# Comparative Analysis of EmbedDiff\_ESM and EmbedDiff\_Dayhoff: A Data-Driven Assessment of Structure-Free De Novo Protein Design Approaches

\*\*Authors:\*\* Melaku Garsamo

\*\*Affiliation:\*\* Independent Researcher

\*\*Corresponding Author:\*\* melaku.garsamo@example.com

\*\*Date:\*\* August 29, 2025

\*\*Journal:\*\* bioRxiv (Preprint)

---

## 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 present a comprehensive comparative analysis of two distinct generative frameworks: EmbedDiff\_ESM, which operates in the latent space of pretrained ESM-2 protein language models, and EmbedDiff\_Dayhoff, which utilizes evolutionary substitution matrices for protein generation. Both approaches enable de novo protein design without structural supervision but employ fundamentally different strategies for capturing biological constraints.

EmbedDiff\_ESM uses ESM-2 embeddings of natural proteins to train a denoising diffusion model that learns the manifold of biologically meaningful sequences, while EmbedDiff\_Dayhoff leverages classical evolutionary substitution patterns to guide sequence generation. We evaluate both approaches on a diverse set of thioredoxin reductases across bacteria, archaea, and fungi, analyzing 240-242 generated sequences per method using multi-tiered validation including t-SNE visualization, cosine similarity metrics, and local BLAST alignment.

Our comparative analysis reveals that EmbedDiff\_Dayhoff generates sequences with significantly higher identity to natural proteins (mean: 58.12% ± 2.89%), while EmbedDiff\_ESM produces more controlled distributions (mean: 42.67% ± 2.47%) with superior classification accuracy (95%). These findings demonstrate complementary strengths: Dayhoff excels in evolutionary conservation modeling, while ESM-2 provides better domain discrimination and controlled novelty. The study identifies critical research gaps in metric standardization and highlights the need for comprehensive evaluation frameworks in protein generation research, laying the groundwork for future hybrid approaches that combine the strengths of both methodologies.

\*\*Keywords:\*\* Protein sequence generation, ESM-2, Dayhoff matrices, diffusion models, computational biology, machine learning, comparative analysis

---

## 1. Introduction

### 1.1 Background and Motivation

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 10^130 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-2, 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 ESM-2.

### 1.2 Two Emerging Approaches: ESM-2 vs Dayhoff

To overcome these limitations, two distinct generative frameworks have emerged, each offering unique advantages for structure-free protein design:

\*\*EmbedDiff\_ESM Approach\*\*: This framework performs protein design in the embedding space of natural sequences using latent denoising diffusion modeling. EmbedDiff\_ESM takes advantage of the ESM-2 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\_ESM to sample entire latent protein representations holistically and generate embeddings that lie within the manifold of plausible protein states.

\*\*EmbedDiff\_Dayhoff Approach\*\*: This alternative framework leverages classical evolutionary substitution matrices to guide protein generation. The Dayhoff approach represents a more traditional evolutionary biology methodology, directly modeling the probability of amino acid changes based on observed substitution patterns across evolutionary time. This method provides a foundation for evolutionary conservation analysis and phylogenetic inference, offering a complementary perspective to the modern deep learning approaches.

### 1.3 Research Objectives and Significance

This study addresses three primary research questions:

1. \*\*Quantitative Comparison\*\*: How do EmbedDiff\_ESM and EmbedDiff\_Dayhoff perform across measurable metrics including sequence identity, classification accuracy, and training characteristics?

2. \*\*Methodological Assessment\*\*: What are the strengths and limitations of each approach based on objective, data-driven analysis?

3. \*\*Standardization Gaps\*\*: What research infrastructure is needed to enable comprehensive comparative analysis in protein generation?

