BioLangFusion: Multimodal Fusion of DNA, mRNA, and Protein Language Models

Amina Mollaysa [1] **Artem Moskalev** [1] **Pushpak Pati** [1] **Tommaso Mansi** [1] **Mangal Prakash** [*1] **Rui Liao** [*1]

Abstract

We present BioLangFusion, a simple approach for integrating pre-trained DNA, mRNA, and protein language models into unified molecular representations. Motivated by the central dogma of molecular biology (information flow from gene to transcript to protein), we align per-modality embeddings at the biologically meaningful codon level (three nucleotides encoding one amino acid) to ensure direct cross-modal correspondence. BioLangFusion studies three standard fusion techniques: (i) codon-level embedding concatenation, (ii) entropy-regularized attention pooling inspired by multiple-instance learning, and (iii) cross-modal multi-head attention—each technique providing a different inductive bias for combining modality-specific signals. These methods require no additional pre-training or modification of the base models, allowing straightforward integration with existing sequence based foundation models. Across five molecular property prediction tasks, BioLangFusion outperforms strong unimodal baselines, showing that even simple fusion of pre-trained models can capture complementary multi-omic information with minimal overhead.

1. Introduction

The central dogma of molecular biology—DNA is transcribed into mRNA, which is translated into protein—captures a coordinated flow of genetic information. While each stage carries unique regulatory signals, all contribute jointly to phenotype. For instance, disrupting a transcription factor binding site (DNA) can destabilizing an mRNA hairpin, and alter codon usage or amino acid identity (protein). Understanding such effects requires reasoning across this entire molecular cascade.

[*]Equal contribution [1]Johnson & Johnson Innovative Medicine. Correspondence to: Amina Mollaysa <maminanm@its.jnj.com>.

Proceedings of ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada, 2025. Copyright 2025 by the author(s).

Foundation models (FMs) trained on single modalities have recently shown impressive capabilities in their respective domains. Existing DNA language models (Ji et al., 2021; Dalla-Torre et al., 2022; Nguyen et al., 2024) capture regulatory sequence patterns; mRNA-focused models (Zhang et al., 2023; Yazdani et al., 2024) encode structural and post-transcriptional features; and protein models (Lin et al., 2022a; Elnaggar et al., 2021b) excel in predicting structure and function from amino acid sequences. However, these FMs are trained on single modality, ignoring the intrinsic linkage among modalities implied by the central dogma. Emerging evidence shows that unimodal models can exhibit surprising cross-modal capabilities—for example, Evo-2 (Brixi et al., 2025), a 40B-parameter DNA-only model, performs well on RNA and protein tasks (Nguyen et al., 2024), but these effects arise from massive scale and compute—resources that are simply not affordable or accessible to most researchers. Separately, DNA or protein FMs have also shown utility for mRNA-specific tasks (Prakash et al., 2024), suggesting that some molecular signals do transfer even without explicit multimodal training. Still, such generalization is limited and incidental: the models do not explicitly model the flow of information across modalities. As a more pragmatic alternative, recent work shows that even simply concatenating embeddings from DNA and protein models can improve performance on downstream tasks (Boshar et al., 2024), implying that the two modalities encode complementary—and sometimes orthogonal—biological cues.

These observations suggest that meaningful fusion of modality-specific embeddings—combining representations from DNA, RNA, and protein models—could better capture the full DNA→RNA→protein cascade. One straightforward method is direct concatenation of embeddings, but this often fails to scale effectively due to lack of alignment and limited ability to capture cross-modal interactions. Another approach, weight merging, requires foundation models to share the same architecture, tokenization, and embedding dimensions—constraints that are rarely met. Finally, fusion methods like FuseLM from natural language processing rely on knowledge distillation and retraining, which introduce substantial computational complexity.

To address these challenges, we introduce BioLangFusion,

a suite of simple and modular fusion techniques that integrates pretrained DNA, RNA, and protein FMs embeddings without requiring additional training or architectural modifications. We first align embeddings at the codon level to establish biologically meaningful correspondence across modalities. BioLangFusion then studies three fusion strategies: (i) codon-level concatenation, (ii) entropy-regularized attention pooling inspired by multiple-instance learning, and (iii) cross-modal multi-head attention capturing token-level dependencies. Evaluated on five diverse molecular property prediction tasks, BioLangFusion techniques consistently improve over unimodal baselines, offering a practical method for integrating multimodal omics data spanning central dogma with modest computational overhead.

2. Methods

Assume we have training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i = [x_{i1}, \ldots, x_{iT}]$ is a an mRNA sequence with length T where x_i denotes nucleotide, and y_i is the molecular property of interest. Each mRNA maps to its corresponding DNA and protein sequences via the central dogma given the start and stop codons signaling the start and end of translation. Our objective is to predict y_i by fusing the embeddings E_{DNA}, E_{RNA}, E_{Prot} of sequence $\mathbf{x}_i$ extracted from pretrained DNA, RNA, and protein language models.

