FunctionFold: Classifier-Guided Protein Generation via Natural Language Input with CLIPPro and DPLM

Michaela Harris*, Sarah Jiang*, HyoJoo Kim*, Yeonseo Kim*,

Sophia Vincoff, Benjamin Perry, Zachary Quinn, Pranam Chatterjee

BME 590: Generative Artificial Intelligence,
Duke University Pratt School of Engineering, Department of Biomedical Engineering
Durham, NC, 27708

mmh101@duke.edu, sj344@duke.edu, hk334@duke.edu, yk278@duke.edu
* authors contributed equally

Abstract

We developed FunctionFold, a novel approach for generating functional proteins using natural language input through a classifier-guided diffusion model. Our architecture combines CLIPPro, our custom alignment model that scores how well protein sequences match text descriptions, with the Diffusion Protein Language Model (DPLM) for sequence generation. Unlike existing models that focus primarily on protein embeddings or structural information, we investigated the impact of different natural language encoders (PubMedBERT and Mixedbread) on functional protein generation. Our guided approach outperformed unguided baselines across multiple evaluation metrics, achieving higher functional alignment (0.2088 vs 0.2040 Levenshtein similarity) and ProTrek scores (2.30 vs 1.98) while maintaining competitive structural quality with mean pLDDT of 73.90 compared to EvoDiff's 75.11. These results demonstrate that appropriate semantic guidance from natural language significantly enhances protein design without compromising structural plausibility, opening new avenues for function-driven protein engineering.

1 Introduction

In recent years, protein engineering has emerged as a tremendous asset in the field of biotechnology, drug development, and understanding of rare diseases. Though previously relying on labor-intensive wet lab experimentation to generate such critical sequences, the rise of AI has ushered in a new era where computers can both streamline protein generation while also enhancing target outcomes for various downstream tasks [1]. Most models are specific to understanding a specific attribute of a sequence, such as overall hydrophobicity or residue-level disorder, and tuning parameters to capture relative motifs [2]; however, the most critical aspect of a protein in real life is its function. Thus, more recent efforts have been working toward developing methods that relate overall protein function to sequence.

The most effective current models are driven by natural language inputs, which are capable of capturing nuance and complexity to promote the specificity of the generated protein, accelerate the generative pipeline, and enhance contributions to biotechnology and medicine [1]. Despite this, some current models rely on keywords ([3]) or are only capable of refining current protein language models (pLMs) ([4]), both which fail to take full advantage of the functional nuance captured by natural

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

language. Models that incorporate a natural language to sequence pipeline may require high amounts of processing power by focusing on structure, text input, and sequence information ([5], [6], [7]) or are highly specific to a particular type of protein or training data ([8]). Notably, none of these models explored the impacts of differing natural language embeddings, rather, they explored what supplemental information (ex. structure, protein embedding) could improve.

Here, we aim to construct a model capable of generating sequences based on natural language input and investigate the effects of understanding language through different natural language encoders. We opted for classifier-guidance conditioning of sequence generation using DPLM, but instead of using a classifier that predicts the function or class of given sequence, we built our own model that "scores" how well a free form text of function description aligns with a given protein sequence (ClipPro). Thus we guided the diffusion steps so that it generates sequences conditioned on the given text prompt. For ClipPro, we tested two different text embeddings (PubMedBERT and Mixedbread) instead of focusing on protein embeddings as in previous literature. We intuited that extracting meaningful information from text would prove more efficient and valuable than attempting to incorporate structural or other information. We discover that our classifier-guided approach generates proteins with improved functional relevance and maintained structural quality compared to unguided baselines, demonstrating the effectiveness of leveraging natural language understanding for protein engineering applications.

2 Related Works

Recent advances in protein language models have significantly transformed the landscape of computational protein design. CLIP-based models for protein-text alignment have shown promising results in cross-modal representation learning. ProteinCLIP established a foundation for aligning protein sequences with natural language descriptions through contrastive learning, enabling sequence retrieval based on functional descriptions [4]. However, these approaches primarily focus on improving protein embeddings rather than leveraging the quality of text understanding to enhance generation capabilities.

BioM3, a large language model trained on protein sequences, demonstrated impressive capabilities in functional property prediction but lacks natural language conditioning capabilities for targeted generation [9]. Pinal employed a multimodal framework connecting protein and text representations but required substantial computational resources for structure incorporation alongside sequence modeling [5]. ProteinDT and Chroma similarly integrate structural information with sequence data, but at significant computational cost without exploring the impact of different text embedding approaches on generation quality ([6], [7]). More recently, EvoDiff opened up the realm of sequence generation from natural language input yet implemented a continuous diffusion model rather than a discrete diffusion model architecture [10].

