BIOBLIP: A MULTIMODAL ARCHITECTURE FOR ZERO-SHOT TASK TRANSFER IN GENOMIC VARIANT INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Developing scientific hypotheses in biology requires integrating heterogeneous evidence from DNA sequence, gene context, protein function, and prior literature. Existing multimodal AI systems expose biological evidence to LLMs through textification or by projecting biological embeddings into fine-tuned language models. However, these models' performance is limited to the specific tasks for which they are fine-tuned. Here we present BioBLIP, a multimodal Q-former based architecture which leverages biological embeddings and a LLM to generalize to unseen genomic tasks without fine-tuning. The key to BioBLIP is a new model architecture that integrates four data modalities – DNA, genes, proteins, and text – through a master Qformer model, which consolidates the modality-specific information into multimodal tokens for a frozen LLM. BioBLIP is pretrained on the task of human single nucleotide variant annotation (85.2% overall accuracy). Without fine tuning and only 66M trained parameters, pretrained BioBLIP achieves 70.2% Top-1 accuracy in variant prioritization, as opposed to 64.9% accuracy for LLMs given matched biological evidence in text and 59.3% for 40B parameter genomic foundation models. In target gene prediction, BioBLIP outperforms matched-text LLM baselines (60.50% versus 43.70%) while producing rich, transparent traces. Overall, BioBLIP demonstrates strong capabilities to leverage learned multimodal representations for robust zero-shot transfer to multi-step inference tasks in genomics.

1 INTRODUCTION

Biology requires integrating and understanding complex datasets containing diverse multimodal entities, such as DNA, RNA, genes, proteins, tissues, and cells. For example, drug development programs shortlist disease-related target genes by integrating signals across genome-wide association studies, regulatory sequence contexts, protein-level changes, and gene expression. Biologists seek to understand the effects of genomic variants in individuals, which are differences in a person's DNA sequence when compared to a standard "reference" genome. Variants may be one or hundreds of base-pairs long and affect human traits through multiple scales of biology. Jointly interpreting different modalities remains a core challenge for AI models.

Reasoning on diverse datasets is also important for scientific hypothesis generation. Ultimately, AI models that understand diverse biological entities can serve as scientific collaborators, surfacing non-obvious connections and accelerating discovery. Large language models (LLMs) have demonstrated reasoning capabilities over textified biological data (Wang et al., 2025a), (Pei et al., 2023), (Taylor et al., 2022). However, textualization discards the rich internal geometric and functional representations of biology encoded by state-of-the-art foundation models (FMs) like AlphaGenome (Avsec et al., 2026), AlphaFold Abramson et al. (2024), and Evo2 Brixi et al. (2026).

The goal of multimodal architectures in biology is to natively integrate multi-step inference capabilities of LLMs with the rich representations contained in biological FMs. Recent works have made progress: BioReason combines an LLM with Evo2 DNA embeddings for variant effect prediction (Fallahpour et al., 2025), BioReason-Pro extends this to proteins via ESM embeddings (Fallahpour et al., 2026). Several GPT-style works connect single-modality FMs to LLMs through lightweight

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.

