1 causes toxic protein aggregation, but designing binders that selectively stabilize or clear the mutant SOD1 could help prevent neurodegeneration, providing a more targeted treatment approach Rosen et al. (1993); Bruijn et al. (2004); Saccon et al. (2013).

While experimental methods for generating mutant-specific binders, such as phage display, yeast display, and high-throughput screening, are often costly and labor-intensive, computational approaches offer a promising avenue for more efficient binder design Chen et al. (2023). Although recent advances have improved the prediction of mutation-induced changes in binding free energy ($\Delta\Delta G$) caused by mutations Wu et al. (2024); Cheng et al. (2024); Jemimah et al. (2020), existing computational framework provides an end-to-end solution for designing binders specifically targeting mutant

proteins. Furthermore, while generative models have made significant strides in *de novo* design within discrete data spaces and property-guided generation Sahoo et al. (2024) Schiff et al. (2024), these methods have not yet been focused on generating binders with a specific mutant-targeting property, leaving a significant gap in our ability to design mutant-specific therapeutics.

To address this gap, in this work, we develop a **mu**tant-specific **PPI t**argeting algorithm, termed **muPPIt**, that enables the design of mutant-specific peptide binders. To enable muPPIt-based generation, we develop **MutBind**, attention-based model differentiating the joint protein language model (pLM) embeddings of binder-mutant and binder-wildtype to predict the relative probabilities of a binder binding to the mutant (p) and to the wildtype $(1-p)$. Considering the limited size of the public SKEMPI dataset Jankauskaitė et al. (2019), we constructed a large dataset, **PPIMut**, containing binders, mutant proteins, wildtype proteins, and the binding affinities of binder-mutant and binder-wildtype complexes. Trained on the combination of PPIMut and SKEMPI dataset, MutBind achieves an over 70% test accuracy. We further trained PepUDLM that generates diverse and biologically plausible peptides, a uniform diffusion language model trained on a custom dataset, comprising peptides from the PepNN, BioLip2, and PPIRef dataset Abdin et al. (2022); Zhang et al. (2024); Bushuiev et al. (2023). muPPIt integrates MutBind into PepUDLM's sampling process, where MutBind's predictions guide PepUDLM to generate binders specifically targeting mutant proteins. We demonstrate muPPIt's efficacy on a diverse set of targets with various mutation levels, as well as disease-related protein variants. Using a combination of AlphaFold3, PyRosetta, and AlphaFold-Multimer, we computationally validate the specificity of muPPIt-designed peptides to the mutants Abramson et al. (2024); Chaudhury et al. (2010); Kim et al. (2024). Our comprehensive approach allows muPPIt to efficiently design highly selective peptide binders that specifically target mutated proteins, paving the way for novel therapeutic strategies.

2 Methods and Results

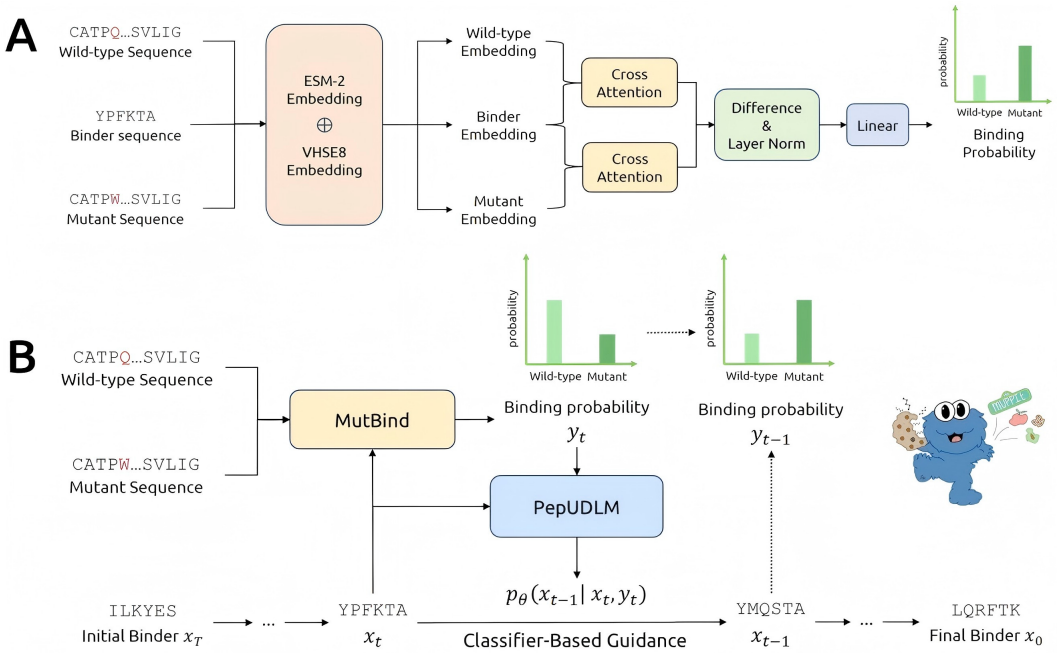
2.1 MutBind Predicts Binding Preferences for wildtype and Mutant Proteins

