SCIENCE CHINA
Information Sciences

• LETTER •

April 2025, Vol. 68, Iss. 4, 149101:1–149101:2
https://doi.org/10.1007/s11432-024-4296-6

TaxDiff: taxonomic-guided diffusion model for protein sequence generation

Zongying LIN[1,3], Hao LI[1,2,3], Liuzhenghao LV[1,3], Yu WANG[1,3], Bin LIN[1], Junwu ZHANG[1], Zijun CHEN[1,3], Calvin Yu-Chian CHEN[1,2,3,4], Li YUAN[1,2,3*] & Yonghong TIAN[1,2,3*]

[1] *School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China*
[2] *Peng Cheng Laboratory, Shenzhen 518000, China*
[3] *AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen 518055, China*
[4] *School of Chemical Biology and Biotechnology, Peking University, Shenzhen 518055, China*

Received 28 May 2024/Revised 28 October 2024/Accepted 8 January 2025/Published online 14 February 2025

Citation Lin Z Y, Li H, Lv L Z H, et al. TaxDiff: taxonomic-guided diffusion model for protein sequence generation. Sci China Inf Sci, 2025, 68(4): 149101, https://doi.org/10.1007/s11432-024-4296-6

Protein design aims to generate protein variants with targeted biological functions, which is significant in multiple biological areas, including enzyme reaction catalysis, vaccine design, and fluorescence intensity. Protein design contains two paradigms: sequence generation and structure generation. Recently, EvoDiff [1] proposed a universal designing paradigm, combining structure and sequence generation using the diffusion framework, which improves the protein design efficiency.

Despite the success of EvoDiff and other sequences generative models [2] that are widely used for designing biologically plausible protein sequences, these protein design models are limited to unconditional generation. In practical scenes, biological researchers need to filter the randomly generated proteins to fulfill the desired criteria, which is time-consuming and labor-intensive. Thus, unconditional protein generation, which cannot control protein properties, is still some way from practical application.

To address the uncontrollable challenge, we propose a taxonomic-guided diffusion model, TaxDiff, to design target proteins with the biological-species control signals, as shown in Figure 1(a). To the best of our knowledge, our TaxDiff is the first controllable protein generation model utilizing guidance from taxonomies. TaxDiff inserts the taxonomic control features into each denoise transformer block to achieve controllable generation. For fine-grained protein sequence generation, we also propose the patchify attention mechanism to capture the protein feature on global and local scales. Our TaxDiff follows the protein design paradigm of EvoDiff. Thus, TaxDiff is capable of generating both protein sequences and structures in a shared space. The details of our model description are provided in Appendix A.

Methods. In recognition of the diversity of amino acids, we introduce an additional dimension D to enrich features at the amino acid level. Through the encoder, feature-augmented x can thus be represented as $x \in \mathbb{R}^{L\times D}$. In the denoise transformer block, three different types of inputs are processed: the data x_T formed by the forward process in diffusion models that gradually adds Gaussian noise, the timestep t, and the protein taxonomic identifier y (tax-id). The timestep t and tax-id y are individually embedded, resulting in two distinct conditional tokens that are concatenated with the x_T.

Our TaxDiff model encodes tax-id y as a condition for controllable generation. The reverse process is formalized as $p_\theta(x_{t-1}|x_t, y)$, with both means μ_θ and variances Σ_θ being conditional by y. In this way, the taxonomic-guided encourages the sampling procedure towards maximizing the conditional log-likelihood $\log p(y|x)$. Score functions guides diffusion model sample x with an increased probability $p(x|y)$ by adjusting noise prediction:

$$\begin{aligned}\hat{\mu}_\theta(x_t, y) &= \mu_\theta(x_t, \emptyset) + s\cdot \nabla_x \log p(x|y)\\ &\propto \mu_\theta(x_t, \emptyset) + s\cdot(\mu_\theta(x_t, y) - \mu_\theta(x_t, \emptyset)),\end{aligned} \tag{1}$$

where the scalar $s > 1$ determines the scale of guidance.

The classification of taxonomic identifiers (tax-id) y within UniProt [3] is highly detailed, so we reclassified the ninth layer of original taxonomic lineages, which corresponds to the family and species levels [4]. This reorganization effectively condenses the categories to a total of 23427. The distribution of different taxonomic lineages is shown in Figure 1(c).

To enhance the model's capacity to extract global features from protein sequences while simultaneously preserving the salient of amino-acid local features, we have implemented patchify attention mechanism upon the input protein sequences. In the experiment, we explored the adoption of five different patchify attention combinations to find the optimal combination. This procedure is delineated in Figure 1(b).