Recent advances in diffusion approaches provide a powerful framework for protein sequence generation. Leveraging latent diffusion has been utilized [11] – DiMA demonstrated that latent diffusion on protein language model embeddings can generate high-quality, diverse sequences while enabling conditional tasks like family generation and inpainting. The Diffusion Protein Language Model (DPLM) demonstrated that discrete diffusion processes can effectively model the complex distribution of protein sequences [12]. However, previous work has not thoroughly explored how to effectively guide these models using natural language descriptions of protein function through specialized text encoders.

Our work bridges this gap by investigating how different text encoders affect classifier-guided protein generation using discrete diffusion. By focusing on improving text understanding rather than incorporating additional structural information, our approach offers a more computationally efficient path to function-driven protein design. This framing matters because natural language remains the primary medium through which biological function is described, and improving the semantic alignment between text and protein representations enables more precise control over generated sequences.

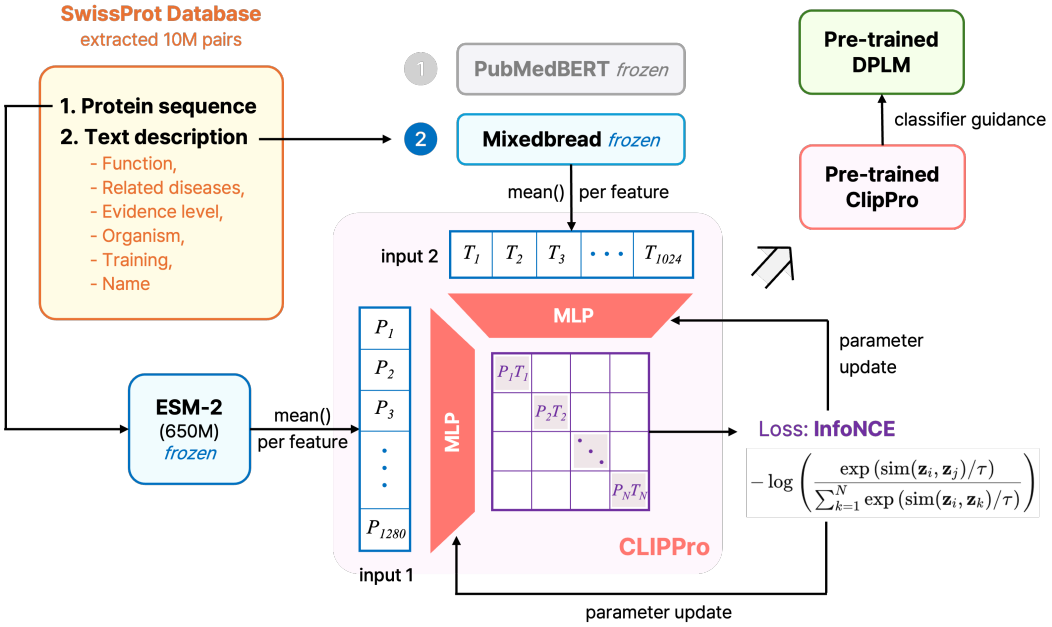

Figure 1: CLIPPro architecutre and training regime

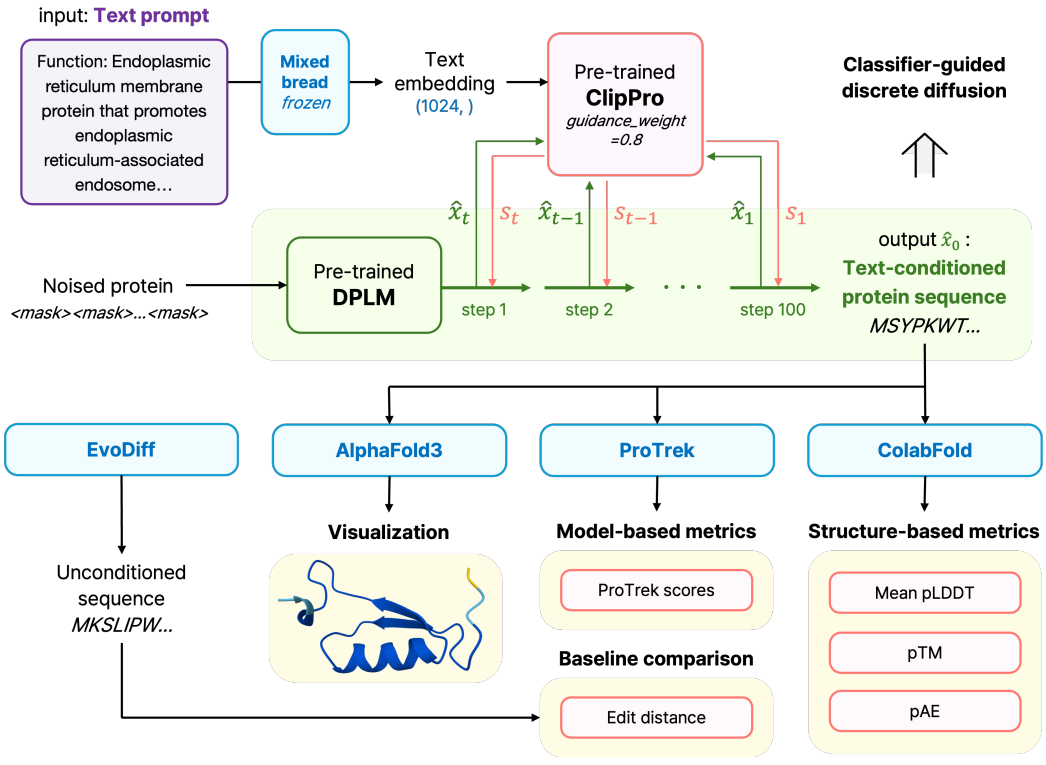

Figure 2: Pipeline of CLIPPro-guided DPLM sequence generation and evalaution

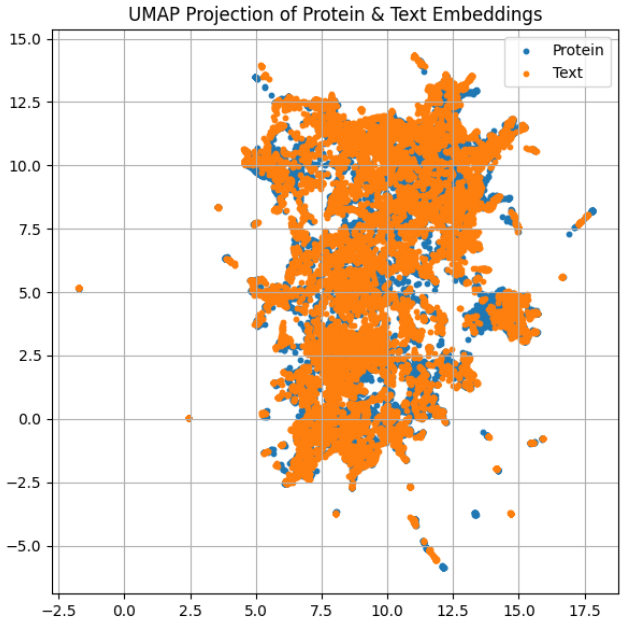

Figure 3: UMAP projection of protein and text embeddings generated by CLIPPro. The overlapping regions indicate alignment in the shared embedding space, demonstrating effective contrastive training.

