ISOALIGN: ISOFORM-AWARE DNA-PROTEIN ALIGNMENT VIA JOINT-EMBEDDING PREDICTIVE ARCHITECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic foundation models (GFMs) have emerged as a powerful paradigm for learning DNA sequence representations, yet existing self-supervised pretraining strategies overlook alternative splicing and underperform compared to specialized models. To address this gap, we introduce IsoAlign, a multimodal framework that grounds genomic pretraining in isoform-aware DNA-protein alignment to learn biologically meaningful representations. IsoAlign aligns transcripts to their corresponding protein isoforms, capturing both coding sequences and broader regulatory contexts, using two Joint-Embedding Predictive Architectures (JEPA): DNA-Protein JEPA (DP-JEPA), which predicts protein representations from DNA using contrastive alignment, and Isoform-Masked JEPA (IM-JEPA), which models relationships among alternative isoforms of the same gene through masked protein prediction. We evaluate the learned DNA representations of these architectures across a multi-metric evaluation suite spanning tasks, including the prediction of pathogenicity, fitness, gene expression, and TF binding. Our results demonstrate that isoform-aware cross-modal alignment using JEPA significantly improves functional representations and downstream performance, establishing a promising direction for GFM pretraining.

1 INTRODUCTION

Genomic foundation models (GFMs) have emerged as a promising paradigm for learning high-dimensional representations of DNA sequences. Inspired by the success of large language models (LLMs) (Brown et al., 2020) and protein foundation models (PFMs) (Lin et al., 2023; Rives et al., 2021), GFMs leverage transformer-based architectures pretrained on large genomic corpora to learn the underlying regulatory syntax shaped by evolutionary constraints, enabling a unified feature space for diverse downstream biological tasks (Zhou et al., 2023; Nguyen et al., 2023; Schiff et al., 2024; Dalla-Torre et al., 2025).

However, current pretraining strategies for GFMs, which primarily leverage masked language modeling (MLM) and next-token prediction (NTP), fail to deliver consistent or meaningful performance gains over specialist models (Marin et al., 2023). In fact, previous studies report that randomly initialized models perform on par with, or even outperform, pretrained GFMs after fine-tuning (Schreiber et al., 2024; Zeng et al., 2023; Tang et al., 2025). This suggests that these standard self-supervised objectives fail to generate representations that adequately capture the biological complexity of genomic data. A key limitation is that they treat genomic sequences as flat token streams, without explicitly modeling the hierarchical and multi-stage processes through which genetic information is expressed.

A central bottleneck in connecting genomic variation to phenotypic function is the diversity introduced by alternative splicing. Although translation of a given coding sequence into its corresponding amino-acid sequence is deterministic under the genetic code, the relationship between a genomic locus and its protein products is not necessarily one-to-one. In eukaryotes, alternative splicing can generate multiple coding transcripts from the same gene, producing protein isoforms with distinct structural properties, subcellular localizations, and biochemical interaction partners (Kelemen et al., 2013). In humans, many genes express multiple transcript isoforms, with some loci producing dozens of distinct proteoforms (Aebersold et al., 2018). Thus, the challenge is not to rediscover the codon mapping itself, but to learn representations that capture biologically meaningful information shared across

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nucleotide and protein sequences despite the diversity of transcript and protein products associated with genomic loci.

Despite this biological centrality, alternative splicing remains a blind spot for existing GFMs. Because most models operate on fixed-length genomic windows or gene-level aggregates, they inherently overlook isoform-specific regulation. Consequently, current architectures struggle to resolve how subtle sequence variations propagate through the splicing machinery to alter functional outcomes. This highlights a critical architectural gap: while GFMs learn representations of genomic sequences and PFMs capture properties of protein sequences, the molecular correspondence between coding transcripts and their protein isoforms remains underexplored.

In this work, we propose ISOALIGN, a multimodal framework designed to learn isoform-aware genomic representations through cross-modal latent alignment. Rather than relying on token reconstruction losses and scaling alone, our approach utilizes transcript-isoform-aligned DNA-protein pairs and **Joint-Embedding Predictive Architectures (JEPA)** to explicitly model the functional consequences of genomic variation. By aligning DNA and protein representations that capture high-level biological properties beyond the deterministic codon-to-amino-acid mapping, we propose a pretraining strategy to move GFMs beyond pattern recognition toward functional discovery and precise variant interpretation.

Major Contributions

1. **ISOALIGN Framework:** We introduce a model-agnostic paradigm that grounds genomic pretraining in multimodal biological alignment rather than flat sequence reconstruction. By semantically mapping transcript representations to their corresponding protein isoforms, ISOALIGN explicitly captures the functional diversity introduced by alternative splicing across coding sequences and broader regulatory contexts.
2. **Isoform-Aware Predictive Architectures (JEPA):** We propose two complementary architectures: **DP-JEPA**, which predicts protein representations from DNA anchors in a shared latent space, and **IM-JEPA**, which predicts masked protein targets from visible isoform context to capture functional relationships among splice variants.
3. **Cross-Modal Objective Benchmarking:** Through architecture-matched controls, we quantify the impact of cross-modal alignment, demonstrating that JEPAs learn more biologically meaningful representations than conventional Masked Language Modeling (MLM) and Contrastive Learning (CL).
4. **Comprehensive Biological Evaluation:** We establish a benchmark spanning general and splicing-sensitive tasks, evaluating representational robustness in fitness prediction, splicing variant effect prediction, and tissue-specific gene expression.

2 PRELIMINARIES AND RELATED WORK

Limitations of Genomic Foundation Models. While scaling laws have dominated the development of PFMs and GFMs (Avsec et al., 2026; Zhou et al., 2023), recent evidence suggests that increasing parameter counts does not monotonically improve performance (Hou et al., 2025; Li et al., 2024). Moreover, large models impose prohibitive training and inference costs, which limits accessibility and iteration. Self-supervised objectives based on token reconstruction, such as Masked Language Modeling (MLM), often bias GFMs toward local sequence statistics at the expense of global functional constraints (Ye et al., 2025). Long-context capabilities have been positioned as a solution for capturing distal genomic dependencies; however, recent studies suggest that current models struggle to effectively exploit this expanded context (Yang et al., 2026; Cheng et al., 2025). These findings highlight a broader limitation: naive DNA-only self-supervised pretraining lacks the grounding necessary to map genomic variation to its functional downstream consequences, often resulting in representations that lack clinical relevance and sensitivity to genomic heterogeneity (Lu et al., 2025).

Moving Beyond Traditional Genomic Pretraining. Current pretraining strategies yield marginal improvements over randomly initialized baselines, with performance frequently influenced by tokenizer choice rather than the pretraining objective itself (Vishniakov et al., 2024; Bai et al., 2024). These limitations have motivated a shift toward objectives that encourage learned representations to capture biologically meaningful features beyond local sequence statistics (Consens et al., 2026; Ye et al., 2025). Moreover, training data composition plays a critical role in representation quality.

The SPACE framework (Zhang et al., 2024) demonstrates that pretraining on functionally annotated profiles improves generalization through a species-aware mixture-of-experts architecture. Contrastive learning (CL) has also emerged as a strong alternative to masking-based objectives. Orthrus leverages decoupled contrastive learning (DCL) with biologically motivated augmentations for RNA modeling (Fradkin et al., 2026), mirroring gains observed in protein sequence–structure alignment (Zhang et al., 2023) and protein–DNA binding prediction (Chen et al., 2025). More recently, joint-embedding predictive architectures (JEPA), including JEPA-DNA (Larey et al., 2026) and GenoJEPA (Wang et al., 2026), move beyond local nucleotide reconstruction toward latent-space prediction, encouraging representations grounded in functional semantics. However, despite these architectural advancements, overall performance gains remain modest, and these models largely fail to demonstrate meaningful improvements on complex, clinically relevant downstream tasks such as variant effect prediction (VEP). Collectively, these developments reflect a broader paradigm shift from conventional sequence-only masked objectives toward biologically informed and latent-aware formulations designed to learn richer and more transferable representations (He et al., 2025; Consens et al., 2026).

Multimodal Integration and Biological Alignment. Biological systems are inherently multimodal, operating across a hierarchical continuum of DNA (regulatory), RNA (transcriptional), and protein (functional) layers. Recent foundation models aim to unify these sequences into a shared latent space through diverse architectural strategies. ProTrek integrates protein language models with natural language and structural tokens (Su et al., 2025), while LucaOne trains a single transformer jointly on nucleotide and amino acid sequences to achieve cross-modal alignment (He et al., 2025). To bridge these modalities beyond simple translation, models like CDBridge utilize adaptive token-merging and tissue-specific conditioning to map genomic sequences to expression levels (Yu et al., 2026), and BioLangFusion aligns embeddings at the codon level for precise property prediction (Mollaysa et al., 2025). Complementing these approaches, RESM-2 leverages cross-modal distillation to transfer evolutionary knowledge from ProtESM-2 to RNA by mapping sequences into pseudo-protein representations, enabling protein-informed RNA modeling (Zhang et al., 2025). Despite these advances, a critical bottleneck persists: many architectures still treat genomic sequences as static inputs or rely on coarse gene-level associations, overlooking how genetic variation propagates through alternative splicing (Mitschka & Mayr, 2022) and isoform-specific regulation to contribute to disease (Scotti & Swanson, 2016).