[*] Corresponding author (email: yuanli-ece@pku.edu.cn, yhtian@pku.edu.cn)

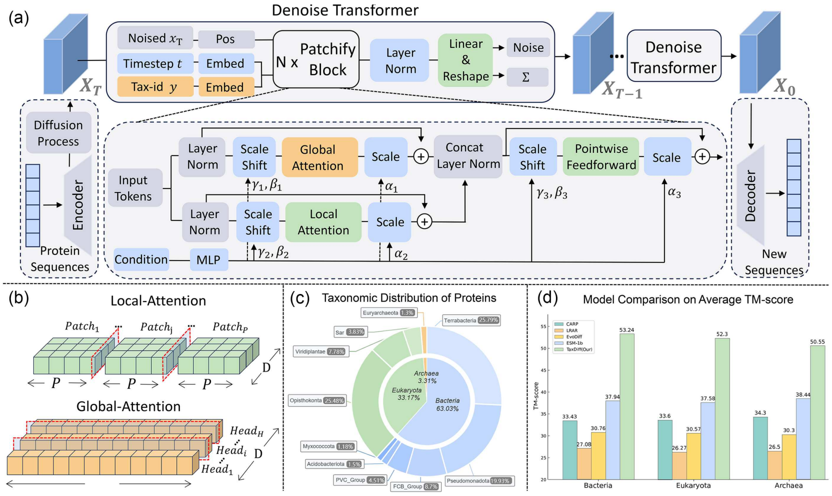

Figure 1 (Color online) (a) This framework delineates how we fuse the denoise transformer into the denoising process of the diffusion model; (b) the division between Local-Attention and Global-Attention; (c) the distribution of different taxonomic lineages, showcasing the top 10 categories; (d) the performance of different models in terms of TM-score under taxonomic-guided.

The attention mechanism operates at both the global and local levels to capture relationships within protein sequences. At the global level, Global-Attention models the intricate interactions between different amino acids using multiple attention heads. For each head i, with a total of H heads, we compute the queries, keys, and values as $Q_i = x \times W_i^Q$, $K_i = x \times W_i^K$ and $V_i = x \times W_i^V$ as linear transformations of the input matrix $x \in \mathbb{R}^{L \times D}$. The scaled dot-product attention mechanism is used to calculate the attention weights:

$$\text{Attention}_{\text{head}}(Q_j, K_j, V_j) = \text{softmax}\left(\frac{Q_j K_j^{\text{T}}}{\sqrt{d_k}}\right) V_j, \quad (2)$$

where d_k is the key's dimension. The final Global-Attention output is obtained by concatenating the outputs of all heads and applying a linear transformation:

$$\text{Global-Attention}(x) = \text{concat}(\text{head}_1, \ldots, \text{head}_H) W^O, \quad (3)$$

where W^O is the output weight matrix.

At the local level, Local-Attention divides the sequence into patches of size P. For each patch j, the queries, keys, values and attention are similarly computed like Global-Attention. The final Local-Attention output is derived by concatenating the outputs of all patches:

$$\text{Local-Attention}(x) = \text{concat}(\text{Patch}_1, \ldots, \text{Patch}_P). \quad (4)$$

Thus, both attention mechanisms capture different levels of relationships within the sequence, offering a comprehensive model for protein interaction analysis. The details of our method are provided in Appendixes B and C.

Experiment. We carry out extensive experiments to evaluate TaxDiff across multiple benchmarks, encompassing both unconditional and taxonomic-guided controllable protein sequence generation. In unconditional protein sequence generation, the sequence-based TaxDiff demonstrated comparable structural modeling capabilities to structure-based protein generation models, even significantly outperforming them in common metrics such as TM-score, RMSD, and Fident, with improvements of 11.93%, 5.4552, and 7.13% respectively. In taxonomic-guided controllable protein sequence generation, the pLDDT scores [5] far surpassed other sequence generation models, nearing the levels of natural protein sequences. The result of the taxonomic-guided controllable protein on average TM-score is shown in Figure 1(d). Empirical studies also indicated that due to the patchify attention mechanism, the efficiency of TaxDiff was markedly enhanced, requiring only 1/4 of the time generated by other models. All experimental results demonstrate that TaxDiff possesses superior capabilities in exploring protein sequence space and producing structurally coherent proteins. More details of our experiment are provided in Appendix D.

Conclusion. While current models can only perform unconditional sequence generation, TaxDiff overcomes this limitation by learning taxonomic-guided over protein sequences space. By combining the Global-Attention to sequences and Local-Attention to amino acids, TaxDiff can effectively generate sequences that are structurally reliable and consistent in sequence. Furthermore, TaxDiff requires only a quarter of the time needed by other diffusion-based models to achieve superior performance. Various experimental results demonstrate its significantly better ability for modeling natural protein sequences.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 62425101, 62088102, 62202014) and Shenzhen Science and Technology Program (Grant No. KQTD20240729102051063).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Alamdari S, Thakkar N, van den Berg R, et al. Protein generation with evolutionary diffusion: sequence is all you need. BioRxiv, 2023. doi: 10.1101/2023.09.11.556673
2 Repecka D, Jauniskis V, Karpus L, et al. Expanding functional protein sequence spaces using generative adversarial networks. Nat Mach Intell, 2021, 3: 324–333
3 The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res, 2017, 45: D158–D169
4 Luby J J. Taxonomic classification and brief history. In: Apples: Botany, Production and Uses. Wallingford UK: Cabi Publishing, 2003. 1–14
5 Wu R D, Ding F, Wang R, et al. High-resolution de novo structure prediction from primary sequence. BioRxiv, 2022. doi: 10.1101/2022.07.21.500999