\*\*Novel Contributions:\*\*

- First systematic comparison of ESM-2 vs Dayhoff approaches in protein generation

- Identification of critical data standardization gaps in the field

- Framework for objective, bias-free comparative analysis

- Quantitative benchmarks for future research

- Recognition of complementary rather than competitive relationship between approaches

### 1.4 Evaluation Framework

To evaluate the effectiveness of both approaches, 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 comparative analysis demonstrates that both approaches are capable of generating diverse, non-trivial protein sequences that are novel yet biologically grounded. Generated sequences occupy realistic regions of their respective spaces, exhibit meaningful similarity to known proteins, and demonstrate complementary strengths in different aspects of protein generation. This study 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 through hybrid approaches.

---

## 2. Methods

### 2.1 Dataset Construction

Thioredoxin reductase homologs were curated from \*\*UniProtKB\*\* across three evolutionary domains: archaea, bacteria, and fungi. Sequences shorter than 200 residues or with >95% identity were removed using \*\*CD-HIT\*\* to ensure non-redundancy. Final dataset contained \*\*N = 480 sequences\*\*, split into \*\*80% training, 10% validation, 10% test\*\*, stratified by domain. All FASTA headers were annotated with domain labels.

### 2.2 Embedding Extraction

#### 2.2.1 ESM-2 Embeddings

\* \*\*Model:\*\* `esm2\_t33\_650M\_UR50D` (Meta AI, HuggingFace implementation).

\* \*\*Embedding size:\*\* 1280 dimensions.

\* \*\*Representation:\*\* [CLS] token concatenated with averaged residue embeddings.

\* \*\*Normalization:\*\* Scaled to [-1, 1] with tanh.

\* \*\*Environment:\*\* PyTorch 2.1.0, CPU execution on Apple M3 Pro.

#### 2.2.2 Dayhoff Embeddings

\* \*\*Model:\*\* `microsoft/Dayhoff-3b-UR90` (Microsoft Protein Atlas).

\* \*\*Embedding size:\*\* 1280 dimensions (auto-detected).

\* \*\*Representation:\*\* Pooled sequence embeddings using Mamba/Jamba-compatible kernel.

\* \*\*Normalization:\*\* Scaled to [-1, 1] to match ESM-2 preprocessing.

\* \*\*Environment:\*\* PyTorch 2.3.1, HuggingFace Transformers 4.42, CPU execution with `use\_mamba\_kernels=False`.

### 2.3 Logistic Regression Probe Analysis

For both pipelines, a \*\*logistic regression classifier\*\* was trained on raw embeddings to assess separability of evolutionary domains.

\* \*\*Implementation:\*\* scikit-learn 1.4, solver = `lbfgs`.

\* \*\*Iterations:\*\* 1000 max.

\* \*\*Evaluation:\*\* 5-fold cross-validation.

\* \*\*Outputs:\*\* Accuracy, recall, F1-score, confusion matrix, per-class recall plots.

This probe quantified how much information about evolutionary domain is linearly accessible from each embedding space.

### 2.4 Latent Diffusion Model

Both pipelines use a \*\*denoising diffusion probabilistic model (DDPM)\*\* with identical architecture and training settings.

#### Architecture

\* \*\*Noise predictor:\*\* MLP with hidden layers:

$$

1280 + d\_\text{label} + d\_t \;\to\; 1024 \;\to\; 1024 \;\to\; 512 \;\to\; 1280

$$

where $d\_\text{label}=3$ (one-hot domain labels) and $d\_t=32$ (timestep encoding).

\* \*\*Activations:\*\* ReLU.

\* \*\*Normalization:\*\* LayerNorm.

\* \*\*Dropout:\*\* 0.2.

#### Diffusion Process

\* \*\*Timesteps:\*\* 1000.

\* \*\*Noise schedule:\*\* Cosine β schedule, β ∈ [1e-4, 0.02].

\* \*\*Forward process:\*\*

$$

q(x\_t | x\_0) = \sqrt{\alpha\_t} x\_0 + \sqrt{1 - \alpha\_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,I)

$$

\* \*\*Reverse process:\*\* Learned denoising distribution $p\_\theta(x\_{t-1}|x\_t)$.

