Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29612

Comparison Between Transformer-based Protein Language Models and Traditional Text-based Language Models from a Computer Science Perspective

Chenglin Xu

School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
23722029@bjtu.edu.cn

Abstract. The Transformer architecture has transformed natural language processing (NLP) by enabling efficient sequence modeling through self-attention and embedding techniques. However, its ability to adapt to domain-specific data, such as protein sequences, introduces both unique computational challenges and opportunities. As the sequence space increases, the understanding of architectural differences is crucial for improving model efficiency and generalization. This study aims to investigate the fundamental differences between protein language models (PLMs) and traditional text-based language models (LLMs), highlighting their modeling principles, embedding structures, and attention mechanisms. By reviewing and analyzing the relevant literature, the methods adopted by PLMs and LLMs are explored, emphasizing their unique features. The results reveal that PLMs, with their sparse attention mechanism and highly linearly separable embeddings, demonstrate superior capabilities in processing long sequences for pattern extraction, while language models focus on semantic dependencies. These differences reveal the potential for cross-domain optimization, helping to improve the application of Transformers in sequence analysis and generation.

Keywords: Natural Language Processing (NLP), Protein Language Models (PLMs), Large Language Models, Transformer

1. Introduction

The function of a protein is dictated by its amino acid sequence, and designing new proteins with specific functions usually relies on computational modeling to predict their structures [1]. However, the vast number of possible sequences makes de novo protein design highly challenging. To address this issue, protein sequences are treated as “text,” and natural language processing (NLP) methods are applied to learn sequence patterns, thus enabling sequence generation and optimization to assist functional design. While protein language models (PLMs) excel in protein sequence analysis and design, their differences from text-based language models in modeling principles, hence embedding representations, and attention mechanisms remain not fully understood [2,3]. Despite partial transfer of NLP pretraining strategies to protein sequences, PLMs capture biological information, structural features, and functional patterns in ways that differ from how text-based language models capture text semantics. As such, this study compares PLMs with text-based language models. Specifically, it

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29612

reviews the principles, pretraining methods, and applications of existing protein language models, including Evolutionary Scale Modeling 2 (ESM2), Protein T5 (ProtT5), Protein GPT-2 (protGPT2), and Protein BERT (ProtBert), as well as text-based language models such as Generative Pre-trained Transformer (GPT), Bidirectional Encoder Representations from Transformers (BERT), and classic Transformers [2,3]. Therefore, the comparison highlights predictive performance, embeddings, and attention, focusing on differences in information capture, interpretability, and use. This study helps support the effective use of PLMs for protein sequence analysis and design, expands Transformer applications to more scenarios, and drives progress toward fully domain-specific intelligent models.

2. Transformer and its applications in different language models

2.1. The mechanism of Transformer in text language models

Based on an encoder-decoder structure, Transformer is a deep learning model that operates through core mechanisms such as self-attention and multi-head attention [4]. The self-attention mechanism allows the model to dynamically adjust weights according to the relationships between words in the sequence, effectively capturing long-range dependencies, especially excelling in NLP tasks. Besides, multi-head attention boosts the ability of the model to attend to diverse features and semantics by performing multiple attention calculations in parallel. This design gives Transformer advantages in sequence tasks like representation, generation, and prediction. Built on the Transformer architecture, BERT and GPT demonstrate unique features in task performance [5]. In particular, BERT adopts an encoder-only architecture that learns sequence representations by leveraging bidirectional context. Its ability to capture both left and right context makes it particularly effective in tasks like question answering and sentiment analysis, enhancing its grasp of word meanings and sentence structure [6].

This design confers upon BERT a distinct advantage in sequence representation and comprehension tasks. In contrast, GPT utilizes a decoder-only architecture, generating text in a left-to-right fashion. It focuses on sequence generation, excelling in tasks such as text generation and language modeling [7]. Due to its unidirectional design, GPT prioritizes sequence generation over understanding word relationships. Thus, BERT's bidirectional encoder and GPT's unidirectional decoder directly impact their performance in sequence modeling.

