Multimodal Learning of Biological Language Models

A multimodal approach that merges DNA, RNA and protein embeddings to improve performance in molecular biology.

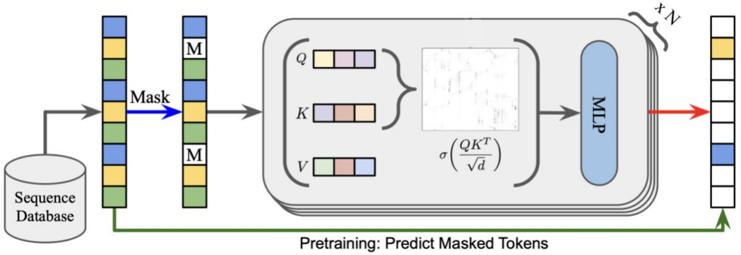
Alireza Noroozi, Sahand Hassanizorgabad, Mustafa Serhat Aydın

Language Models for Biological Sequences

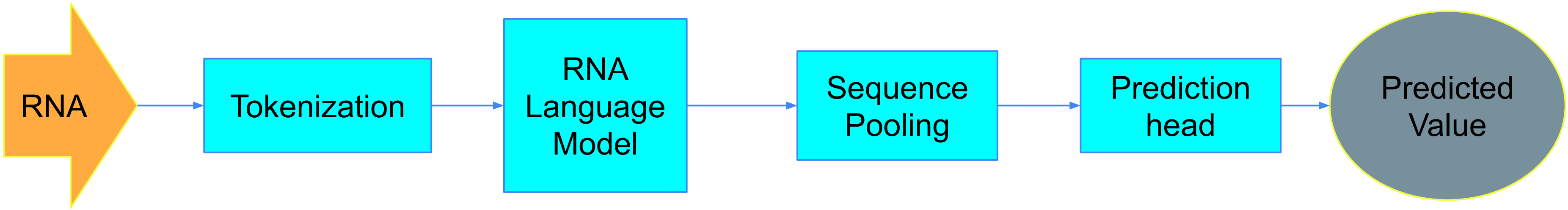
- BioLangFusion by Mollaysa et al. ICML 2025 workshop

Example RNA sequence: AUGCCAGUGACUUCAGGGACGAAUGACUUA (vocabulary: A, U, G, C)

- Many LMs trained on biological sequences use a masked language modeling objective (RNA-FM, ESM etc.)
- Predicting missing tokens (nucleotide, amino acids, k-mers etc.) corresponds to learning structural and evolutionary constraints.
- We want to use the embeddings produced by these LMs!


Sequence Database
Mask
M
M
Q
K
V
$\sigma\left(\frac{QK^T}{\sqrt{d}}\right)$
MLP
×N
Pretraining: Predict Masked Tokens

Example: predict protein expression level from RNA sequences.



RNA

Tokenization

RNA Language Model

Sequence Pooling

Prediction head

Predicted Value

Problem Definition

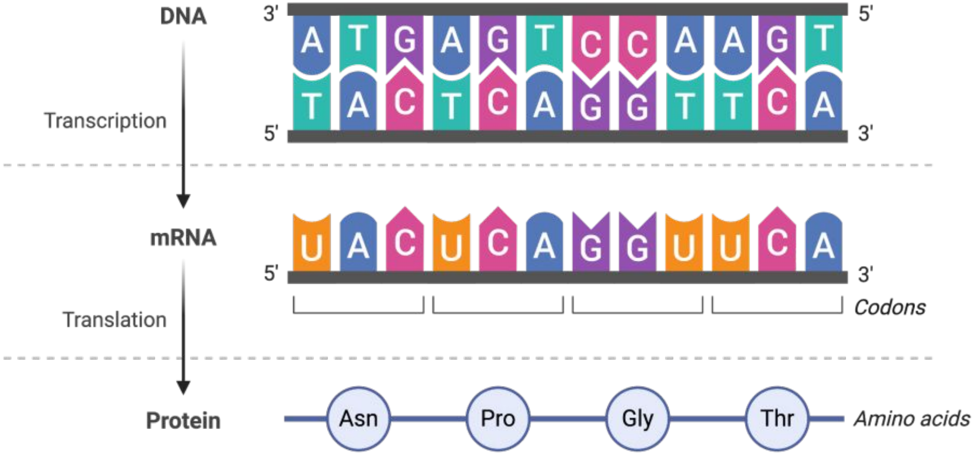
Modern biological language models, like Nucleotide Transformer (DNA), RNA-FM (RNA), and ESM-2 (proteins), capture rich modality-specific information, yet they operate independently. However, real biology is inherently multimodal:

- DNA provides regulatory context,
- mRNA reflects transcription and stability,
- proteins determine functional outcome.