#### Training

\* \*\*Loss:\*\*

$$

L(\theta) = \mathbb{E}\_{x\_0,\epsilon,t}\big[\|\epsilon - \epsilon\_\theta(x\_t,t,y)\|^2\big]

$$

\* \*\*Optimizer:\*\* Adam (lr=1e-4, β=(0.9, 0.999)).

\* \*\*Batch size:\*\* 32.

\* \*\*Epochs:\*\* 300 for both (with Dayhoff showing slightly earlier stabilization but trained to match).

\* \*\*Hardware:\*\* Apple M3 Pro CPU, 32 GB RAM.

### 2.5 Sampling Synthetic Embeddings

At inference, Gaussian noise was denoised over 1000 timesteps to produce novel embeddings. Each run yielded ~240 synthetic embeddings per pipeline, conditioned on evolutionary domain.

### 2.6 Transformer Decoder

#### Dataset

For both pipelines, decoder training used paired (embedding, sequence) data from the real set.

#### Architecture

\* \*\*Model:\*\* Transformer decoder (encoder–decoder variant).

\* \*\*Input projection:\*\* Embeddings mapped to 256-d model space.

\* \*\*Layers:\*\* 4, hidden size=512, 8 attention heads.

\* \*\*Loss:\*\* Cross-entropy with label smoothing (ε=0.1).

\* \*\*Optimization:\*\* Adam, lr=1e-4, batch size=32.

#### Decoding strategy

\* \*\*Hybrid stochastic decoding:\*\*

\* 60% of positions sampled stochastically from softmax output.

\* 40% reference-guided (nearest real sequence by cosine similarity).

\* \*\*Rationale:\*\* Balances novelty (fully stochastic) with biological plausibility (reference-guided).

\* \*\*Configurable:\*\* Ratio adjustable in `transformer\_decode\_[esm2|dayhoff].py`.

### 2.7 Validation

1. \*\*t-SNE Visualization\*\*

\* sklearn TSNE (perplexity=30, learning rate=200, n\_iter=2000, random\_state=42).

\* Real vs generated embeddings plotted jointly.

2. \*\*Cosine Similarity\*\*

\* Pairwise cosine distances: real–real, real–generated, generated–generated.

3. \*\*Entropy Analysis\*\*

\* Per-sequence Shannon entropy:

$$

H = -\sum\_{a \in A} p(a) \log p(a)

$$

where $p(a)$ = residue frequency.

4. \*\*Sequence Identity (BLAST)\*\*

\* Local `blastp` against SwissProt (release 2024\_05).

\* Outputs: % identity, alignment length, bit score, E-value.

\* Threshold for plausibility: ≥30% identity.

5. \*\*Classification of Generated Embeddings\*\*

\* Logistic regression retrained on real+generated embeddings.

\* Measures whether synthetic embeddings maintain domain-separable structure.

### 2.8 Reporting

Each pipeline produced an \*\*interactive HTML report\*\* containing:

\* Training curves.

\* Logistic regression confusion matrices.

\* t-SNE plots (real vs generated).

\* Cosine similarity histograms.

\* Identity and entropy distributions.

\* BLAST validation summaries.

Reports for both pipelines are available online:

\* [EmbedDiff-ESM2 Report](https://mgarsamo.github.io/EmbedDiff\_ESM/embeddiff\_esm2\_summary\_report.html)

\* [EmbedDiff-Dayhoff Report](https://mgarsamo.github.io/EmbedDiff-Dayhoff/embeddiff\_dayhoff\_summary\_report.html)

---

## 3. Results

### 3.1 Embedding Space Visualization (t-SNE)

\*\*Figure 1: t-SNE Visualization of Thioredoxin Reductase Sequences\*\*

\*2D t-SNE projections of protein sequences from both approaches, colored by biological domain (bacteria: green, fungi: orange, archaea: blue), illustrating the clustering and separation quality in their respective embedding spaces.\*

\*\*EmbedDiff\_ESM t-SNE Characteristics:\*\*

- \*\*Title\*\*: "Thioredoxin Reductase (ESM-2)"

- \*\*Axes Range\*\*: t-SNE 1: -60 to 60, t-SNE 2: -60 to 60

- \*\*Clustering Quality\*\*: Clear separation between domains with minimal overlap

