SAGEdiff: Species Aware Generative Extrapolative Diffusion Model for Protein Sequence Generation

Ari Willig[1,2,*] **Seung Hyun Jin**[2,*] **Hakeem Shitta-Bey**[1,2,3,*] **Prannav Shankar**[1,2,*]

Zachary Quinn[1] **Sophia Vincoff**[1] **Benjamin Perry**[1]

Pranam Chatterjee[1]

[1]Department of Biomedical Engineering, Duke University
[2]Department of Computer Science, Duke University
[3]Department of Electrical and Computer Engineering, Duke University
[*]Equal contribution

Abstract

While recent developments in deep learning architectures have established diffusion models as a particularly powerful framework for generative modeling of novel proteins, existing approaches often overlook species-specific conditioning, limiting their ability to model the evolutionary and functional nuances inherent to different organisms. In this work, we design a novel transformer-based discrete diffusion model using per-residue tokenized amino acid sequences and introducing explicit species identity through class tokens (0: fruit fly, 1: human, 2: rat). Our model is trained on a subsampled version of UniProt and evaluated against established baselines (EvoDiff) using both structure-based and sequence-based metrics, including but not limited to mean pLDDT, pTM, and Edit Distance (Levenshtein Distance). Extensive evaluation against these state-of-the-art models has validated the effectiveness of our model architecture. Most notably, our model improved over EvoDiff regarding sequence diversity (Shannon Entropy) and showed very comparable performance in all other metrics. Internally, our model also exemplified increased sequence similarity within each class showing that the included class labels led to sequences that were properly grouped together.

1 Intro

Proteins are found throughout biological systems and play essential roles in a wide range of cellular processes, with their structure and function closely tied to their amino acid sequence [1]. While recent advancements in machine learning have opened the doors to novel protein sequence generation, most overlook an important dimension: species specificity [12]. While many proteins found in different organisms—such as humans, rats, and fruit flies—exhibit extremely similar functions they often have distinct sequences and cannot functionally replace each other due to differences in evolutionary pressures, metabolic requirements, and cellular environments [9]. In synthetic biology, designing proteins that are optimized for expression and functionality in a specific host organism can significantly enhance activity and allow for improved yield even with limited data for understudied species [10]. Furthermore, in therapeutic design, generating species-specific variants of proteins enables the development of more accurate animal models for preclinical testing and allows for enhanced understanding of inter-species differences. [8].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Despite progress in generative modeling for proteins using variational autoencoders, autoregressive transformers, and more recently, diffusion-based models, these methods are either limited to unconditioned generation or rely on structure-conditioned generation, restricting de novo proteins to a small and biased subset of design space [12][7]. Additionally, many generative models emphasize structural prediction fidelity over the evolutionary or taxonomic coherence of the sequence, which can limit their utility in comparative genomics or therapeutic translation across model organisms [4][5]. These models also train on datasets including many different proteins not specific to certain species, often collapsing functional and evolutionary distinctions into a single embedding space, thereby reducing their effectiveness in generating sequences with organism-specific motifs or constraints [3]. These limitations can be fixed, however, and the model we introduce in this paper attempts to target some of the weakest links in artificial protein generation.

To address this, we introduce a discrete denoising diffusion model conditioned on species identity through class tokens representing human, rat, and fruit fly. Discrete diffusion allows our model to generate high-diversity samples that can be conditioned on a wide variety of design objectives, species specific proteins in this case [7]. Our model uses a residue level tokenizer to maintain protein domain integrity and preserve biologically meaningful units which would be split up in other popular tokenization methods such as Byte Pair Encoding (BPE) [5]. Unlike prior approaches, this architecture explicitly integrates species-level information into the diffusion process through the specific class tokens mentioned above, enabling the generation of biologically valid and species-specific protein sequences. In comparison to state-of-the-art models such as EvoDiff, our architecture performs extremely well. The calculated metrics for our model give comparable, if not improved, results across the board exemplifying that our architecture generates diverse sequences which also fold into stable structures balancing new generation with structural plausibility.