3 DATASET

We construct DNA-protein training pairs from the Broad's Gene-to-Protein (G2P) database (Kwon et al., 2024), linking Ensembl transcripts to UniProt protein isoforms. For each gene, we extract genomic DNA (retaining intron/exon/UTR annotations), transcript regions (with CDS boundaries), and all associated protein isoforms. Pairs are generated via two alignment strategies outlined in Appendix A.1. We train each of our architectures using both alignment strategies and report the best-performing configuration.

4 METHOD

4.1 DOUBLE ENCODER FRAMEWORK

Our architecture employs a dual-encoder framework to jointly model DNA and protein sequences in a shared embedding space. Two modality-specific encoders generate fixed-dimensional representations, which are aligned through projection heads for cross-modal learning.

DNA Encoder (X-Encoder). The DNA encoder f_θ is a trainable transformer-based model that processes tokenized DNA sequences. Given an input sequence $\mathbf{x} \in \mathcal{V}_{\text{DNA}}^L$, it produces token-level hidden states $\mathbf{H}_x = f_\theta(\mathbf{x}) \in \mathbb{R}^{L \times D_x}$. We evaluate two backbone architectures: **NTv2-500m** (Dalla-Torre et al., 2025), a transformer pretrained on multi-species genomic data, and **AlphaGenome** (Avsec et al., 2026), a CNN-based model that operates on one-hot encoded sequences and produces chunk-level embeddings.

Protein Encoder (Y-Encoder). The protein encoder g_ϕ processes protein sequences into fixed representations. We employ a pretrained protein language model, **ESM2-150M** (Lin et al., 2023), trained on millions of protein sequences with masked language modeling. Given a protein sequence

$\mathbf{y}$, the encoder produces token-level hidden states $\mathbf{H}y = g_\phi(\mathbf{y}) \in \mathbb{R}^{M \times D_y}$. By default, the protein encoder is frozen (ϕ fixed) to provide stable prediction targets.

Pooling and Projection. Both encoders produce variable-length token representations that are aggregated via multi-head attention pooling with a learnable query vector:

$$\mathbf{z} = \text{MultiHeadAttn}(\mathbf{q}, \mathbf{H}, \mathbf{H}), \tag{1}$$

where $\mathbf{q} \in \mathbb{R}^D$ is a learnable query, $\mathbf{H} \in \mathbb{R}^{L \times D}$ are token representations, and we use 8 attention heads. Pooled representations are passed through 3-layer MLP projection heads with batch normalization and ReLU activations:

$$\mathbf{z}_x = \pi_x(\text{pool}(\mathbf{H}_x)) \in \mathbb{R}^{512}, \quad \mathbf{z}_y = \pi_y(\text{pool}(\mathbf{H}_y)) \in \mathbb{R}^{512}, \tag{2}$$

with hidden dimension 256. Embeddings are ℓ_2-normalized before similarity computation.

4.2 IsoAlign JEPA Architectures

4.2.1 DNA-Protein JEPA (DP-JEPA)

Architecture. We introduce DNA-Protein Joint-Embedding Predictive Architecture (DP-JEPA), a cross-modal self-supervised framework that learns to predict protein representations from DNA sequences in a shared latent space, outlined in Figure 1a. It consists of: (1) a trainable DNA encoder f_θ, (2) a protein encoder g_ϕ (frozen by default), and (3) a TransformerPredictor h_ψ that predicts protein embeddings from DNA representations.

The TransformerPredictor is a stack of transformer encoder layers (default: 4 layers, 8 heads) that processes the full sequence of DNA hidden states with attention masking, then mean-pools across valid positions:

$$\hat{\mathbf{z}}_y = \pi_{\text{pred}}\big(\text{LN}(\text{MeanPool}(h_\psi(\mathbf{H}_x, \mathbf{m}_x)))\big), \tag{3}$$

where $\mathbf{m}_x$ is the attention mask excluding padding tokens.

The target protein embedding applies stop-gradient to the encoder output while keeping the projection trainable:

$$\mathbf{z}_y = \pi_y\big(\text{sg}(\text{pool}(g_\phi(\mathbf{y})))\big). \tag{4}$$

Both π_{pred} and π_y are multi-layer projection heads mapping to a shared embedding space.

Training Objective. DP-JEPA employs contrastive alignment in the shared projection space. For a batch of genes, each potentially having multiple protein isoforms, we compute the loss per isoform slot and average:

$$\mathcal{L}_{\text{DP-JEPA}} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \mathcal{L}_{\text{contrast}}(\hat{\mathbf{Z}}_y, \mathbf{Z}_{y_j}; \tau), \tag{5}$$

where $\mathcal{V}$ denotes valid isoform slots (non-padded entries with ≥ 2 samples for contrastive discrimination). The target projection π_y is frozen under stop-gradient to provide stable targets.

4.2.2 Isoform-Masked JEPA (IM-JEPA)

Architecture. While DP-JEPA aligns each transcript-isoform pairs through cross-modal prediction, it does not explicitly model relationships among alternative isoforms of the same gene. To incorporate this intra-gene structure, we introduce Isoform-Masked JEPA (IM-JEPA), which jointly processes multiple transcript–isoform pairs and predicts masked protein isoforms from their DNA anchors and the remaining visible isoform context. Specifically, for genes with multiple transcript-isoform pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, we extend DP-JEPA with object-level masking inspired by Causal-JEPA (Nam et al., 2026), outlined in Figure 1b. Approximately 50% of isoforms are randomly masked per sample while all DNA anchors remain visible, ensuring at least one isoform stays visible for context. The predictor operates on *pooled* encoder representations and employs three types of learnable embeddings: (1) modality embeddings distinguishing DNA from protein isoforms, (2) slot embeddings indexing each pair, and (3) a learnable mask token $\mathbf{m}$. For masked isoforms, we inject *identity queries* derived from their DNA anchors via a learned projection ρ:

$$\mathbf{s}_i^{\text{iso}} = \begin{cases} \pi_{\text{iso}}(\mathbf{z}_{\mathbf{y}_i}) + \mathbf{e}_{\text{iso}} + \mathbf{e}_i & \text{if visible} \\ \mathbf{m} + \rho(\pi_{\text{dna}}(\mathbf{z}_{\mathbf{x}_i})) + \mathbf{e}_{\text{iso}} + \mathbf{e}_i & \text{if masked} \end{cases}, \tag{6}$$

where $\mathbf{z}_{\mathbf{x}_i} = \text{pool}(f_\theta(\mathbf{x}_i))$ and $\mathbf{z}_{\mathbf{y}_i} = \text{pool}(g_\phi(\mathbf{y}_i))$ are pooled encoder outputs, and π_{dna}, π_{iso} are input projections with layer normalization.

The full input sequence concatenates all DNA representations with isoform representations, processed by a bidirectional Transformer encoder:

$$\hat{\mathbf{Z}} = h_\psi\big([\pi_{\text{dna}}(\mathbf{z}_{\mathbf{x}}^{(1:N)}) + \mathbf{e}_{\text{dna}} + \mathbf{e}_{\text{slot}};\ \mathbf{s}^{(1:N)}]\big). \tag{7}$$

Training Objective and Regularization. For IM-JEPA, the loss is computed on masked and valid (non-padded) positions using ℓ_2-normalized embeddings:

$$\mathcal{L}_{\text{IM-JEPA}} = \frac{1}{|\mathcal{M} \cap \mathcal{V}|} \sum_{i \in \mathcal{M} \cap \mathcal{V}} \left\| \bar{\hat{\mathbf{z}}}_i - \bar{\mathbf{z}}_i \right\|_2^2 + \lambda_{\text{reg}} \mathcal{L}_{\text{SIGReg}}, \tag{8}$$

where $\bar{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$, $\mathcal{M}$ denotes masked positions, $\mathcal{V}$ denotes valid (non-padded) pairs, and $\mathcal{L}_{\text{SIGReg}}$ is an anti-collapse regularization term.

Specifically, SIGReg (Balestriero & LeCun, 2025) encourages predictions to follow an isotropic Gaussian $\mathcal{N}(0, \mathbf{I})$ via characteristic function matching. Predictions are projected onto K random unit vectors $\{\mathbf{p}_k\}_{k=1}^K$ (resampled each iteration) and compared to the Gaussian characteristic function $\varphi(t) = e^{-t^2/2}$:

$$\mathcal{L}_{\text{SIGReg}} = \frac{1}{K} \sum_{k=1}^{K} \int_0^{t_{\max}} w(t) \left[\left(\bar{c}_k(t) - e^{-t^2/2} \right)^2 + \bar{s}_k(t)^2 \right] dt, \tag{9}$$

where $\bar{c}_k(t) = \mathbb{E}[\cos(t \cdot \mathbf{p}_k^\top \hat{\mathbf{z}})]$ and $\bar{s}_k(t) = \mathbb{E}[\sin(t \cdot \mathbf{p}_k^\top \hat{\mathbf{z}})]$ are empirical characteristic function components, and $w(t) \propto e^{-t^2/2}$ weights errors by target magnitude. We use $K{=}1024$ projections with trapezoidal integration over $t \in [0, 3]$.