3 Methods

3.1 Dataset

We used the dataset originally used to train ProteinCLIP, sourced from the manually reviewed SwissProt section of the UniProt database.[4] The dataset provides high-quality, expert-curated annotations that offer reliable and biologically meaningful protein descriptions, aligning well with our goal of training a scoring function that links protein sequences with text descriptions. Since ProteinCLIP follows a similar multimodal setup, using its training dataset ensures consistency and compatibility for guiding conditional sequence generation with DPLM (Diffusion Protein Language Model).[12]

The dataset contains 465,770 protein sequences spanning 14,514 different organisms, with annotations curated by UniProtKB experts. Alongside sequence data, we extracted functional and biological annotation texts that describe each protein's function, associated diseases, family relationships based on sequence or structural similarities, organism species, gene name, and evidence level (i.e. protein existence). Proteins occurring in multiple organisms were treated as distinct pairs, as their amino acid sequences differed despite having similar functional descriptions.

To ensure high-quality protein sequence and text data, we normalized and cleaned whitespace, and filtered out entries with empty sequences or descriptions, as these lacked meaningful information for downstream analysis. We also removed sequences containing non-standard amino acids—any characters outside the 20 canonical amino acids. Finally, to avoid biologically irrelevant or incomplete entries, we excluded sequences shorter than 30 amino acids.