- \*\*Domain Distribution\*\*:

- \*\*Archaea (blue)\*\*: Dense cluster in upper-left quadrant (-40 to -20, 20 to 45)

- \*\*Bacteria (green)\*\*: Clusters in lower-left (-40 to -10, -40 to -10) and central-right (10 to 20, -40 to -10)

- \*\*Fungi (orange)\*\*: Concentrated in right half (0 to 60, -20 to 40) with dense regions around (30, 10) and (50, -10)

- \*\*Separation Strength\*\*: Strong distinction between archaea and fungi, bacteria shows intermediate distribution

\*\*EmbedDiff\_Dayhoff t-SNE Characteristics:\*\*

- \*\*Title\*\*: "Thioredoxin Reductase (Dayhoff)"

- \*\*Axes Range\*\*: t-SNE 1: -60 to 60, t-SNE 2: -60 to 60

- \*\*Clustering Quality\*\*: Good separation with some intermingling between bacteria and archaea

- \*\*Domain Distribution\*\*:

- \*\*Fungi (orange)\*\*: Most distinct cluster in lower half (-25 to 10, -70 to -20) with secondary cluster around (35, -15)

- \*\*Bacteria (green)\*\*: Upper-left quadrant (-30 to -10, 20 to 40) with isolated cluster around (-50, -15)

- \*\*Archaea (blue)\*\*: Upper half distribution with clusters in upper-right (20 to 60, 0 to 20) and upper-middle (-5 to 15, 20 to 40)

- \*\*Separation Strength\*\*: Fungi shows clearest separation, bacteria-archaea overlap in upper-middle region

\*\*Comparative Analysis of Embedding Quality:\*\*

| Aspect | EmbedDiff\_ESM | EmbedDiff\_Dayhoff | Observation |

|--------|---------------|-------------------|-------------|

| \*\*Overall Separation\*\* | Excellent | Good | ESM-2 provides cleaner domain boundaries |

| \*\*Fungi Isolation\*\* | Strong | Strong | Both approaches excel at fungi separation |

| \*\*Bacteria-Archaea\*\* | Clear separation | Some overlap | ESM-2 better distinguishes closely related domains |

| \*\*Cluster Density\*\* | High | Moderate | ESM-2 produces more compact, well-defined clusters |

| \*\*Inter-domain Mixing\*\* | Minimal | Moderate | ESM-2 maintains cleaner domain separation |

\*\*Biological Interpretation:\*\*

- \*\*ESM-2 Advantage\*\*: Modern language model captures more nuanced evolutionary relationships, enabling better domain discrimination

- \*\*Dayhoff Limitation\*\*: Classical substitution matrices may miss complex evolutionary patterns, leading to some domain overlap

- \*\*Consistent Patterns\*\*: Both approaches show fungi as the most distinct domain, reflecting genuine biological separation

- \*\*Evolutionary Insight\*\*: Bacteria-archaea overlap in Dayhoff suggests shared evolutionary history that ESM-2 can better resolve

### 3.2 Training Dynamics and Convergence

\*\*Figure 2: Training Loss Comparison Between EmbedDiff\_ESM and EmbedDiff\_Dayhoff\*\*

\*Training loss curves showing convergence patterns and learning dynamics for both approaches over 300 epochs. Both use Mean Squared Error (MSE) loss, with raw and smoothed lines for train and validation.\*

\*\*EmbedDiff\_ESM Training Characteristics:\*\*

- \*\*Training Duration\*\*: 300 epochs

- \*\*Initial Loss\*\*: ~1.000 at epoch 0

- \*\*Final Smoothed Training Loss\*\*: ~0.860 at epoch 300

- \*\*Final Smoothed Validation Loss\*\*: ~0.865 at epoch 300

- \*\*Convergence Pattern\*\*: Rapid initial decrease in the first 50-100 epochs, then gradual decline; raw lines show oscillations, smoothed reveal steady trend

- \*\*Generalization\*\*: Validation loss tracks training closely without divergence, indicating no overfitting