4.3 Architecture-Matched Pretraining Controls

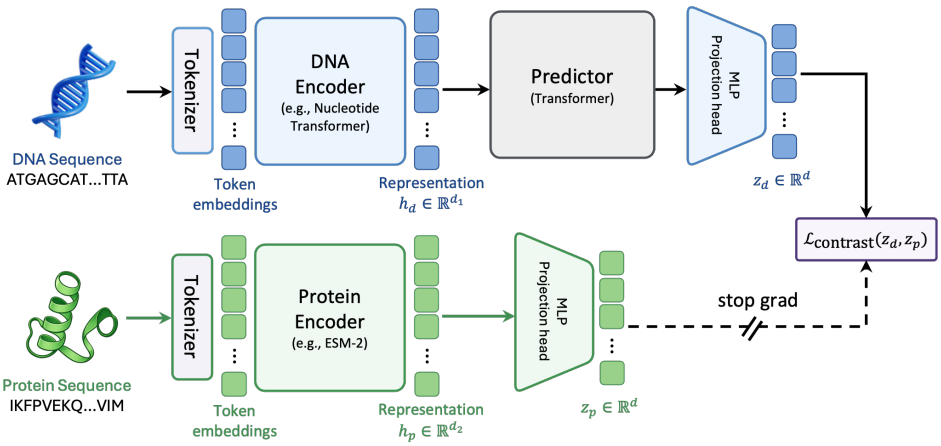
To isolate the performance gains of our JEPA objectives from the underlying dual-encoder architecture and increased parameter count, we implement several architecture-matched controls. These controls utilize the exact same DNA encoder, frozen protein encoder, and projection heads, but replace the JEPA objective with MLM or CL:

Masked Language Modeling (MLM). We implement four MLM configurations with a 15% masking rate. Cross-modal conditioning utilizes either **sequence-level** (global pooling) or **token-level** (position-specific) multi-head cross-attention (8 heads), with linear projections to align hidden dimensions between modalities.

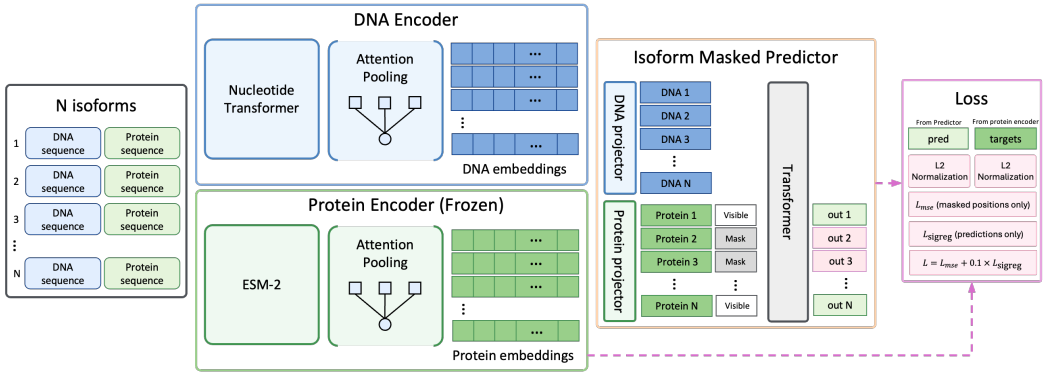
1. **Standard Joint:** Independent MLM with combined loss: $\mathcal{L}_{\text{Joint}} = \lambda_D \mathcal{L}_{\text{MLM}}^D + \lambda_P \mathcal{L}_{\text{MLM}}^P$.
2. **P→D:** Unmasked protein predicts masked DNA: $\hat{\mathbf{H}}_D = \text{LN}(\text{Attn}(\mathbf{H}_D, \mathbf{W}_P \mathbf{H}_P) + \mathbf{H}_D)$.
3. **D→P:** Unmasked DNA predicts masked protein: $\hat{\mathbf{H}}_P = \text{LN}(\text{Attn}(\mathbf{H}_P, \mathbf{W}_D \mathbf{H}_D) + \mathbf{H}_P)$.
4. **Bidirectional:** Both modalities are masked; combines (2) and (3) for mutually enhanced reconstruction.

Contrastive Learning. We align DNA-protein embeddings using three loss variants, governed by a biological correctness matrix $\mathbf{C} \in \{0, 1\}^{N \times N}$. All variants utilize a 8,192-entry momentum memory bank ($m = 0.999$) (He et al., 2020).

- **InfoNCE:** Maximizes matched-pair similarity using learnable temperature τ (init 0.5).
- **DCL:** Decoupled loss (Yeh et al., 2022) separating positive and negative terms for independent weighting: $\mathcal{L}_{\text{DCL}} = \lambda_{\text{pos}}(-\bar{s}_{\text{pos}}) + \lambda_{\text{neg}} \log \sum \exp(s_{\text{neg}})$.
- **CLIP:** Symmetric bidirectional NCE with low initial temperature ($\tau_0 = 0.1$, learnable) for sharp distribution alignment Radford et al. (2021).



(a) **DP-JEPA (DNA-Protein Joint-Embedding Predictive Architecture):** A dual-encoder framework where DNA sequences are processed by a trainable DNA encoder (e.g., NTv2) and projected through a Transformer predictor to align with protein representations from a frozen protein encoder (e.g., ESM2). The model learns cross-modal alignment via contrastive loss in a shared latent space.



(b) **IM-JEPA (Isoform-Masked JEPA):** An extension of DP-JEPA that jointly models multiple transcript-isoform pairs from the same genes. DNA sequences for all isoforms are encoded and pooled, while protein isoforms are randomly masked (∼50%). The Transformer Predictor uses DNA anchors as identity queries to predict masked protein embeddings, with ℓ_2-normalized MSE loss computed only on masked positions and SIGReg regularization to prevent collapse.

Figure 1: **IsoAlign JEPA Architectures**

5 Results

5.1 Deep Mutation Scan Fitness Prediction

We evaluated linear probing for fitness prediction on five DMS datasets from four human genes—*CBS*, *P53*, *PDE3A*, and *GDI1*—following Brixi et al. (2026) (Appendix A.2). Each dataset contains wild-type and mutant sequences with experimentally measured fitness scores. GFMs and IsoAlign are evaluated using DNA representations, while PFMs are evaluated separately using protein representations for comparison. We use group K-fold cross-validation stratified by amino acid mutation identity to prevent codon leakage, where synonymous codons encoding identical substitutions could otherwise appear across splits and artificially inflate GFM performance. We evaluate IsoAlign variants combining different GFMs (NTv2/AlphaGenome) with ESM2 under multiple pretraining objectives and initialization schemes. The best-performing models use our JEPA objectives (Figure 2), achieving median Spearman correlations of 0.557 ± 0.21 (DP-JEPA) and 0.554 ± 0.22 (IM-JEPA). Notably, DP-JEPA performs better with NTv2, while IM-JEPA favors AlphaGenome, and both substantially outperform their respective GFM baselines (DP-JEPA: +0.094 over NTv2; IM-JEPA: +0.136 over AlphaGenome) as well as architecture-matched cross-modal MLM and contrastive con-

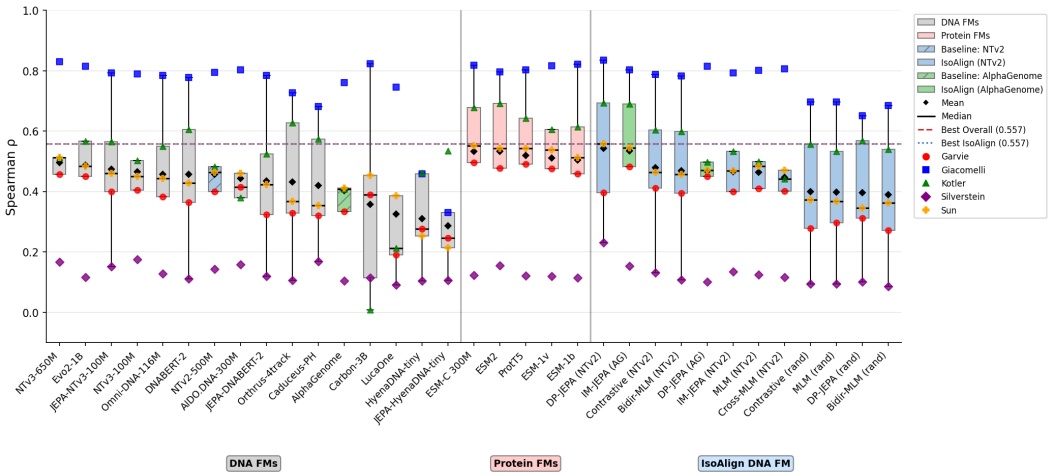


Figure 2: **DMS Fitness Prediction.** Spearman correlation between embedding-based variant scores and fitness scores on Deep Mutational Scanning (DMS) datasets. Models are grouped by category: DNA FMs (gray), Protein FMs (red), and ISOALIGN models (blue). ISOALIGN consistently outperforms unimodal DNA baselines, demonstrating that cross-modal pretraining with protein-derived inductive biases improves variant effect prediction without task-specific fine-tuning.

trols, demonstrating that isoform-aware alignment improves DNA representations. Across 16 GFMs and 5 PFMs, our best ISOALIGN model outperforms all baselines. Per-model setup and regression protocol are detailed in Appendix D.

5.2 PROBING DNA REPRESENTATIONS FOR SPLICING VARIANT EFFECTS

We evaluate variant effect prediction (VEP) from DNA representations, with the goal of distinguishing pathogenic variants (those that disrupt normal gene function) from benign ones. We focus on two splicing-related single-nucleotide variant (SNV) datasets: ClinVar Splicing, which contains clinically annotated splice-site variants, and MFASS (Massively Functional Assay of Splicing SNVs), a functional assay of splicing disruption (Chong et al., 2019); dataset details are provided in Appendix A.3. For each variant, we compare the wild-type and mutant representations, quantify the resulting perturbation using `cosine` and ℓ_2 distances, and measure performance using AUPRC under leave-one-chromosome-out (LOCO) cross-validation. Additional representation-based evaluations are reported in Appendix B, and per-model setup and evaluation protocols are detailed in Appendix D.