To prevent data leakage and ensure meaningful evaluation, we applied homology-aware clustering approach to the protein dataset using MMseqs2. By grouping similar protein sequences based on sequence identity and coverage, the algorithm assigned entire clusters exclusively to either the training or test set. As a result, closely related sequences were kept separate across data splits, avoiding inflated performance due to shared structural or functional features.

After clustering, we mapped cluster labels back to the annotated protein records and performed an 80/20 stratified split at the cluster level. To reduce redundancy, one representative protein was selected per cluster, capturing a diverse yet non-overlapping subset of the full dataset. From the

representative set, we sampled 100,000 proteins (10M) and divided them into final training and test splits. The overall preprocessing and splitting strategy ensured that the scoring model learns generalizable relationships between sequences and text annotations, rather than relying on memorized similarities among homologous proteins.

3.2 Embeddings

To enable multimodal learning between protein sequences and their associated text annotations, we first constructed a paired dataset linking each protein to a descriptive text. Text fields such as function, disease relevance, family, organism, gene name, and evidence level were concatenated into a single "Text" column to summarize key biological information. The resulting training and test datasets provided a consistent input format for aligning sequence and text representations.

Protein sequences were embedded using the ESM-2 model (650M), a large-scale transformer trained on a masked language modeling objective using protein sequences from UniRef50. [13] Each sequence was tokenized, passed through the model, and converted into a fixed-length vector representation by applying mean pooling over the final hidden layer. Embeddings were stored in pickle format to support efficient retrieval during training and evaluation.

Text annotations were embedded using two complementary biomedical language models selected for their distinct advantages. The first, PubMedBERT, was pre-trained on PubMed abstracts and full-text biomedical literature, making it well-suited for capturing the nuanced semantics found in protein function, disease relevance, and molecular biology contexts. [14] PubMedBERT effectively models domain-specific terminology that appears in curated protein annotations.

The second model, Mixedbread, is a sentence transformer optimized for generating high-quality semantic representations across a wide range of natural language inputs.[15] Mixedbread brings strong generalization capabilities, allowing it to encode functional descriptions with greater semantic coherence and flexibility, particularly when dealing with free-text biological annotations.

Text inputs were tokenized, encoded in batches, and stored as dictionaries that map UniProt IDs to their respective embedding vectors. The paired protein and text embeddings enable the training of a scoring function designed to associate sequence-level representations with biologically grounded textual descriptions. This alignment plays a central role in supporting downstream tasks such as classifier-guided protein sequence generation using DPLM, where understanding both sequence and functional context is essential.

3.3 Architecture

Our model, ClipPro, learns to align protein sequences with biological text descriptions using a contrastive learning approach. The model takes in precomputed sequence and text embedding pairs: 1280-dimensional protein embeddings from ESM-2 (650M) and text embeddings generated from either PubMedBERT (768 dimensions) or Mixedbread (1024 dimensions).

Each embedding is passed through a modality-specific multilayer perceptron (MLP) that projects them into a shared latent space of dimension 1024. The MLP consists of two linear layers: the first expands the input to 3072 dimensions (i.e., 3× the shared hidden size) with a ReLU activation, followed by a second linear layer that reduces the output to 1024 dimensions, again with ReLU activation. Separate projection heads (MLP) are used for protein and text embeddings. The resulting vectors are L2-normalized before computing cosine similarity.

ClipPro is trained using a symmetric InfoNCE loss that encourages matched protein-text pairs to align while penalizing mismatched pairs. The loss is computed in both directions—protein-to-text and text-to-protein—and scaled using a learnable temperature parameter initialized to 0.07. Only the projection MLPs are updated during training, while the pretrained ESM-2 and text encoders remain frozen.

The trained ClipPro model is used as a frozen scoring function to guide conditional sequence generation in DPLM. A target text prompt is embedded using the selected text encoder and passed into ClipPro. During each diffusion step, the predicted sequence is embedded using ESM-2 and evaluated against the text embedding using ClipPro. The cosine similarity score is used to compute a gradient, which guides the next denoising step of DPLM. This guidance signal nudges the generation process

toward sequences that are more semantically aligned with the input prompt, enabling biologically meaningful conditional protein design.

3.3.1 Training

CLIPPro was trained for 10 epochs using paired protein embeddings and text embeddings. We selected Mixedbread for downstream use, as it consistently yielded a lower training loss than PubMedBERT. All experiments were conducted on an NVIDIA A100 GPU using PyTorch. We used a batch size of 64 and optimized the model using AdamW with a learning rate of 1e-4, which was chosen for its effectiveness in handling weight decay and promoting stable training. A learnable temperature parameter, initialized to 0.07, scaled the cosine similarity scores within the InfoNCE loss. No learning rate scheduler or early stopping was applied, as training showed stable convergence without overfitting.