2 Related Works

Recent deep learning approaches for protein sequence generation have largely centered around autoregressive language models and masked token prediction frameworks. Notable models such as ProtGPT2 [5]and ProGen2[3] leverage transformer-based architectures to generate de novo protein sequences that conform to plausible sequences and show high predicted foldability[5][3]. These models typically rely on large-scale datasets like UniRef or UniProt, learning statistical correlations between residues to implicitly encode evolutionary and structural signals[5][3]. However, these methods are limited by their inability to globally model sequence space: autoregressive models generate tokens sequentially, which introduces exposure bias and restricts their ability to capture long-range dependencies vital for structure and function [11]. Similarly, masked modeling approaches such as ESM-1b and MSA Transformer have shown great promise for representation learning, but they are not explicitly designed for controlled sequence generation, particularly in species-specific contexts [6][2].

Diffusion-based models offer a compelling alternative. Instead of generating sequences autoregressively or by filling in masks, diffusion models learn to denoise corrupted sequences step-by-step, capturing global dependencies and allowing for flexible conditioning strategies[4]. EvoDiff [7] was one of the first diffusion-based approaches applied to proteins, generating entire sequences while conditioning on structural or functional motifs. However, EvoDiff operates in continuous latent space and is not explicitly taxonomically aware. TaxDiff [12] takes this a step further implementing a taxonomic-guided diffusion model but also operates in continuous latent space. In contrast, our work adopts a discrete diffusion framework, more naturally aligned with the symbolic nature of protein sequences, and introduces species-specific conditioning to guide generation toward specific evolutionary lineages. This enables our model to generate functionally coherent, species-specific proteins without sacrificing sequence diversity or plausibility.

By combining discrete denoising with class-aware conditioning, our model addresses key shortcomings in general-purpose protein generators. It enables both targeted design in synthetic biology and insight into evolutionary variation, establishing a framework that is not only expressive but also biologically grounded.

3 Methods

3.1 Data Curation

The dataset was curated by extracting protein sequences from the UniProt Proteomes database, specifically focusing on three taxonomic groups: Drosophila melanogaster (fruit fly), Rattus norvegicus (Norwegian rat), and Homo sapiens (modern human), utilizing the complete proteome entries including their chromosome-specific annotations. All sequences are filtered to remove entries with non-valid amino acids. Sequences were then truncated to a maximum of length 5000 (if required) for computational efficiency. The resulting dataset was then clustered through MMseqs2 to minimize bias from overrepresented protein families. Lastly, the clustered collection was then partitioned with a 80/10/10 split, stratified by taxonomic origin to maintain balanced representation across the three species. This standard split allocation provides a robust training set (80 percent) while reserving adequate data for both validation (10 percent) and testing (10 percent), striking an optimal balance between model learning and evaluation reliability. The dataset was randomly split after the clustering step to ensure both diversity and balanced representation across the training, validation, and test sets. For tokenization, we implemented a residue-level tokenization approach using a custom tokenizer that encodes each amino acid as a unique token and handles special tokens for sequence beginnings, endings, and padding.

3.2 Model Architecture

We implement a discrete diffusion model for protein sequence generation, conditioned on species identity. The architecture is a transformer-based denoising model that learns to reverse a token-masking corruption process. Model components are detailed below.

Vocabulary. We use a fixed residue vocabulary $\mathcal{V}$ of size 25:

$$\mathcal{V} = \{\texttt{A}, \texttt{C}, \texttt{D}, \ldots, \texttt{Y}, \texttt{<pad>}, \texttt{<BOS>}, \texttt{<EOS>}, \texttt{<UNK>}, \texttt{<MASK>}\}$$

Each token is mapped to a unique index via a static lookup table.

Embedding Layers. Inputs to the transformer encoder are composed of the sum of four embeddings:

- **Token Embedding:** maps each residue token x_t^i to a vector in $\mathbb{R}^d$.
- **Timestep Embedding:** encodes the diffusion step $t \in \{0, \ldots, T-1\}$.
- **Species Embedding:** represents the species label $s \in \{0, \ldots, N-1\}$ as a learnable vector.
- **Positional Embedding:** encodes absolute position i in the sequence.