As shown in Table 1, cross-modal DNA-protein pretraining can increase the functional discriminability of the underlying DNA representation, although the magnitude of the improvement depends on the backbone and training objective. On ClinVar Splicing, the large-scale baselines Evo2 and Carbon achieve strong performance, consistent with their larger model capacity and extensive pretraining. Despite its smaller backbone, DP-JEPA performs best among our objectives, substantially improving NTv2-500M from 0.32 to 0.63 `cosine` AUPRC and AlphaGenome from 0.61 to 0.63. In contrast, existing JEPA-DNA adaptations do not consistently improve their parent DNA models and can even degrade performance, indicating that gains are not inherent to JEPA-based pretraining alone. MFASS presents a substantially more challenging setting, with only 1,050 pathogenic variants among 28,972 (3.6%), and correspondingly near-chance AUPRC values across the baseline models, including the larger models. Nevertheless, DP-JEPA produces substantial gains over its baseline representations: with NTv2-500M, ℓ_2 AUPRC increases from 0.04 to 0.17, while with AlphaGenome, `cosine` AUPRC increases from 0.29 to 0.36. IM-JEPA provides little improvement over NTv2-500M and underperforms the parent AlphaGenome representation. Overall, DP-JEPA consistently improves or maintains the functional sensitivity of its underlying DNA representations across these settings, whereas the effectiveness of IM-JEPA is more backbone-dependent.

5.3 SUPERVISED FINETUNING

We evaluate our model on two primary biological tasks: i) TF-DNA binding specificity (TFLink) (Liska et al., 2022) and ii) tissue-specific transcript expression (GTEx), using two fine-tuning

Table 1: Variant effect prediction from DNA representations, evaluated by AUPRC using two metrics. Values in parentheses indicate changes relative to the parent DNA model; underlined values indicate the best performance per column, and **bold** values are the best ISOALIGN model per backbone.

		ClinVar Splicing		MFASS	
Model	Params	`Cosine`	ℓ_2	`Cosine`	ℓ_2
DNABERT-2 (Zhou et al., 2023)	117M	0.14	0.15	0.04	0.04
JEPA-DNABERT-2 (Larey et al., 2026)	117M	0.14	0.15	0.04	0.04
HyenaDNA-tiny (Nguyen et al., 2023)	0.44M	0.16	0.15	0.04	0.04
JEPA-HyenaDNA-tiny	0.44M	0.15	0.16	0.04	0.04
Caduceus-PH (Schiff et al., 2024)	7.7M	0.20	0.18	0.04	0.04
LucaOne (He et al., 2025)	1.8B	0.15	0.16	0.05	0.06
Orthrus-4track (Fradkin et al., 2026)	10M	0.18	0.16	0.04	0.04
Orthrus-MLM	10.7M	0.18	0.19	0.04	0.04
Omni-DNA-116M (Li et al., 2025)	116M	0.15	0.15	0.04	0.04
AIDO.DNA-300M (Ellington et al., 2024)	300M	0.52	0.58	0.05	0.09
NTv3-100M (Boshar et al., 2025)	100M	0.51	0.66	0.10	0.12
JEPA-NTv3-100M	100M	0.47	0.55	0.12	0.09
Evo2-1B-base (Brixi et al., 2026)	1B	0.38	0.85	0.05	0.16
Carbon (Allal et al., 2026)	3B	0.63	0.90	0.05	0.08
NTv2-500M (Dalla-Torre et al., 2025)	500M	0.32	0.73	0.04	0.04
+ DP-JEPA (CLIP)	520M	**0.63** (+.31)	0.67	0.08 (+.04)	**0.17** (+.13)
+ DP-JEPA (DCL)	520M	0.39 (+.07)	0.47	0.05 (+.01)	0.06 (+.02)
+ DP-JEPA (InfoNCE)	520M	0.58 (+.26)	0.55	**0.07** (+.03)	0.12 (+.08)
+ IM-JEPA	520M	0.35 (+.03)	**0.74** (+.01)	0.04	0.05 (+.01)
AlphaGenome (Avsec et al., 2026)	450M	0.61	0.57	0.29	0.31
+ DP-JEPA (CLIP)	470M	**0.63** (+.02)	**0.60** (+.03)	**0.36** (+.07)	0.31
+ IM-JEPA	470M	0.33	0.30	0.06	0.09

strategies: i) Linear Probing and ii) Parameter Efficient Fine-Tuning (LoRA), described in Appendix C. Per-baseline fine-tuning configurations are detailed in Appendix D.

TFLink. We benchmark on the TFLink dataset using a hard negative split with 35% within-family negatives, specifically designed to assess discrimination between structurally similar TFs with overlapping sequence motifs. Details of the dataset construction, sampling procedure, and TF-family composition are provided in Appendix A.4. As shown in Table 2, frozen representations alone underperform, suggesting TF binding specificity requires task adaptation beyond generic sequence embeddings. In contrast, LoRA fine-tuning substantially improves performance, outperforming prior GFM and specialist baselines. These results suggest that isoform-aware representations can capture regulatory sequence features highly predictive of TF-DNA interactions.

Table 2: TFLink transcription factor binding prediction (mean $\pm$ SD over 4 training seeds).

Model	**Training**	**Params**	**AUROC**	**AUPRC**
LucaOne (He et al., 2025)	LoRA	1.8B	0.785 ± 0.001	0.778 ± 0.002
BPNet (Avsec et al., 2021)	Full	0.12M	0.861 ± 0.004	0.851 ± 0.003
ChromBPNet (Pampari et al., 2023)	Full	6M	0.867 ± 0.005	0.851 ± 0.007
AlphaGenome (Avsec et al., 2026)	Full	450M	0.892 ± 0.002	0.879 ± 0.003
ISOALIGN	LoRA	650M	$\mathbf{0.934 \pm 0.001}$	$\mathbf{0.928 \pm 0.001}$

GTEx. We use expression data from five tissues in GTEx (Lonsdale et al., 2013): Brain, Heart, Kidney, Liver, and Stomach (N=445,888 records across 19,587 unique genes). We implement gene-level splitting (8:1:1), ensuring that all transcripts from a given gene remain within the same train, validation, or test partition to prevent data leakage. As shown in Table 3, ISOALIGN achieves the best overall performance, outperforming both LucaOne and AlphaGenome across most tissues. ISOALIGN attains the highest overall Spearman correlation ($\rho = 0.725$) and Pearson correlation ($r = 0.738$), with particularly strong gains in Kidney and Liver tissue prediction, demonstrating the benefit of protein-informed genomic representations for tissue-specific expression modeling.

Table 3: GTEx tissue-specific expression prediction.

Model	Overall		Brain		Heart		Kidney		Liver		Stomach	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
LucaOne	0.432	0.430	0.448	0.437	0.456	0.455	0.383	0.380	0.426	0.417	0.453	0.453
AlphaGenome	0.717	0.732	0.702	0.729	0.732	0.743	0.740	0.742	0.683	0.704	0.719	0.737
IsoAlign	**0.725**	**0.738**	0.716	0.739	0.735	0.744	0.745	0.747	0.697	0.715	0.722	0.736

5.4 Sequence Ablation Analysis

To show that IsoAlign learns beyond a codon lookup table, we systematically ablated input sequences and evaluated cross-modal retrieval (Figure 3; Appendix E). If IsoAlign relied on deterministic codon-to-amino-acid translation, codon-shuffled sequences (which preserve the encoded protein while altering nucleotide positions) should match or exceed the CDS baseline, while frame-shifted sequences (which produce an entirely different amino acid sequence) should fail entirely. We observe the opposite: codon shuffling degrades retrieval by 19%, whereas frame shifting degrades only 7%, demonstrating that IsoAlign captures positional codon dependencies, including codon usage bias, translational efficiency signals, and local sequence context, beyond simple translation lookup. Training exclusively on untranslated regions (UTRs), which contain no coding information, retains 65–78% of CDS-baseline performance far above the random DNA baseline (6%), confirming that IsoAlign learns regulatory sequence features that correlate with protein function independently of the genetic code. Finally, isoform-swapped negatives (wrong isoform, same gene) retain 79% of baseline retrieval, whereas homolog-swapped negatives (protein from a related gene family) degrade to 39%, indicating that IsoAlign distinguishes subtle isoform-level differences while recognizing that homologs, despite sequence similarity, have distinct functional contexts.

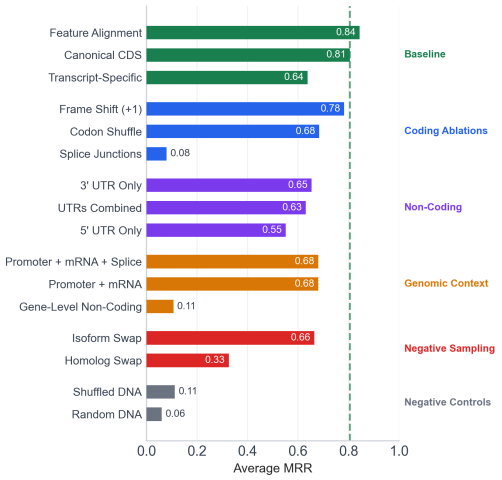


Figure 3: Cross-modal retrieval (Avg MRR) by ablation category. Dashed line: CDS baseline.