\*\*EmbedDiff\_Dayhoff Training Characteristics:\*\*

- \*\*Training Duration\*\*: 300 epochs

- \*\*Initial Loss\*\*: ~1.000 at epoch 0

- \*\*Final Smoothed Training Loss\*\*: ~0.830 at epoch 300

- \*\*Final Smoothed Validation Loss\*\*: ~0.835 at epoch 300

- \*\*Convergence Pattern\*\*: Similar rapid initial drop, but steeper late-stage improvement (epochs 200-300); oscillations comparable to ESM-2

- \*\*Generalization\*\*: Validation slightly higher than training but tracks well, suggesting robust generalization

\*\*Key Observations:\*\*

1. \*\*Similar Loss Scales\*\*: Both operate in MSE range from ~1.000 to ~0.850, with Dayhoff achieving ~3-4% lower final loss

2. \*\*Training Efficiency\*\*: Both converge over 300 epochs, but Dayhoff shows better late-stage optimization

3. \*\*Generalization Quality\*\*: No signs of overfitting in either; validation follows training trends

4. \*\*Convergence Stability\*\*: Smoothed lines show smooth downward trajectories for both

5. \*\*Learning Rate\*\*: Initial drops similar, but Dayhoff continues improving more in later epochs

\*\*Training Efficiency Comparison:\*\*

- \*\*Early Learning (Epochs 0-100)\*\*: Rapid improvement in both

- \*\*Mid Training (Epochs 100-200)\*\*: Gradual decline, with Dayhoff slightly steeper

- \*\*Late Training (Epochs 200-300)\*\*: Dayhoff edges out with continued reduction

- \*\*Overall Efficiency\*\*: Dayhoff demonstrates marginally superior convergence, potentially due to simpler encoding

\*\*Methodological Implications:\*\*

- \*\*Loss Differences\*\*: Dayhoff's lower final loss suggests better fit to data manifold

- \*\*Convergence Speed\*\*: Comparable, but Dayhoff's late gains indicate robustness

- \*\*Training Duration\*\*: Standardized to 300 epochs for fair comparison

### 3.3 Classification Performance Analysis

\*\*Figure 3: Logistic Regression Classification Performance Comparison\*\*

\*Per-class recall scores and confusion matrices for both approaches, demonstrating the discriminative power of their respective embeddings for biological domain classification.\*

\*\*EmbedDiff\_ESM Classification Results:\*\*

- \*\*Overall Accuracy\*\*: 95% (based on confusion matrix analysis)

- \*\*Per-Class Recall\*\*:

- Archaea: 90.0% (18/20 correct)

- Bacteria: 95.0% (19/20 correct)

- Fungi: 100.0% (20/20 correct)

- \*\*Confusion Patterns\*\*: Primary misclassification between archaea and bacteria (10% archaea→bacteria, 5% bacteria→archaea)

- \*\*Fungi Separation\*\*: Perfect classification with no cross-domain confusion

\*\*EmbedDiff\_Dayhoff Classification Results:\*\*

- \*\*Overall Accuracy\*\*: 91.7% (55/60 total correct)

- \*\*Per-Class Recall\*\*:

- Archaea: 89.0% (~0.89, approximately 18/20 correct)

- Bacteria: 84.0% (~0.84, approximately 17/20 correct)

- Fungi: 99.0% (~0.99, approximately 20/20 correct)

- \*\*Confusion Patterns\*\*: Similar archaea-bacteria confusion (10% archaea→bacteria, 15% bacteria→archaea)

- \*\*Fungi Separation\*\*: Near-perfect classification with minimal cross-domain confusion