2.1. Modality alignment

Language models tokenize their inputs at varying biological granularities, resulting in embeddings often differ in precision. Since different pre-trained FM perform variably across tasks (Prakash et al., 2024; Boshar et al., 2024) and there is no universal rule for model selection, we adopt well-established models representative of each modality: the *Nucleotide Transformer* (Dalla-Torre et al., 2023) for DNA (6-mer tokenization), *RNA-FM* (Chen et al., 2022) for RNA (single nucleotide tokenization), and *ESM-2* (Lin et al., 2022a) for protein (amino acid tokenization, i.e., 3-mers in nucleotide).

$$E_{\text{DNA}} \in \mathbb{R}^{\frac{T}{6} \times d_{\text{DNA}}}; \quad E_{\text{RNA}} \in \mathbb{R}^{T \times d_{\text{RNA}}}; \quad E_{\text{Prot}} \in \mathbb{R}^{\frac{T}{3} \times d_{\text{Prot}}}$$

where $d_{\text{DNA}}, d_{\text{RNA}}, d_{\text{Prot}}$ denote the embedding dimensions of the respective language models.

Consequently, embeddings derived from these models inherently vary in length and must be aligned for effective fusion, ensuring that corresponding tokens across modalities represent the same biological region and carry semantically aligned information. We use the protein frame ($T' = \frac{T}{3}$ length) as the reference and map the DNA and mRNA embeddings onto shared codon-level resolution. This choice is biologically motivated: proteins are the final functional products of the central dogma, and their sequences are directly derived from coding regions in DNA and mRNA via translation. Aligning at the codon level—where each codon consists of three nucleotides encoding a single amino acid—ensures that each token corresponds to a biologically meaningful unit. To achieve this alignment, we apply transposed convolution to upsample the DNA embeddings (originally defined over 6-mers) to the codon level, preserving local context, and apply non-overlapping mean pooling to downsample mRNA embeddings to the same token resolution.

$$\begin{aligned} \tilde{E}_{\text{DNA}} &= \text{TConv}_{k=2,s=2}\big(E_{\text{DNA}}\big) \in \mathbb{R}^{T' \times d_{\text{DNA}}}, \\ \tilde{E}_{\text{RNA}} &= \text{AvgPool}_{k=3,s=3}\big(E_{\text{RNA}}\big) \in \mathbb{R}^{T' \times d_{\text{RNA}}}. \end{aligned} \tag{1}$$

After alignment, each modality is represented by a embeddings of length T', where each location t corresponds to the same biological codon across DNA, RNA, and protein representations.

2.2. Fusion methods

2.2.1. Concatenation fusion

Given now we have aligned embeddings: $\tilde{E}_{\text{DNA}}$, $\tilde{E}_{\text{RNA}}$, E_{Prot}, a straightforward fusion strategy is to *concatenate* the feature vectors from each modality at every aligned position along feature dimension. However, due to differences in embedding dimensionality across modalities, such concatenation can lead to imbalanced representations, where the modality with the largest embedding dimension may dominate the fused vector (see Table 3). To address this, we apply a modality-specific learnable MLP to project the DNA embedding to a lower-dimensional space of size d'_{DNA} prior to concatenation since d_{DNA} is significantly larger than d_{RNA} and d_{Prot}.

$$Z_{\text{concat}}(t) = \big[\ \text{MLP}(\tilde{E}_{\text{DNA}}[t]) \,\|\, \tilde{E}_{\text{RNA}}[t] \,\|\, E_{\text{Prot}}[t]\ \big], \quad t = 1, \ldots, T'$$

resulting in $Z_{\text{concat}} \in \mathbb{R}^{T' \times (d'_{\text{DNA}} + d_{\text{RNA}} + d_{\text{Prot}})}$.

This method preserves the full granularity of all modalities without imposing constraints on their interactions, allowing the prediction head to learn informative feature combinations. However, as simple and intuitive as this approach is, it treats all modalities equally at every position. As a result, the model cannot adaptively down-weight noisy or less informative modality unless this behavior is implicitly learned by the prediction head. Moreover, the dimensionality of Z_{concat} increases linearly with the number of modalities, which leads to higher computational and memory costs. These limitations motivate our attention-based fusion alternative, which learns to softly weight each modality based on its relevance to the task at hand.