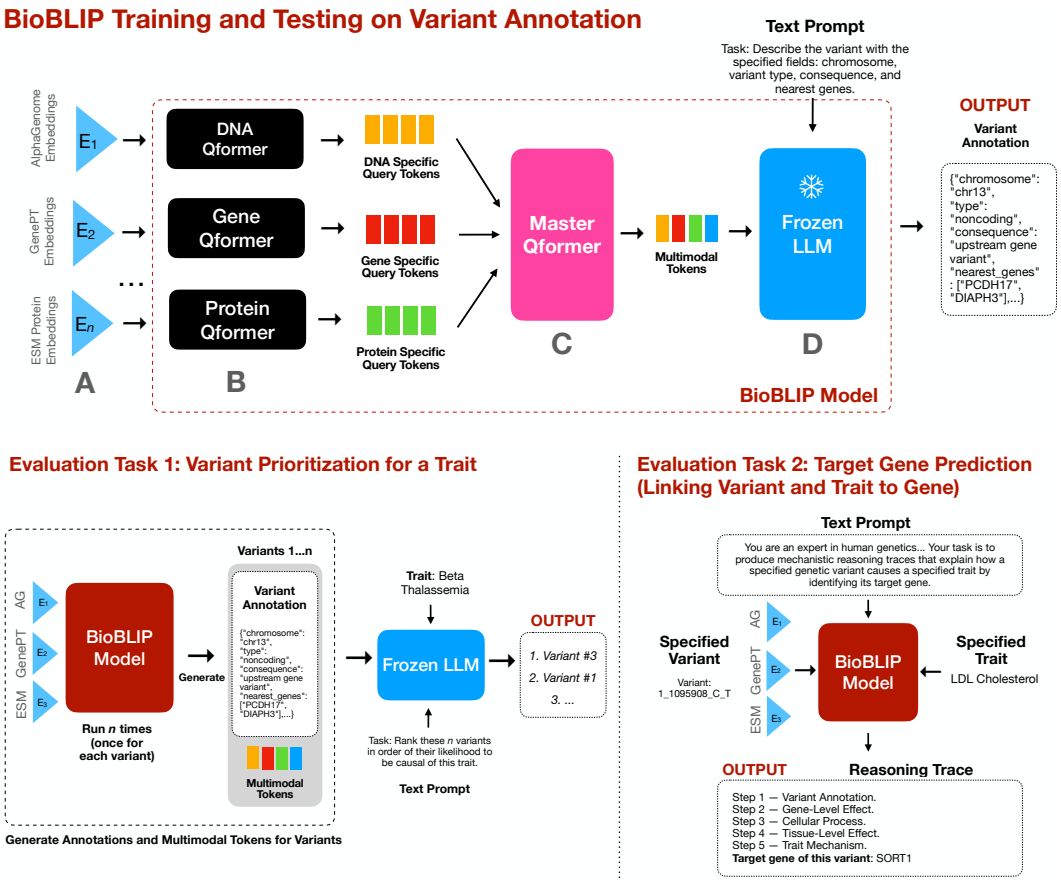


Figure 1: **(a)** BioBLIP training and testing on variant annotation. AlphaGenome, GenePT, and ESM protein embeddings (A) are encoded by modality-specific Q-Formers (B) into DNA-, gene-, and protein-specific query tokens, which a Master Q-Former (C) fuses into multimodal tokens passed to a frozen LLM (D) that generates a structured variant annotation. **(b)** Downstream evaluation tasks. *Task 1:* BioBLIP generates annotations and multimodal tokens for each of n candidate variants, and a frozen LLM ranks them by likelihood of being causal for a specified trait. *Task 2:* given a variant and trait, BioBLIP produces a mechanistic reasoning trace that identifies the variant's target gene.

projections and supervised fine tuning (Xiao et al., 2024; Abdine et al., 2024; Tsou et al., 2025; Shoshan et al., 2024; de Almeida et al., 2025). However, these models are constrained to the specific sets of tasks for which they are trained. Biological FM embeddings have demonstrated broad utility across biological predictions (Benegas et al., 2025b; Brixi et al., 2026; Marquet et al., 2022; Yadav et al., 2024). Theoretically, the general reasoning capabilities of the LLM should be able to leverage Biological FM embeddings to complete varied tasks. However, few multimodal models have shown task transfer without further fine-tuning.

Here we introduce BioBLIP (Fig. 1), a new transformer-based architecture for multimodal inference that enables zero-shot task transfer without further fine-tuning. BioBLIP introduces three key contributions. First, it takes in four input modalities simultaneously – DNA sequence embeddings, gene neighborhood information, protein sequences, and text. Second, the architecture introduces a new master Q-former model in Stage 2, which integrates modality-specific tokens through cross-attention, producing a fixed-length set of multimodal tokens which form a soft prompt for the LLM. Finally, the LLM in BioBLIP is never fine-tuned with SFT or RL, which allows us to maintain reasoning capabilities of the LLM backbone. This is because the unique training recipe of BioBLIP keeps the LLM frozen and only backpropagates next-token loss to the Q-formers. Together, these mechanisms capture task-specific knowledge encoded in Biological FMs, while leveraging the LLM's multi-step inference capabilities.