Central Dogma implies;

Given a DNA sequence, you can obtain the corresponding mRNA and protein sequences.

Given an mRNA sequence, you can obtain the corresponding DNA and protein sequences.



Step 1 — Extract embeddings from pretrained models

For each input mRNA sequence:

- **DNA model → Nucleotide Transformer embeddings**
- **RNA model → RNA-FM embeddings**
- **Protein model → ESM-2 embeddings**

These three embeddings differ in:

- token resolution
- sequence length
- embedding dimensionality

Step 2 — Codon-level Modality Alignment

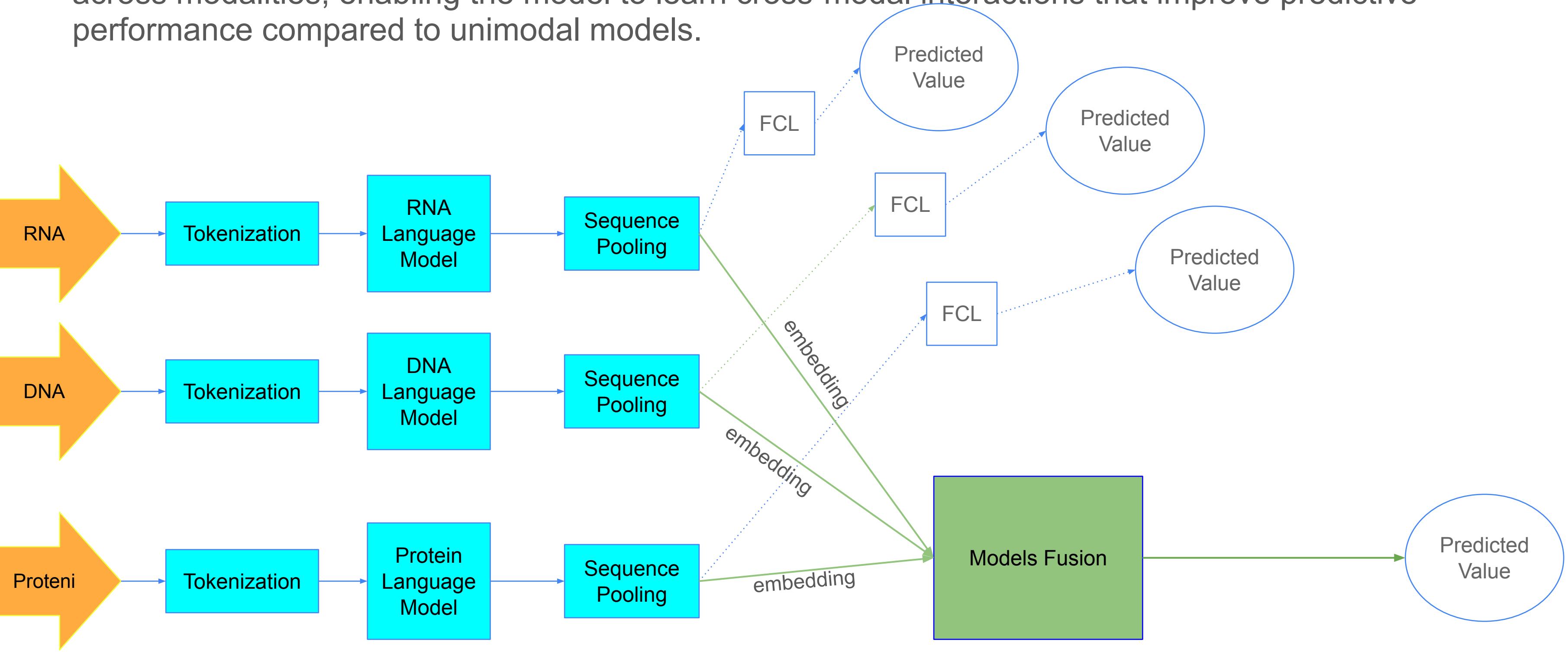
Because DNA, RNA, and Proteins are linked through the central dogma, we align embeddings at **codon resolution (3 nucleotides → 1 amino acid)**:

- DNA (6-mer tokens) → upsample through transposed convolution
- RNA (single nucleotide) → downsample through mean pooling
- Protein (amino acids) → used as natural codon reference

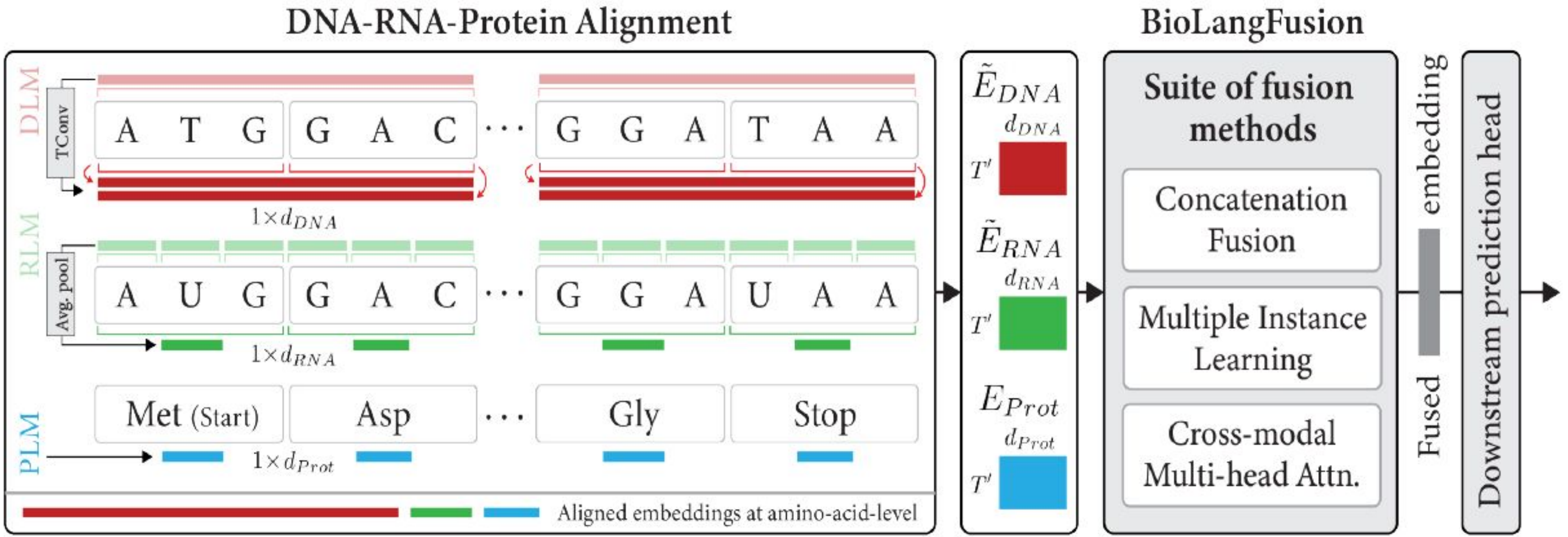
n 6-mers = 6n nucleotides = 2n amino acids

Multimodal Fusion in Biology

the multimodal approach jointly leverages DNA, RNA, and protein embeddings by fusing their pretrained model outputs into a shared representation. This fusion captures complementary biological information across modalities, enabling the model to learn cross-modal interactions that improve predictive performance compared to unimodal models.