To enable the generation of mutant-specific binders, we developed **MutBind**, a model designed to predict the binding preference of a binder between wildtype and mutant proteins (Figure 1A-B). Specifically, MutBind takes as input a binder sequence, a wildtype protein sequence, and a mutant protein sequence, and outputs the binding probability p for the binder interacting with the wildtype protein and $1-p$ for the binder interacting with the mutant protein. The three input sequences are first embedded using the pre-trained ESM-2-650M model Lin et al. (2022). These embeddings are then concatenated with VHSE8 embeddings, which encode essential physical and chemical properties critical for biomolecular interactions Mei et al. (2005). A multi-head cross-attention module is employed to capture interaction information between the binder and both the wildtype and mutant proteins. The difference between the resulting representations is mapped to binding probabilities via a linear layer. To validate the contribution of VHSE8 embeddings, we trained an ablated version of the model without VHSE8 embeddings, which demonstrated inferior performance compared to the full model (Figure 1C).

Given that the publicly available SKEMPI dataset contains only 1,058 entries after processing, which is insufficient for training MutBind, we constructed a novel, large-scale dataset, **PPIMut**, comprising 19,704 entries. This dataset includes binder, wildtype, and mutant sequences, along with binding affinities between the binder and both wildtype and mutant proteins Jankauskaitė et al. (2019) (Appendix A.1). Trained on the combined PPIMut and SKEMPI datasets, MutBind demonstrated strong performance on the test set, achieving an accuracy exceeding 0.7. This is particularly notable given that over 50% mutants in the test data only have less than 15% amino acid difference relative to the total sequence length compared with the wildtype proteins.

2.2 PepUDLM Generates Diverse and Biologically Plausible Peptides

To enable the efficient generation of peptide binders, we developed an unconditional peptide generator, **PepUDLM**, based on the Uniform Diffusion Language Model (UDLM) Schiff et al. (2024). UDLMs can reverse random token perturbations and continuously edit discrete data, making them highly suitable for guided generation. We trained PepUDLM on a custom dataset that includes all peptides from the PepNN and BioLip2 datasets, as well as sequences from the PPIRef dataset with lengths



C

Test Metrics	w/ VHSE8	w/o VHSE8
Loss	0.0126	0.0154
Accuracy	0.7071	0.6972

D

Test Metrics	PepUDLM	PepAR
Loss	2.51	4.20
BPD	3.77	6.86
NLL	2.61	4.75
PPL	13.62	116.10

Figure 1: **(A) Overview of the architecture of MutBind.** MutBind predicts the relative probabilities of a binder interacting with the wildtype protein versus the mutant protein. **(B) Schematic of muPPIt.** muPPIt employs a pre-trained PepUDLM model to sample peptides that bind specifically to the mutant protein in a diffusion process guided by MutBind. **(C)** Performance comparison of MutBind with and without VHSE8 embedding. **(D)** Test performance metrics of PepUDLM and an auto-regressive model trained on the same dataset.

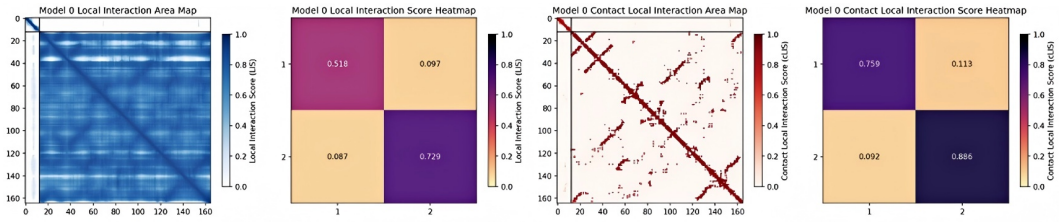
ranging from 6 to 49 amino acids Abdin et al. (2022); Zhang et al. (2024); Bushuiev et al. (2023). PepUDLM demonstrates superior performance compared to autoregressive generators across multiple evaluation metrics, including lower Bits Per Dimension (BPD), reduced Negative Log-Likelihood (NLL), and significantly improved perplexity (PPL) (Figure 1D). Furthermore, PepUDLM generates peptides with substantially high Hamming distances from the test set, indicating a great degree of diversity and novelty in the generated sequences (Figure 3). Additionally, the Shannon entropy of the generated peptides closely matches that of the test set, highlighting the model's capability to produce biologically plausible peptides with diverse sequence lengths (Figure 3).