We pretrain BioBLIP on variant annotation on the OpenTargets Genetics database and observe strong zero-shot transfer to unseen genomic tasks. For example, zero-shot Top-1 accuracy on variant prioritization is 70.2% on Traitgym and 60.5% on target gene prediction in non-closest gene cases. For comparison, LLM-only baselines that receive each of the biological evidence sources as text only achieve 64.9% Top-1 accuracy on variant prioritization and 43.7% Top-1 accuracy on target gene prediction when the target gene is not the closest gene. We also observe that Bio-BLIP outperforms alignment-free genomic language models GPN-Promoter and Evo2-40B on regulatory variant prioritization. Further, due to its frozen LLM, BioBLIP has only 66M trainable parameters for efficient pretraining. These findings establish the BioBLIP architecture as capable in (1) integrating heterogenous sources of data, (2) leveraging learned multimodal representations, (3) moving beyond pretraining tasks to complex downstream objectives.

2 BioBLIP Model

BioBLIP (Fig. 1) is a hierarchical architecture that leverages multiple types of embeddings for zero shot transfer to downstream tasks. The training of BioBLIP is guided in two stages: **Stage 1** focuses on individual *modality-specific* text alignment and information extraction, while **Stage 2** integrates modality-specific information through the master Q-former model and makes it available to the LLM for inference.

2.1 Training and Task Transfer Approach.

We pretrain BioBLIP on variant annotation: a task which allows the model to align biological embeddings with structural and functional features of human single-nucleotide variants. We choose this pretraining task because if the model can learn variant annotation well, this would serve as a solid foundation to generalize to more complex variant-centered tasks.

Our goal in BioBLIP is not generalization to entirely unseen genomic regions. BioBLIP is pretrained broadly across the genome so that it can learn accurate representations of genomic variants.

2.2 Pretraining Regime.

We train the BioBLIP model (Fig. 1) in two stages. These are preceded by an input processing stage which precomputes and stores embeddings from various biological foundation models, here: encoder outputs from AlphaGenome, GenePT-based neighborhood embeddings, and nearest-protein ESM2 embeddings. Precomputing embeddings provides significant memory savings, time savings, and greater modularity. This also ensures that we never update the weights of the foundation models. Each embedding type receives its own modality-specific *Q-former*.

Qformer Definition. A Q-Former (Querying Transformer) is a lightweight transformer module that extracts a fixed-size set of learned representations from variable-length input embeddings. It does so via a set of K learnable query vectors $\mathbf{Q} = \{q_k\}_{k=1}^{K}$, which are randomly initialized and updated during training. These query vectors interact with the input embeddings through cross-attention, effectively "querying" the input for the most task-relevant information. The query vectors also interact with the text caption tokens via cross-attention Li et al. (2023). The outputs are *modality-specific query tokens*: K fixed-dimensional vectors that extract relevant information from the input.

Stage 1 of training simultaneously trains each modality-specific Q-former model to align its input embeddings with shared textual captions, allowing each to learn modality-specific attention patterns that are relevant to the shared text in different ways. Following the BLIP-2 style of fusing vision and language Li et al. (2023), each Q-former has three losses to optimize in this stage:

- *Image–text contrastive* (ITC) loss maximizes the cosine similarity between each query representation and its paired text caption while minimising similarity to unpaired captions in the batch:

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{2}\mathbb{E}\left[\log \frac{\exp(\hat{q}^\top \hat{t}/\tau)}{\sum_j \exp(\hat{q}^\top \hat{t}_j/\tau)} + \log \frac{\exp(\hat{t}^\top \hat{q}/\tau)}{\sum_j \exp(\hat{t}^\top \hat{q}_j/\tau)}\right] \tag{1}$$

where $\hat{q}$ and $\hat{t}$ are ℓ_2-normalised query and text projections respectively, and τ is a learnable temperature.

- *Image–text matching* (ITM) loss trains the model to identify, given a batch of embeddings and captions, whether a given embedding-caption pair is a match. A binary cross-entropy is applied over positive pairs and negatives mined via ITC similarity scores
- *Image-grounded text generation* (ITG) loss, a causal language-modeling objective in which the text decoder attends to the output query tokens as soft visual prompts and is trained to generate the paired caption autoregressively.