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



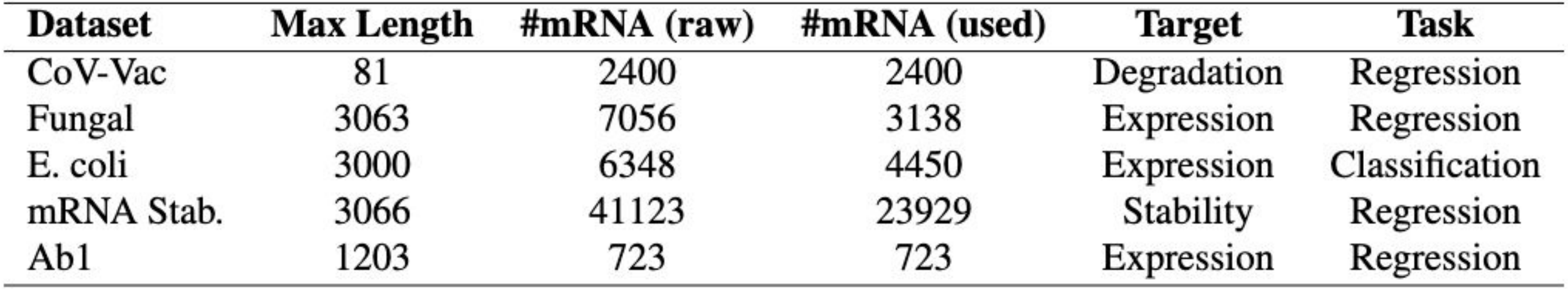
BioLangFusion predicts molecular properties by integrating information from three biological modalities DNA, mRNA, and protein using pretrained language models.

Feature Pipeline Summary Table

Step	Component	Input Type	Output Shape	Description
1	Raw sequence	`str`	length varies	Raw mRNA nucleotide sequence
2	Clean + truncate	`str`	≤1022 bases	Prepares sequence for model
3	Tokenizer	string → tokens	`(1, 1024)`	Adds CLS/EOS; produces integers
4	RNA-FM model	token IDs	`(1024, 640)`	Transformer contextual embeddings
5	Align length	embeddings	`(1022, 640)`	Trim or pad to fixed length
6	Mean pooling	`(1022, 640)`	`(640,)`	Produces sequence representation
7	Batch stack	list of vectors	`(B, 640)`	B = batch size
8	MLP head	`(B, 640)`	`(B,)`	Final regression output

Preliminary Results – Baselines

Experimental Dataset

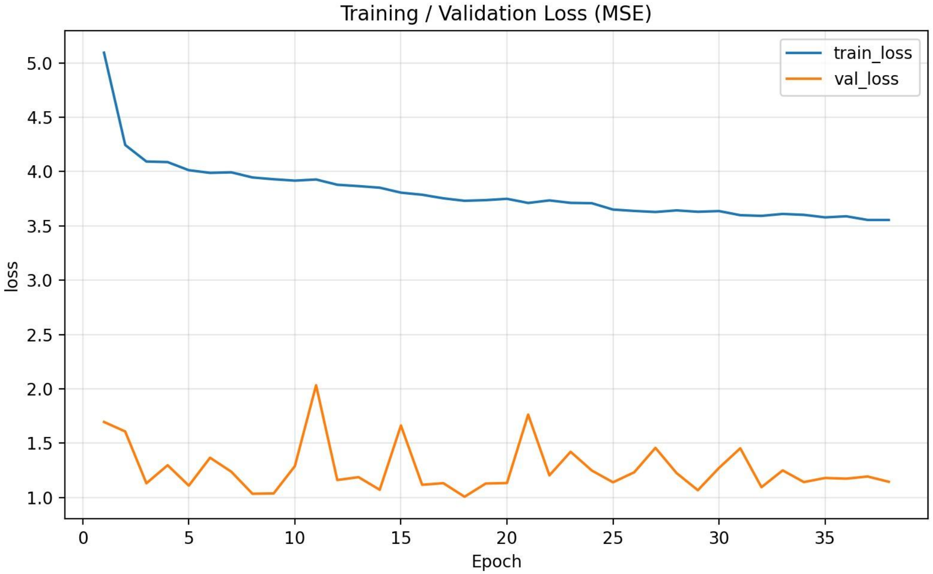


Dataset	Max Length	#mRNA (raw)	#mRNA (used)	Target	Task
CoV-Vac	81	2400	2400	Degradation	Regression
Fungal	3063	7056	3138	Expression	Regression
E. coli	3000	6348	4450	Expression	Classification
mRNA Stab.	3066	41123	23929	Stability	Regression
Ab1	1203	723	723	Expression	Regression

Metrics: Spearman Correlation for regression tasks and accuracy for classification task.