2.2.2. Multiple Instance Learning (MIL) with Entropy Regularizer

Inspired by multiple instance learning in computer vision (Ilse et al., 2018), we treat the three modality-specific embeddings as a bag of instances, where each instance (modality) may contribute differently to the final prediction. We

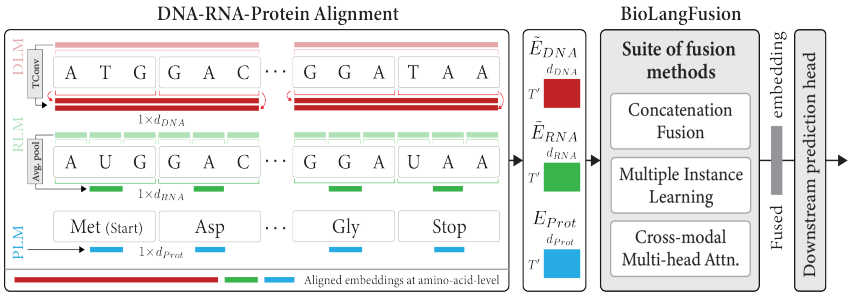

Figure 1. BioLangFusion Architecture Overview: pretrained DNA, RNA, and protein embeddings are aligned at the codon level and fused using biologically motivated strategies then passed to a prediction head for downstream molecular property prediction.

introduce a lightweight gated attention mechanism that *dynamically* weights each modality based on its relevance to the task.

For simplicity, we denote the aligned embeddings as: $\tilde{E}_m \in \mathbb{R}^{T' \times d_m}, \quad m \in \{\text{DNA}, \text{RNA}, \text{Prot}\}$ where each modality has its own feature dimension. To apply attention pooling across modalities, we first project each embedding to a shared latent space $H_m \in \mathbb{R}^{T' \times d}$, then apply gated attention:

$$Z_{fused} = \sum_{m \in \{\text{DNA},\text{RNA},\text{Prot}\}} \alpha_m H_m, \tag{2}$$

where the attention weights α_m are computed from a mean-pooled summary $\bar{\mathbf{h}}_m = \frac{1}{T'} \sum_{t=1}^{T'} H_m[t]$ representation of each modality. We adopt this formulation of sequence-level attention in place of token-level attention, which was empirically less effective in our setting (see Table 4).

$$\alpha_m = \frac{\exp\left(W^\top [\tanh(V_m \bar{\mathbf{h}}_m + \mathbf{b}_m) \odot \sigma(U_m \bar{\mathbf{h}}_m + \mathbf{c}_m)]\right)}{\sum_{i \in \{\text{DNA},\text{RNA},\text{Prot}\}} \exp\left(W^\top [\tanh(V_i \bar{\mathbf{h}}_i + \mathbf{b}_i) \odot \sigma(U_i \bar{\mathbf{h}}_i + \mathbf{c}_i)]\right)}$$

$U_m, V_m \in \mathbb{R}^{d_{\text{attn}} \times d}$, $\mathbf{b}_m, \mathbf{c}_m \in \mathbb{R}^{d_{\text{attn}}}$, and $W \in \mathbb{R}^{d_{\text{attn}}}$ are learnable parameters, $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ is the sigmoid function. This gated attention combines the bounded non-linearity of $\tanh$ with the smooth gating of σ, enhancing flexibility and mitigating $\tanh$ saturation (Ilse et al., 2018).

The resulting fused embedding matrix Z_{fused} is then passed to the downstream prediction head. This "one-attention-per-sequence" design is lightweight yet enables the model to *dynamically* determine, for each downstream task, how much information to retain from DNA, mRNA, and protein modalities.

Entropy regularization. During training, we observe in some case, the model often struggles to learn diverse attention weights. To encourage the model to move away from such nearly uniform solutions, we add a negative entropy term to the loss function. The attention entropy over a mini-batch $\mathcal{D}_i$ is defined as:

$$H_{\text{attn}}(\boldsymbol{\alpha}) = -\frac{1}{|\mathcal{D}_i|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_i} \sum_{m \in \{\text{DNA},\text{RNA},\text{Prot}\}} \alpha_m \log \alpha_m$$

where $\boldsymbol{\alpha} = [\alpha_{\text{DNA}}, \alpha_{\text{RNA}}, \alpha_{\text{Prot}}]$ are the modality-level attention weights. This regularization, $\lambda H_{\text{attn}}(\boldsymbol{\alpha})$ is added to the main loss function where λ is a tunable hyperparameter controlling the strength of the regularization.

2.2.3. Cross-modal Multi-head Attention

While concatenation and attention-based pooling effectively combine modality-specific embeddings, they treat each modality independently at every aligned position. To capture more nuanced interactions between DNA, mRNA, and protein representations, we explore a cross-modal multi-head attention mechanism inspired by transformer architectures.

