Multimodal Learning of Biological Language Models

Alireza Noroozi Sahand Hassanizorgabad Mustafa Serhat Aydın

Abstract

Building on BioLangFusion, which demonstrated that simple fusion of pretrained omics models can outperform unimodal approaches, we propose to reimplement BioLangFusion in PyTorch and extend it toward multimodal transfer learning. Our approach will integrate pretrained sequence models (Nucleotide Transformer, RNA-FM, and ESM-2) through fusion mechanisms. We aim to deliver multimodal molecular representations and demonstrate their utility in predicting sequence-level properties such as RNA stability, protein solubility, and expression efficiency. Also in addition to this implantation we are going to add another modality as texts into our embedding presentations and also using parameter-efficients training algorithms (LoRA) for better performance.

1. Introduction

Recent advances in biological foundation models have enabled powerful single-modality representations of DNA, RNA, and protein sequences (Dalla-Torre et al., 2022; Chen et al., 2022; Yazdani et al., 2024). However, biological processes operate as an interconnected cascade across these molecular layers (Liu et al., 2025a). We aim to investigate deep learning approaches that can integrate different types of sequential biological data to determine whether multimodal modeling yields improved performance on various molecular property prediction tasks. Deep-learning applications to biological sequences differ from natural-language tasks, since it is often useful to introduce biological priors (Hinton & Salakhutdinov, 2006); in this case, we aim to mimic the *central dogma* of biology while integrating different modalities. We intend to build multimodal molecular representations and demonstrate their utility in predicting sequence-level properties such as RNA stability, protein solubility, and expression efficiency (Liu et al., 2025a). Following this implementation, we plan to explore how such embeddings can be combined with other biological data (e.g., gene expression, medical imaging) to predict outcomes beyond sequence-level tasks (Liu et al., 2025a; Garau-Luis et al., 2024). To achieve this, we will implement the approach presented in *BioLangFusion* (Mollaysa et al., 2025) and extend it with additional modalities.

2. Related Work

2.1. Biological Language Models

Foundational models such as DNABERT (Ji et al., 2021), the Nucleotide Transformer (Dalla-Torre et al., 2022), RNA-FM (Chen et al., 2022), and ESM-C (ESM Team, 2024) have established self-supervised frameworks for DNA, RNA, and protein sequences. These models capture modality-specific patterns—regulatory motifs, structural stability, and protein function—but lack cross-modal awareness (Prakash et al., 2024).

2.2. Multimodal Fusion in Biology

BioLangFusion (Mollaysa et al., 2025) introduced biologically motivated alignment and fusion methods—concatenation, entropy-regularized attention, and cross-modal attention—that integrate embeddings without retraining base models. Meanwhile, *Ceviche* (Garau-Luis et al., 2024) and *MOFS* (Liu et al., 2025b) demonstrated that multimodal integration across omics and imaging modalities yields more robust biological insights.

2.3. Adaptation strategies for knowledge transfer

We freeze the language model (LM), extract embeddings from the last layer, and train a downstream head to map these embeddings to task-specific labels. We use TextCNN(Cai & Xia, 2015) head for all experiments. This approach allows us to assess the transferability of learned representations without modifying the original backbone LMs.

2.4. Supervised fine-tuning

We also explore supervised fine-tuning of the bio-LMs on each labeled dataset. We evaluate only the parameter-efficient finetuning using Low-Rank Adaptation (Hu et al., 2022) method due to small access to resources.

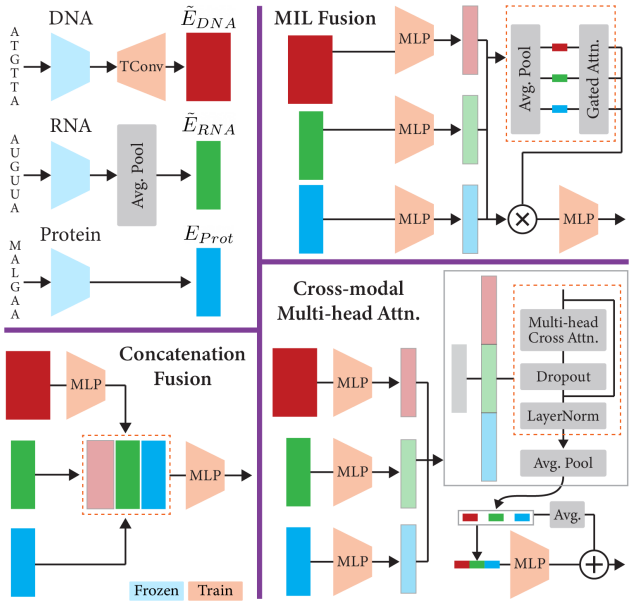

Figure 1. Model architectures for the three BioLangFusion fusion strategies. Codon-aware concatenation: DNA embeddings are upsampled via transposed convolution and mRNA embeddings downsampled via non-overlapping mean pooling to align with the protein framing; the three aligned embeddings are then concatenated at each codon position and projected before entering the TextCNN head. Entropyregularized gated attention (MIL): After alignment, each modality is projected to a shared latent space; a gated attention mechanism with an entropy regularizer computes modality weights per sequence, producing a weighted sum fused embedding that feeds into the TextCNN head. Cross-modal multi-head attention: All modality embeddings are projected, concatenated along the temporal axis to form a joint context, and each modality attends to this context via multi-head attention; the updated embeddings are merged with residual averaging and layer normalization before the TextCNN head.

3. The Approach

3.1. Model Overview

We will reimplement *BioLangFusion*'s core modules (Mollaysa et al., 2025) in PyTorch, focusing on three fusion strategies:

- **Codon-Aligned Concatenation:** direct embedding fusion after codon-level alignment.
- **Entropy-Regularized MIL Attention:** modality-weighted pooling that learns task-specific attention scores (Garau-Luis et al., 2024).
- **Cross-Modal Multi-Head Attention:** capturing token-level dependencies between modalities (Prakash et al., 2024).