2.3 muPPIt Generates Mutant-Specific Binders

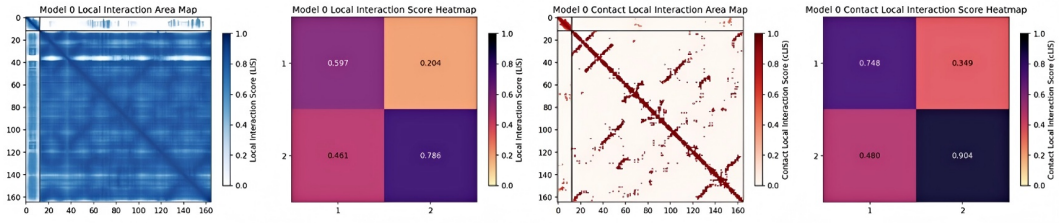
With MutBind for predicting mutant-binding probabilities and PepUDLM for peptide generation, we developed the **mu**tant-specific **PPI t**argeting algorithm (**muPPIt**) to generate mutant-specific peptide binders based solely on mutant and wildtype protein sequences. Instead of filtering random sequences through PepUDLM, we adopted a classifier-guided diffusion approach, where mutant-binding probabilities predicted by MutBind guide PepUDLM to generate binders specific to the mutant protein (Figure 1B).

muPPIt begins with a randomly initialized peptide sequence of a defined length. Applying a classifier-guided diffusion approach, it iteratively refines the sequence by sampling from a tempered distribution:

$$p^\gamma(z_s|z_t, y) \propto p_\phi(y|z_s)^\gamma p_\theta(z_s|z_t), \tag{1}$$



(a) binder – wildtype SOD1 complex: LIS = 0.092 LIA = 336



(b) binder – mutant SOD1 complex: LIS = 0.332 LIA = 1135

Figure 2: Comparison of the Local Interaction Area (LIA) and Local Interaction Score (LIS) between binder-mutant and binder-wildtype complexes of the SOD1 protein. The LIA map, LIS heatmap, contact LIA map, and contact LIS heatmap are presented for both complexes.

where $p_\theta(z_s|z_t)$ represents the pre-trained PepUDLM diffusion prior, and $p_\phi(y|z_s)$ is the pre-trained MutBind providing mutant-specific guidance which is the probability of the binder binding with the user-defined mutant protein. The parameter γ controls the strength of classifier guidance. From Eq1, the guidance is derived as:

$$\nabla_{\mathbf{z}_s} \log p^\gamma(\mathbf{z}_s \mid y, \mathbf{z}_t) = \gamma \nabla_{\mathbf{z}_s} \log p_\phi(y \mid \mathbf{z}_s) + \nabla_{\mathbf{z}_s} \log p_\theta(\mathbf{z}_s \mid \mathbf{z}_t). \tag{2}$$

This ensures that mutant-binding specificity is reinforced throughout the diffusion process. By iteratively refining sequences under this framework, PepUDLM generates peptide binders that are highly likely to interact with the mutant protein, but not with the wildtype protein.

To evaluate muPPIt in a well-controlled setting, we designed binders for eight mutant proteins randomly selected from the SKEMPI dataset, each differing by a single amino acid from their wildtype counterparts (Table 3). Using AlphaFold3, we calculated the ipTM scores which represent confidence in interface formation for the predicted peptide-protein complexes, comparing binder-wildtype and binder-mutant complexes Abramson et al. (2024). We observed that muPPIt-designed binders formed complexes with superior ipTM scores for mutant proteins than for wildtype proteins, indicating greater stability with mutant targets. We further assessed the free energy of both binder-wildtype and binder-mutant complexes using PyRosetta, which reflects the binding affinities between molecules Chaudhury et al. (2010). All muPPIt-designed binders exhibited improved free energy when interacting with mutant proteins, confirming high mutant-binding specificity. Additionally, we analyzed local interaction areas (LIA) and local interaction scores (LIS) based on the complex structures using AlphaFold-Multimer Kim et al. (2024). Most binder-wildtype complexes displayed low LIA and LIS, indicating weak interactions, while high LIA and LIS scores for binder-mutant complexes further underscored the binders' mutant-specificity.