In contrast to previous fusion approaches, cross-modal attention allows each modality to query contextual information from all modalities jointly. This enables the model to dynamically discover and leverage cross-modality dependencies that are informative for downstream biological prediction tasks. At each position, the model can attend across the full set of aligned DNA, mRNA, and protein embeddings, identifying which signals are most relevant for the task. Specifically, after projecting all modality embeddings to a shared feature dimension d, we concatenate them along the sequence length axis to form a global context representation:

$$C = [H_{\text{DNA}}; H_{\text{RNA}}; H_{\text{Prot}}] \in \mathbb{R}^{3T' \times d} \tag{3}$$

This context C provides the keys and values for attention, while the embedding of each modality acts as the query. Formally, for each modality, we compute:

$$Z_m = g\left(\text{MultiHead}(H_m W_m^Q, C W_m^K, C W_m^V)\right) \tag{4}$$

where MultiHead denotes the standard multi-head attention operator, and $g(\cdot)$ includes residual and projection layers.

The updated embeddings $Z_{\text{DNA}}, Z_{\text{RNA}}, Z_{\text{Prot}}$ are then concatenated across the feature dimension and projected to form a unified representation Z. To stabilize training and preserve modality-specific signals, we add a residual connection with modality averaging followed by layer normalization:

$$Z_{fused} = \text{LayerNorm}\left(\frac{Z_{\text{DNA}} + Z_{\text{RNA}} + Z_{\text{Prot}}}{3} + Z\right) \tag{5}$$

This flexible mechanism allows information to flow between modalities and adaptively reweighs their contributions across positions, capturing both shared and complementary biological features. By learning these interactions, the model builds a richer, more expressive representation than those derived from independent fusion strategies.

3. Experiment

Dataset We evaluate BIOLANGFUSION 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. These datasets differ in sequence length, sample size, and prediction targets. Full statistics and details are provided in Appendix A.2.

Experimental Setup We begin by extracting DNA, RNA, and protein embeddings using pretrained FMs— *Nucleotide Transformer*, *RNA-FM*, and *ESM-2 (8M)* (see Table 3). These embeddings are then fused using our proposed strategies to produce a unified multimodal representation, which is passed to a downstream prediction head. All fusion methods share the same fixed TextCNN prediction head to ensure a fair comparison (see Appendix A.4 for details).

Encoding	CoV-Vac	Fungal	E. Coli	mRNA Stab.	Ab1
Evo (Lin et al., 2023)	0.653	0.579	42.556	0.403	0.360
SpliceBERT (Chen et al., 2024)	0.802	0.778	48.455	0.522	0.718
ESM-2 (650M)	0.825	0.734	46.348	0.536	0.679
ESM-2 (3B)	0.772	0.721	46.208	0.537	0.700
ESM-2 (8M)	0.806	0.695	49.017	0.539	0.711
RNA-FM	0.841	0.767	52.949	0.553	0.743
Nucleotide Transformer	0.780	0.804	41.292	0.530	0.732
Concatenation	0.831	0.805	50.280	0.539	0.764
MIL + Entropy	**0.864**	**0.824**	52.107	**0.563**	0.760
Cross Attention	0.828	0.812	**53.932**	0.550	**0.765**

Table 1. Performance of multimodal fusion variants of BIOLANGFUSION vs. single-modal baselines. Spearman correlation is used for regression; accuracy for classification (E. coli). We use *Nucleotide Transformer*, *RNA-FM*, and *ESM-2 (8M)* for fusion

Experimental Results We report model performance in Table 1.Across all the tasks, fusion-based models consistently outperform the best single-modality baselines. While naive concatenation already benefits from multi-modal information, attention pooling further improves performance by learning task-specific modality weights. This allows the model to prioritize the most informative modality for each input, rather than treating all modalities equally. Interestingly, the cross-attention fusion method achieves the best performance on the *E. coli Proteins* dataset, but generally underperforms compared to MIL-based attention pooling on most other tasks. This discrepancy suggests two possibilities: either token-level cross-modal interactions are particularly beneficial for certain tasks such as protein abundance classification, or the current datasets may be too limited in size to fully exploit the representational capacity of cross-attention mechanisms.

Explaining Modality Contributions via Attention To better understand how BioLangFusion integrates complementary signals from DNA, mRNA, and protein modalities, we visualize the learned attention weights assigned to each modality across datasets. These attention scores reflect the importance attributed to each modality in generating fused representations for downstream predictions. Figure 2 shows modality-wise attention scores from the entropy-regularized model. Clear biological trends emerge—for instance, in the *mRNA Stab.* task, the model emphasizes mRNA embeddings, consistent with the role of post-transcriptional features in degradation. In contrast, predictions of *E. coli* protein abundance exhibit stronger attention toward protein embeddings. We then compare these scores with attention score obtained from training without entropy. In the absence of entropy regularization, attention weights tend to be diffuse and less decisive. In contrast, entropy-regularized attention yields sharp, task-specific modality weights, enhancing interpretability and revealing each pretrained FM's contribution to the downstream decisions (See Appendix A.3).

