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.

3. The Approach

3.1. Unimodal Representation Learning

We begin with unimodal sequence modeling, where each biological modality—DNA, RNA, and protein—is processed independently using pretrained foundation models. Given an input sequence $s^{(m)}$ from modality $m \in \{\text{DNA}, \text{RNA}, \text{Protein}\}$, a modality-specific encoder maps the sequence to a contextual embedding:

$$\mathbf{E}^{(m)} = f^{(m)}(s^{(m)}) \in R^{L_m \times d_m},$$

where L_m denotes the sequence length in tokens and d_m is the embedding dimension.

For DNA sequences, we use **Nucleotide Transformer v2 (100M, multi-species)**, a transformer model pretrained on large-scale genomic data across multiple organisms. This model captures regulatory and evolutionary patterns in nucleotide sequences and produces high-dimensional token-level embeddings suitable for downstream genomic prediction tasks.

For RNA sequences, we employ **RNA-FM T12**, a transformer-based RNA foundation model pretrained on diverse RNA sequences with a focus on secondary structure and functional motifs. RNA-FM outputs contextual embeddings that encode both sequence-level and structural information relevant to expression, stability, and localization tasks.

For protein sequences, we use **ESM-2 T6 (8M parameters, UR50D)**, a lightweight protein language model pretrained on the UniRef50 database. Despite its relatively small size, ESM-2 T6 provides strong residue-level representations that capture evolutionary and biochemical properties of proteins while remaining computationally efficient.

To reduce computational cost and avoid overfitting on small biological datasets, all pretrained encoders are frozen during training. Each unimodal embedding is subsequently projected into a shared latent space:

$$\mathbf{H}^{(m)} = \mathbf{E}^{(m)}\mathbf{W}^{(m)}, \quad \mathbf{W}^{(m)} \in R^{d_m \times d},$$

ensuring that all modalities lie in a common d-dimensional representation space prior to fusion.

Table 1. Pretrained unimodal encoders used in this work.

Model	Modality	# Params	d_m
RNA-FM T12	RNA	∼120M	640
Nucleotide Transformer v2	DNA	100M	4107
ESM-2 T6 UR50D	Protein	8M	320

3.2. Multimodal Learning and Fusion

Multimodal learning aims to integrate complementary biological information across DNA, RNA, and protein sequences. Due to differences in sequence length and biological framing, we first perform codon-aware alignment such that all modalities are mapped to a shared codon index $c = 1, \ldots, C$. Let $\tilde{\mathbf{H}}^{(m)} \in R^{C \times d}$ denote the aligned embeddings for modality m.

We explore three fusion strategies.

3.2.1. Codon-Aligned Concatenation

In the simplest fusion scheme, aligned embeddings from all modalities are concatenated at each codon position:

$$\mathbf{Z}_c = \left[\tilde{\mathbf{H}}_c^{(\text{DNA})} \parallel \tilde{\mathbf{H}}_c^{(\text{RNA})} \parallel \tilde{\mathbf{H}}_c^{(\text{Protein})}\right] \in R^{3d}.$$

The concatenated representation is linearly projected:

$$\mathbf{F}_c = \mathbf{Z}_c\mathbf{W}_f, \quad \mathbf{W}_f \in R^{3d \times d},$$

yielding a fused embedding sequence $\mathbf{F} \in R^{C \times d}$. This strategy preserves all modality-specific information but assumes equal importance across modalities.

3.2.2. Entropy-Regularized MIL Attention

To enable adaptive modality weighting, we employ a multiple-instance learning (MIL) attention mechanism. For each modality, a pooled representation is computed:

$$\mathbf{u}^{(m)} = \frac{1}{C}\sum_{c=1}^{C} \tilde{\mathbf{H}}_c^{(m)}.$$

Attention scores are then computed using a gated mechanism:

$$a_m = \frac{\exp(\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{u}^{(m)}))}{\sum_{m'} \exp(\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{u}^{(m')}))}.$$