6 Conclusion

We introduced IsoAlign, a model-agnostic framework for genomic pretraining driven by isoform-aware DNA-protein alignment. By explicitly modeling the functional diversity of alternative splicing via two complementary JEPA architectures (DP-JEPA and IM-JEPA), the framework moves beyond flat sequence reconstruction. Our evaluations demonstrate that isoform-aware cross-modal prediction improves functional genomic representations, with DP-JEPA providing consistent gains across fitness and variant effect prediction, while IM-JEPA shows more task- and backbone-dependent improvements. Supervised evaluations further demonstrate strong performance in transcription factor binding and tissue-specific expression prediction. Crucially, sequence ablations confirm that these learned representations capture true regulatory syntax rather than merely memorizing deterministic codon mappings. Ultimately, this establishes biologically grounded, cross-modal latent alignment as a highly effective paradigm for GFMs.

Limitations and Future Work. IsoAlign is limited by the completeness of DNA-protein databases and curated benchmarks, which may not capture context-specific splicing or real-world patient variation. Future work will evaluate patient-derived cohorts and incorporate epigenomic and structural information. Clinical translation will additionally require safeguards against population bias and privacy vulnerabilities.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with code correction and improvement of the flow and composition of text originally written by the authors, including checking grammar, spelling, and consistency of terminology and notation across sections. We have not used generative AI tools for the conception of the research, formulation of theoretical arguments or derivations. All AI-assisted code was reviewed and tested by the authors for correctness, and no AI-assisted text was incorporated into the manuscript without being read, edited, and verified by a human co-author. All references were independently checked against their primary sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work relies exclusively on publicly available datasets, pretrained models, and computational methods. It does not involve human subjects, personally identifiable information, or private or sensitive data.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility of our experiments. For the datasets used in our experiments, Appendix A provides a detailed description of training and evaluation dataset construction and preprocessing, including DNA-protein pair alignment and dataset-specific filtering procedures. For model training and evaluation, Appendix C describes the supervised fine-tuning protocols, and Appendix D provides model-specific configurations and evaluation procedures for all benchmarked models. For the ablation experiments, Appendix E describes the construction of each ablated sequence and its evaluation procedure. Finally, Appendix F reports the computational resources, training configurations, hyperparameters, and compute requirements used in our experiments.

REFERENCES

Ruedi Aebersold et al. How many human proteoforms are there? *Nature Chemical Biology*, 14(3): 206–214, 2018.

Loubna Ben Allal, Qiuyi Li, Maurizio Fiusco, Lewis Tunstall, Kashif Rasul, Ed Beeching, Dana Aubakirova, Carlos Patiño, Thibaud Frere, Anton Lozhkov, et al. Carbon: Decoding the language of life. *bioRxiv*, pp. 2026–05, 2026.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics*, 53(3):354–366, 2021.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218, 2026. doi: 10.1038/s41586-025-10014-0.

Weiqiang Bai, Lifeng Qiao, Yuchen Ren, Jianle Sun, Peng Ye, Hongliang Yan, Xinzhu Ma, Wangmeng Zuo, Wanli Ouyang, et al. Rethinking the bert-like pretraining for dna sequences. 2024.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Gonzalo Benegas, Gökcen Eraslan, and Yun S. Song. Benchmarking dna sequence models for causal regulatory variant prediction in human genetics. *bioRxiv*, 2025.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K. Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, et al. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, pp. 2025–12, 2025.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652:1349 – 1361, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Leo Tianlai Chen, Rishab Pulugurta, Pranay Vure, and Pranam Chatterjee. DPAC: Prediction and design of protein-dna interactions via sequence-based contrastive learning. *bioRxiv*, pp. 2025–05, 2025.

Wenduo Cheng, Zhenqiao Song, Yang Zhang, Shike Wang, Danqing Wang, Muyu Yang, Lei Li, and Jian Ma. DNALONGBENCH: a benchmark suite for long-range dna prediction tasks. *Nature Communications*, 16(1):10108, 2025.

Rockie Chong, Kimberly D Insigne, David Yao, Christina P. Burghard, Jeffrey Wang, Yun-Hua E. Hsiao, Eric M. Jones, Daniel B. Goodman, Xinshu Xiao, and Sriram Kosuri. A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. *Molecular cell*, 73(1):183–194, 2019.

Micaela Elisa Consens, Kevin K. Yang, Jimmy Hall, Ashley Mae Conard, Bo Wang, Lorin Crawford, Alan Moses, and Alex X. Lu. Predicting evolutionary rate as a pretraining task improves genome language model representations. *bioRxiv*, pp. 2026–02, 2026.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. De Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

Caleb N. Ellington, Ning Sun, Nicholas Ho, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P Xing. Accurate and general dna representations emerge from genome foundation models at scale. *bioRxiv*, pp. 2024–12, 2024.

Philip Fradkin, Ruian "Ian" Shi, Taykhoom Dalal, Keren Isaev, Brendan J. Frey, Leo J. Lee, Quaid Morris, and Bo Wang. Orthrus: toward evolutionary and functional rna foundation models. *Nature Methods*, pp. 1–11, 2026.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, et al. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, 7(6):942–953, 2025.

Chao Hou, Di Liu, Aziz Zafar, and Yufeng Shen. Understanding language model scaling on protein fitness prediction. *bioRxiv*, 2025.

Ondrej Kelemen et al. Function of alternative splicing. *Gene*, 514(1):1–30, 2013.

Seulki Kwon, Jordan Safer, Duyen T. Nguyen, David Hoksza, Patrick May, Jeremy A. Arbesfeld, Alan F. Rubin, Arthur J. Campbell, Alex Burgin, and Sumaiya Iqbal. Genomics 2 Proteins portal: a resource and discovery tool for linking genetic screening outputs to protein sequences and structures. *Nature Methods*, 21(10):1947–1957, 2024.

Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2018.

Ariel Larey, Elay Dahan, Amit Bleiweiss, Raizy Kellerman, Guy Leib, Omri Nayshool, Dan Ofer, Tal Zinger, Dan Dominissini, Gideon Rechavi, et al. Jepa-dna: Grounding genomic foundation models through joint-embedding predictive architectures. *arXiv preprint arXiv:2602.17162*, 2026.

Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex X. Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, pp. 2024–02, 2024.

Zehui Li, Vallijah Subasri, Yifei Shen, Dongsheng Li, Yiren Zhao, Guy-Bart Stan, and Caihua Shan. Omni-dna: A unified genomic foundation model for cross-modal and multi-task learning. *arXiv preprint arXiv:2502.03499*, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Orsolya Liska, Balázs Bohár, András Hidas, Tamás Korcsmáros, Balázs Papp, Dávid Fazekas, and Eszter Ari. TFLink: an integrated gateway to access transcription factor–target gene interactions for multiple species. *Database*, 2022:baac083, 2022.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature genetics*, 45(6):580–585, 2013.

Baiyu Lu, Xueshen Liu, Po-Yu Lin, and Nadav Brandes. Genomic heterogeneity inflates the performance of variant pathogenicity predictions. *bioRxiv*, pp. 2025–09, 2025.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. Bend: Benchmarking dna language models on biologically meaningful tasks. *arXiv preprint arXiv:2311.12570*, 2023.

Sibylle Mitschka and Christine Mayr. Context-specific regulation and function of mrna alternative polyadenylation. *Nature Reviews Molecular Cell Biology*, 23(12):779–796, 2022.

Amina Mollaysa, Artem Moskale, Pushpak Pati, Tommaso Mansi, Mangal Prakash, and Rui Liao. Biolangfusion: Multimodal fusion of dna, mrna, and protein language models. *arXiv preprint arXiv:2506.08936*, 2025.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent interventions. *arXiv preprint arXiv:2602.11389*, 2026.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in Neural Information Processing Systems*, 36:43177–43201, 2023.

Anusri Pampari, Anna Shcherbina, Surag Nair, Jacob Schreiber, Aman Patel, Austin Wang, Soumya Kundu, Avanti Shrikumar, and Anshul Kundaje. Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *Zenodo*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alexander Rives, Joshua Meier, Tom Sbiroli, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Deshpande, Myle Ott, Mark Slowik, Baptiste Almeida, et al. Biological structure and function emerged from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences (PNAS)*, 118(15):e2016239118, 2021.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proceedings of Machine Learning Research*, 235:43632, 2024.

Jacob Schreiber et al. GENA-LM: A family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 2024.

Marina M. Scotti and Maurice S Swanson. Rna mis-splicing in disease. *Nature Reviews Genetics*, 17 (1):19–32, 2016.

Jin Su, Yan He, Shiyang You, Shiyu Jiang, Xibin Zhou, Xuting Zhang, Yuxuan Wang, Xining Su, Igor Tolstoy, Xing Chang, et al. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, pp. 1–7, 2025.

Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K Koo. Evaluating the representational power of pre-trained dna language models for regulatory genomics. *Genome Biology*, 26(1):203, 2025.

Kirill Vishniakov, Karthik Viswanathan, Aleksandr Medvedev, Praveen K. Kanithi, Marco AF. Pimentel, Ronnie Rajan, and Shadab Khan. Genomic foundationless models: Pretraining does not promise performance. *bioRxiv*, pp. 2024–12, 2024.

Chengsen Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Bo He, Siying Liu, Jianxin Liao, and Jingyu Wang. From nucleotides to semantics: genomic representation learning via joint-embedding predictive architecture. *bioRxiv*, pp. 2026–04, 2026.