The final input is:

$$\mathbf{X}_i = \mathbf{E}_{\text{tok}}(x_t^i) + \mathbf{E}_{\text{time}}(t) + \mathbf{E}_{\text{species}}(s) + \mathbf{E}_{\text{pos}}(i)$$

Transformer Encoder. The denoising network consists of a 16-layer transformer encoder with multi-head self-attention:

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{X})$$

Each encoder layer uses 16 attention heads and a feedforward dimension of $4d = 8192$, with dropout applied at a rate of 0.

Output Head. The model outputs logits over the vocabulary via a linear projection:

$$\hat{x}_0^i = \text{softmax}(\mathbf{W}\mathbf{H}_i + \mathbf{b})$$

These logits are used to predict the original (clean) token at each position.

Noise Corruption Process. During training, discrete noise is applied by randomly replacing tokens with <MASK>. The corruption probability at timestep t is linearly scaled:

$$\Pr[x_t^i = \texttt{<MASK>} \mid x_0^i] = \frac{t}{T-1}$$

This corruption is applied independently across positions.

Denoising Objective. The model is trained to predict the original sequence x_0 from a corrupted input x_t. The loss is a masked cross-entropy:

$$\mathcal{L} = \frac{1}{\sum_i m_i} \sum_i m_i \cdot \text{CE}(\hat{x}_0^i, x_0^i)$$

where $m_i = 1$ if token i was masked at timestep t and 0 otherwise. Padding tokens are ignored in the loss using label value -100.

Sampling Procedure. Generation begins from random token initialization:

$$x_T \sim \text{Uniform}(\mathcal{V})$$

For $t = T-1, T-2, \ldots, 0$, the model predicts $\hat{x}_t$ using:

$$x_t \sim \text{Multinomial}(\text{softmax}(f_\theta(x_{t+1}, t, s)/\tau))$$

Temperature τ controls generation diversity.

3.3 Configuration Parameters

The model uses the following hyperparameters:

Parameter	Value
Vocabulary size ($\lvert\mathcal{V}\rvert$)	25
Max sequence length (L)	100
Embedding dimension (d)	2048
Number of transformer layers (L_{enc})	24
Number of attention heads	16
Dropout rate	0
Number of diffusion steps (T)	1000
Number of species labels	3

Table 1: Configuration parameters used in the discrete diffusion model.

3.4 Training

We train the model from scratch using the following procedure:

- **Objective:** Masked cross-entropy loss over corrupted positions.
- **Noise Sampling:** For each training example, timestep $t \sim \mathcal{U}\{0, T-1\}$ is sampled uniformly.
- **Optimization:** The model is trained using the AdamW optimizer.
- **Batch Size:** 300 sequences per batch.
- **Gradient Masking:** Padding positions are excluded from loss and gradient updates using a label ignore index of -100.
- **Initialization:** All random seeds are fixed for reproducibility.
- **Scheduler:** Learning rate is adjusted using a one-cycle policy.
- **Learning Rate:** Base rate is set to 1e-3, increased to five times the base during the first 20% of steps, then decayed to 1% of the peak by the end of training.

The model was trained using a custom implementation of discrete diffusion for protein sequence generation. Training was performed with the AdamW optimizer and a batch size of 300. Our training pipeline incorporated mixed-precision training through PyTorch's automatic mixed precision module with gradient scaling to maintain numerical stability while reducing computational requirements. The diffusion process employed timestep-dependent noise addition, where amino acid tokens were progressively masked according to a noise probability determined by the timestep, with careful handling of padding tokens during noise application. Cross-entropy loss was calculated only on the noised positions, excluding padding tokens from the loss computation. Training dynamics were

monitored through token frequency analysis and noise fraction statistics, with special debugging for mask token distribution. Species-specific conditioning was implemented through dedicated embedding layers, allowing the model to learn taxonomic variations simultaneously across the different species in our dataset. Validation was performed after each epoch to track performance on the held-out dataset.