The total Stage 1 loss aggregates all three objectives across each of N modality-specific Q-Formers:

$$\mathcal{L} = \sum_{i=1}^{N} \left(\mathcal{L}_{\text{ITC}}^{(i)} + \mathcal{L}_{\text{ITM}}^{(i)} + \mathcal{L}_{\text{ITG}}^{(i)} \right) \tag{2}$$

Each Q-former Q_i outputs K_i modality-specific tokens at the end of Stage 1 training.

Stage 2 of training introduces the Master Q-former model attached to a frozen LLM (Qwen3-4B or Qwen3-32B Yang et al. (2025)). The master Q-former model fuses the Stage-1 modality specific tokens and produces 32 multimodal tokens that are fed to the LLM as a soft prompt.

The Stage 2 training loss is next-token-prediction with the LLM's weights frozen, so the loss only updates the projection weights, W_{proj}, the Master Qformer weights, and the Stage 1 Qformer weights. Notably, the LLM is not trained or fine-tuned during any point in the training regime. Freezing the LLM supports transfer to downstream tasks without additional LLM fine-tuning and is a key difference over existing multimodal models.

Input Representations. The architecture as designed supports any number of input modalities. Here we describe the specific inputs to BioBLIP for genomic variant interpretation:

- **DNA Sequence encoder.** Pre-computed AlphaGenome Avsec et al. (2026) embeddings for the reference and alternate alleles of a genomic variant, each represented as 256 context-window bins of hidden dimension 3072 and interleaved to form a sequence of length 512.
- **Gene-neighborhood embeddings.** GenePT Chen & Zou (2024) embeddings for the up to 15 protein-coding genes nearest to the variant. Each embedding is concatenated with log-distance of the gene from the variant and a boolean column which encodes whether the gene physically overlaps the variant or not.
- **Protein-sequence encoder.** ESM-2 Lin et al. (2023) embeddings for the reference and alternate amino-acid sequences of the nearest protein-coding gene to the genomic variant.

Textual variant captions. Structured JSON objects for each single nucleotide variant containing chromosome, variant type (`coding`/`noncoding`), most severe consequence (e.g. intron variant, missense variant), and all nearest protein-coding genes.

3 Experiments

3.1 Training and Testing: Variant Annotation

Task Definition. Given biological embeddings, the model must produce a structured JSON annotation for the variant. Ground-truth labels come from the OpenTargets Genetics portal Buniello et al. (2025). The annotation fields are as follows:

1. **Chromosome identification.** Predict which of the 23 human chromosomes the variant resides on.
2. **Variant type** (coding vs. non-coding). Classify whether the variant is within a protein-coding region or not. Coding variants may alter protein sequence, while non-coding may affect gene regulation.

3. **Most severe consequence.** Predict the most deleterious annotated consequence of the variant (e.g. missense, splice-acceptor, stop-gained). Consequence severity is a primary signal used in clinical variant prioritisation and reflects the likely mechanism of pathogenicity.

4. **Nearest genes.** Identify the genes in proximity to the variant. For non-coding variants in particular, nearest genes are a key indicator of which genes may be disregulated.

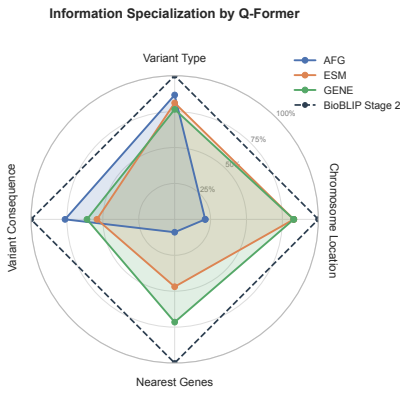


Figure 2: **Modality Diversity.** AFG = AlphaGenome Embeddings; ESM = ESM-2 Protein Embeddings; GENE = GenePT Neighborhood Embedding. The Qformers for each embedding type show differential per-field accuracy in variant annotation. With the Master Qformer, BioBLIP outperforms any specific modality's model.