Zhao Yang, Yi Duan, Jiwei Zhu, Ying Ba, Chuan Cao, and Bing Su. Extending sequence length is not all you need: Effective integration of multimodal signals for gene expression prediction. *arXiv preprint arXiv:2602.21550*, 2026.

Chengzhong Ye, Gonzalo Benegas, Carlos Albors, Jianan Canal Li, Sebastian Prillo, Peter D. Fields, Brian Clarke, and Yun S. Song. Predicting functional constraints across evolutionary timescales with phylogeny-informed genomic language models. *bioRxiv*, 2025.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European conference on computer vision*, pp. 668–684. Springer, 2022.

Chang Yu, Siyuan Li, Zicheng Liu, Jingbo Zhou, Xianglong Guo, Kai Yu, Yuqing Zhou, Ken Li, Zelin Zang, Zhen Lei, et al. CDBridge: A cross-omics post-training bridge strategy for context-aware biological modeling. In *The Fourteenth International Conference on Learning Representations*, 2026.

Wei Zeng et al. Are genomic foundation models truly better? a systematic evaluation of pretraining strategies. *bioRxiv*, 2023. doi: 10.1101/2023.10.25.564012.

Y. Zhang et al. SPACE: Pretraining on functionally annotated genomic profiles with mixture-of-experts. *Nature Machine Intelligence*, 2024.

Yikun Zhang, Hao Zhang, Guo-Wei Li, He Wang, Xing Zhang, Xu Hong, Tingting Zhang, Liangsheng Wen, Yu Zhao, Jiuhong Jiang, et al. RESM: Capturing sequence and structure encoding of rnas by mapped transfer learning from ESM (evolutionary scale modeling) protein language model. *bioRxiv*, pp. 2025–08, 2025.

Zuobai Zhang, Minghao Xu, Aurelie Chen, Jiarui Lu, Riheng Chen, Xinyu Liu, and Jian Tang. Structure-aware protein self-supervised learning. *Bioinformatics*, 39(Supplement_1):i43–i52, 2023.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V. Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023. doi: 10.48550/arXiv.2306.15006.

APPENDIX

A DATASET CONSTRUCTION AND STATISTICS

A.1 TRAINING DATA ALIGNMENT

We construct DNA-protein training pairs using two complementary alignment strategies that differ in the granularity of sequence correspondence.

Transcript-Specific. Pairs spliced transcripts (exons and UTRs) with CDS-sliced protein subsequences. DNA chunks of 4,608 bp include explicit markers denoting 5'UTR, CDS, and 3'UTR boundaries, with only the corresponding protein region extracted for each chunk.

Feature-Aligned (CDS-Only). Pairs coding sequences directly with protein counterparts using strict 3:1 codon-to-amino-acid mapping. Exon boundaries are preserved across both modalities to facilitate learning of splice junction patterns. DNA chunks of 3,072 bp are paired with up to 1,024 amino acid tokens.

A.2 DEEP MUTATIONAL SCANNING (DMS)

The DMS evaluation dataset measures mutation effects on protein function using experimental fitness scores. Each sample contains a mutated protein sequence, its corresponding DNA sequence, and a fitness score derived from high-throughput functional assays.

Table 4: Deep Mutational Scanning dataset statistics.

Study	Protein	Samples	Protein (aa)	DNA (bp)	Value Range
Kotler et al. 2018	TP53	2,967	393	1,182	-4.99–1.51
Giacomelli et al. 2018	TP53	7,448	393	1,182	-4.37–2.84
Sun et al. 2020	CBS	6,757	551	1,656	0.00–1.00
Garvie et al. 2021	PDE3A	7,893	1,141	3,426	-3.83–2.47
Silverstein et al. 2021	GDI1	4,365	447	1,344	-7.33–3.09
Total		**29,430**			

A.3 VARIANT EFFECT PREDICTION

ClinVar is a widely-used clinical variant database containing expert-curated pathogenicity labels for human genetic variants Landrum et al. (2018). We focus on ClinVar SNVs within splice-site regions in the immediate vicinity of exon–intron junctions, where conserved splice donor (5') and acceptor (3') motifs regulate RNA splicing. Even single-nucleotide perturbations near these sites can disrupt splicing fidelity and lead to severe functional consequences, including exon skipping or intron retention. Following Avsec et al. Avsec et al. (2026), we define the splice-site region as the interval spanning 6 bp into the intron and 3 bp into the exon at each annotated junction, using GENCODE v46 as the transcript reference. We include both donor- and acceptor-proximal variants and restrict the dataset to autosomal and sex-chromosome SNVs with unambiguous ClinVar annotations (Pathogenic/Likely pathogenic or Benign/Likely benign), excluding variants with conflicting or uncertain interpretations. Variants are filtered from the ClinVar GRCh38 release, resulting in a dataset containing 7,028 pathogenic and 41,535 benign SNVs.

MFASS (Massively Functional Assay of Splicing SNVs) provides experimentally grounded measurements of splicing disruption, making it a particularly reliable benchmark for VEP. Unlike clinically annotated datasets, MFASS variants were directly assayed using a massively parallel minigene reporter system (SORT-seq) Chong et al. (2019), which quantifies the splicing consequences of thousands of single-nucleotide variants in a controlled cellular context. Variants are labeled as positive if they induce a strong loss-of-function splicing effect, and negative otherwise. Following the benchmark protocol of Avsec et al. Avsec et al. (2026), we retain only SNVs with unambiguous reference and alternate alleles and exclude any sequences containing ambiguous nucleotides. The dataset is derived from the publicly available MFASS repository, enabling a direct evaluation of whether models can detect functionally disruptive splicing variants independent of clinical or population-level annotations.

A.4 TFLINK: TF-DNA BINDING PREDICTION

We construct an evaluation dataset from TFLink Liska et al. (2022), a database of 8.8M experimentally-validated human transcription factor binding sites. The task is binary classification: given a TF protein sequence and a 1024 bp genomic DNA sequence centered on a potential binding site, predict whether binding occurs.

Dataset Construction. Positive examples pair each TF with its validated binding sites. Negative examples pair each TF with binding sites of other TFs—including structurally similar TFs from the same family—creating hard negatives that require the model to learn TF-specific binding preferences rather than generic DNA accessibility. We apply chromosome-stratified sampling with a minimum 10 kb distance between sites to ensure sample independence. The dataset contains 200,000 samples (100,000 positive, 100,000 negative) across 20 TFs from 10 TF families. Table 5 summarizes the per-TF statistics.

Table 5: TFLink Evaluation Dataset

Stratification	Categories	Samples
Binding Status	Positive	100,000
	Negative (same-family)	100,000
Binding Site Length	Short ($\leq$12 bp)	111,982
	Medium (13–20 bp)	78,807
	Long ($>$20 bp)	9,211
TF Family	TF_bZIP	70,000
	zf-GATA	20,000
	zf-C2H2	20,000
	Fork_head	20,000
	bHLH	20,000
	Other (5 families)	50,000

B ADDITIONAL VARIANT EFFECT PREDICTION EVALUATIONS

To evaluate the predictive power of dna representations, we compare wildtype and mutant embeddings through four representation-based metrics. This includes two **scalar metrics**: `cosine` distance at mutated positions and the ℓ_2 norm of the embedding difference, assessed using AUPRC. Additionally, we employ two **vector metrics**: element-wise embedding products (`Mul`) Benegas et al. (2025) and (`CLS`) token differences, evaluated through leave-one-chromosome-out (LOCO) cross-validation logistic regression. For models without a `[CLS]` token, the `CLS` feature is taken from the final position for autoregressive models and from the first position otherwise.

C SUPERVISED FINETUNING

Linear Probing Both the DNA encoder and Protein encoder are frozen. Features are extracted via mean pooling and concatenated into a $d = 1664$ representation.

Parameter-Efficient Fine-Tuning Low-Rank Adaptation (LoRA) was applied to the DNA encoder's query and value attention layers ($r = 4$). This configuration introduces ~ 1.4M trainable adapter parameters while maintaining a frozen Protein encoder, totaling ~ 2.4M trainable parameters including the head.

Prediction Head For both tasks, we utilize a standardized MLP head:

$$f(\mathbf{h}) = \mathbf{W}_3\phi\left(\mathbf{W}_2\phi\left(\mathbf{W}_1\text{LN}(\mathbf{h})\right)\right) \tag{10}$$

where LN is Layer Normalization and ϕ denotes GELU activation followed by Dropout ($p = 0.2$). For GTEx regression, a 64-dimensional tissue embedding is concatenated to the pooled features before the MLP. Ultra-long sequences are managed via 6-mer tokenization with a 512-token truncation cutoff.

Table 6: Variant effect prediction evaluation on ClinVar Splicing and MFASS using four metrics. Underlined values indicate the best performance per column, and **bold** values are the best IsoAlign model per backbone.