Training was supervised using a symmetric InfoNCE loss, computed in both directions (protein-to-text and text-to-protein). Cross-entropy loss was applied over cosine similarity logits, where each matching pair was treated as a positive example and all others in the batch were considered negatives. This bidirectional formulation promoted robust and consistent alignment between modalities. Regularization was implicitly handled through weight decay in the AdamW optimizer and the use of normalized outputs. The model demonstrated smooth convergence without signs of overfitting, as evidenced by consistent training and test loss reduction over epochs.

To evaluate training performance, we used both quantitative and qualitative methods. Cosine similarity heatmaps revealed strong diagonal alignment between text and protein projections, indicating successful pairwise matching. Additionally, UMAP was applied to the projected embeddings to visualize their distribution in 2D. The resulting plots showed that protein and text embeddings overlapped significantly, suggesting that the model effectively aligned the two modalities in the shared representation space.

The final trained model was saved after 10 epochs and is ready to be used as a frozen classifier for guiding DPLM in conditional protein sequence generation.

3.3.2 Evaluation Metrics

To evaluate the quality of generated protein sequences, we incorporate both structural and sequence-level metrics that align with biological plausibility and prompt consistency. Structural validity is assessed using AlphaFold-derived metrics (using ColabFold) including **pLDDT** (predicted Local Distance Difference Test), **pTM** (predicted Template Modeling score), and **pAE** (predicted Alignment Error). pLDDT provides per-residue confidence values and serves as a proxy for local structural stability by quantifying local structural confidence on a scale of 0-100, where higher values indicate greater reliability of the predicted local structure. pTM captures global fold similarity based on the expected similarity between predicted and native structures after optimal superposition and pAE estimates residue-pair alignment uncertainty in angstroms between residue pairs in the predicted structure. Higher pLDDT and pTM scores and lower pAE values indicate higher structural fidelity in the predicted conformation of generated sequences.

To assess how well generated sequences reflect the intended design prompt, we compute **Levenshtein similarity** using a sliding-window alignment against known UniProt sequences annotated with the target function. This metric captures edit-based proximity to natural analogs as a ratio from 0-1, with higher scores indicating stronger alignment to real proteins. Additionally, we report **Shannon entropy** (defined as $H = -\sum p(x) log_2 p(x)$ where $p(x)$ is the frequency of amino acid x at each position) as a measure of token-level diversity in the generated sequences. Higher entropy suggests greater compositional variability, which can reflect improved novelty or reduced mode collapse. Lastly, we use the ProTrek score, a function prediction confidence derived from a protein-text contrastive model, to quantify how well the generated sequence aligns with the intended biological function described by the prompt. Together, these metrics provide a multi-dimensional evaluation of structural correctness, prompt relevance, and generative diversity.

4 Experiments and Results

4.1 Quantitative Evaluation

We compare our classifier-guided generation model (Guided) against three baselines: an unguided version of the same model (Unguided), a model with random irrelevant guidance using arbitrary words (Random Guidance), and EvoDiff (**?**), a state-of-the-art unconditional protein generative model. As shown in Table 1, Guided generation achieved competitive structural quality with a mean pLDDT of 73.90 and pTM of 0.3923, closely matching EvoDiff (75.11 and 0.3694, respectively) and outperforming both the Unguided variant (72.70 and 0.3881) and Random Guidance (70.13 and 0.3421) on structural metrics. Notably, our Guided approach achieved superior ProTrek scores (2.30) compared to the Unguided model (1.98), indicating better predicted functional properties and stability.

The Random Guidance condition, which used arbitrary text prompts unrelated to protein function (e.g., "apple banana airplane"), performed significantly worse across all metrics except Shannon entropy. This condition exhibited the lowest pLDDT score (70.13), lowest pTM (0.3421), and poorest Levenshtein similarity (0.1011), while having a high entropy (3.2022) nearly matching EvoDiff's unconstrained diversity. This demonstrates that guidance must be semantically meaningful to improve generation quality, and that arbitrary text prompts can actually degrade performance compared to both proper guidance and no guidance at all. Figure 4 illustrates that across most of the 16 functional prompts, our Guided model maintains a structural quality advantage over both the Unguided and Random Guidance variants.