2.2. The architecture of Transformer in protein language models

In a similar way to NLP, protein language models (PLMs) depend on the Transformer architecture, which can be divided into encoder-only, decoder-only, and encoder-decoder models, depending on the specific task. The design of each architecture directly determines its effectiveness in modeling protein sequences. First, Evolutionary Scale Modeling 2 (ESM2) is an encoder-only protein large language model trained with a masked language modeling (MLM) objective, similar to BERT, and employs 65 million protein sequences for training [8]. Through large-scale unsupervised learning, it extracts structural insights from evolutionary data and processes protein sequences with multi-head self-attention and feed-forward networks to capture essential features. This enables ESM2 to excel in sequence representation and feature extraction. In contrast, ProGen is a decoder-only protein generation model, with its first version containing 1.2 billion parameters, which is increased to 6.4 billion parameters in the ProGen2 series [9,10]. Similar to decoder-only models like GPT, ProGen focuses on protein generation tasks and is trained through unsupervised learning with the goal of predicting the next amino acid. As a result, ProGen has a significant advantage in protein generation, especially for generating novel protein sequences. Additionally, ProtT5 adopts an encoder-decoder

DOI: 10.54254/2755-2721/2026.TJ29612

structure, based on Google's T5 model design. The encoder processes protein sequences to extract contextual features, while the decoder is used for generating sequences or performing downstream tasks (such as predicting variant effects) [11]. As a bidirectional encoding architecture, ProtT5 fully leverages contextual information to boost its sequence generation and task prediction capabilities.

3. Predictive performance metrics and their applicability

Perplexity (PPL) is an important metric for evaluating the predictive ability of language models, reflecting how well a model fits the probability distribution of sequences. Autoregressive models, like the GPT series, can directly compute standard PPL through cross-entropy, which indicates the model's ability to predict the next token [7]. However, for models trained with a MLM objective, such as BERT or certain PLMs like ESM-2 and ProtT5, standard PPL cannot be directly computed because they do not have a full generative probability distribution.

To overcome this issue, the pseudo-log-likelihood (PLL) method can be applied, in which the model sequentially predicts the probability of each masked token in the sequence and accumulates the log-likelihood [12]. Formally, it is defined as Equation 1.

$$\mathrm{PLL}(\mathrm{w}_1,\ldots,w_\mathrm{n})=\sum_{\mathrm{i}=1}^{\mathrm{n}}\log\mathrm{P}(w_\mathrm{i}|w_{\backslash\mathrm{i}}) \tag{1}$$

where $w_{\backslash i}$ represents the sequence with the i-th token replaced by [MASK]. Based on the PLL, the pseudo-perplexity (pseudo-PPL) can be further computed as Equation 2.

$$\mathrm{pseudo{-}PPL}=\exp\left(-\tfrac{1}{n}\,\mathrm{PLL}\left(w_1,\ldots,w_n\right)\right) \tag{2}$$

This metric is similar to standard PPL but serves as an approximation, hence making it useful for evaluating the performance of MLM models in sequence modeling tasks [8,13].

The size of the vocabulary and the choice of tokenization method greatly influence a model's capacity to identify sequence patterns. In NLP, methods such aa WordPiece or Byte-Pair Encoding (BPE) typically produce vocabularies of tens of thousands of subwords. For example, BERT has a vocabulary of approximately 30k, while GPT-2 has around 50k [6,14]. Larger vocabularies enhance the model's ability to capture rare patterns but come at the cost of higher computational complexity, whereas smaller vocabularies are computationally more efficient but may miss nuanced details. In protein sequences, each residue is typically treated as a token, resulting in a small vocabulary of around 20-33 tokens, including special symbols [8-11]. When comparing PPL across domains, it is important to focus on relative changes or normalized results rather than raw values. This is because perplexity metrics not only reflect how well a model fits the training sequence distribution but also correlate with downstream tasks such as sequence generation, classification, and prediction. In fact, better sequence pattern capture, indicated by lower PPL or pseudo-PPL, generally leads to improved performance in subsequent tasks.

4. Embedding features and their differences in protein language models

There are notable gaps in the embedding features between PLMs and text-based language models. These variations are evident in semantic granularity, separability, scalability, and also in how they incorporate physical constraints and biological contexts.