\*\*Quantitative Performance Comparison:\*\*

| Metric | EmbedDiff\_ESM | EmbedDiff\_Dayhoff | Difference |

|--------|---------------|-------------------|------------|

| \*\*Overall Accuracy\*\* | 95.0% | 91.7% | +3.3% |

| \*\*Archaea Recall\*\* | 90.0% | 89.0% | +1.0% |

| \*\*Bacteria Recall\*\* | 95.0% | 84.0% | +11.0% |

| \*\*Fungi Recall\*\* | 100.0% | 99.0% | +1.0% |

| \*\*Mean Recall\*\* | 95.0% | 90.7% | +4.3% |

\*\*Key Classification Insights:\*\*

1. \*\*ESM-2 Superiority\*\*: EmbedDiff\_ESM demonstrates consistently higher performance across all metrics

2. \*\*Bacteria Classification Gap\*\*: Largest performance difference observed in bacteria classification (+11.0%)

3. \*\*Fungi Excellence\*\*: Both approaches achieve near-perfect fungi classification

4. \*\*Archaea-Bacteria Confusion\*\*: Common challenge for both approaches, indicating genuine biological similarity

5. \*\*Domain Discrimination\*\*: ESM-2 embeddings provide better separation between closely related domains

\*\*Biological Interpretation:\*\*

- \*\*ESM-2 Advantage\*\*: Modern language model embeddings capture more nuanced evolutionary relationships

- \*\*Dayhoff Limitation\*\*: Classical substitution matrices may miss complex, non-linear evolutionary patterns

- \*\*Consistent Patterns\*\*: Both approaches struggle with archaea-bacteria distinction, indicating genuine biological similarity

- \*\*Fungi Specificity\*\*: Both approaches excel at fungi classification, suggesting distinct evolutionary trajectory

### 3.4 Sequence Identity Distribution Analysis

#### 3.4.1 ESM-2 Performance

The EmbedDiff\_ESM pipeline generated 240 sequences with identity scores ranging from 36.86% to 48.86% (Table 1). The distribution demonstrates remarkable consistency with a standard deviation of only 2.47%, indicating highly controlled sequence generation.

\*\*Table 1: ESM-2 Sequence Identity Statistics\*\*

| Metric | Value |

|--------|-------|

| \*\*Sample Size\*\* | 240 sequences |

| \*\*Range\*\* | 36.86% - 48.86% |

| \*\*Mean\*\* | 42.67% |

| \*\*Median\*\* | 43.14% |

| \*\*Standard Deviation\*\* | 2.47% |

| \*\*Coefficient of Variation\*\* | 5.79% |

#### 3.4.2 Dayhoff Performance

The EmbedDiff\_Dayhoff approach generated 242 sequences with identity scores ranging from 37.43% to 64.86% (Table 2). This represents a significantly broader distribution with higher mean values compared to ESM-2.

\*\*Table 2: Dayhoff Sequence Identity Statistics\*\*

| Metric | Value |

|--------|-------|

| \*\*Sample Size\*\* | 242 sequences |

| \*\*Range\*\* | 37.43% - 64.86% |

| \*\*Mean\*\* | 58.12% |

| \*\*Median\*\* | 58.57% |

| \*\*Standard Deviation\*\* | 2.89% |

| \*\*Coefficient of Variation\*\* | 4.97% |

#### 3.4.3 Comparative Analysis

\*\*Table 3: Direct Comparison of Sequence Identity Metrics\*\*

| Comparison Metric | ESM-2 | Dayhoff | Difference | Ratio |

|-------------------|--------|---------|------------|-------|

| \*\*Mean Identity\*\* | 42.67% | 58.12% | +15.45% | 1.36x |

| \*\*Identity Range\*\* | 12.00% | 27.43% | +15.43% | 2.29x |

| \*\*Standard Deviation\*\* | 2.47% | 2.89% | +0.42% | 1.17x |

| \*\*Coefficient of Variation\*\* | 5.79% | 4.97% | -0.82% | 0.86x |

\*\*Key Findings:\*\*

- Dayhoff generates sequences with 36% higher mean identity

- Dayhoff explores 2.29x broader identity range

- ESM-2 shows 17% higher variance in identity scores

- ESM-2 demonstrates more controlled generation (higher CV indicating relative variability but tighter absolute control)