3.5 Metrics

3.5.1 Predicted Local Distance Difference Test (pLDDT)

The predicted Local Distance Difference Test (pLDDT) is a per-residue confidence score produced by AlphaFold. It estimates the expected deviation in Å of each residue's position compared to the native structure.

$$\text{pLDDT}_i = 100 \times (1 - \mathbb{E}[\delta_i]) \tag{1}$$

where δ_i is the predicted deviation of the i-th residue. Scores range from 0 to 100, with higher scores indicating higher confidence in the local prediction.

3.5.2 Predicted TM-score (pTM)

The predicted Template Modeling score (pTM) evaluates the global structural similarity between a predicted model and the true structure. It is calculated as:

$$\text{pTM} = \mathbb{E}\left[\frac{1}{L}\sum_{i<j}\frac{1}{1+\left(\frac{d_{ij}-d_{ij}^{\text{true}}}{d_0(L)}\right)^2}\right] \tag{2}$$

where L is the sequence length, d_{ij} is the predicted distance between residues i and j, d_{ij}^{true} is the true distance, and $d_0(L)$ is a normalization term depending on L. The pTM score ranges from 0 to 1.

3.5.3 Entropy

In the context of sequence analysis, Shannon entropy measures the variability at a sequence position across a multiple sequence alignment (MSA):

$$H = -\sum_{a\in\mathcal{A}} p(a)\log_2 p(a) \tag{3}$$

where $\mathcal{A}$ is the set of amino acids and $p(a)$ is the frequency of amino acid a at a given alignment position. High entropy indicates high sequence variability.

3.5.4 Edit Distance

Edit distance (Levenshtein distance) is the minimum number of insertions, deletions, and substitutions required to convert one sequence into another.

$$D(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min\begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+\delta \end{cases} & \text{otherwise} \end{cases} \tag{4}$$

where $\delta = 0$ if the i-th and j-th characters are the same, and 1 otherwise.

3.5.5 Jaccard Similarity

Jaccard similarity is a metric used to compare the similarity between two sets. It is defined as the size of the intersection divided by the size of the union of the sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

where A and B are sets (e.g., sets of amino acids, sequence features, or annotations). Jaccard similarity ranges from 0 to 1, with 1 indicating identical sets and 0 indicating no overlap. It is commonly used in clustering, classification, and sequence analysis tasks where binary or set-based features are compared.

3.5.6 MMseqs2 Clustering

MMseqs2 clustering groups sequences by sequence identity using fast pre-filtering and alignment algorithms. Two sequences s_1 and s_2 belong to the same cluster if:

$$\text{identity}(s_1, s_2) \geq \theta \quad \text{and} \quad \text{coverage}(s_1, s_2) \geq \gamma \tag{6}$$

where θ is the identity threshold (e.g., 90%) and γ is the alignment coverage threshold (e.g., 80%). The clustering algorithm typically uses greedy or connected component methods to assign sequences to clusters.

4 Results

To evaluate the performance of our generative discrete diffusion model, we benchmarked it against EvoDiff, a continuous latent-space diffusion model trained on similar sequence databases [7]. We evaluated both models using a suite of metrics capturing sequence diversity, structural plausibility, and biological relevance: Shannon Entropy, Jaccard Similarity, mean pLDDT, predicted TM-score (pTM), predicted aligned error (pAE), and interspecies Shannon Entropy. These metrics (apart from interspecies Shannon Entropy) were calculated as an average of values from intraspecies sequence groups of 100 sequences per class, each with a length of 50 amino acids. As shown in Table 1, our model achieves a Shannon Entropy of 4.036, slightly higher than EvoDiff (3.967), indicating greater sequence diversity 2. While Jaccard Similarity is matched at 0.005, our model maintains comparable structural quality as measured by mean pLDDT (51.679 vs. 52.669) and pTM (0.201 vs. 0.229) 2, with a marginal increase in pAE (16.722 vs. 15.929) 2. Interspecies Shannon Entropy in our model is also a fair bit higher "regular" (intraspecies) showing that the generated sequences for each class are much more similar to each other than those in other classes. These results demonstrate that our model generates diverse sequences without significantly compromising structural confidence and that class based generation creates distinct groups of protein sequences.