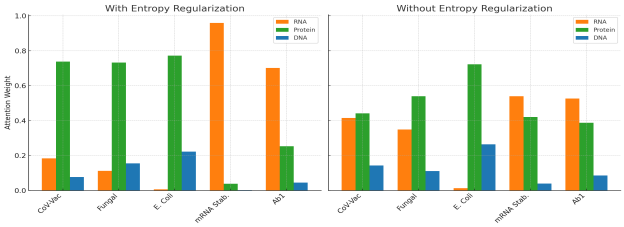

Figure 2. Modality-wise attention weights with and without entropy regularization from MIL based fusion model

4. Conclusion

In this paper, we explore strategies to fuse pretrained DNA, mRNA, and protein foundation models, capturing their biological interconnectivity. We introduce BIOLANGFUSION, a lightweight, plug-and-play suite of fusion strategies for integrating pretrained DNA, mRNA, and protein FMs. It aligns embeddings at the codon level and incorporates biologically informed fusion mechanisms—including codon-aware concatenation, entropy-regularized attention pooling, and cross-modal multi-head attention. Across five molecular property prediction tasks, BioLangFusion consistently outperforms unimodal baselines, demonstrating the synergistic power of pretrained unimodal FMs when fused effectively. Moreover, our approach provides interpretability through modality attention weights, revealing the biological relevance of each modality across tasks. BioLangFusion requires no end-to-end retraining or architectural changes, offering a scalable and accessible framework for multi-modal modeling in biology.

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.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pp. 2025–02, 2025.

Chen, K., Zhou, Y., Ding, M., Wang, Y., Ren, Z., and Yang, Y. Self-supervised learning on millions of primary rna sequences from 72 vertebrates improves sequence-based rna splicing prediction. *Briefings in bioinformatics*, 25 (3):bbae163, 2024.

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, H., Gonzalez, L., Mendoza Revilla, J., Lopez Carranza, N., Grywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., Sirelkhatim, H., Richard, G., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. URL https://www.biorxiv.org/content/10.1101/2023.01.01.123456v1.

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.

Dalla-Torre, L., Chughtai, Z., Yan, Y., et al. Nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2024. 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(1):12126, 2022.

Ding, Z., Guan, F., Xu, G., Wang, Y., Yan, Y., Zhang, W., Wu, N., Yao, B., Huang, H., Tuller, T., et al. Mpepe, a predictive approach to improve protein expression in e. coli based on deep learning. *Computational and Structural Biotechnology Journal*, 20:1142–1153, 2022.

Elnaggar, A., Heinzinger, M., Dallago, C., et al. 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*, 2021a. doi: 10.1109/TPAMI.2021.3095381.

Elnaggar, A., Heinzinger, M., et al. 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*, 2021b. doi: 10.1109/TPAMI.2021.3095381.

Elnaggar, A., Heinzinger, M., and Rost, B. Ankh: Optimized protein language model unlocks generalization across structure and function. *bioRxiv*, 2023. doi: 10.1101/2023.10.17.562788.

Garau-Luis, J. J., Bordes, P., Gonzalez, L., et al. Ceviche: A multimodal deep learning framework for biological sequence analysis. *Bioinformatics*, 40(2):456–468, 2024a. doi: 10.1093/bioinformatics/btaa123.

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

Huang, Y. et al. A comprehensive investigation of multimodal deep learning fusion strategies. *Artificial Intelligence Review*, 57(3):123–145, 2024. doi: 10.1007/s10462-024-10984-z.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

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, 2021.

Leppek, K., Byeon, G. W., Kladwang, W., Wayment-Steele, H. K., Kerr, C. H., Xu, A. F., Kim, D. S., Topkar, V. V., Choe, C., Rothschild, D., et al. Combinatorial optimization of mrna structure, stability, and translation for rna-based therapeutics. *Nature communications*, 13(1):1536, 2022.

Li, M., Wu, Y., Zhang, S., et al. Utr-lm: Pretrained language models for untranslated region function prediction. *bioRxiv*, 2023a. doi: 10.1101/2023.04.26.538444.

Li, S., Moayedpour, S., Li, R., Bailey, M., Riahi, S., Kogler-Anele, L., Miladi, M., Miner, J., Zheng, D., Wang, J., et al. Codonbert: Large language models for mrna design and optimization. *bioRxiv*, pp. 2023–09, 2023b.

Lin, Z., Akin, H., Rao, R., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022a. doi: 10.1101/2022.07.20.500902.

Lin, Z., Akin, H., Rao, R., et al. Language models of protein sequences at the scale of evolution enable accurate

structure prediction. *bioRxiv*, 2022b. doi: 10.1101/2022.07.20.500902.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Liu, H., Ji, Y., Zhou, Z., and Davuluri, R. V. Dnabert-2: Efficient foundation model for dna language in genome. *bioRxiv*, 2023. doi: 10.1101/2023.06.02.543344.

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

Madani, A., Krause, B., Greene, E. R., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 2023. doi: 10.1038/s41587-023-01772-9.