		ClinVar Splicing				MFASS			
Model	Params	Cosine	ℓ_2	CLS	Mul	Cosine	ℓ_2	CLS	Mul
Baselines									
DNABERT-2 (Zhou et al., 2023)	117M	0.14	0.15	0.18	0.18	0.04	0.04	0.05	0.04
JEPA-DNABERT-2 (Larey et al., 2026)	117M	0.14	0.15	0.27	0.19	0.04	0.04	0.08	0.04
HyenaDNA (Nguyen et al., 2023)	6.6M	0.17	0.18	0.25	0.18	0.04	0.04	0.05	0.04
HyenaDNA-tiny	0.44M	0.16	0.15	0.22	0.17	0.04	0.04	0.04	0.04
JEPA-HyenaDNA-tiny	0.44M	0.15	0.16	0.19	0.17	0.04	0.04	0.04	0.03
Caduceus-PH (Schiff et al., 2024)	7.7M	0.20	0.18	0.19	0.19	0.04	0.04	0.05	0.05
Caduceus-PS	7.7M	0.20	0.45	0.42	0.42	0.04	0.08	0.13	0.10
LucaOne (He et al., 2025)	1.8B	0.15	0.16	0.18	0.17	0.05	0.06	0.09	0.07
LucaOne-gene	1.8B	0.15	0.17	0.18	0.17	0.05	0.06	0.08	0.06
Orthrus-4track (Fradkin et al., 2026)	10M	0.18	0.16	0.14	0.16	0.04	0.04	0.04	0.04
Orthrus-MLM	10.7M	0.18	0.19	0.15	0.19	0.04	0.04	0.04	0.04
Omni-DNA-116M (Li et al., 2025)	116M	0.15	0.15	0.55	0.21	0.04	0.04	0.10	0.06
Omni-DNA-300M	300M	0.15	0.14	0.49	0.21	0.04	0.04	0.13	0.06
AIDO.DNA-300M (Ellington et al., 2024)	300M	0.52	0.58	0.52	0.52	0.05	0.09	0.12	0.07
NTv3-650M (Boshar et al., 2025)	650M	0.35	0.57	0.34	0.28	0.04	0.07	0.08	0.07
NTv3-100M	100M	0.51	0.66	0.22	0.21	0.10	0.12	0.07	0.08
JEPA-NTv3-100M	100M	0.47	0.55	0.25	0.20	0.12	0.09	0.08	0.08
Evo2-1B-base (Brixi et al., 2026)	1B	0.38	0.85	0.59	0.35	0.05	0.16	0.16	0.10
Evo2-7B-base	7B	0.57	0.88	0.72	0.36	0.05	0.09	0.13	0.13
Carbon (Allal et al., 2026)	3B	0.63	0.90	0.65	0.75	0.05	0.08	0.09	0.06
NTv2-500M (Dalla-Torre et al., 2025)	500M	0.32	0.73	0.52	0.42	0.04	0.04	0.05	0.04
+ DP-JEPA (CLIP)	520M	**0.63**	0.67	0.47	**0.46**	**0.08**	**0.17**	**0.18**	**0.08**
+ DP-JEPA (DCL)	520M	0.39	0.47	0.49	0.24	0.05	0.06	0.10	0.06
+ DP-JEPA (InfoNCE)	520M	0.58	0.55	**0.59**	0.25	0.07	0.12	**0.18**	0.06
+ IM-JEPA	520M	0.35	**0.74**	**0.59**	**0.46**	0.04	0.05	0.07	0.04
AlphaGenome (Avsec et al., 2026)	450M	0.61	0.57	0.63	0.30	0.29	0.31	0.32	0.23
+DP-JEPA (CLIP)	470M	**0.63**	**0.60**	0.61	0.29	**0.36**	0.31	**0.36**	0.21
+IM-JEPA	470M	0.33	0.30	0.57	0.20	0.06	0.09	0.21	0.13
Architecture-matched pretraining controls (NTv2-500M)									
MLM	520M	0.27	0.73	0.59	0.39	0.04	0.04	0.06	0.04
Conditioned MLM	520M	0.27	0.73	0.53	0.41	0.04	0.04	0.05	0.04
Bidirectional TCA	520M	0.50	0.61	0.63	0.49	0.04	0.04	0.04	0.04
Cross-Modal	520M	0.52	0.48	0.42	0.43	0.04	0.05	0.05	0.05
Contrastive (CLIP)	520M	0.40	0.70	0.47	0.50	0.04	0.07	0.06	0.06
Contrastive (DCL)	520M	0.47	0.64	0.39	**0.57**	0.04	0.05	0.05	0.05
Contrastive (InfoNCE)	520M	0.39	0.69	0.46	0.52	0.05	0.09	0.06	0.05

D Implementation Details for Benchmarked Models

This section catalogues per-model setup decisions that diverge from a naïve Hugging Face load, plus the scoring conventions used across DMS, VEP, TFLink, and GTEx. Every divergence below is forced by an architectural constraint of the model in question (context cap, tokenizer choice, missing head, attention kernel), not by a tunable hyperparameter.

D.1 Per-Model Setup

AIDO.DNA-300M (Ellington et al., 2024). Loaded via the `modelgenerator` package (the model's reference RNABert backbone is not exposed through the standard `AutoModel` routing). DMS and all VEP tracks use the model's native context L=3998 (`max_pos_embeddings`=4000); because RNABert uses eager attention and materialises the full $[B, h, L, L]$ probability tensor, we run with batch size 8.

Caduceus-PH / Caduceus-PS (Schiff et al., 2024). Bidirectional Mamba SSMs requiring a separate environment (`mamba_ssm` 2.2.4, `causal_conv1d` 1.5.0). Single-nt character tokenizer with no `[CLS]` token; pooled representations are mean-pooled across positions.

ChromBPnet (Pampari et al., 2023). The reference TensorFlow/Keras stack pins old versions incompatible with our PyTorch environment, so we load weights through `bpnet-lite`, a PyTorch

port that reproduces the Keras `.h5` forward bit-equivalently. The K562-pretrained checkpoint underperformed by ∼0.10 AUROC across all fine-tuning recipes we tried; we report the same architecture trained from scratch.

BPNet (Avsec et al., 2021). A 64-filter, 8-layer dilated CNN trained from scratch. We follow the original architecture (motif kernel 21, dilated residual kernels 3 with dilations $2^1, \ldots, 2^8$) and pair the DNA trunk with the same frozen ESM2-150M protein encoder and MLP head used for ChromBPnet. The BPNET (FULL) row matches the small-CNN supervised baseline of Avsec et al. (2021).

DNABERT-2 (Zhou et al., 2023). Run with the vanilla-attention kernel (Triton kernel disabled): the Triton path produces non-identical representations under the same weights, so we standardise on the vanilla implementation for all reported numbers. The model's BPE tokenizer means a single-nucleotide variant changes the token identity at the variant position rather than a single position within a token, which precludes masked-marginal LLR scoring at single-nt resolution.

Evo2-1B / 7B (Brixi et al., 2026). The two variants need different installs: Evo2-7B uses the Light path (`vtx` 1.0.8, bf16, no `transformer_engine`), while Evo2-1B sets `use_fp8_input_projections=True` and requires `transformer_engine` $\geq$ 2.3, which we build from source in a separate environment (PyPI's TE meta-wheel pins a `transformer_engine_cu12` version not published on PyPI). DMS scoring routes through the SDK's `score_sequences(reduce_method='sum')` for whole-sequence LLR because the vortex `CharLevelTokenizer` does not expose an HF-style `__call__`.

HyenaDNA (Nguyen et al., 2023). Causal / autoregressive; DMS scoring uses whole-sequence LLR rather than masked-marginal, since the model has no mask token. L=5994.

LucaOne / LucaOne-gene (He et al., 2025). Trained with `max_pos_embeddings`=4096; we set VEP_WINDOW_SIZE=4094 to stay within the trained range. The shared gene–protein vocabulary is disambiguated by a `seq_type` keyword (`'gene'` / `'prot'`) supplied at tokenisation time. The upstream `batch_encode_plus` reads from the wrong argument slot, so inputs are tokenised one record at a time through a thin adapter. Joint DNA+protein evaluation (TFIso) instantiates two LucaOne forward graphs sharing the same weights but distinct `seq_type` tags.

NTv3-650M / 100M (Boshar et al., 2025). The U-Net architecture requires sequence length divisible by 128 (seven downsample stages); we pad with the N-token (a real token in NTv3's alphabet) rather than `[PAD]`. VEP uses L=4096 (next multiple of 128 below the trained context), no leading or trailing special tokens, and we walk `embed_dim` (rather than `hidden_size`) to retrieve the representation width.

Omni-DNA-116M / 300M (Li et al., 2025). Hard 250 BPE-token context cap. For VEP, we set L=1200 bp (about 244 BPE tokens, leaving room for special tokens). Three of the five DMS sequences exceed the 250-token budget, so DMS scoring uses a *mutation-centred* 246-token window: we locate the BPE token covering the variant nucleotide, slide a 246-token window centred on it (clamped if it would overshoot the sequence), and compute LLR on that window only. Wildtype log-likelihood is cached per (sequence, mutation-position) pair, giving roughly 3× WT re-use under DMS's average of 3 alleles per site. The same windowing logic is applied in the embedding-suite path so that all DMS metrics for Omni-DNA are computed on identical token windows.

Orthrus-4track (Fradkin et al., 2026). A contrastive Mamba SSM trained with one-hot inputs (no learned token embedding). We wrote a tokenizer adapter that maps single nucleotides to integer ids and a model wrapper that converts those ids to a one-hot tensor inside the forward pass. A small `Block` class shim bridges the `mamba_ssm` 1.x API the released checkpoint expects to the 2.x runtime in our environment. The model is encoder-only with no language-model head, so DMS-LLR and VEP-LLR are undefined; only embedding-based metrics are reported.

JEPA-DNABERT-2 (Larey et al., 2026). DNABERT-2-117M (Zhou et al., 2023) continually pretrained with a joint-embedding predictive objective, token-level MLM, and VICReg regularisation. We load the backbone via the standard path and apply the JEPA checkpoint weights. Vanilla attention is used (Triton kernel disabled) to match the base DNABERT-2 configuration, and the checkpoint includes the MLM head, which was updated during JEPA training. VEP uses L=5994.