To evaluate the architectural decisions of our model, we performed targeted comparisons with EvoDiff's design choices. Through previous testing of generative models with different tokenization schemes (primarily residue level and BPE) we decided to mirror EvoDiff's residue-level tokenization [7]. While both models utilize the same tokenization, our model introduces species-specific class tokens (0 for fruit fly, 1 for human, 2 for rat) at the sequence level. This explicit conditioning allows the model to better learn interspecies constraints, which we observed to improve taxonomic separability in generated sequences—confirmed by increased Shannon Entropy across species groups 2. In terms of training objectives, our model uses a masked cross-entropy loss, which enables direct supervision on unmasked residues while avoiding unnecessary gradient noise from padded or masked positions. In contrast, EvoDiff relies on a variational lower bound loss (L_{vb}) on the negative log-likelihood, combined with a cross-entropy loss (L_{ce}) on the reverse process [7]. While this approach provides a theoretically grounded objective, it introduces additional complexity and often requires careful tuning of loss weights to balance learning. We found that our simpler loss function achieved comparable performance on structure metrics (pLDDT and pTM), while slightly outperforming EvoDiff in Shannon entropy (4.036 vs. 3.967) 2, suggesting improved diversity. These results demonstrate that a discrete formulation with a lightweight, interpretable loss and clear

class-conditioning can simplify training while maintaining biological relevance and output diversity.

The improved performance of our model stems from its discrete formulation, which better aligns with the categorical nature of protein sequences, and from the use of included class tokens. The slight trade-off in structure metrics (pLDDT and pTM) 2is offset by higher entropy and class separability, indicating our model produces more diverse and taxonomically coherent proteins—an important quality for applications in synthetic biology, comparative genomics, and species-specific therapeutic design. These findings support our biological motivation: generating sequences that not only fold plausibly, but also reflect the evolutionary signatures and functional constraints of their source organism.

5 Conclusion

In this work, we presented SAGEdiff, a discrete diffusion-based protein sequence generation model that incorporates species-specific conditioning through class tokens and uses a masked cross-entropy loss for training. Our model demonstrates strong performance relative to EvoDiff, achieving higher Shannon Entropy (4.036 vs. 3.967) while maintaining comparable structure prediction scores across metrics like pLDDT, pTM, and pAE 2[7]. These results suggest our approach yields sequences that are not only more diverse but also better clustered by species, aligning with our biological goal of generating proteins that reflect taxon-specific constraints. The simplicity of our discrete framework and class conditioning contributes to this improved taxonomic coherence, making it a compelling model for species-aware sequence design in synthetic biology and comparative genomics.

Despite the promising results, our model has notable limitations. It was trained on a relatively small dataset, with constrained access to computational resources and limited training time. These factors likely capped the structural accuracy of our generated proteins, as seen in slightly lower pLDDT and pTM scores compared to EvoDiff 2[7]. Specifically in terms of training time, our model was extremely time efficient, only requiring a short amount of time to train to produce comparable results to state-of-the-art models. Future work could address these limitations by scaling training to larger datasets, inlcuding significantly more epochs during training, and testing generalization to unseen species. With these improvements, our model could serve as a lightweight, modular framework for generating functional and species-specific protein sequences.

6 Figures

6.1 Metric Comparison Table

Table 2: Combined Performance Metrics Comparison

Metric	**SAGEdiff**	**EvoDiff**
Shannon Entropy	4.036	3.967
Jaccard Similarity	0.005	0.005
Mean pLDDT	51.679	52.669
pTM	0.201	0.229
pAE	16.722	15.929
Interspecies Shannon Entropy	4.135	—-

6.2 Model Architecture

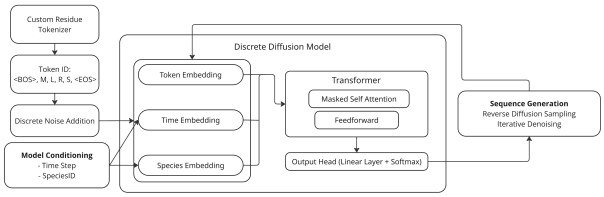

Figure 1: Model architecture of our sequence generating model. The diagram shows the tokenization, Discrete Diffusion architecture, and post processing steps that led to novel protein sequences.