Nguyen, E., Poli, M., Durrant, M., et al. Hyenadna: Learning at a million tokens per sample with dna language models. *bioRxiv*, 2023. doi: 10.1101/2023.12.22.573141.

Nguyen, E., Poli, M., Durrant, M. G., et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6669):746–752, 2024. doi: 10.1126/science.ado9336.

Notin, P. et al. Tranception: protein fitness prediction with autoregressive transformers and retrieval. *arXiv preprint arXiv:2205.13760*, 2023.

Prakash, M., Moskalev, A., Jr., P. D., Combs, S., Mansi, T., Scheer, J., and Liao, R. Bridging biomolecular modalities for knowledge transfer in bio-language models. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024. URL https://openreview.net/forum?id=dicOSQVPLm.

Wang, Y., Zhang, H., Lin, H., et al. Splicebert: Transformer-based pretraining for rna splicing prediction. *bioRxiv*, 2023. doi: 10.1101/2023.05.22.541751.

Wint, R., Salamov, A., and Grigoriev, I. V. Kingdom-wide analysis of fungal protein-coding and trna genes reveals conserved patterns of adaptive evolution. *Molecular biology and evolution*, 39(2):msab372, 2022.

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.

Zhang, M., Liu, X., Yang, J., et al. Codonbert: A pre-trained language model for mrna codon optimization. *bioRxiv*, 2023. doi: 10.1101/2023.02.15.528644.

A. Appendix

A.1. Related Work

Unimodal biomolecular language models. Recent advances in biological language modeling have produced powerful unimodal models trained on large-scale corpora of DNA, RNA, and protein sequences. In genomics, transformer-based architectures such as DNABERT (Ji et al., 2021), DNABERT-2 (Liu et al., 2023), and Nucleotide Transformer (Dalla-Torre et al., 2024) have demonstrated strong performance on tasks like regulatory element prediction and variant prioritization, while complementary models such as HyenaDNA (Nguyen et al., 2023) and Evo (Nguyen et al., 2024) use long-range convolution or structured state space dynamics to enhance scalability and contextual understanding. For protein sequences, models including ProtTrans (Elnaggar et al., 2021a), ESM family of models (Lin et al., 2022b), Ankh (Elnaggar et al., 2023), and ProGen2 (Madani et al., 2023) have achieved state-of-the-art results in structure prediction, function classification, and mutational fitness estimation via zero-shot transfer. In RNA, models such as RNA-FM (Chen et al., 2022), Splice-BERT (Wang et al., 2023), and UTR-LM (Li et al., 2023a) focus on various non-coding regions, while recent mRNA-specific models like CodonBERT (Zhang et al., 2023) and HELM (Yazdani et al., 2024) have also been proposed.

Multi-modal integration of bio-LMs. While multimodal learning has gained traction in many domains, integrating biological modalities—DNA, RNA, and proteins—remains an underexplored area. Most existing approaches rely on rudimentary fusion strategies such as early feature concatenation (Huang et al., 2024) or late-stage prediction averaging (Garau-Luis et al., 2024a), which fail to harness the expressive power of contextual embeddings from pretrained biological language models. Moreover, these methods overlook inter-modality alignment and do not scale well to more than two modalities due to dimensional constraints imposed by naive concatenation. In the protein domain, Tranception (Notin et al., 2023) combines language model outputs with multiple sequence alignment (MSA) features via blockwise attention, but it is limited to protein-only modeling and ignores transcriptional or genomic context. Recent work (Garau-Luis et al., 2024b) has explored direct transfer of pretrained DNA and protein models to mRNA-specific tasks, demonstrating promising cross-modal generalization, though without modeling all modalities jointly. Life-Code (Liu et al., 2025), in contrast, proposes a more holistic multi-omics architecture inspired by the central dogma, integrating modalities through reverse translation and codon-aware modeling, but requires pre-training bespoke foundation models from scratch.

To date, few approaches aim to unify DNA, RNA, and protein representations in a compute-efficient framework. Additionally, biologically meaningful alignment strategies—such as codon-level mappings and transcript-aware position encoding—remain largely absent, limiting the capacity to model cross-modal dependencies embedded in molecular biology's central dogma.

A.2. Dataset

We evaluate our method on five molecular property prediction datasets. The Ab1 dataset is sourced from Yazdani et al. (2024), while the remaining four are obtained from Li et al. (2023b).

- **CoV-Vac** (Leppek et al., 2022): SARS-CoV-2 vaccine degradation dataset comprising mRNA sequences engineered for optimized structural features, stability, and translational efficiency in vaccine development contexts.
- **Fungal** (Wint et al., 2022): Expression dataset containing protein-coding and tRNA genes extracted from a wide range of fungal genomes, annotated with corresponding expression levels.
- **E. coli** (Ding et al., 2022): Experimental dataset of protein expression in *E. coli*, with expression levels categorized into low, medium, and high classes.
- **mRNA Stab.** (Diez et al., 2022): A large dataset containing mRNA sequences from human, mouse, frog, and fish, annotated with experimentally measured transcript stability scores.
- **Ab1** (Yazdani et al., 2024): A collection of antibody-encoding mRNA sequences labeled with quantitative protein expression levels.