To further assess muPPIt's performance, we designed peptide binders for four mutant proteins randomly selected from the PPIMut dataset, each with more amino acid differences compared to their wildtype counterparts (Table 4). These mutants were derived from wildtype proteins in the PPIRef dataset by replacing all binding site residues with those possessing the most distinct properties, thereby increasing the challenge of designing mutant-specific binders (Appendix A.1). Remarkably, we observed comparable improvements in AlphaFold3 ipTM scores and PyRosetta free energy when muPPIt-designed binders interacted with mutant proteins versus wild-type proteins. Moreover, binder-wildtype complexes exhibited very low local interaction areas (LIA) and local interaction scores (LIS), while binder-mutant complexes showed significantly higher LIA and LIS, further validating the mutant-specificity of muPPIt-designed binders.

Table 1: muPPIt-designed binders specifically target disease-related mutant proteins, exhibiting higher ipTM scores, LIA (Local Interaction Area), and LIS (Local Interaction Score), as well as lower free energy compared to their binding with wildtype counterparts. 'WT' denotes binding to the wildtype, 'MUT' denotes binding to the mutant, and '# muts' indicates the number of mutations in the wildtype sequence.

UniProt	Name	Type	ipTM score	free energy	LIA	LIS	# muts
P68871	HBB	WT	0.54	-446.99	1309	0.351	1
		MUT	0.63	-456.98	1631	0.4	
P00441	SOD1	WT	0.31	-476.94	336	0.092	2
		MUT	0.6	-497.52	1135	0.332	
P01112	H-Ras	WT	0.54	-519.45	2134	0.25	2
		MUT	0.63	-610.24	2536	0.308	
Q99497	PARK7	WT	0.56	-621.05	1858	0.243	2
		MUT	0.59	-635.28	2283	0.273	

The mission of muPPIt is to design highly specific therapeutics targeting disease-related mutations. To this end, we applied muPPIt to design mutant-specific binders for HBB (associated with sickle cell anemia), H-Ras (linked to various cancers), SOD1 (implicated in amyotrophic lateral sclerosis, ALS), and PARK7 (associated with Parkinson's disease) (Table 1). Specific mutations on these wildtype proteins and their corresponding muPPIt-designed binders are detailed in Table 2. All binders exhibited superior performance in ipTM scores, free energy, local interaction areas (LIA), and local interaction scores (LIS) when interacting with their mutant targets, achieving maximum improvements of 0.29 in ipTM score, over 90 in free energy, 799 in LIA, and 0.24 in LIS. These results underscore the exceptional mutant-specificity of muPPIt-designed binders.

To further illustrate this specificity, we visualized the LIA maps and LIS heatmaps for SOD1 binder-mutant and binder-wildtype complexes (Figure 2). Each map is divided into four quadrants: the upper left and lower right show intra-molecular interactions, while the upper right and lower left depict binder-target interactions. The binder-wildtype complex showed minimal contact in the inter-molecular interaction quadrants in the LIA map, whereas the binder-mutant complex exhibited extensive interactions. The LIS heatmaps further confirmed stronger interactions between the binder and mutant, with scores of 0.461 and 0.204, compared to 0.087 and 0.097 for the binder-wildtype complex. Additionally, the contact LIA map and contact LIS heatmap highlighted more residue contact points with stronger interactions in the binder-mutant complex compared to the binder-wildtype complex. These findings demonstrate muPPIt's robust capability to design mutant-specific binders, particularly for disease-related mutations.

3 Discussion

Designing highly mutant-specific peptide binders for targets driven by single-point mutations or complex mutational landscapes has long posed a significant challenge in therapeutic development. In this work, we have presented muPPIt, a purely sequence-based approach that tackles this challenge by enabling the design of mutant-specific binders independent of the mutation's complexity. Leveraging attention-based deep learning and conditional uniform discrete diffusion, muPPIt generates peptides that exhibit strong binding specificity to a broad range of mutant proteins while minimizing interactions with their wildtype counterparts.

We believe muPPIt has the potential to be effective across a broad spectrum of protein targets. To prove this, our next steps will include a comprehensive experimental validation of muPPIt, evaluating the specificity of designed binders to the mutant proteins. This will involve biochemical binding affinity assays and leveraging our chimeric peptide-E3 ubiquitin ligase ubiquibody (uAb) architecture for target degradation studies Bushuiev et al. (2023); Chen et al. (2024); Bhat et al. (2025). Importantly, muPPIt's capability to target mutants specifically could be particularly valuable in developing active and safe therapeutics by only targeting related protein variants, thus minimizing off-target effects and

maximizing treatment efficacy. Overall, these capabilities could prove invaluable for both detection and therapeutic applications. As we move forward with experimental validation, we anticipate that muPPIt will contribute significantly to advancing the field of precision biotherapeutics.

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.