### 3.5 Distribution Shape Analysis

#### 3.5.1 ESM-2 Distribution Characteristics

- \*\*Shape\*\*: Bimodal-like distribution with peaks around 40% and 45%

- \*\*Consistency\*\*: 95% of sequences fall within 38-47% identity range

- \*\*Outliers\*\*: Minimal extreme values, suggesting robust quality control

#### 3.5.2 Dayhoff Distribution Characteristics

- \*\*Shape\*\*: Broader, more uniform distribution

- \*\*Variability\*\*: 95% of sequences fall within 52-64% identity range

- \*\*Outliers\*\*: Some sequences achieve very high identity (60%+)

---

## 4. Discussion

### 4.1 Interpretation of Key Findings

#### 4.1.1 Sequence Identity Patterns

The observed differences in sequence identity distributions reveal fundamental differences in approach philosophy:

\*\*ESM-2 Conservative Strategy:\*\*

- Mean identity of 42.67% suggests a conservative approach prioritizing biological plausibility over novelty

- Low variance (2.47%) indicates highly controlled generation with consistent quality

- Narrow range (12%) suggests careful exploration of sequence space

\*\*Dayhoff Exploration Strategy:\*\*

- Mean identity of 58.12% indicates a more aggressive approach to sequence generation

- Higher variance (2.89%) suggests greater exploration of diverse sequence properties

- Broad range (27.43%) indicates extensive sequence space exploration

#### 4.1.2 Biological Implications

\*\*ESM-2 Advantages:\*\*

- \*\*Controlled Novelty\*\*: Generates sequences with predictable, moderate novelty

- \*\*Quality Consistency\*\*: Uniform sequence quality across the dataset

- \*\*Domain Discrimination\*\*: 95% classification accuracy suggests excellent evolutionary relationship preservation

\*\*Dayhoff Advantages:\*\*

- \*\*Evolutionary Conservation\*\*: Higher identity scores suggest better preservation of natural protein properties

- \*\*Sequence Diversity\*\*: Broader exploration of sequence space

- \*\*Evolutionary Focus\*\*: Direct integration of substitution matrices yields lower final training loss

### 4.2 Methodological Considerations

#### 4.2.1 Data Standardization Gaps

\*\*Critical Issues Identified:\*\*

1. \*\*Inconsistent Metric Reporting\*\*: Different projects report metrics in different formats

2. \*\*Binary Data Accessibility\*\*: Training curves stored in inaccessible binary formats

3. \*\*Visual-Only Data\*\*: Many metrics exist only as images without numerical values

4. \*\*Missing Comparative Frameworks\*\*: No standardized evaluation protocols

#### 4.2.2 Bias Prevention Measures

\*\*Analysis Safeguards:\*\*

- \*\*Quantitative Focus\*\*: Analysis limited to accessible numerical data

- \*\*Transparent Limitations\*\*: Clear documentation of what cannot be concluded

- \*\*Balanced Interpretation\*\*: Equal consideration of both approaches' strengths

- \*\*Methodological Transparency\*\*: Explicit reporting of analysis constraints

### 4.3 Research Implications

#### 4.3.1 Application Domain Recommendations

\*\*Use ESM-2 for:\*\*

- Applications requiring consistent, controlled sequence generation

- Projects needing high domain classification accuracy

- Research requiring predictable sequence properties

- Quality-focused protein engineering applications

\*\*Use Dayhoff for:\*\*

- Evolutionary conservation studies

- Applications requiring higher sequence identity to natural proteins

- Research focused on substitution pattern modeling

- Projects needing extensive sequence space exploration

#### 4.3.2 Future Research Directions

\*\*Immediate Priorities:\*\*

1. \*\*Data Standardization\*\*: Implement consistent metric reporting across projects

2. \*\*Numerical Extraction\*\*: Develop tools for analyzing binary training data

3. \*\*Visual Analysis\*\*: Create quantitative analysis tools for image-based metrics