Dataset. We train BioBLIP's Qformers on a set of 278k single nucleotide variants (SNVs) from the OpenTargets genetics database (Buniello et al., 2025). The data processing pipeline samples 300k SNVs from OpenTargets, mapping each to a consequence label and extracting reference and mutant DNA sequences. We apply a genome build match to filter out variants where the gene coordinate lookup tool, genelocator, does not agree with the closest gene labeled by OpenTargets Lab (2019). This filtering step leaves 278k/300k SNVs, which we split into 80% training and 20% validation, with approx. 1k variants held-out as a test set for model evaluation.

Baselines and Comparative Models. BioBLIP architecture and pretraining details are provided in Appendix B. We compare with prior multimodal models: BioReason using Evo2-7B embeddings and Qwen3-4B (Fallahpour et al., 2025). We retrain BioReason for its original task of Variant Effect Prediction (Bioreason-VEP) using the authors' provided datasets and training code. We also train BioReason using SFT on the same OpenTargets Genetics dataset on which BioBLIP is trained, producing BioReason-OT for a fair architecture comparison.

We evaluate Qwen3-32B LLMs on this task as a floor to how well an LLM can perform on this task without any multimodal embeddings, but just with the variant ID. We also explicitly provide an LLM with relevant AlphaGenome information by including the top AlphaGenome predictions for each variant with a 131,072 bp window, capturing local regulatory context across chromatin accessibility, transcription, and transcription factor binding signals.

Results and Observations. Variant Annotation Results (Table 1) demonstrate that BioBLIP exhibits strong performance across fields, demonstrating effective learning of variant features from biological embeddings alone. BioBLIP achieves $\geq 30\%$ boost on overall accuracy over existing prior methods in describing functional and structural variant features.

Held-out Loci and Neighborhoods. In order to account for random splitting of variants for the annotation task, we therefore also report accuracy on two stricter subsets of the 999-variant test set in Table 7. *Held-out loci* are test variants whose chromosome and position do not occur in the training split, regardless of reference and alternate allele. *Held-out neighborhoods* additionally require that a test variant lie more than 1 kb from every training variant and share at most half of its annotated genes with the five closest training variants within 10 kb. A variant's annotated genes are its overlapping genes and nearest protein-coding genes from OpenTargets. The gene criterion removes variants whose nearest genes the model could recover from an adjacent training variant. The results demonstrate that BioBLIP performance gains hold on held-out loci or neighborhoods. They ensure that the exact evaluation variants were not annotation-training examples or close to annotation-training examples, rather than claiming independence from all nearby training sequence context.

Fig. 2 demonstrates that modality-specific Qformers perform better on different aspects of variant annotation. For instance, AlphaGenome's Qformer performs best on variant consequence generation. With the Master Qformer, BioBLIP gains on all of the axes after integrating modality-specific outputs.

Table 1: **Variant annotation accuracy.** Comparison of BioBLIP against BioReason and LLM baselines. Best results in each column are shown in bold. Gray shading denotes the primary BioBLIP model. AG Text: Textified AlphaGenome Outputs.

Model	LLM	Overall	Per-Field Accuracy (%)			
		Accuracy (%)	Variant Type	Consequence	Chromosome	Nearest Gene
BioBLIP	Qwen3-32B	**85.16**	*91.39*	**74.37**	**97.80**	**77.09**
BioBLIP	Qwen3-4B	78.09	89.49	69.07	87.59	66.19
BioReason–OT	Qwen3-4B	41.85	**91.89**	54.35	15.22	5.91
BioReason–VEP	Qwen3-4B	14.54	31.63	18.22	8.31	0.00
LLM + VarID + AG Text*	Qwen3-32B	37.26	66.60	44.50	– *	0.68
LLM + VarID*	Qwen3-32B	38.11	68.27	46.05	– *	0.00
Majority Class	–	30.93	68.40	46.30	8.30	0.70

* LLM + VarID receives Variant ID with Chromosome in the prompt.

3.2 Evaluation Task: Regulatory Variant Prioritization.

Dataset and Task Definition The TraitGym Mendelian traits benchmark defines a task for regulatory prioritization in human genetics (Benegas et al., 2025b). A Mendelian trait is controlled by a single gene locus and follows simple dominance patterns for inheritance. The benchmark provides 338 groups of candidate regulatory variants with matched location and functions. Exactly one variant per group is labeled "causal" for the given Mendelian trait. The task for the models is to produce a ranked list of all genetic variants from most to least likely to be causal. All models were evaluated on n=319 groups of held-out loci from the OpenTargets training data.

