**EmbedDiff: Structure-Free De Novo Protein Design via Latent Diffusion in ESM2 Embedding Space**

**Abstract**

The ability to design novel protein sequences that are structurally viable and biologically plausible is a central goal in synthetic biology and protein engineering. Traditional approaches often rely on structure-based modeling or autoregressive sequence generation, which are constrained by prior templates, alignment requirements, or limited context modeling. Here, we introduce EmbedDiff, a modular generative framework that enables de novo protein design by operating directly in the latent space of pretrained protein language models. EmbedDiff uses ESM2 embeddings of natural proteins to train a denoising diffusion model that learns the manifold of biologically meaningful sequences without any structural supervision. Synthetic embeddings are sampled by reversing a noise corruption process and decoded into amino acid sequences using a Transformer-based decoder trained to reconstruct sequences from latent representations. The decoder supports reference-guided decoding, enabling controllable generation anchored to specific classes or motifs.

We apply EmbedDiff to a diverse set of thioredoxin reductases across bacteria, archaea, and fungi, and evaluate generated sequences using a multi-tiered validation pipeline, including t-SNE visualization, cosine similarity metrics, and local BLAST alignment. Generated sequences are evolutionarily coherent, diverse, and distinct from the training data—occupying realistic regions of embedding space and exhibiting moderate-to-high identity to known proteins. These results demonstrate that diffusion modeling in embedding space offers a scalable, structure-free approach to protein generation, laying the groundwork for new pipelines in template-free design and protein function exploration.

Introduction

The rational design of de novo proteins that are both functional and structurally viable remains one of the grand unsolved problems in computational biology. Proteins are the fundamental machines of life, catalyzing reactions, transmitting signals, maintaining structure, and enabling complex cellular behavior. The ability to engineer novel proteins—beyond those sampled by natural evolution—has the potential to revolutionize biotechnology, from sustainable manufacturing to personalized medicine. However, the challenge lies in the astronomical size and sparsity of protein sequence space: even small proteins of 100 residues can have more than possible sequences, but only a vanishingly small fraction of these are expected to fold correctly or perform any biological function.

Historically, protein design has relied on a mix of structure-based modeling, residue-level heuristics, and evolutionary conservation. Techniques like Rosetta and directed evolution have made substantial progress in creating or optimizing proteins for specific tasks. Yet these methods often require significant prior knowledge, structural templates, or experimental screening, making them time-consuming and difficult to scale. Moreover, most structure-based pipelines generate sequences by solving inverse folding problems—designing a sequence to fit a predefined 3D structure—rather than learning to model the broader statistical and functional landscape of natural protein families.

In parallel, the rise of deep learning has led to a new generation of protein models trained on large, unlabeled sequence datasets. Among these, protein language models (PLMs) such as ESM, ProtTrans, and TAPE have shown remarkable capacity to encode information about structure, function, and phylogeny in dense, learned representations. These models, trained on millions of sequences without alignment or labels, produce high-dimensional embeddings that capture biologically meaningful signals. Importantly, these embeddings are derived in a structure-agnostic fashion, offering an alternative view of protein space that is implicitly shaped by evolution.

Despite this progress, the application of such models to sequence generation remains limited. Most existing generative protein frameworks operate in token space, using autoregressive transformers or masked language modeling to predict sequences residue-by-residue. While powerful, these approaches are constrained by local context windows, generation order, and limited global coherence. They also lack control mechanisms for guiding generation toward specific biological classes or functions. Furthermore, because they operate in raw sequence space, they struggle to directly leverage the rich latent structures captured by embedding models like ESM2.

To overcome these limitations, we introduce EmbedDiff, a generative framework that performs protein design in the embedding space of natural sequences using latent denoising diffusion modeling. EmbedDiff takes advantage of the ESM2 model to map amino acid sequences into high-dimensional latent vectors that reflect structural, evolutionary, and functional constraints. A denoising diffusion process is then trained to model the distribution of these natural embeddings by learning to reverse a progressive noise corruption process. Unlike autoregressive decoders, this allows EmbedDiff to sample entire latent protein representations holistically and generate embeddings that lie within the manifold of plausible protein states.

Once new embeddings are sampled, a Transformer-based decoder is used to translate them into full-length amino acid sequences. This decoder is trained to reconstruct the original sequence from its ESM2 embedding, optionally using reference-guided decoding where a portion of the input is clamped to a known sequence. This feature allows users to steer generation toward a specific family, motif, or organism class—striking a flexible balance between novelty and biological plausibility.