Allistair A Abraham and John F Tisdale. Gene therapy for sickle cell disease: moving from the bench to the bedside. *Blood, The Journal of the American Society of Hematology*, 138(11):932–941, 2021.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Suhaas Bhat, Kalyan Palepu, Lauren Hong, Joey Mao, Tianzheng Ye, Rema Iyer, Lin Zhao, Tianlai Chen, Sophia Vincoff, Rio Watson, Tian Z. Wang, Divya Srijay, Venkata Srikar Kavirayuni, Kseniia Kholina, Shrey Goel, Pranay Vure, Aniruddha J. Deshpande, Scott H. Soderling, Matthew P. DeLisa, and Pranam Chatterjee. De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *Science Advances*, 11(4), January 2025. ISSN 2375-2548. doi: 10.1126/sciadv.adr8638. URL http://dx.doi.org/10.1126/sciadv.adr8638.

Lucie I Bruijn, Timothy M Miller, and Don W Cleveland. Unraveling the mechanisms involved in motor neuron degeneration in als. *Annu. Rev. Neurosci.*, 27(1):723–749, 2004.

Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, et al. Learning to design protein-protein interactions with enhanced generalization. *arXiv preprint arXiv:2310.18515*, 2023.

Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics*, 26(5):689–691, 2010.

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.

Tianlai Chen, Madeleine Dumas, Rio Watson, Sophia Vincoff, Christina Peng, Lin Zhao, Lauren Hong, Sarah Pertsemlidis, Mayumi Shaepers-Cheu, Tian Zi Wang, et al. Pepmlm: target sequence-conditioned generation of therapeutic peptide binders via span masked language modeling. *ArXiv*, pp. arXiv–2310, 2024.

Peng Cheng, Cong Mao, Jin Tang, Sen Yang, Yu Cheng, Wuke Wang, Qiuxi Gu, Wei Han, Hao Chen, Sihan Li, et al. Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering. *Cell Research*, 34(9):630–647, 2024.

William A Eaton and H Franklin Bunn. Treating sickle cell disease by targeting hbs polymerization. *Blood, The Journal of the American Society of Hematology*, 129(20):2719–2726, 2017.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

Sherlyn Jemimah, Masakazu Sekijima, and M Michael Gromiha. Proaffimuseq: sequence-based method to predict the binding free energy change of protein–protein complexes upon mutation using functional classification. *Bioinformatics*, 36(6):1725–1730, 2020.

Ah-Ram Kim, Yanhui Hu, Aram Comjean, Jonathan Rodiger, Stephanie E Mohr, and Norbert Perrimon. Enhanced protein-protein interaction discovery via alphafold-multimer. *bioRxiv*, pp. 2024–02, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

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.

HU Mei, Zhi H Liao, Yuan Zhou, and Shengshi Z Li. A new set of amino acid descriptors and its application in peptide qsars. *Peptide Science: Original Research on Biomolecules*, 80(6):775–786, 2005.

Linus Pauling, Harvey A Itano, Seymour J Singer, and Ibert C Wells. Sickle cell anemia, a molecular disease. *Science*, 110(2865):543–548, 1949.

Ian A Prior, Paul D Lewis, and Carla Mattos. A comprehensive survey of ras mutations in cancer. *Cancer research*, 72(10):2457–2467, 2012.

Daniel R Rosen, Teepu Siddique, David Patterson, Denise A Figlewicz, Peter Sapp, Afif Hentati, Deirdre Donaldson, Jun Goto, Jeremiah P O'Regan, Han-Xiang Deng, et al. Mutations in cu/zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, 362 (6415):59–62, 1993.

Rachele A Saccon, Rosie KA Bunton-Stasyshyn, Elizabeth MC Fisher, and Pietro Fratta. Is sod1 loss of function involved in amyotrophic lateral sclerosis? *Brain*, 136(8):2342–2358, 2013.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*, 2024.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024.

Dhirendra K Simanshu, Dwight V Nissley, and Frank McCormick. Ras proteins and their regulators in human disease. *Cell*, 170(1):17–33, 2017.

Lirong Wu, Yijun Tian, Haitao Lin, Yufei Huang, Siyuan Li, Nitesh V Chawla, and Stan Z Li. Learning to predict mutation effects of protein-protein interactions by microenvironment-aware hierarchical prompt learning. *arXiv preprint arXiv:2405.10348*, 2024.

Li C Xue, João Pglm Rodrigues, Panagiotis L Kastritis, Alexandre Mjj Bonvin, and Anna Vangone. Prodigy: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, 32(23):3676–3678, 2016.