Experimental Setup. To extend BioBLIP for multiple variants, we ran BioBLIP for each variant $1...n$ to compute that variants' multimodal tokens and variant annotation, as shown in Figure 1b. The multimodal tokens and variant annotations were interleaved, and fed to a frozen LLM for ranking.

Baselines. We evaluate BioReason-OT as an existing multimodal model, which has been fine-tuned on the OpenTargets dataset. To isolate the effect of representing biological evidence through BioBLIP's learned multimodal tokens rather than through text, we construct a Qwen3-32B baseline with matched textified biological evidence as what is received by BioBLIP. For each candidate variant, the baseline receives AlphaGenome predictions rendered as text, identities of the 15 nearest genes and their distances (corresponding to textualized GenePT and ESM). We do not textify the entire protein coding sequence since Qwen3-32B cannot parse long amino acid sequences. The comparison tests whether learned multimodal representations provide an advantage over a Qwen3-32B baseline that receives each of BioBLIP's evidence sources as text. We also compare against Qwen3-32B with only access to the variant ID.

We also compare against specialized methods used for variant prioritization: Genomic Foundation Models (GFMs) and Genomic Language Models (GLMs) including Evo2 (7B and 40B) and GPN-Promoter. GPN-MSA is a genomic language model which requires whole genome alignment inputs to make predictions (Benegas et al., 2025a). For these models, log-likelihood-ratio (LLR) between variant and the reference sequence is used to predict variant deleteriousness (Benegas et al., 2025b). As a strong lightweight biological baseline, we scored each variant with AlphaGenome and rank variants by their maximum effect size across all tracks.

Results and Observations. Table 2 demonstrate that BioBLIP shows strong zero-shot performance and outperforms alignment-free genomic language models and genomic foundation models including Evo2-40B and GPN-Promoter - across the board. BioBLIP also slightly surpasses GPN-MSA in Top 1 accuracy without requiring whole genome alignments as input. BioBLIP also beats the non-generative AlphaGenome effect scores baseline by a wide margin.

We seek to understand whether multimodal tokens as used in BioBLIP provide an advantage over textified inputs. We compare BioBLIP with an LLM given the nearest genes, their distances, and textified AlphaGenome tracks, so it has access to the same evidence sources as BioBLIP. Against

Table 2: **Regulatory variant prioritization on held-out loci ($n = 319$). Bold** marks the best result excluding models that receive whole genome alignments. GLM = Genomic Language Model; GFM = Genomic Foundation Model. AG = AlphaGenome.

			Top-k acc. (%) ↑		
Category of Method	**Model**	**Inputs to Model**	$k = 1$	$k = 2$	$k = 3$
Chance	Random	–	10.00	20.00	30.00
LLM only	Qwen3-32B	Variant ID	6.90	12.23	19.44
LLM + matched text	Qwen3-32B + matched text	AG, GenePT, ESM	64.89	65.83	66.46
Multimodal	**BioBLIP**	AG, GenePT, ESM	**70.22**	**76.49**	**79.62**
Multimodal	BioReason-OT	DNA sequence + text	11.60	22.57	31.66
Non-generative bio.	AG effect scores	AG	39.50	50.78	58.62
Generative bio.	GLM: GPN-Promoter	DNA sequence	38.87	51.72	60.19
	GFM: Evo2-7B	DNA sequence	36.68	47.34	56.11
	GFM: Evo2-40B	DNA sequence	59.25	71.79	78.06
Generative bio. (WGA)	GPN-MSA[†]	DNA sequence + WGA	68.03	79.31	85.89

[†]Alignment-based models explicitly require whole genome alignments (WGAs) and are shown as a reference.

this strong baseline, BioBLIP gains 5.3% in Top-1 accuracy, 10.7% in Top-2 accuracy, and 13.2% in Top-3 accuracy. BioBLIP produces human-interpretable inference traces (illustrative examples in App. H.1). The model identifies nearby genes of all variants listed, links genes to the Mendelian trait provided, and incorporates variant consequence.