4. \*\*Benchmark Development\*\*: Establish standardized evaluation protocols

\*\*Long-term Goals:\*\*

1. \*\*Hybrid Approaches\*\*: Combine strengths of both methods

2. \*\*Comprehensive Evaluation\*\*: Multi-dimensional performance assessment

3. \*\*Biological Validation\*\*: Experimental testing of generated sequences

4. \*\*Scalability Analysis\*\*: Performance assessment across different scales

---

## 5. Conclusion

### 5.1 Summary of Key Findings

This data-driven comparative analysis reveals complementary strengths between EmbedDiff\_ESM and EmbedDiff\_Dayhoff approaches:

\*\*Quantitative Evidence:\*\*

- \*\*Sequence Identity\*\*: Dayhoff generates higher-identity sequences (58.12% vs 42.67%)

- \*\*Distribution Control\*\*: ESM-2 provides more controlled, consistent generation

- \*\*Classification Performance\*\*: ESM-2 achieves 95% accuracy vs Dayhoff's 91.7%

- \*\*Training Dynamics\*\*: Dayhoff reaches lower final MSE loss (~0.830 vs ~0.860)

- \*\*Exploration Breadth\*\*: Dayhoff explores 2.29x broader identity range

\*\*Qualitative Assessment:\*\*

- Both approaches generate biologically reasonable sequences

- Each method demonstrates unique advantages for specific applications

- No single approach demonstrates universal superiority

### 5.2 Research Contributions

\*\*Methodological Advances:\*\*

- First systematic comparison of ESM-2 vs Dayhoff approaches

- Identification of critical data standardization gaps

- Framework for objective, bias-free comparative analysis

- Recommendations for future research infrastructure

\*\*Practical Insights:\*\*

- Clear application domain guidance for both approaches

- Recognition of complementary rather than competitive relationship

- Identification of research gaps requiring immediate attention

### 5.3 Limitations and Future Work

\*\*Current Limitations:\*\*

- Training dynamics analyzed visually; quantitative extraction limited

- Some metrics unavailable for direct comparison due to format

- Inconsistent data formats limit comprehensive comparison

- Lack of experimental biological validation

\*\*Future Research Requirements:\*\*

- Development of standardized evaluation protocols

- Creation of quantitative analysis tools for visual data

- Implementation of consistent metric reporting

- Comprehensive biological validation studies

### 5.4 Broader Impact

This study contributes to the broader field of computational protein generation by:

- \*\*Establishing Benchmarks\*\*: Providing quantitative performance baselines

- \*\*Identifying Gaps\*\*: Highlighting critical infrastructure needs

- \*\*Promoting Standards\*\*: Advocating for consistent evaluation protocols

- \*\*Guiding Applications\*\*: Informing choice of methods for specific use cases

The findings suggest that rather than competing approaches, ESM-2 and Dayhoff represent complementary methodologies, each excelling in different aspects of protein sequence generation. Future research should focus on developing hybrid approaches that leverage the strengths of both methods while addressing the standardization gaps identified in this study.

---

## References

[References would be added here following journal-specific formatting requirements, including citations for ESM-2, Dayhoff matrices, diffusion models, and related tools like UniProtKB, CD-HIT, scikit-learn, and BLAST.]

---

## Supplementary Information

\*\*Data Availability:\*\*

- ESM-2 sequence identity data: Available in project repository

- Dayhoff sequence identity data: Available in project repository

- Classification results: Available in reports

- Training curves: Analyzed from visual plots; raw data in repositories

\*\*Code Availability:\*\*

- Both projects are open-source and available on GitHub

- Analysis scripts used in this study available upon request

\*\*Author Contributions:\*\*

- M.G.: Conceptualization, methodology, investigation, analysis, writing

\*\*Funding:\*\*

- Self-funded

\*\*Conflicts of Interest:\*\*

- The authors declare no conflicts of interest

---

\*\*Document Version:\*\* 5.0 (Updated with Plot Analysis and Corrections)

\*\*Last Updated:\*\* August 29, 2025

\*\*Analysis Completeness:\*\* Comprehensive based on available data

\*\*Peer Review Status:\*\* Preprint ready

\*\*Target Journal:\*\* bioRxiv (initial), potentially Nature Computational Science or PLOS Computational Biology