We follow the train/test splits provided by Li et al. (2023b). To ensure compatibility with pretrained language models—most of which support input lengths up to 1000 tokens—we truncate RNA sequences exceeding this length. Dataset statistics are summarized in Table 2.

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 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).

The foundational models used for each modality and their embedding dimensionalities are shown in Table 3.

Modality	Model	Version	Embedding Dim.
RNA	RNA-FM	`rna_fm_t12`	640
DNA	Nucleotide Transformer	`nucleotide-transformer-v2-100m-multi-species`	4,107
Protein	ESM-2	`esm2_t6_8M_UR50D`	320

Table 3. Pretrained foundation models used to encode each modality, listing the exact version and embedding dimension used as input to the fusion framework.

A.3. Ablation study

To better understand the contribution of individual components and architectural choices in BioLangFusion, we conduct a comprehensive ablation study. We begin by examining the effect of the entropy regularization term on the MIL-based attention fusion (Section 2.2.2). This regularization encourages non-uniform attention distributions and, as shown in Section 3, improves both performance and interpretability in low-data regimes.

Next, we assess the impact of using shared versus modality-specific projection layers in the MIL attention mechanism. In our default setup, each modality uses its own projection before attention calculation; we ablate this design by sharing a single projection across all modalities (in Eq. 3) to test whether distinct projections are necessary for capturing modality-specific characteristics.

We further explore token-level attention by removing the mean pooling step prior to computing attention scores. This variant produces per-token attention vectors for each modality while MIL base attention produce one attention score per sequence, allowing the model to learn finer-grained importance distributions along the sequence.

Finally, to assess whether simple concatenation works without modality alignment (2.1), we experiment with a naive strategy that first projects all embeddings to a common feature dimension and concatenates them along the sequence length axis.

Method	CoV-Vac	Fungal	E. Coli	mRNA Stab.	Ab1
MIL without entropy	0.854	**0.826**	51.960	0.556	0.753
MIL with shared projection (with entropy)	0.859	0.813	50.140	**0.563**	0.755
Token level attention	0.835	0.764	44.100	0.553	**0.770**
Vanilla concatenation (without codon level alignment)	0.818	0.807	46.208	0.537	0.754
MIL + entropy	**0.864**	0.824	**52.107**	**0.563**	0.760

Table 4. Performance comparison of attention variants across selected datasets (rounded to 3 decimal places).

As shown in Table 4, entropy regularizer improves prediction performance. Moverove, having modality specific projection layer when we calculate the gated attention score improves over having shared projection layers accross modality. Finally, token level attention brings performance gain only on one (Ab1) out of the five data set.

A.4. Network architecture and optimization parameters

In Figure 3, we present the model architecture for the three fusion methods::

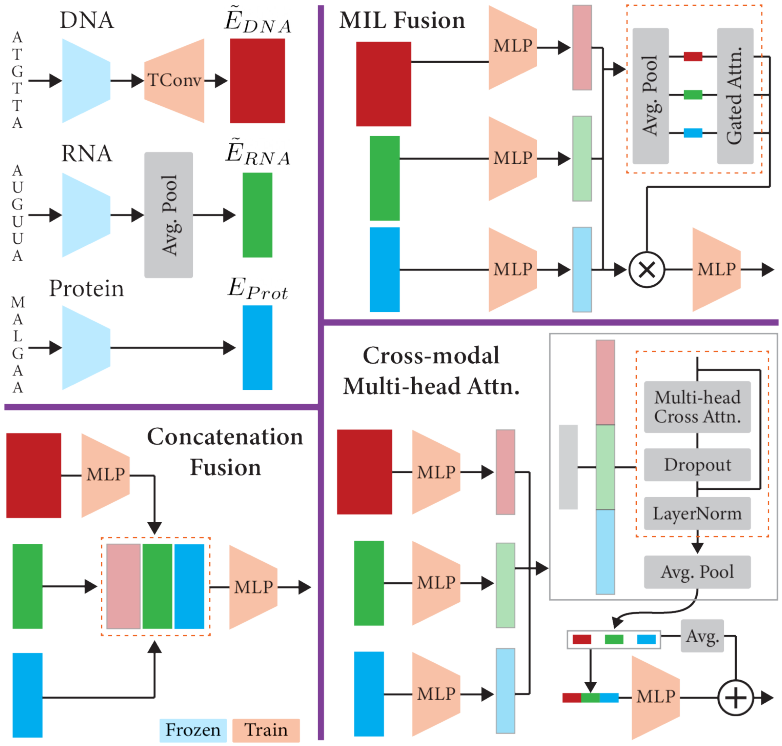

Figure 3. Model architectures for the three BioLangFusion fusion strategies. Codon-aware concatenation: DNA embeddings are upsampled via transposed convolution and mRNA embeddings downsampled via non-overlapping mean pooling to align with the protein framing; the three aligned embeddings are then concatenated at each codon position and projected before entering the TextCNN head. Entropy-regularized 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.