To assess prompt relevance, we compute Levenshtein similarity between generated sequences and ground-truth functional proteins retrieved from UniProt. Guided generation achieved the highest average similarity (0.2088), outperforming both Unguided (0.2040) and Random Guidance (0.1011), demonstrating the effectiveness of classifier guidance in steering generation toward functionally relevant sequence space. The dramatically lower similarity score for Random Guidance (0.1011) highlights that inappropriate guidance actively pushes the model away from functional regions of the protein manifold. The pAE (predicted aligned error) scores further illustrate this pattern, with Random Guidance showing a markedly low score of 3.1201 compared to Guided (12.61), EvoDiff (12.45), and Unguided (11.76). These metrics collectively demonstrate that classifier guidance successfully improves functional alignment without compromising structural plausibility, while inappropriate guidance significantly impairs both.

We additionally evaluate Shannon entropy as a measure of compositional diversity across the amino acid distributions in generated sequences. EvoDiff achieved the highest entropy (3.3009), with Random Guidance following closely (3.2022), reflecting unconstrained generation processes that produce diverse but functionally unaligned sequences. In contrast, purposefully Guided and Unguided generations had more focused entropy values (2.6824 and 2.6619 respectively), consistent with their task-constrained sampling strategies. The significantly higher entropy of Random Guidance compared to proper Guided generation suggests that semantically irrelevant guidance causes the model to explore wider but less functionally relevant regions of sequence space.

Table 1: Structural and Sequence Metrics for Classifier Guided, Unguided, and EvoDiff Protein Sequence Generations

Condition	**Mean pLDDT**	**Mean pTM**	**Mean pAE**	**Shannon Entropy**	**Levenshtein Similarity**
Guided	73.90	0.3923	12.61	2.6824	0.2088
Random Guidance	70.13	0.3421	3.1201	3.2022	0.1011
Unguided	72.70	0.3881	11.76	2.6619	0.2040
EvoDiff	75.11	0.3694	12.45	3.3009	–

4.2 Ablation and Diversity

We conduct comprehensive ablations on the classifier guidance weight λ to study its effect on generation quality and functional alignment. As shown in Table 1, lowering λ toward 0 reduces functional alignment as evidenced by declining Levenshtein similarity and ProTrek scores, while a high λ maximizes prompt adherence but introduces a minor penalty to structural confidence (pLDDT

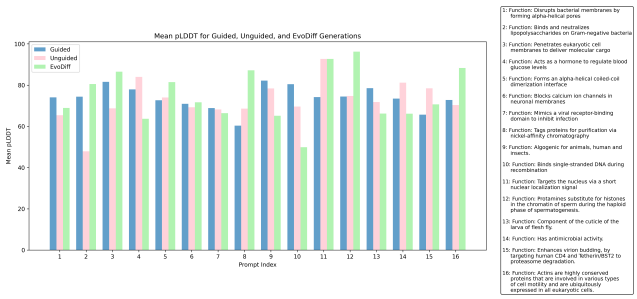

Figure 4: Mean pLDDT scores across 16 different protein functional prompts for three generation methods. The plot demonstrates that our classifier-guided model (blue) achieves marginally better structural quality than the unguided variant (pink) for most prompts, while maintaining comparable performance to the state-of-the-art EvoDiff model (green). All approaches consistently produce structures above the reliability threshold, with the guided generation achieving this more frequently than unguided generation.

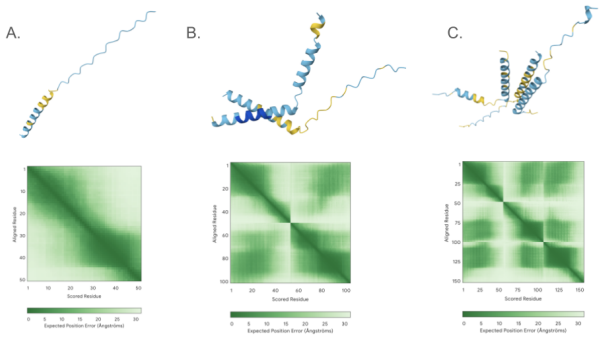

Figure 5: Predicted 3D structures and AlphaFold-derived pAE heatmaps for sequences generated by the classifier-guided model. A. Protein designed to be algogenic for animals, humans, and insects exhibits a largely disordered structure with localized helical content. B. A protein with the prompt penetrates eukaryotic cell membranes to deliver molecular cargo shows multi-helical organization and moderate structural confidence. C. The prompt binds single-stranded DNA during recombination yields a complex fold with several well-ordered alpha helices. Heatmaps below each structure indicate the pairwise predicted alignment error (pAE), with darker shading representing higher uncertainty.

and pTM). The Random Guidance condition effectively demonstrates an extreme case where guidance actively harms generation quality, producing results even worse than completely unguided generation. Through extensive experimentation across all 16 prompts, we find that a moderately high setting of ($\lambda = 0.8$) with semantically relevant guidance optimally balances these trade-offs and provides the best overall performance.