To evaluate the effectiveness of EmbedDiff, we constructed a benchmark dataset of thioredoxin reductase homologs spanning bacteria, archaea, and fungi. We performed extensive validation of the generated sequences using a multi-stage evaluation pipeline, which includes (i) t-SNE visualization of the latent space to examine clustering and manifold interpolation, (ii) cosine similarity analysis to quantify diversity and proximity to natural sequences, and (iii) local BLAST alignment to assess evolutionary plausibility via sequence identity and alignment significance.

Our results demonstrate that EmbedDiff is capable of generating diverse, non-trivial protein sequences that are novel yet biologically grounded. Generated sequences occupy realistic regions of latent space, exhibit meaningful similarity to known proteins. The model is modular, lightweight, structure-free, and adaptable to various protein classes—offering a scalable and interpretable platform for protein generation. We believe EmbedDiff provides a compelling new direction for embedding-based generative design and opens the door to future applications in therapeutic discovery, enzyme engineering, and de novo protein function prediction.

Overview of the EmbedDiff Pipeline

To enable the de novo generation of evolutionarily plausible proteins, we designed EmbedDiff as a modular framework that decouples the generative process from raw sequence tokens and instead operates in the biologically informed latent space of pretrained language models. Figure 1a illustrates the complete generative pipeline. Natural protein sequences are first embedded using ESM2, producing high-dimensional representations that encode both structural and evolutionary constraints. These embeddings serve as input to a latent denoising diffusion model, which learns to reverse a corruption process that adds Gaussian noise over a fixed number of timesteps. The trained model thus models the true distribution of protein embeddings and enables the generation of new synthetic embeddings from pure noise.

To translate sampled embeddings back into discrete sequences, we employ a Transformer-based decoder trained to map latent vectors to amino acid sequences. Importantly, the decoder supports reference-guided decoding, allowing conditioning on organism class or anchor motifs. This provides flexible control over novelty while preserving class-specific priors, enabling both stochastic exploration and family-targeted generation.

Figure 1b outlines the evaluation framework applied to the generated sequences. This pipeline includes:  
(i) embedding-level metrics, such as cosine similarity and t-SNE clustering, to assess plausibility and diversity,  
(ii) local sequence alignment via BLAST, to quantify evolutionary proximity, and  
(iii) density analyses, such as pairwise identity distributions and entropy, to evaluate generative diversity.

Together, these steps offer a multi-layered interrogation of generated sequences, ensuring that they are not only diverse, but also biologically realistic and evolutionarily interpretable.

Latent Space Structure and Model Training

To determine whether the embedding space effectively captures biologically meaningful structure, we projected ESM2 representations of natural thioredoxin reductase sequences into two dimensions using t-SNE. As shown in Figure 2a, the sequences form distinct clusters by organism domain (bacteria, archaea, and fungi). This confirms that ESM2 embeddings preserve high-level phylogenetic and functional relationships without explicit supervision—providing a rich manifold for generative sampling.

Next, we evaluated the learning dynamics of the latent diffusion model. Figure 2b displays the training and validation mean squared error (MSE) loss curves over 300 epochs. Both raw and smoothed curves demonstrate consistent convergence, with the validation loss plateauing after ~250 epochs. This indicates that the model successfully learned to denoise corrupted embeddings back toward the manifold of natural proteins. Importantly, no overfitting was observed, suggesting that the learned trajectory generalizes well to unseen latent points during sampling.

Embedding Similarity and Latent Diversity

To assess the fidelity and diversity of EmbedDiff-generated embeddings, we computed pairwise cosine similarity matrices across three conditions: natural–natural, generated–generated, and natural–generated.

* In Figure 3a, natural–natural comparisons exhibit strong block structure, reflecting subfamily conservation and shared motifs.
* In Figure 3b, generated–generated comparisons show a broader and more uniform distribution, indicating that EmbedDiff does not simply replicate training embeddings but produces diverse outputs spanning the learned latent manifold.
* Figure 3c (natural–generated) reveals moderate but significant similarity, supporting the idea that sampled embeddings reside within plausible regions of the natural protein space—near, but not identical to known proteins.

These trends are quantified in Figure 3d, where cosine similarity densities are plotted by pair type. Natural–natural pairs peak near 0.97, generated–generated around 0.91, and natural–generated around 0.82. This stratification provides strong evidence that EmbedDiff generates embeddings that are distinct from the training set, yet constrained within the bounds of biological plausibility.