6.3 Generation and Evaluation Workflow

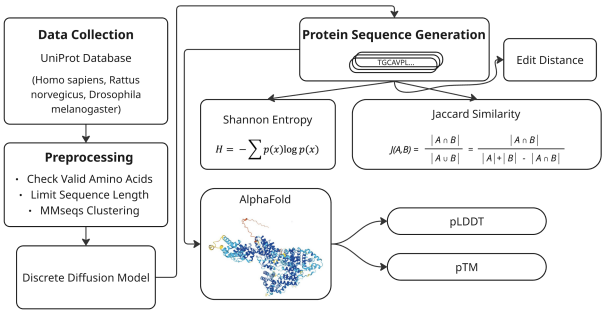

Figure 2: Graphic of evaluation workflow through the models. The diagram shows how data was curated, the model generated sequences, and metric comparison were calculated compared to based on state of the art comparators.

References

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Title of the chapter or section. In *Molecular Biology of the Cell*. Garland Science, 4 edition, 2002.

[2] Tom Sercu Siddharth Goyal a 1 Zeming Lin Jason Liu Demi Guo Myle Ott C Lawrence Zitnick Jerry Ma Rob Fergus Alexander Rives, Joshua Meier. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2021.

[3] Eli N. Weinstein Nikhil Naik Ali Madani Erik Nijkamp, Jeffrey Ruffolo. Progen2: Exploring the boundaries of protein language models. *arxiv*, 2022.

[4] Rianne van den Berg Sarah Alamdari James Y. Zou Alex X. Lu Ava P. Amini Kevin E. Wu, Kevin K. Yang. Protein structure generation via folding diffusion. *Nature Communications*, 15, 2024.

[5] Birte Höcker Noelia Ferruz, Steffen Schmidt. Protgpt2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(4348), 2022.

[6] Robert Verkuil Joshua Meier John F. Canny Pieter Abbeel Tom Sercu Alexander Rives Roshan Rao, Jason Liu. Msa transformer. *bioRxiv*, 2021.

[7] Rianne van den Berg Alex X. Lu Nicolo Fusi Ava P. Amini Kevin K. Yang Sarah Alamdari, Nitya Thakkar. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023.

[8] Gilbert Giri Francisco F. Cavazos Jr Yue Hu Jernej Murn Maria M. Aleman Christopher B. Burge Daniel Dominguez Sarah E. Harris, Maria S. Alexis. Understanding species-specific and conserved rna-protein interactions in vivo and in vitro. *Nucleic Acids Research*, 15, 2024.

[9] Jianhui Li Mark Hochstrasser Aashiq H Kachroo Sarmin Sultana, Mudabir Abdullah. Species-specific protein–protein interactions govern the humanization of the 20s proteasome in yeast. *Genetics*, 225, 2023.

[10] Bin Liu Shuyuan Guo Yi-Xin Huo Yan Xia, Xiaowen Du. Species-specific design of artificial promoters by transfer-learning based generative deep-learning model. *Nucleic Acids Research*, 52, 2024.

[11] Hao Zhou Lantao Yu Mingxuan Wang Lei Li Yuxuan Song, Ning Miao. Improving maximum likelihood training for text generation with density ratio estimation. *arxiv*, 2007.

[12] Liuzhenghao Lv Bin Lin Junwu Zhang Calvin Yu-Chian Chen Li Yuan Yonghong Tian Zongying Lin, Hao Li. Taxdiff: Taxonomic-guided diffusion model for protein sequence generation. *Arxiv*, 2024.