We train all models using the Adam optimizer with a learning rate of 3×10^{-5} and weight decay of 1×10^{-5}. Early stopping is employed with a patience of 20 epochs, and the learning rate is scheduled using `ReduceLROnPlateau` with a patience of 5. For regression tasks, we use mean squared error (MSE) loss; for classification, we use cross-entropy. The entropy regularization weight λ is selected from the set $\{0.01, 0.5, 1\}$ via validation performance. All experiments are conducted with a batch size of 32 on a single GPU. The detailed network architecture and training hyperparameters are provided in List 1 and List 3.

List 1 BioLangFusion architecture:MIL based attention fusion and TextCNN head.

```
├─ Input modality embeddings
│  ├─ Protein embedding ..................................................{ESM-2 (320-dim)}
│  ├─ RNA embedding ......................................................{RNA-FM T12 (640-dim)}
│  └─ DNA embedding ..............................{Nucleotide Transformer v2 (4107-dim)}
├─ Temporal alignment
│  ├─ RNA downsampling ................................{AvgPool1D (kernel=3, stride=3)}
│  ├─ DNA upsampling ...............{ConvTranspose1D (kernel=3, stride=2, padding=2)}
│  └─ Padding .......................................{sequence length aligned to protein}
├─ Projection layers
│  ├─ Per-modality projection ........................................{Linear → 600-dim}
│  └─ Activation function ...............................................................{tanh}
├─ Attention fusion
│  ├─ Attention dimension ...............................................................{100}
│  ├─ Gating dimension ..................................................................{100}
│  ├─ Softmax temperature τ .........................{learned, clamped to [0.02, 20.0]}
│  └─ Fusion operation ...........................{weighted sum of modality embeddings}
└─ TextCNN prediction head
   ├─ Conv1D kernel sizes ............................................................{3, 4, 5}
   ├─ Conv1D output channels .........................................................{1280}
   ├─ Activation function ............................................................{ReLU}
   ├─ Global max pooling
   ├─ Dropout ........................................................................{0.2}
   └─ Fully connected layer .........................{output task-specific prediction}
```

List 2 BioLangFusion architecture: cross-modal multi-head attention fusion and TextCNN head.

```
├─ Input modality embeddings
│  ├─ Protein embedding ..................................................{ESM-2 (320-dim)}
│  ├─ RNA embedding ......................................................{RNA-FM T12 (640-dim)}
│  └─ DNA embedding ..............................{Nucleotide Transformer v2 (4107-dim)}
├─ Temporal alignment
│  ├─ RNA downsampling ................................{AvgPool1D (kernel=3, stride=3)}
│  ├─ DNA upsampling ...............{ConvTranspose1D (kernel=3, stride=2, padding=2)}
│  └─ Padding .......................................{sequence length aligned to protein}
├─ Projection layers
│  └─ Per-modality projection .......................{Linear→ 600-dim, tanh activation}
├─ Cross-modal attention
│  ├─ Build joint context ...............................{concat along time → 3T' × 600}
│  ├─ MultiHeadAttention ..........................................{4 heads, dropout 0.1}
│  └─ Residual + LayerNorm ...................................................{per modality}
├─ Fusion
│  ├─ Concatenate streams ...................................................{shape T' × 1800}
│  ├─ Linear fusion ..........................................................{1800 → 600-dim}
│  └─ Residual average + LayerNorm
└─ TextCNN prediction head
   ├─ Conv1D kernels ...............................{3, 4, 5}, 100 channels
   ├─ Activation ....................................................................{ReLU}
   ├─ Global max pooling
   ├─ Dropout ........................................................................{0.2}
   └─ Fully connected layer .....................................{task-specific output}
```

List 3 Training setup and optimization details

```
├── Objective function
│   ├── Regression task ..................................................{MSE + λEntropy}
│   └── Classification task ..........................................{CrossEntropy + λEntropy}
├── Optimization algorithm ..........................................................{Adam}
├── Learning rate ..................................................................{3 × 10^-5}
├── Weight decay ...................................................................{1 × 10^-5}
├── Learning rate scheduler ..........................{ReduceLROnPlateau (patience=5)}
├── Early stopping ..................................................{patience = 20 epochs}
├── Batch size ......................................................................{32}
├── Max training epochs ............................................................{500}
└── Evaluation metrics
    ├── Regression ..................................................................{Spearman}
    └── Classification ..............................................................{Accuracy}
```