Chengxin Zhang, Xi Zhang, Peter L Freddolino, and Yang Zhang. Biolip2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 52(D1): D404–D412, 2024.

MEANINGFULNESS STATEMENT

We present a novel method to design mutant-specific peptide binders. By combining MutBind's accurate prediction of binding preference with a diffusion-based peptide generator, muPPIt reliably designs binders with high mutant specificity. muPPIt provides a tangible framework for developing mutation-specific therapeutics with the potential to reduce off-target effects, demonstrating clear, data-driven progress in precision biotherapeutics.

ACKNOWLEDGMENTS

We thank Yinuo Zhang for advice on mutatnt binding dataset preparation.

A SUPPLEMENTARY MATERIAL

A.1 DATASET CURATION

PPIMut Curation. We randomly selected around 19000 data from the PPIRef dataset, a large protein-protein interaction dataset containing interacting sequences and binding interface positions Bushuiev et al. (2023). For each selected entry, we created two data for MutBind training by selecting binder or target and mutating all the residues on the binding sites into their least likely counterparts according to the BLOSUM62 matrix, thus creating the mutant sequence Henikoff & Henikoff (1992). All binder-wildtype and binder-mutant complex structures are predicted by ESMFold and their binding affinities are predicted by PRODIGY Lin et al. (2022); Xue et al. (2016).

SKEMPI Processing. The SKEMPI dataset was rigorously processed by removing the following data: duplicate entries, entries containing N/A values, entries where the mutant and wildtype sequences differed in length, entries with negative binding affinities for either the mutant or wildtype, and entries with sequences containing ambiguous amino acids ('X', 'Z', or 'B'). Additionally, entries with conflicting binding affinities for cases where the same binder had switched wildtype and mutant sequences were excluded. Following this comprehensive filtering process, the refined SKEMPI dataset comprises 1058 entries.

Training Set Curation. We combined the PPIMut dataset with the filtered SKEMPI dataset to construct the complete dataset for MutBind training. The merged dataset was split into training, validation, and test sets at an 80/10/10 ratio. Additionally, we recorded the indices of the SKEMPI data within the training set for use in MutBind's curriculum training.

Since both datasets provide only binding affinities for binder-mutant and binder-wildtype interactions, we converted these affinities into binding probabilities for MutBind training using the following equations:

$$P_{wt} = \frac{log(A_{wt})}{log(A_{mut}) + log(A_{wt})}, \tag{3}$$

$$P_{mut} = \frac{log(A_{mut})}{log(A_{mut}) + log(A_{wt})}, \tag{4}$$

where A_{wt} and A_{mut} represent the binding affinities of the binder to the wildtype and mutant, respectively, P_{wt} and P_{mut} denote the corresponding binding probabilities. A base-10 logarithm was applied to the binding affinities to normalize their magnitudes.

A.2 MUTBIND ARCHITECTURE

MutBind takes three sequences as input: binder, mutant, and wildtype sequences. These sequences will first be transformed into embeddings using a pre-trained ESM-2-650M model with its weights fixed. These embeddings will then be concatenated with VHSE8 embeddings. Then the binder and mutant embeddings, so as the binder and wildtype embeddings, will go through a multi-head cross attention module, where binder embeddings are used as the query and mutant/wildtype embeddings are used as key and value. The attention results, namely the joint embedding of binder-mutant and binder-wildtype will be input into a layer normalization module before computing their difference. A linear model comprises of a linear layer which maps down half of the model dimension, followed by a SiLU layer and another linear layer mapping down another half of the dimension with SiLU layer, and a linear layer that finally maps the joint embedding difference to two classes. A final softmax layer converts the predicted logits to probabilities.

A.3 MUTBIND TRAINING

Loss Function. In each training step, two forward passes are executed, with the mutant and wildtype inputs swapped in the second pass. The loss function is designed to incorporate the Kullback-Leibler (KL) divergence between the predicted and true probabilities for each pass, as well as a symmetry-enforcing term that ensures consistency when the mutant and wildtype inputs are interchanged.

Specifically, the loss function is defined as:

$$L = L_1 + L_2 + L_{diff}, \tag{5}$$

where

$$L_1 = KL([Pred^1_{wt}, Pred^1_{mut}] \,||\, [P_{wt}, P_{mut}]), \tag{6}$$

$$L_2 = KL([Pred^2_{wt}, Pred^2_{mut}] \,||\, [P_{mut}, P_{wt}]), \tag{7}$$

$$L_{diff} = |Pred^1_{wt} - Pred^2_{mut}|. \tag{8}$$

Here, $Pred^i$ represents the predicted probability of wildtype/mutant in the ith forward pass.