The fused representation is given by:

$$\mathbf{F} = \sum_m a_m \mathbf{u}^{(m)}.$$

To prevent mode collapse and encourage balanced modality utilization, an entropy regularization term is added:

$$\mathcal{L}_{\text{ent}} = -\sum_m a_m \log a_m.$$

3.2.3. Cross-Modal Multi-Head Attention

To capture fine-grained token-level interactions, we employ cross-modal multi-head attention. Aligned embeddings are concatenated along the temporal axis:

$$\mathbf{H}_{\text{joint}} = \left[\tilde{\mathbf{H}}^{(\text{DNA})}; \tilde{\mathbf{H}}^{(\text{RNA})}; \tilde{\mathbf{H}}^{(\text{Protein})}\right].$$

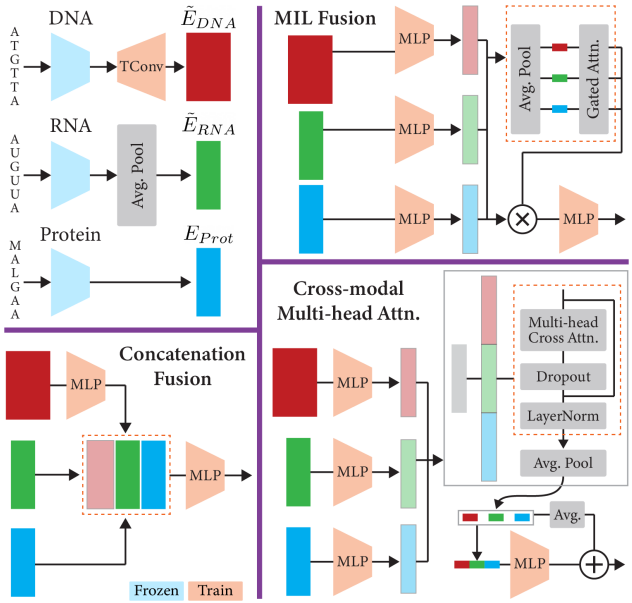

Figure 1. Model architectures for the three BioLangFusion fusion strategies. Codon-aware concatenation: DNA embeddings are upsampled via transposed convolution and mRNA embeddings downsampled via non-overlapping mean pooling to align with the protein framing; the three aligned embeddings are then concatenated 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.

Each modality attends to this joint context:

$$\text{Attn}(\mathbf{Q}^{(m)}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}^{(m)}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V},$$

where queries $\mathbf{Q}^{(m)}$ originate from modality m, and keys and values are derived from $\mathbf{H}_{\text{joint}}$. The updated embeddings are merged using residual averaging and layer normalization, producing the final fused sequence $\mathbf{F}$.

3.3. Prediction Head

The fused representation $\mathbf{F}$ is passed to a lightweight prediction head, either: (i) a TextCNN with multiple kernel sizes for local motif extraction, or (ii) a shallow MLP for global regression or classification tasks. This design ensures expressive modeling while maintaining computational efficiency.

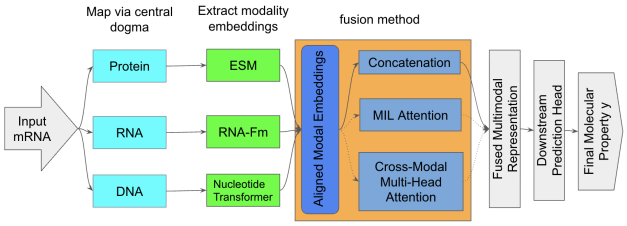

Figure 2. Overview of the proposed multimodal representation learning and fusion pipeline. Input mRNA sequences are mapped through the central dogma to derive DNA, RNA, and protein modalities. Each modality is encoded independently using pretrained foundation models (Nucleotide Transformer for DNA, RNA-FM for RNA, and ESM for proteins). Codon-aligned modality embeddings are then fused using one of three strategies: concatenation, entropy-regularized MIL attention, or cross-modal multi-head attention. The resulting fused multimodal representation is passed to a downstream prediction head to estimate the target molecular property y.