3.3 Evaluation Task: Target Gene Prediction.

Task Definition and Dataset. The OpenTargets Genetics Gold Standard dataset contains 397 unique SNVs mapped to their target genes and phenotypes. We request 5-step chain-of-thought reasoning traces from the model linking GWAS variants from the Gold Standard Dataset to target genes and to the observed phenotype. Ground truth labels are OpenTargets expert-curated target genes. The task definition is: Given a variant ID, a disease, and a neighborhood of nearby genes, can the model identify the target gene that links the variant to the disease?

Baselines. A strong baseline is simply picking the closest gene by distance to the variant. In the majority of cases in OpenTargets, the closest gene is the correct target gene. We also specifically measure performance on traces where the closest gene is not the target gene ("Non-Closest Gene Correct"). Multimodal Baselines are BioReason-OT, which has been fine-tuned on variant annotation on the same dataset as BioBLIP (Sec. 3.1). For a matched comparison, Qwen3-32B and BioReason are provided the same candidate-gene neighborhood information in text-format as BioBLIP receives.

Results and Observations. Results for all reasoning traces and non-closest gene traces are shown in Table 3. We evaluate only on held-out loci from the training data and we additionally report results on non-closest-gene loci more than 1 kb from any training variant. This is done to prevent direct reuse of annotation-training examples. BioBLIP leverages its multimodal architecture to achieve notable wins over Qwen3-32B on the more challenging set of non-closest gene traces (16.8 percentage points). We used an LLM-as-a-judge to characterize the non-closest gene traces where BioBLIP wins over Qwen3-32B (illustrative traces are shown in App. 9).

4 Ablation Studies

In this section, we ablate various components of the BioBLIP architecture as we seek to understand the observed performance improvements.

Decoy Variants / Embedding Permutations We perform experiments to test the model's reliance on the multimodal tokens. We generated decoys of inputs by randomly switching the AlphaGenome embeddings, GenePT embeddings, and ESM embeddings between variants to generate decoy variants.

Table 3: **Target gene prediction accuracy.** Performance is evaluated on held-out loci that have no overlap from the training set, held-out loci where the causal gene is not the closest gene. The "Distance Cutoff" set refers to held-out loci that are also >1kb away from any training variant by distance. All models are provided matched gene neighborhood context (gene identities and distances). Best results in each column are shown in bold.

Model	**Held-Out Loci** ($n = 369$)	**Non-Closest Gene & Held-Out Loci** ($n = 119$)	**Non-Closest & Distance Cutoff** ($n = 92$)
Baseline – Closest Gene	67.75	0.00	0.00
BioBLIP Qwen3-32B	**83.47**	**60.50**	**59.78**
Qwen3-32B + matched text	81.30	43.70	43.48
Qwen3-4B + matched text	70.73	39.50	40.22
BioReason – OT + matched text	34.42	18.49	17.39
BioReason – VEP + matched text	0.51	0.00	0.00

Evaluation Task	**Permuted Embeddings**
Variant Annotation	-88.75%
Target Gene Prediction	-78.41%
Variant Prioritization	-70.56%

Table 4: **Performance with permuted multimodal embeddings.** Relative change in performance after randomly permuting AlphaGenome, GenePT, and ESM embeddings between variants to generate decoy inputs.

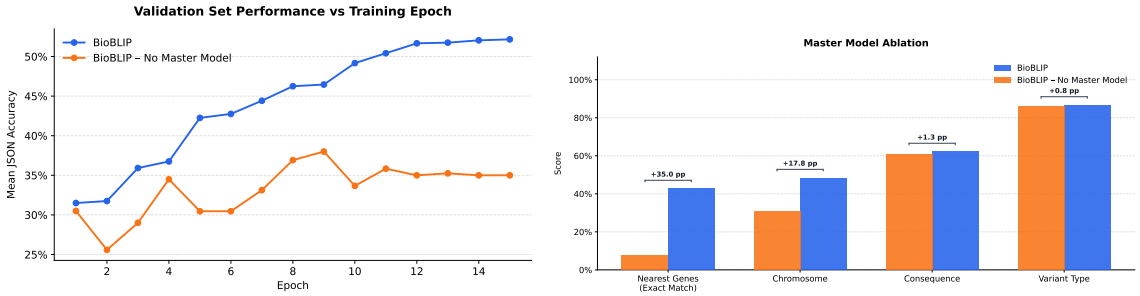
We re-ran the evaluation tasks with these decoys and the relative drops in accuracy are shown in Table 4 - we observe a drastic drop in performance on every task, ranging from 70% to 88%. This illustrates that BioBLIP relies on the correct multimodal tokens to drive performance.