Evolutionary Validation via Sequence Alignment

We next sought to evaluate the biological realism of generated sequences in token space. Using local BLAST alignment against the original training set, we computed both percent identity and E-value metrics for each generated sequence.

Figure 4a shows the distribution of percent identities between generated sequences and their top BLAST hits. The distribution centers near 59%, with most sequences falling between 50–65%. This indicates that generated sequences are not trivial reconstructions, but maintain moderate homology to real proteins. Very few sequences exceed 80% identity (indicative of redundancy), and only a small minority fall below 35% (suggesting functional divergence).

In Figure 4b, we visualize how generated sequences relate to natural ones in embedding space. t-SNE projection shows that synthetic sequences (black) form distinct but adjacent clusters, often interpolating between known domains. This suggests that EmbedDiff does not merely memorize training data but explores novel, evolutionarily accessible regions.

Finally, Figure 4c plots top BLAST hit identity versus −log₁₀(E-value). A clear linear trend is observed, with higher identity sequences yielding lower E-values—further confirming that EmbedDiff-generated sequences are biologically meaningful, not random artifacts. The spread in identity-E-value space also highlights the model’s capacity to balance novelty with realism.

Summary

Together, these results demonstrate that EmbedDiff can generate protein sequences that are novel, diverse, and biologically coherent. The model faithfully reconstructs the manifold of natural protein embeddings, samples effectively from this space, and decodes to sequences that are statistically and evolutionarily grounded. Importantly, all of this is achieved without using structural supervision—offering a scalable, interpretable, and structure-free approach to de novo protein design. EmbedDiff thus lays the groundwork for future protein engineering pipelines that are biologically informed yet unconstrained by existing templates or alignments.

Methods

Dataset Curation

We constructed a benchmark dataset of thioredoxin reductase sequences spanning three phylogenetic domains: bacteria, archaea, and fungi. Sequences were sourced from UniProt and curated to remove duplicates, partial entries, and low-complexity regions. Only full-length sequences with known domain annotations were retained. Sequences were clustered using CD-HIT at a 90% identity threshold to reduce redundancy, and a final set of 540 sequences was selected to ensure broad taxonomic and functional diversity across classes.

Protein Embedding with ESM2

All sequences were embedded using the ESM2 protein language model (650M parameters), pretrained on the UniRef50 database. The embedding of each sequence was computed by averaging the final-layer hidden states across all amino acid residues, yielding a fixed-length 1280-dimensional vector. These embeddings served as the input for both training the diffusion model and as targets during decoding. Embeddings were standardized (z-scored) across the dataset prior to training.

Latent Diffusion Model Training

To learn the distribution of natural protein embeddings, we trained a denoising latent diffusion model. The model follows the standard DDPM (Denoising Diffusion Probabilistic Model) framework, wherein a Gaussian noise process corrupts embeddings over a sequence of T timesteps. A multilayer perceptron (MLP) with LayerNorm and Dropout was trained to predict the noise added at each timestep, conditioned on the timestep index and optional class label (e.g., organism type).

* Loss Function: Mean squared error (MSE) between predicted and true noise.
* Optimizer: AdamW with a learning rate of 1e-4.
* Epochs: 300
* Batch size: 32
* Input dimension: 1280
* Noise schedule: Linear β schedule over 1000 steps.

To ensure stability, the model was trained using both raw and smoothed loss curves, and early stopping was applied based on validation performance.

Transformer Decoder for Sequence Reconstruction

To decode embeddings into amino acid sequences, we trained a Transformer-based decoder. This model receives either a real or denoised latent vector as input and generates the corresponding protein sequence. The decoder architecture consists of:

* Positional embedding projection of latent vector
* 4-layer Transformer encoder
* Output softmax over 20 amino acid tokens
* Optional reference-guided decoding, where a portion of the target sequence is clamped to a known anchor.

The decoder was trained on embedding–sequence pairs using cross-entropy loss with label smoothing. We employed teacher forcing during training and autoregressive sampling during inference. Conditioning vectors (organism class) were appended to the latent embeddings when training class-aware decoders.

Sequence Generation and Sampling

To generate synthetic sequences:

1. We sampled pure Gaussian noise vectors from ℝ¹²⁸⁰.
2. Each vector was denoised using the trained latent diffusion model over 1000 reverse steps.
3. The resulting embeddings were decoded into sequences using the Transformer decoder.
4. For reference-guided decoding, the decoder was conditioned on 40% of a known input sequence.