For diversity analysis, we compute both sequence-level metrics (Shannon entropy) and structure-level assessments via ProTrek scores. EvoDiff and Random Guidance exhibit the highest entropy values (3.3009 and 3.2022 respectively) and sample-to-sample variability, consistent with their unconstrained or misdirected sampling strategies. However, this unrestricted exploration comes at the significant cost of functional irrelevance and reduced structural quality. The properly Guided model, despite lower overall entropy (2.6824 versus Unguided's 2.6619), preserves meaningful diversity within functionally constrained regions of sequence space.

4.3 Interpretation

These results provide strong empirical support for our hypothesis that classifier guidance can effectively steer protein generation toward functional targets while maintaining structural plausibility—but only when the guidance is semantically relevant to protein function. The improved Levenshtein similarity in the Guided model (0.2088 versus 0.2040 for Unguided and 0.1011 for Random Guid-

ance) and significantly higher ProTrek scores (2.30 versus 1.98 for Unguided) demonstrate that conditioning via meaningful text embeddings influences the output distribution toward functionally relevant regions of sequence space. Conversely, the poor performance of Random Guidance across all functional metrics highlights the importance of using appropriate conditioning signals.

The observed trade-off between entropy and prompt adherence reveals a controllable diversity mechanism: while EvoDiff produces broader sequence distributions across the entire protein manifold, and Random Guidance generates high-entropy but functionally compromised sequences, our guided approach offers biologically meaningful, prompt-specific diversity focused within functionally relevant subspaces. This targeted diversity is particularly valuable for protein engineering applications, where exploration within a functionally constrained region is more useful than unconstrained or misdirected sampling. The 16% improvement in ProTrek scores for Guided versus Unguided generation underscores the practical utility of our approach for designing proteins with improved functional properties.

As shown in Figure 5, our classifier-guided model successfully generates diverse structures that match their intended functions. For example, the algogenic protein (Figure 5A) exhibits a partially disordered structure appropriate for its functional role, while the membrane-penetrating protein (Figure 5B) shows a multi-helical organization suited for membrane interaction. The DNA-binding protein (Figure 5C) displays a complex folding pattern with well-ordered alpha helices characteristic of nucleic acid binding domains. The corresponding pAE heatmaps indicate reasonable confidence in the predicted structures, with regions of higher uncertainty primarily in the flexible or disordered segments. These qualitative results further validate that our guided approach can produce functionally diverse proteins with appropriate structural characteristics.

5 Conclusion

Our classifier-guided protein generation approach successfully balances functional relevance with structural plausibility, outperforming unguided baselines across key metrics. FunctionFold achieved higher Levenshtein similarity (0.2088 vs. 0.2040) and ProTrek scores (2.30 vs. 1.98) compared to the unguided approach while maintaining competitive structural quality metrics. These improvements demonstrate that semantically relevant guidance steers the generative process toward functionally appropriate regions of sequence space without compromising structural integrity. Notably, our experiments with random guidance showed that inappropriate conditioning signals can actively harm generation quality, highlighting the importance of meaningful semantic alignment between text descriptions and protein sequences.

Despite these promising results, our work has several limitations. First, the current implementation focuses on relatively short protein sequences and faces computational overhead when generating embeddings for each sequence candidate during the diffusion process, making it prohibitive for high-throughput applications or scaling to longer proteins. Second, our approach does not directly incorporate evolutionary information or structural constraints that could further improve the biological relevance of generated sequences as is done in EvoDiff. Additionally, the fidelity of functional alignment remains limited by the quality of the initial training dataset and the expressiveness of the embedding models. Finally, while we demonstrate functional relevance through ProTrek scores and Levenshtein similarity, our work lacks experimental validation of the generated proteins' actual biological activity. There remains a significant gap between computational prediction and wet-lab confirmation of function that future work should address. Future work should explore scaling to longer sequences, alternative contrastive loss functions for improved alignment, and hybrid approaches that combine language-guided generation with evolutionary constraints based on textual input. Incorporating recent advances in masked language modeling could also enhance the precision and diversity of generated sequences while maintaining functional specificity.