What we have done so far

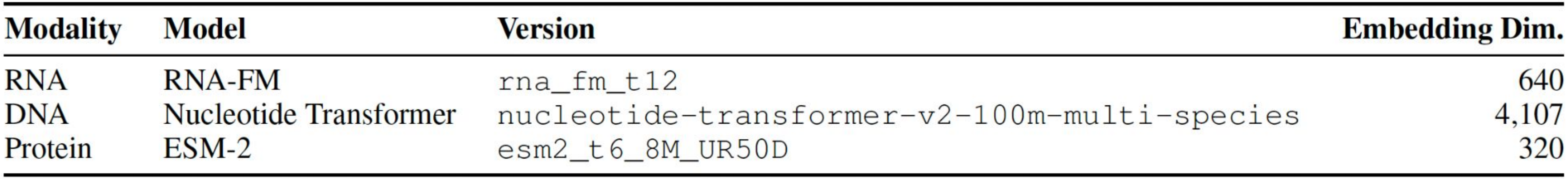
- Processing the dataset
- Calculating the embeddings for each dataset and modality for efficiency
- Training unimodals for baseline



Training loss for fungal_expression with ESM2 (protein) model.
spearman correlation: ~0.50

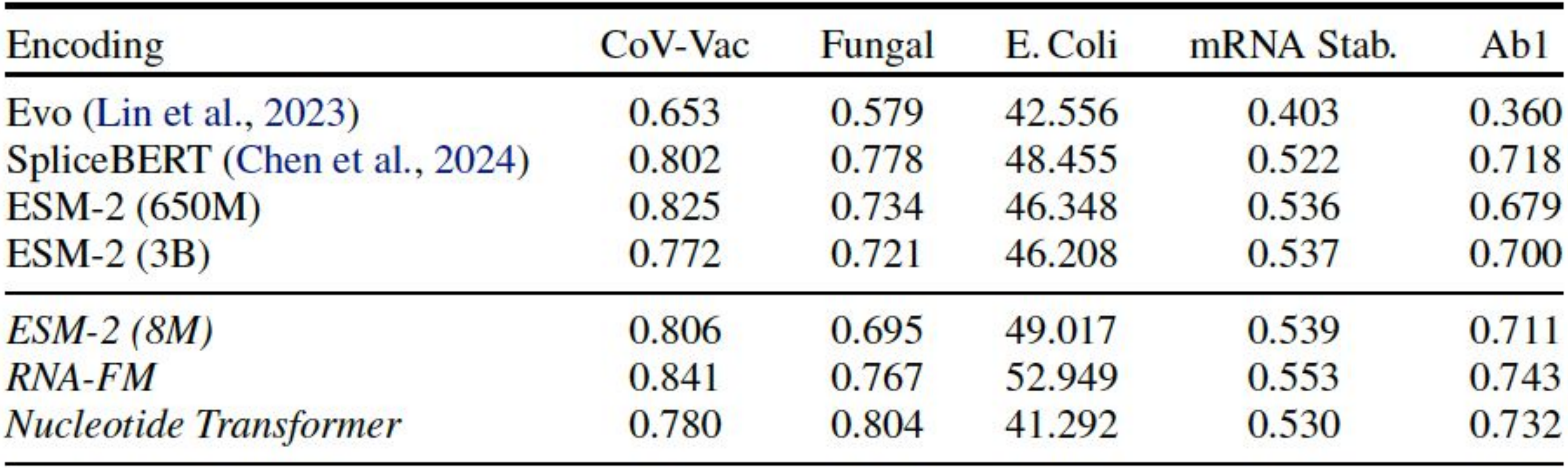
Baseline metrics from unimodals

Models



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

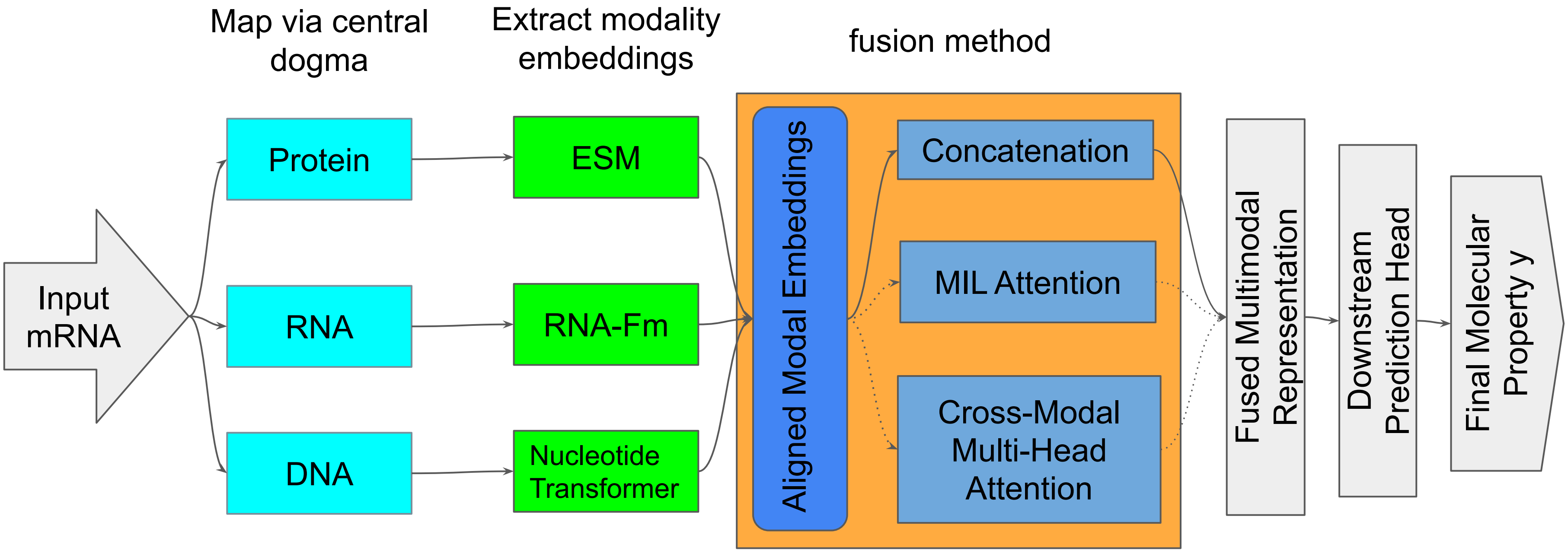
Metrics



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

Future Steps:

Step 1 — End-to-End Multimodal Training


Map via central dogma
Extract modality embeddings
fusion method
Input mRNA
Protein
RNA
DNA
ESM
RNA-Fm
Nucleotide Transformer
Aligned Modal Embeddings
Concatenation
MIL Attention
Cross-Modal Multi-Head Attention
Fused Multimodal Representation
Downstream Prediction Head
Final Molecular Property y

Fusion methods

Concatenation Fusion:

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

Multiple Instance Learning (MIL) with Gated Attention

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

attention weights

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

meanpooled

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

attention entropy

Cross-Modal Multi-Head Attention

$$Z_m = g\left(\mathrm{MultiHead}(H_m W_m^Q, CW_m^K, CW_m^V)\right)$$

keys and values

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

$$C = [H_{\text{DNA}}; H_{\text{RNA}}; H_{\text{Prot}}]$$

Fusion Mechanisms

We explore three alternative fusion strategies:

1. Concatenation Fusion

Combine DNA, RNA, and protein embeddings position-wise → simple but high-dimensional.

2. MIL-based Attention + Entropy Regularization

Model learns modality importance weights:

- high weight → modality is informative
- low weight → modality is noisy

Entropy regularization encourages **decisive** attention.