Curriculum Training. While PPIMut dataset comprises of mutants with multiple amino acid differences from the wildtype proteins, most mutants in the SKEMPI dataset only differ from their wildtypes with one amino acid. And due to the different size between PPIMut and SKEMPI dataset, we employed curriculum training to support gradual and effective model training. Specifically, we evenly split the SKEMPI data in the training set into 27 batches. For the first 3 training epochs, we trained muPPIt only using the PPIMut data in the training set. In the following epochs, we gradually add one batch to the training data. This curriculum training enabled MutBind gradually learn to differentiate binder-mutant and binder-wildtype joint embeddings.

Hyper-parameter configurations. MutBind was trained on one H100 NVIDIA NVL GPU system with 94 GB of VRAM for 30 epochs. The learning rate was set to 1e-3, batch size to 4, model dimension to 32, number of attention heads to 4, and gradient accumulation steps to 4. The AdamW optimizer was used with weight decay of 1e-5, beta1 of 0.9, beta2 of 0.99. A learning rate scheduler with linear warming up and cosine decay was employed to optimize training, where the minimum learning rate was set to 1e-4 and warm-up epochs was set to 3 epochs.

A.4 PepUDLM Training and Evaluation

Dynamic Batching. To enhance computational efficiency and manage variable-length token sequences, we implemented dynamic batching. Drawing inspiration from ESM-2's approach Lin et al. (2023), input peptide sequences were sorted by length to optimize GPU memory utilization, with a maximum token size of 100 per GPU.

Hyper-parameter Configurations. PepUDLM employed a DDIT backbone model with a hidden layer size of 768, 12 blocks, 12 attention heads, and a dropout rate of 0.1. Training was conducted on a 2xH100 NVIDIA NVL GPU system with 94 GB of VRAM for 100 epochs. The AdamW optimizer was employed with a learning rate of 1e-5, weight decay of 1e-4, beta1 of 0.9, beta2 of 0.999, and epsilon of 1e-8. Gradient clipping was set to 1, and a learning rate scheduler with 10 warm-up epochs and cosine decay was used, with initial and minimum learning rates of 1e-5 and 1e-6, respectively.

Evaluation Settings. To evaluate the Hamming distance and the Shannon entropy of PepUDLM's unconditionally sampled peptides to the peptides in the test set, we randomly sampled 1000 peptides from PepUDLM for each length ranging from 6 to 49. The random seed was set to 42. The Hamming distance and Shannon entropy were evaluated based on each peptide length (Figure 3).

A.5 Mutant-Specific Binder Sampling details

The pre-trained MutBind model with VHSE8 embeddings and pre-trained PepUDLM were used in muPPIt to sample peptide candidates for in-silico benchmarking. The gamma hyper-paramerter that controls the guidance strength was set to 2.0 during sampling. The total sampling steps was set to 128. Various peptide lengths and random seeds were tried to generate optimal mutant-specific binders.