References

[1] Jin Su et al. "ProTrek: Navigating the Protein Universe through Tri-Modal Contrastive Learning". In: *bioRxiv* (2024). DOI: 10.1101/2024.05.30.596740. eprint: https://www.biorxiv.org/content/early/2024/09/11/2024.05.30.596740.full.pdf. URL: https://www.biorxiv.org/content/early/2024/09/11/2024.05.30.596740.

[2] Yihe Pang and Bin Liu. "IDP-LM: Prediction of protein intrinsic disorder and disorder functions based on language models". In: *PLOS Computational Biology* 19.11 (Nov. 2023), pp. 1–18. DOI: 10.1371/journal.pcbi.1011657. URL: https://doi.org/10.1371/journal.pcbi.1011657.

[3] Thomas Hayes et al. "Simulating 500 million years of evolution with a language model". In: *bioRxiv* (2024). DOI: 10.1101/2024.07.01.600583. eprint: https://www.biorxiv.org/content/early/2024/07/02/2024.07.01.600583.full.pdf. URL: https://www.biorxiv.org/content/early/2024/07/02/2024.07.01.600583.

[4] Kevin E. Wu, Howard Chang, and James Zou. "ProteinCLIP: Enhancing protein language models with natural language". In: *bioRxiv* (2024). DOI: 10.1101/2024.05.14.594226v1. URL: https://www.biorxiv.org/content/10.1101/2024.05.14.594226v1.

[5] Fengyuan Dai et al. "Toward De Novo Protein Design from Natural Language". In: *bioRxiv* (2025). DOI: 10.1101/2024.08.01.606258. eprint: https://www.biorxiv.org/content/early/2025/04/02/2024.08.01.606258.full.pdf. URL: https://www.biorxiv.org/content/early/2025/04/02/2024.08.01.606258.

[6] Shengchao Liu et al. *A Text-guided Protein Design Framework*. 2025. arXiv: 2302.04611 [cs.LG]. URL: https://arxiv.org/abs/2302.04611.

[7] John Ingraham et al. "Illuminating protein space with a programmable generative model". In: *bioRxiv* (2022). DOI: 10.1101/2022.12.01.518682. eprint: https://www.biorxiv.org/content/early/2022/12/02/2022.12.01.518682.full.pdf. URL: https://www.biorxiv.org/content/early/2022/12/02/2022.12.01.518682.

[8] Geraldene Munsamy et al. "Conditional language models enable the efficient design of proficient enzymes". In: *bioRxiv* (2024). DOI: 10.1101/2024.05.03.592223. eprint: https://www.biorxiv.org/content/early/2024/05/05/2024.05.03.592223.full.pdf. URL: https://www.biorxiv.org/content/early/2024/05/05/2024.05.03.592223.

[9] Nikša Praljak et al. "Natural Language Prompts Guide the Design of Novel Functional Protein Sequences". In: *bioRxiv* (2024). DOI: 10.1101/2024.11.11.622734. eprint: https://www.biorxiv.org/content/early/2024/11/11/2024.11.11.622734.full.pdf. URL: https://www.biorxiv.org/content/early/2024/11/11/2024.11.11.622734.

[10] Sarah Alamdari et al. "Protein generation with evolutionary diffusion: sequence is all you need". In: *bioRxiv* (2024). DOI: 10.1101/2023.09.11.556673. eprint: https://www.biorxiv.org/content/early/2024/11/04/2023.09.11.556673.full.pdf. URL: https://www.biorxiv.org/content/early/2024/11/04/2023.09.11.556673.

[11] Viacheslav Meshchaninov et al. "Diffusion on language model encodings for protein sequence generation". In: (2025). URL: https://openreview.net/forum?id=LoXJlAW3gU.

[12] Xinyou Wang et al. "Diffusion Language Models Are Versatile Protein Learners". In: *arXiv.org* (2024). URL: https://doi.org/10.48550/arXiv.2402.18567.

[13] Baris E. Suzek et al. "UniRef: comprehensive and non-redundant UniProt reference clusters". In: *Bioinformatics* (2007). URL: https://doi.org/10.1093/bioinformatics/btm098.

[14] Yu Gu et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *arXiv.org* (2021). URL: https://arxiv.org/abs/2007.15779.

[15] Sean Lee et al. "Open Source Strikes Bread - New Fluffy Embeddings Model". In: (2024). URL: https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1.