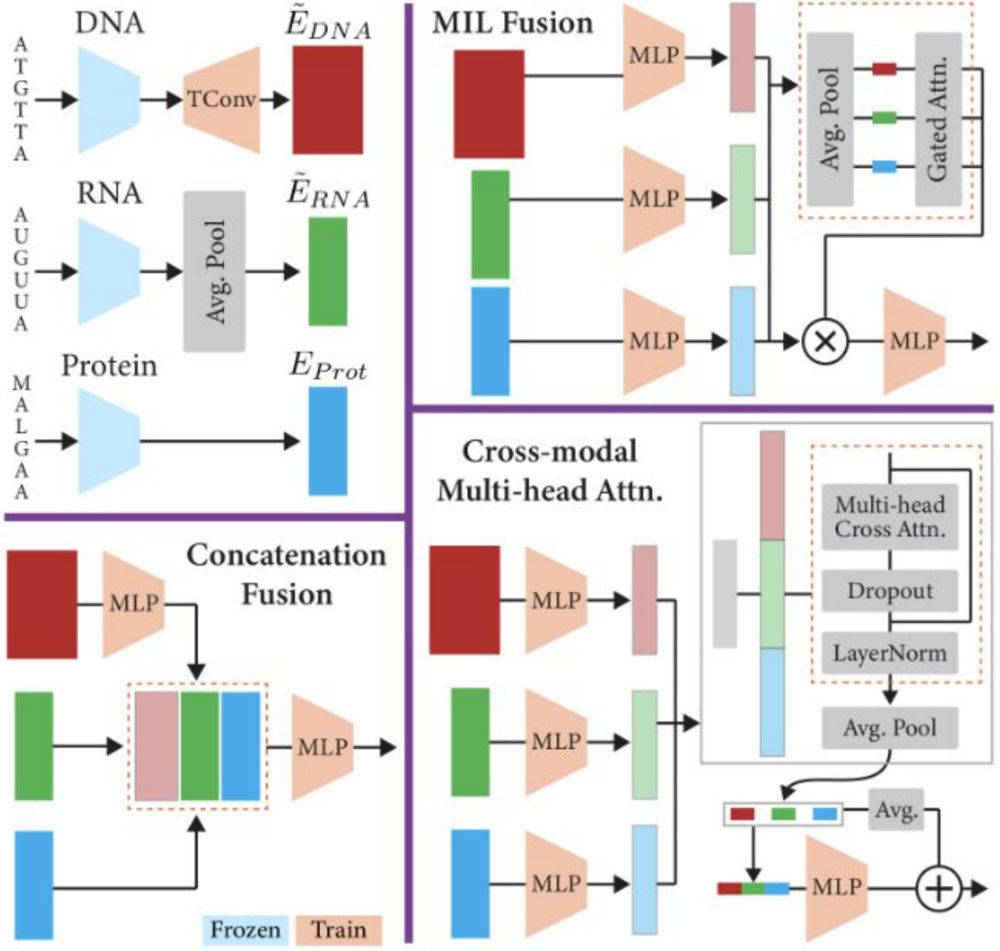
Inspired by multiple instance learning in computer vision (Ilse et al., 2018)

3. Cross-modal Multi-Head Attention

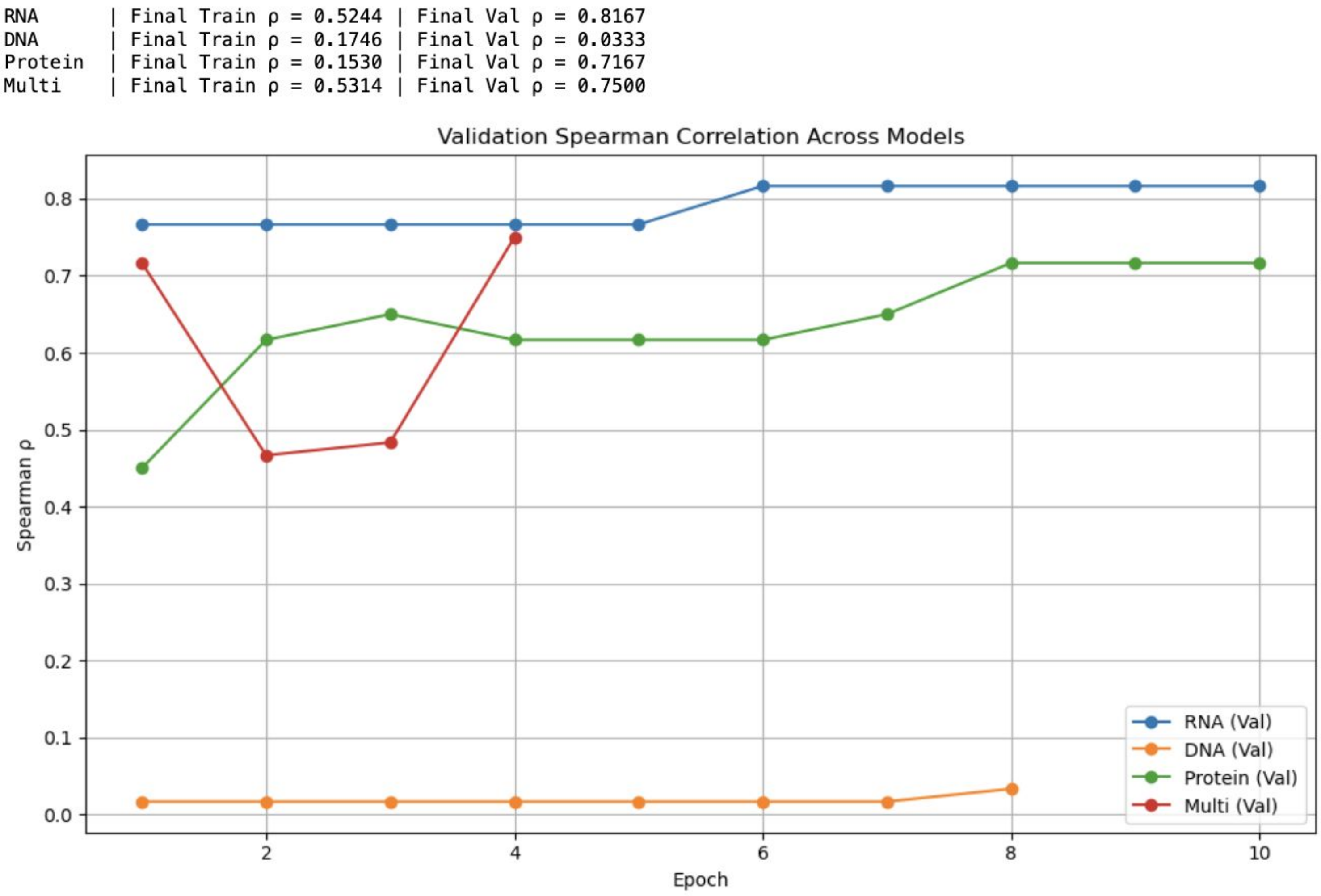
Each modality queries information from all others, learning:

- DNA ↔ RNA regulatory dependencies
- RNA ↔ protein translation relationships

inspired by transformer architectures



Uni Modals
and
MultiModals with
Concatenation
Fusion:



```
RNA      | Final Train ρ = 0.5244 | Final Val ρ = 0.8167
DNA      | Final Train ρ = 0.1746 | Final Val ρ = 0.0333
Protein  | Final Train ρ = 0.1530 | Final Val ρ = 0.7167
Multi    | Final Train ρ = 0.5314 | Final Val ρ = 0.7500
```

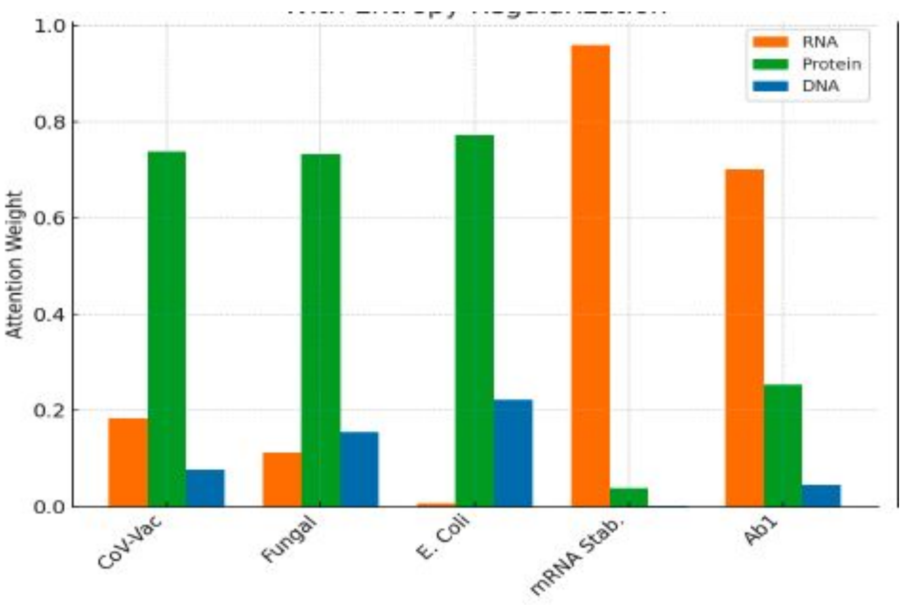
Step 2 — Add Interpretability Analysis

Show:

- modality weights
- codon-level heatmaps
- task-specific contribution patterns

Step 3 — Fine Tuning models

- Parameter-efficient training (LoRA)



Thanks for your attention!

Any questions?