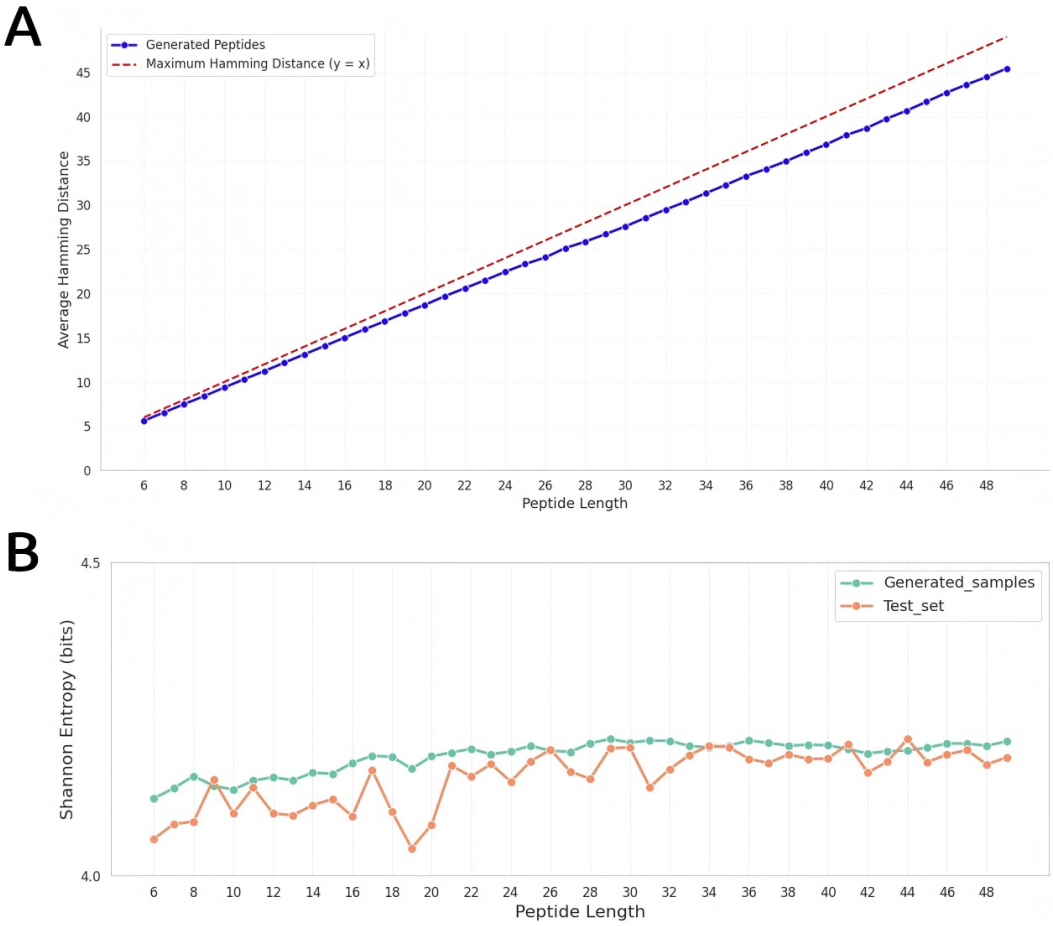


Figure 3: **(A)** The Hamming distance of sampled peptides of different lengths to the peptides of the same length in the test set. **(B)** The Shannon Entropy of sampled peptides of different lengths to the peptides of the same length in the test set.

Table 2: Specific mutations on the wildtype proteins and the muPPIt-designed binder sequences are presented for each disease-related protein.

UniProt	Name	mutations	binder
P68871	HBB	E6K	VGTVSAEKSQAQPD
P00441	SOD1	A4V, H46R	EAAADAEAMQAE
P01112	H-Ras	G12V, Q61L	RAAKKAEAQQAEYDEAQNV
P37840	PARK7	M26I, L166P	GSLEKPLTAMTLLFSISPVLLR

Table 3: muPPIt-designed binders specifically target mutants from SKEMPI, exhibiting higher ipTM scores, LIA (Local Interaction Area), and LIS (Local Interaction Score), as well as lower free energy compared to their binding with wildtype counterparts. 'WT' denotes binding to the wildtype, 'MUT' denotes binding to the mutant, and '# muts' indicates the number of mutations in the wildtype sequence.

SKEMPI ID	Type	ipTM score	free energy	LIA	LIS	# muts
1BRS_A_D	WT	0.3	-306.33	680	0.248	1
	MUT	0.53	-343.96	1243	0.333	
1CBW_FGH_I	WT	0.37	-350.19	349	0.083	1
	MUT	0.47	-353.69	822	0.221	
1A4Y_A_B	WT	0.53	-478.39	917	0.28	1
	MUT	0.68	-480.20	1171	0.41	
4UYQ_A_B	WT	0.55	-227.8	922	0.442	1
	MUT	0.67	-231.7	1085	0.538	
2B0U_AB_C	WT	0.45	-778.26	1749	0.157	1
	MUT	0.56	-796.78	2480	0.202	
3BT1_A_U	WT	0.48	-60.82	656	0.447	1
	MUT	0.53	-72.66	498	0.498	
1R0R_E_I	WT	0.37	-150.35	550	0.34	1
	MUT	0.63	-151.06	597	0.595	
1FCC_A_C	WT	0.38	-212.82	669	0.365	1
	MUT	0.55	-218.06	827	0.539	

Table 4: muPPIt-designed binders specifically target mutants from PPIMut, exhibiting higher ipTM scores, LIA (Local Interaction Area), and LIS (Local Interaction Score), as well as lower free energy compared to their binding with wildtype counterparts. 'WT' denotes binding to the wildtype, 'MUT' denotes binding to the mutant, and '# muts' indicates the number of mutations in the wildtype sequence.

PPIMut ID	Type	ipTM score	free energy	LIA	LIS	# muts
4Q2P_A_B	WT	0.54	-306.81	710	0.318	9
	MUT	0.61	-316.64	799	0.453	
7KBE_A_E	WT	0.24	-228.84	5	0.032	15
	MUT	0.41	-233.16	715	0.195	
2X83_C_D	WT	0.48	-450.33	583	0.188	17
	MUT	0.57	-470.65	825	0.319	
6U3A_A_B	WT	0.24	-430.83	37	0.052	23
	MUT	0.73	-443.76	422	0.389	