Each modality embedding (DNA, RNA, protein) will be projected to a shared latent space. Fusion representations are fed into a lightweight TextCNN or MLP prediction head (Cai & Xia, 2015).

3.2. Multimodal Transfer Learning Extension

Building on *MOFS* principles (Liu et al., 2025b), we will incorporate cross-task transfer learning:

- Fine-tune RNA-Protein fusion models on small datasets for expression or solubility prediction.
- Use adapter layers for modality-specific adaptation (Boshar et al., 2024).
- Explore contrastive pre-alignment between modalities using cosine similarity and codon-level matching (Diez et al., 2022).

3.3. Computational Efficiency

To ensure low GPU demand, we will use frozen pretrained embeddings (no backpropagation through large foundation models), parameter-efficient adapters (5–10M trainable parameters), mixed-precision training (fp16), and sequence truncation (< 500 tokens per sequence). The full system can be trained efficiently on a single A40 GPU.

4. Experimental Evaluation

4.1. Dataset

We will evaluate *BioLangFusion* (Mollaysa et al., 2025) on five biologically diverse datasets covering both regression and classification tasks. Each dataset consists of mRNA sequences paired with phenotype-relevant labels, spanning applications such as vaccine design, expression profiling, stability analysis, and antibody prediction (Liu et al., 2025b; Yazdani et al., 2024).

Dataset	Max Length	#mRNA (raw)	#mRNA (used)	Target	Task
CoV-Vac	81	2400	2400	Degradation	Regression
Fungal	3063	7056	3138	Expression	Regression
E. coli	3000	6348	4450	Expression	Classification
mRNA Stab.	3066	41123	23929	Stability	Regression
Ab1	1203	723	723	Expression	Regression

Table 1. Summary of the datasets used in our experiments, including the maximum input sequence length, number of available mRNA sequences before and after preprocessing, the type of molecular property being predicted, and the corresponding task type (regression or classification).

5. Work Plan

For this project we will first have to gather the datasets and implement the *BioLangFusion* modules, and train them for

the first report. After this, we want to explore the ways we can utilize these fused embeddings by integrating them with other types of data (gene expression, medical images etc.), and then test our models on datasets that are relevant to the added modality.

Activity	Deadline
Gathering Data and implement BioLang	11/30/2025
train the model	12/7/2025
Project Report	12/21/2025
Add other modality and fine-tuning code	01/4/2026
Train the other modality and fine-tuning	01/11/2026
Prepare final report and presentation	01/25/2026

Table 2. Timeline of major project activities

References

Boshar, S., Trop, E., de Almeida, B. P., and Pierrot, T. Are genomic language models all you need? exploring genomic language models on protein downstream tasks. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024. URL https://openreview.net/forum?id=B60QZ0R2Zw.

Cai, G. and Xia, B. Convolutional neural networks for multimedia sentiment analysis. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 159–167. Springer, 2015.

Chen, X., Sun, Z., Lin, Y., et al. Rna-fm: A foundation model and benchmark for functional rna embeddings. *bioRxiv*, 2022. doi: 10.1101/2022.11.17.516915.

Dalla-Torre, L., Chughtai, Z., Yan, Y., et al. Nucleotide transformer: Building and evaluating robust foundation models for dna. *bioRxiv*, 2022. doi: 10.1101/2022.11.15.516627.

Diez, M., Medina-Muñoz, S. G., Castellano, L. A., da Silva Pescador, G., Wu, Q., and Bazzini, A. A. icodon customizes gene expression based on the codon composition. *Scientific Reports*, 12:12126, 2022. doi: 10.1038/s41598-022-16448-4.

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian.

Garau-Luis, J. J., Bordes, P., Gonzalez, L., et al. Multimodal transfer learning between biological foundation models. *bioRxiv*, 2024. doi: 10.1101/2024.06.14.123456.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313: 504 – 507, 2006.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

Liu, Z., Li, S., Chen, Z., et al. Life-code: Central dogma modeling with multi-omics sequence unification. *arXiv preprint arXiv:2502.07299*, 2025a.

Liu, Z., Wu, Y., Xu, H., Wang, M., Weng, S., Pei, D., Chen, S., Wang, W., Yan, J., Cui, L., Duan, J., Zhao, Y., Wang, Z., Ma, Z., Li, R., Duan, W., Qiu, Y., Su, D., Li, S., Liu, H., Li, W., Ma, C., Yu, M., Yu, Y., and Zhang, Z. Multimodal fusion of radio-pathology and proteogenomics identify integrated glioma subtypes with prognostic and therapeutic opportunities. *Nature Communications*, 16: 3510, 2025b. doi: 10.1038/s41467-025-58675-9.

Mollaysa, A., Moskale, A., Pati, P., Mansi, T., Prakash, M., and Liao, R. Biolangfusion: Multimodal fusion of dna, mrna, and protein language models, 2025. URL https://arxiv.org/abs/2506.08936.

Prakash, M., Moskalev, A., DiMaggio, P. A., Combs, S., Mansi, T., Scheer, J., and Liao, R. Bridging biomolecular modalities for knowledge transfer in biolanguage models. In *NeurIPS 2024 Workshop on Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024. doi: 10.1101/2024.10.15.618385. URL https://openreview.net/forum?id=dicOSQVPLm. Preprint.

Yazdani, A., Roy, S., Wang, F., et al. Helm: Hierarchical embeddings for language modeling of mrna sequences. *bioRxiv*, 2024. doi: 10.1101/2024.01.05.574223.