3.4. Multimodal Transfer Learning Extension

Building on multimodal transfer learning principles, we fine-tune pretrained fusion models across related biological tasks. Adapter layers are inserted after projection layers:

$$\mathbf{H}' = \mathbf{H} + \phi(\mathbf{H}; \theta_{\text{adapter}}),$$

allowing task-specific adaptation with minimal additional parameters. We further explore contrastive pre-alignment between modalities using cosine similarity objectives to encourage semantically consistent representations at the codon level.

3.5. Computational Efficiency

To ensure scalability, we employ frozen pretrained encoders, parameter-efficient adapters (5–10M trainable parameters), mixed-precision training (fp16), and sequence truncation (< 1024 tokens).

4. Experimental Results

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). As our progress, we report our results on unimodal training over the fungal expression dataset.

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 2. Summary of the datasets used in the 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).

4.2. Unimodal Performance on Fungal Dataset

We first evaluate the unimodal baselines for protein, DNA, and RNA sequence representations on the fungal dataset. Each modality is trained independently using frozen pretrained encoders and a lightweight prediction head. Performance is assessed using training and validation loss, mean squared error (MSE), and Spearman rank correlation across epochs. We use TextCNN as the architecture for the unimodal models, which is also used as the prediction head in the multimodal case. The unimodal architecture was not specified in the paper we are reimplementing, but the alternative, a shallow MLP, seems to perform worse, so we go with the TextCNN.

4.2.1. RNA Modality

In the unimodal RNA model (Figure 3), we see that there is some learning over the first 30 epochs, decreasing training loss from around 4.0 to around 3.3. The improvement in validation loss curve is more modest, going down from 1.5 to stabilizing around 1.1 at the end of training, showcasing a spiky regime over the earlier epochs. In fact, we can see that the model seems to overfit beyond the first few epochs, since the validation loss curve does not consistently go down.
Our second metric, Spearman correlation, seems to improve greatly on the training dataset, going from around 0.1 to 0.45. We also see a modest increase in the Spearman correlation on the validation set, and unlike the validation loss curve, it stabilizes early and does not show any spiking. This shows that

4.2.2. Protein Modality

The unimodal protein model (Figure 4) demonstrates a similar loss and Spearman correlation regime over the epochs, with a slightly better performance than in the RNA model. In fact, we see the training loss going down to 2.5, and the validation loss going below 1.0. Note that both models use the exact same training and validation splits, so we can say that the protein model performed better on this dataset. We also see a training Spearman correlation of above 0.6, and a validation Spearman correlation of around 0.57.

4.2.3. DNA Modality

The unimodal DNA model achieves the best performance on the training set compared to other single-modality models. Validation performance is not too different to protein and RNA models, both in terms of loss and Spearman correlation. However, the model is able to achieve a training Spearman correlation of above 0.8 and training loss of around 1.7.

4.2.4. Comparative Analysis and Challenges

We see from the plots that the TextCNN prediction head performs the best on DNA embeddings, then protein embeddings, and the worst on RNA embeddings. The token sequences for the RNA modality have the longest length, since they make one token per nucleotide. Since the protein model tokenizes every sequence of three nucleotides and the DNA model makes one token out of every sequence of six nucleotides, these result in smaller number of tokens per sequence. We suspect that the performance differences between the modalities might be due to these different tokenizations.

Our results are below the reported single-modality metrics in the BioLangFusion paper. We have two possible explanations for this: (i) The paper does report the training hyperparameters, which we use in these experiments, but it is not clear if they are only for fusion models or also for the unimodal baselines. Moreover, they do not specify the model details for the unimodal cases, like the TextCNN kernel sizes. A more extensive hyperparameter search might improve our model's performance on single modality. (ii) They do not use the entirety of the datasets (Table 2). In particular, less than half of the samples in the fungal expression dataset, which we experimented with, were used.

There is also another challenge that we faced: Since the RNA sequences from the covid vaccine degradation datasets are synthetic RNA sequences, they do not correspond to real proteins. The paper still reports a Spearman correlation of above 0.8 for the covid vaccine dataset, which is counterintuitive given the biological incompatibility.

5. Conclusion

In this work, we investigate unimodal and multimodal representation learning for molecular property prediction by integrating DNA, RNA, and protein sequence information under a biologically grounded framework. Using frozen pretrained foundation models and lightweight prediction heads, we aim to establish strong unimodal baselines on the fungal dataset, demonstrating that the modalities yield stable convergence and strong rank-based predictive capability.