A total of 180 sequences were generated per class (bacteria, fungi, archaea), balancing guided and unguided decoding strategies.

Evaluation Metrics

We employed a multi-tiered evaluation pipeline to assess the quality and realism of generated sequences:

Embedding-Based Analyses

* t-SNE projection of both real and synthetic embeddings to visualize manifold overlap and diversity.
* Cosine similarity matrices computed within (real–real, gen–gen) and across (real–gen) groups to evaluate novelty and biological plausibility.
* Histogram of cosine distances to characterize similarity distribution.

Sequence-Level Analyses

* Local BLAST+ alignment of each generated sequence against the curated dataset.
  + Metrics: Top hit percent identity, E-value, and alignment length.
* Entropy and percent identity histograms to assess diversity and redundancy.
* Scatter plot of top hit identity vs. −log₁₀(E-value) to evaluate evolutionary plausibility.

All analyses were conducted using Python (NumPy, SciPy, PyTorch, Biopython) and plotted with Matplotlib and Seaborn.

Discussion

In this work, we introduced EmbedDiff, a novel generative framework for protein design that leverages latent diffusion in pretrained protein language model embedding space. Our approach departs from traditional sequence generation paradigms by avoiding direct autoregressive decoding or explicit structure conditioning. Instead, EmbedDiff models the manifold of biologically meaningful protein embeddings and learns to sample new representations through a denoising process—subsequently decoding these into sequences using a Transformer model trained on real embedding–sequence pairs.

Our results demonstrate that this latent-space modeling strategy can produce diverse, biologically plausible, and evolutionarily grounded protein sequences. Generated embeddings maintain structured proximity to real proteins, as evidenced by cosine similarity matrices and t-SNE clustering, while their decoded sequences exhibit moderate-to-high identity to natural proteins without collapsing into memorized examples. The model achieves this while remaining fully structure-free, highlighting the potential of embedding-guided generative design as a scalable alternative to resource-intensive structure-based pipelines.

Critically, our approach benefits from the representational power of pretrained models like ESM2. These embeddings implicitly encode structural and functional constraints derived from large-scale natural protein data. By learning to operate in this space, EmbedDiff inherits these priors without requiring structural labels, annotations, or explicit folding models. This opens the door to class-aware, template-free, and computationally lightweight protein design pipelines, particularly for under-characterized or novel protein families.

The success of EmbedDiff also underscores the value of diffusion models in protein generation. Compared to variational autoencoders or GANs, diffusion models offer more stable training, better mode coverage, and finer control over the generation process. In our case, the diffusion process learns to denoise corrupted ESM2 embeddings—effectively modeling the geometry of natural protein space. Importantly, this denoising process generalizes well: synthetic embeddings are not just reconstructions, but span novel regions within the evolutionary landscape, as shown by the separation and interpolation observed in t-SNE plots and sequence identity histograms.

Nonetheless, several limitations remain. First, while our cosine similarity and BLAST-based analyses provide strong indirect validation of biological realism, structural foldability of generated sequences was not exhaustively evaluated due to computational constraints. Integrating fast predictors like ESMFold or training foldability classifiers based on pLDDT scores would provide a valuable next step. Second, the Transformer decoder currently treats sequence generation as a single-step mapping from embedding to sequence. Incorporating autoregressive refinement, sequence diffusion, or structure-conditioned decoding could enhance fidelity and enable more nuanced control over generation.

Moreover, although reference-guided decoding improves biological relevance, it introduces a tradeoff between novelty and conservation that warrants further exploration. Future work could explore adaptive decoding schemes, where the degree of conditioning is dynamically adjusted based on confidence or target diversity.

Looking ahead, EmbedDiff opens several promising avenues. It could be used to expand protein families, generate distant homologs, or scaffold design around existing functional motifs. By conditioning generation on additional features such as domain annotations, enzyme class, or subcellular localization, the framework could support goal-directed design in industrial and therapeutic settings. Furthermore, coupling EmbedDiff with experimental screening pipelines or wet-lab feedback loops may enable rapid design–test–learn cycles, particularly in high-throughput expression platforms.

In conclusion, EmbedDiff represents a new direction in protein generation—grounded not in explicit structural rules or sequence templates, but in the latent evolutionary signatures captured by deep unsupervised models. By combining these representations with the flexibility of diffusion processes and the expressiveness of Transformer decoders, we demonstrate that structure-free generative models can produce diverse and biologically meaningful sequences. We believe EmbedDiff lays the foundation for future advancements in scalable, interpretable, and data-efficient protein engineering.