In particular, embeddings map tokens like words in language models or amino acids in protein models) as dense vectors, thus capturing both their semantic meaning and contextual information. In

DOI: 10.54254/2755-2721/2026.TJ29612

text-based language models, embeddings represent semantic and syntactic information, enabling direct application to tasks such as text classification, sentiment analysis, and question answering. In contrast, embeddings in PLMs encode evolutionary information, secondary and tertiary structure cues, and functional site signals, resulting in different application contexts and task objectives [15,16].

When comparing the embeddings of these two types of models, the first point to note is their differences in semantic granularity. Embeddings in text-based language models primarily focus on encoding grammatical and semantic information, enabling them to be directly applied to a variety of text-based tasks. However, embeddings in PLMs represent biological concepts, such as domains or functional regions, thus establishing more intricate semantic relationships at the biological level. Prior studies have shown that protein sequences in the embedding space often form clusters based on families, structural categories, or functional labels, while such clustering patterns are less common in text model embeddings [16]. Moreover, the differences in linear separability are worth highlighting. In PLMs, linear probing approaches can effectively recover complex biological characteristics from embeddings, hence indicating a stronger linear separability in protein model embeddings [11,15]. In contrast, linear separability in text-based language models is primarily suited for simpler tasks like sentiment analysis, whereas more complex reasoning tasks still require nonlinear decoding layers. In addition, as the scale of the model increases, embeddings exhibit new "emergent capabilities." For instance, when ESM-2's parameter scale reaches billions, it is able to capture more detailed sequence representations, improving the accuracy of downstream predictions [8]. Similarly, ProGen2 shows improved protein design abilities as its parameter scale increases, fueled by the physical foundation and constraints of PLMs, whereas text-based models rely on statistical patterns, ignoring physical structural constraints [10].

Furthermore, PLMs and text-based language models differ in scalability and interpretability. As the model scale grows, protein models like ESM-2 and ProGen2 capture more detailed sequence information, supporting more accurate predictions [8,10]. In contrast, text-based language models focus on modeling semantic richness and long-range dependencies. For example, GPT-3/4 models, when scaled to hundreds of billions of parameters, demonstrate few-shot learning and chain-of-thought reasoning abilities [17,18]. However, these abilities are captured primarily through abstract patterns at the language level, rather than inferred at the physical structure level. PLMs have a physical foundation and natural constraints, with embeddings that carry real physical meaning and are governed by physical laws. In contrast, text-based models lack such "natural physical constraints." The embeddings of large-scale LLMs focus on internalizing statistical patterns, rather than geometric laws.

5. The attention mechanism of natural language and protein language models

The attention mechanism improves model performance by assigning weights to different parts of the input data, dynamically focusing on the parts most relevant to the task [4]. It enables the model to selectively focus on important information by calculating the correlation (weight) between each element in the input sequence and the target output. While attention plays a central role in both text language modeling and protein sequence modeling, its application varies because of variations in sequence length and structural traits

In terms of context length, text-based language models like the BERT and GPT series typically process shorter text sequences, with input lengths generally ranging from 512 to 2048 tokens. Their attention mechanisms are mainly used to model grammatical and semantic relationships, such as subject-verb-object dependencies or cross-sentence reasoning [6,17]. In contrast, protein sequences

DOI: 10.54254/2755-2721/2026.TJ29612

are usually much longer than natural language sentences, averaging 300 to 1000 residues, with some even exceeding 2000. Handling these long sequences poses a significant challenge for both the platform running the model and the model itself. Thus, PLMs commonly utilize sparse or local attention mechanisms to minimize quadratic complexity and maintain local dependencies [19,20].

Moreover, attention modes differ. The attention layers of text-based language models have been experimentally shown to exhibit a hierarchical function: lower layers primarily capture lexical and local dependencies, while higher layers gradually encode semantics and discourse relations [20]. In contrast, attention in PLMs captures both local dependencies along the sequence and more complex global patterns specific to the domain. Previous research has shown that the attention mechanism in protein language models can exhibit structural awareness beyond linear token dependencies. For example, certain attention heads correspond to higher-order sequential patterns, indicating that PLMs capture both local context and broader, domain-specific regularities [11,19].