JEPA-HyenaDNA-tiny (Larey et al., 2026). HyenaDNA-tiny-16k-d128 (Nguyen et al., 2023) (0.44M parameters, 16 kb context) continually pretrained with the JEPA objective. The architecture is autoregressive, so DMS scoring uses whole-sequence LLR rather than masked-marginal scoring. We benchmark the corresponding base model (HYENADNA-TINY) alongside; this is distinct from the larger HYENADNA variant (large-1m, 6.6M parameters) reported elsewhere. VEP uses L=5994.

JEPA-NTv3-100M (Larey et al., 2026). NTv3-100M (`NTv3_100M_pre`) (Boshar et al., 2025) continually pretrained with the JEPA objective. The U-Net architecture requires sequence lengths divisible by 128, so we pad with the N-token (a real token in NTv3's vocabulary) rather than `[PAD]`. We benchmark the corresponding base model (NTV3-100M) alongside; this is distinct from the larger NTV3-650M variant reported elsewhere. VEP uses L=4096.

Carbon-3B. A Llama-style autoregressive DNA LM, `HuggingFaceBio/Carbon-3B` (Allal et al., 2026), with a hybrid tokenizer comprising a complete 6-mer vocabulary (4096 DNA 6-mers) plus Qwen3-BPE for English, switched by a `<dna>` tag that we prepend automatically. Tiling is fixed and non-overlapping ($1 + \lceil L/6 \rceil$ tokens), so a SNV flips exactly one token and whole-sequence LLR is well-defined. The model has no mask token, so DMS routes to whole-sequence LLR and VEP to the embedding suite, following the same handling as Evo2. Loaded in bf16; VEP uses L=5994.

ESM-1b / ESM-1v (Rives et al., 2021). 1024-position context cap (1022 usable amino acids). Of the five DMS studies, only Garvie 2021 (PDE3A, 1141 aa) overflows; we tail-crop to the last 1022 aa. Because every PDE3A variant in this dataset sits at residue ≥ 668, tail-cropping is equivalent to a variant-centred window for this study while preserving per-WT amortisation.

ESM-C-300M / 600M. Loaded through the official EvolutionaryScale SDK (`esm` 3.2.1) in a dedicated environment (`torch` 2.3.1). A thin HF-compatible adapter wraps the SDK's forward pass so that the rest of the evaluation pipeline (logits, `last_hidden_state`, hidden-states tuple) is shape-compatible with the encoder-MLM path used for all other protein models.

ProtT5. An encoder-only T5 trained with span-corruption rather than single-token masked language modelling, so masked-marginal DMS-LLR is undefined; we report the embedding-based DMS metrics only.

D.2 DMS SCORING

DMS embedding-based regression uses 5-fold cross-validation with `GroupKFold` grouped by amino-acid-level mutation identity. The codon-level JSONL rows in four of the five DMS studies have $1.66\times$ to $3.02\times$ codon expansion (synonymous codons mapping to the same mutation), and a random 5-fold split would leaks codon twins between train and test, inflating Spearman by 0.05–0.27 per study. The probe is a `StandardScaler` followed by Ridge regression with α=10^{-3} and the Cholesky solver.

D.3 VARIANT EFFECT PREDICTION

Per-model context window. The default benchmark window is $L = 5,994$ bp (a 6-mer-aligned $\pm$2,997 nt window centred on the variant), used for DNABERT-2, JEPA-DNABERT-2, HyenaDNA, HyenaDNA-tiny, JEPA-HyenaDNA-tiny, Caduceus-PH/PS, NTv2-500M, Orthrus, Evo2, Carbon-3B, and AlphaGenome. Models with stricter caps use:

Table 7: Per-model VEP context-window deviations from the default.

Model	VEP window (L)	Reason
LucaOne / LucaOne-gene	4,094	`max_pos_embeddings`=4096
AIDO.DNA-300M	3,998	`max_pos_embeddings`=4000
NTv3-650M / 100M, JEPA-NTv3-100M	4,096	U-Net 7-downsample (must be $\div 128$)
Omni-DNA-116M / 300M	1,200	250 BPE-token context cap

Strand averaging. VEP scores are computed independently on the forward strand and the reverse complement and averaged. Both strands carry identical biological information at any chromosomal

location, and for autoregressive models the reverse-complement pass also recovers "downstream-context" information that the causal forward pass lacks.

Linear-probe configuration. The two vector-valued metrics (`CLS`, `Mul`) are evaluated through leave-one-chromosome-out L2-regularised logistic regression (C=1, standardized features, balanced class weights, Newton-Cholesky solver) and reported as AUPRC. The two scalar metrics (`cosine`, ℓ_2) are reported as direct AUPRC of the metric against the binary pathogenicity label.

D.4 TFLink and GTEx Supervised Fine-Tuning

Both tasks use the same fine-tuning regime as the IsoAlign rows: LoRA with $r{=}4$ on the DNA encoder's query and value attention projections (about 1.4M trainable adapter parameters), a frozen protein encoder, and the standardised MLP head described in Appendix E. The LucaOne baseline uses the same LoRA configuration applied to its query and value projections. Because LucaOne is a single transformer that handles both nucleotide and amino-acid inputs, we instantiate two separate forward graphs sharing the same backbone weights but with distinct `seq_type` tags (`'gene'` for DNA, `'prot'` for protein) so that the DNA-side LoRA adapters do not modify the protein representation.

E Sequence Ablation Details

E.1 Experimental Setup

We train DP-JEPA variants on systematically modified DNA-protein pairs to isolate the contribution of different sequence features. Each ablation uses identical architecture (NTv2-500M DNA encoder + ESM2-150M protein encoder) and training configuration, varying only the input sequences. Models are trained for 50 epochs with early stopping based on validation loss. We evaluate bidirectional retrieval using DNA-to-Protein (D2P) and Protein-to-DNA (P2D) Mean Reciprocal Rank (MRR) on a held-out validation set.

E.2 Ablation Categories

Baselines.

- **Feature-Aligned (CDS):** Coding sequences with strict 3:1 codon-to-amino-acid alignment preserving exon boundaries.
- **Canonical CDS:** Standard CDS extraction without explicit feature alignment.
- **Transcript-Specific:** Full spliced transcripts including UTRs with CDS boundaries marked.

Coding Ablations.

- **Frame Shift (+1):** Shift reading frame by 1 nucleotide, disrupting codon boundaries and producing an entirely different amino acid sequence while preserving nucleotide composition.
- **Codon Shuffle:** Replace each codon with a randomly selected synonymous codon encoding the same amino acid, preserving the protein sequence while altering codon usage patterns.
- **Splice Junctions Only:** Intronic sequences flanking annotated splice sites ($\pm$50 bp), containing only splice donor and acceptor motifs.

Non-Coding Ablations.

- **5′ UTR Only:** 5′ untranslated regions only, excluding all coding sequence.
- **3′ UTR Only:** 3′ untranslated regions only, excluding all coding sequence.
- **UTRs Combined:** Concatenated 5′ and 3′ UTRs without CDS.

Genomic Context.

- **Promoter + mRNA:** 1 kb upstream promoter region concatenated with the transcript.
- **Promoter + mRNA + Splice:** Promoter, transcript, and explicit splice junction annotations.
- **Gene-Level Non-Coding:** Distal regulatory sequences only, excluding all coding and proximal elements.

Negative Sampling Strategies.

- **Isoform Swap:** Pair each DNA sequence with a different protein isoform from the same gene.
- **Homolog Swap:** Pair DNA with a protein from a homologous gene family.

Negative Controls.

- **Shuffled DNA:** Randomly permute nucleotides within each biological DNA sequence, destroying all sequence structure while preserving nucleotide composition.
- **Random DNA:** Pair proteins with randomly generated DNA sequences matching overall nucleotide frequencies.

F Compute Resources

All experiments were conducted on an internal SLURM-managed GPU cluster.

Pretraining. Each training job used 4× NVIDIA H100 GPUs (80GB) with DeepSpeed ZeRO Stage 2 and bf16 mixed precision. Table 8 summarizes the training hyperparameters for each objective, and Table 9 reports the total compute used. The reported totals include hyperparameter searches and ablation studies; average run duration was ∼8 hours (∼32 GPU-hours).

Supervised Fine-tuning. SFT experiments used a single NVIDIA A40 GPU (48GB). GTEx fine-tuning on 10% downsampled data (∼35k training samples) took ∼2 hours per epoch. We used LoRA adapters with learning rate 1e-4, batch size 16, and early stopping (patience 4). Training typically converged within 5–10 epochs.

Evaluation. Zero-shot evaluations used a single NVIDIA A100/H100 GPU (80GB) with batch size 256. We evaluated 105 model configurations across 4 benchmark tasks (DMS, GTEx, TF-Iso). Each model evaluation completed within ∼1–2 hours depending on task complexity.

Table 8: Training hyperparameters by objective.

Parameter	DP-JEPA	IM-JEPA	Contrastive	MLM
Batch size (per GPU)	16–32	8	16	16
Gradient accumulation	4	8	4	4
Effective batch size	256	256	256	256
Learning rate	1e-5	1e-5	1e-5	1e-4
Max epochs	150	150	150	150

Table 9: Total compute used for pretraining experiments.

Objective	Runs	Wall-Clock (h)	GPU-Hours
DP-JEPA	34	274	1,097
IM-JEPA	12	136	543
Contrastive	48	368	1,474
MLM	11	63	254
Total	**71**	**568**	**2,270**