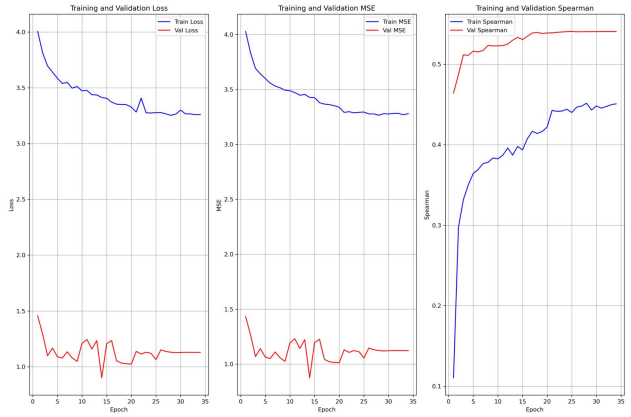

Figure 3. Training and validation curves for unimodal RNA model on the fungal dataset. From left to right: loss, mean squared error (MSE), and Spearman rank correlation across epochs.

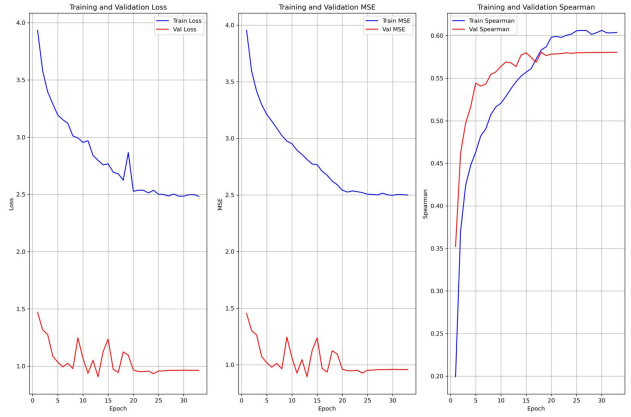

Figure 4. Training and validation curves for unimodal protein model on the fungal dataset. From left to right: loss, mean squared error (MSE), and Spearman rank correlation across epochs.

Building on these findings, we aim to implement a multimodal fusion pipeline that aligns modality-specific embeddings at the codon level and integrate them using complementary fusion strategies, including concatenation, entropy-regularized MIL attention, and cross-modal multi-head attention. These designs enable the model to leverage both modality-specific signals and cross-modality interactions while maintaining computational efficiency through parameter-efficient training.

Our experiments are designed to highlight the complementary nature of molecular modalities and underscore the limitations of unimodal learning when biological information is distributed across multiple sequence types. The proposed framework provides a flexible and scalable foundation for multimodal biological modeling and can be readily extended

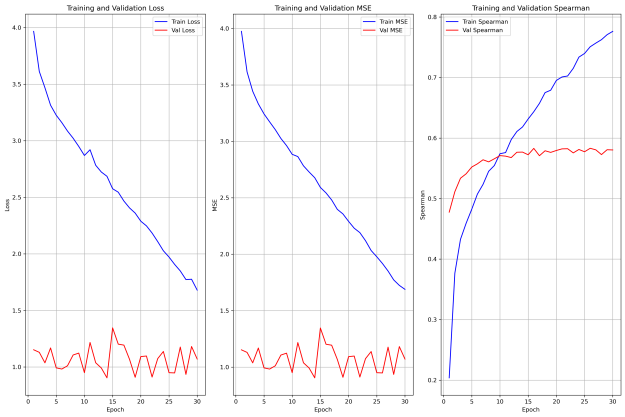

Figure 5. Training and validation curves for unimodal DNA model on the fungal dataset. From left to right: loss, mean squared error (MSE), and Spearman rank correlation across epochs.

to additional tasks, datasets, and training paradigms such as federated or privacy-preserving learning. Future work will explore adaptive modality weighting, cross-task transfer learning, and multimodal pretraining to further enhance generalization and robustness.

References

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.

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 bio-language 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.