In terms of inductive bias and interpretability, attention in text-based language models is often used to explain linguistic structures, like pronoun resolution preferences in BERT [21]. In contrast, the biological interpretability of attention in protein language models is much more prominent. It often corresponds to both local dependencies and higher-order sequence structures, highlighting important sequence elements. For example, larger PLMs like ESM-2 exhibit attention distributions that capture increasingly rich structural information as the model scale grows [22]. Thus, text-based and PLMs both rely on the Transformer's self-attention to capture global dependencies.

6. Conclusion

The study reveals that both PLMs and text-based large language models are built upon a common foundation in the Transformer architecture. However, there are many differences in their objectives, representations, and interpretability. Text-based language models focus on capturing semantic and syntactic patterns in human language, but PLMs encode the evolutionary, structural, and functional properties of proteins according to physical and biological laws. These differences are evident in tokenization approaches, evaluation criteria such as perplexity and semantic embeddings, and the biological plausibility of attention mechanisms. In contrast to text-based language models that learn abstract statistical patterns, PLMs' embeddings and attention mechanisms typically align with real physical constraints. The clarification of differences reveals application pathways for Transformer models across diverse fields and steers the creation of more biology-oriented models. Furthermore, comparative analysis of embeddings, tokenization, evaluation metrics, and attention mechanisms uncovers challenges and opportunities in protein modeling, with insights being transferable to other atypical language models. Moving forward, the integration of NLP advancements with Transformer adaptations for sequential data is anticipated to drive progress in computational sequence modeling, embedding optimization, and attention-based architectures across various domains, thereby fueling innovation in data-driven computing.

References

[1] Zhao, H. (2013). Synthetic biology: tools and applications. Academic Press, 13-18.
[2] Heinzinger, M., & Rost, B. (2025). Teaching AI to speak protein. Current Opinion in Structural Biology, 91, 102986.
[3] Wang, X., Luo, J., Cai, X., et al. (2025). DeepHVI: A multimodal deep learning framework for predicting human-virus protein-protein interactions using protein language models. Biosafety and Health, 7(4), 257-266.
[4] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 5998-6008.

DOI: 10.54254/2755-2721/2026.TJ29612

[5] Wang, L., Li, X., Zhang, H., et al. (2025). A comprehensive review of protein language models. arXiv. https: //arxiv.org/abs/2502.06881v1
[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 4171-4186.
[7] Radford, A., Narasimhan, K., Salimans, T., et al. (2018). Improving language understanding by generative pre-training. https: //cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[8] Lin, Z., Akin, H., Rao, R., et al. (2023). Language models of protein sequences at the scale of evolution enable accurate structure prediction. Science, 379(6637), 1123-1130.
[9] Madani, A., McCann, B., Naik, N., et al. (2020). ProGen: Language modeling for protein generation. bioRxiv. doi: 10.1101/2020.03.07.982272.
[10] Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., et al. (2023). ProGen2: exploring the boundaries of protein language models. Cell systems, 14(11), 968-978.
[11] Elnaggar, A., Heinzinger, M., Dallago, C., et al. (2020). ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44, 7112-7127.
[12] Wang, A., Singh, A., Michael, J., et al. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 7th International Conference on Learning Representations (ICLR). Available at: https: //arxiv.org/abs/1804.07461 doi: 10.48550/arXiv.1804.07461.
[13] Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked language model scoring. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2699-2712.
[14] Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
[15] Rao, R., Bhattacharya, N., Thomas, N., et al. (2019). Evaluating protein transfer learning with TAPE. Advances in neural information processing systems, 32.
[16] Rives, A., Meier, J., Sercu, T., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS, 118(15).
[17] Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS), 33, 1877-1901.
[18] Achiam, J., Adler, S., Agarwal, S., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv: 2303.08774.
[19] Rao, R., Liu, J., Verkuil, R., et al. (2021). MSA Transformer: Modeling protein sequences with evolutionary data. Proceedings of the 38th International Conference on Machine Learning (ICML).
[20] Zaheer, M., Guruganesh, G., Dubey, K. A., et al. (2020). Big Bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33, pp. 17283-17297.
[21] Clark, K., Khandelwal, U., Levy, O., et al. (2019). What does BERT look at? An analysis of BERT's attention. In Proceedings of the 2019 ACL Workshop BlackboxNLP. Florence, Italy: ACL, 276-286.
[22] Lin, Z., Akin, H., Rao, R., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130.