Individual Modality Ablations. We performed variant annotation experiments to probe the impact of the various components in the model, including single modality ablations, dual modality ablations, and noise only inputs (Table 5). Each ablation was sampled three times with different random seeds. We find that (1) adding modalities leads to a nonlinear effect on performance, and (2) different data inputs contribute to different fields of variant annotation. AlphaGenome as a modality contributes highly to Chromosome, Variant Type, and Variant Consequence. GenePT contributes the most to the nearest genes field. ESM protein embeddings contribute the least, which is consistent with the majority of these variants being noncoding so they do not affect protein sequence.

Table 5: Accuracy (%) across modality ablations. Shading shows each configuration's gain over the noise baseline, relative to the best ablation in that row (darker = larger contribution). The full model outperforms every ablation and is shaded darkest. Three-seed averages shown.

	3 modality (Full)	**2 modality**			**1 modality**			**None**
	AFG+ESM+GENE	AFG+GENE	AFG+ESM	ESM+GENE	AFG	ESM	GENE	Noise
Chromosome	**97.60**	22.49	21.25	12.63	13.18	9.47	8.67	8.00
Variant type	**91.09**	88.32	79.31	70.07	81.51	53.87	57.00	38.87
Variant consequence	**73.60**	58.19	55.43	40.20	53.02	33.20	28.67	21.43
Nearest genes	**76.56**	7.43	0.24	2.74	0.00	0.08	0.48	0.01
Overall Accuracy	**84.71**	44.11	39.06	31.41	36.93	24.16	23.71	17.08

Master Q-former Model Ablation. Next, we seek to understand the effect of the Master Qformer model in BioBLIP. To probe this, we remove the master model and prepend the tokens from each modality-specific Qformer to the LLM prompt. Results (Fig. 3 and Table 6) demonstrate that performance gains from the master model start with Epoch 2 and lead to an overall $\sim$20% increase in performance by the end of the training regime.



(a) Validation accuracy over training epochs.

(b) Per-field performance at convergence.

Figure 3: **Master Q-former Model ablation.** (a) Mean accuracy on the validation set across training epochs for BioBLIP with and without Master Q-former. (b) Per-field breakdown of final performance, with point-improvement (pp) gains. The master model provides substantial gains on fields requiring knowledge from one specific encoder in Stage 1, such as Nearest Genes (+35.0 pp). Fields handled by every encoder (Consequence, Variant Type) demonstrate less difference.

5 Related Work

LLaVA-style encoder-projection-LLM architectures (Liu et al., 2024) have been adapted for biological modalities. BioReason (Fallahpour et al., 2025) and BioReason-Pro (Fallahpour et al., 2026) project DNA and protein embeddings through linear layers into an LLM trained with SFT and GRPO, but linear projections provide no mechanism to actively distill task-relevant information from biological embeddings. This incentivizes such models to rely on textual prompt cues. (Xiao et al., 2024), (Abdine et al., 2024), (Tsou et al., 2025), and (Shoshan et al., 2024) connect biological encoders to LLMs through lightweight projections, but do not evaluate zero-shot transfer to reasoning. ChatNT (de Almeida et al., 2025) fine-tunes the DNA encoder jointly with the projection layer, but does not assess whether pretrained biological representations contribute to performance or whether the system learns task-specific features end-to-end. TxGemma (Wang et al., 2025a), BioT5 (Pei et al., 2023), and Galactica (Taylor et al., 2022) operate on text-serialized biological inputs, but discards the rich geometric and functional representations learned by specialized biological FMs. Agentic systems such as GeneGPT (Jin et al., 2024) equip LLMs with database retrieval at inference time, but can only surface information already in existing knowledge bases.

6 Discussion and Conclusion

BioBLIP demonstrates strong zero-shot transfer to unseen genomics tasks with only $\sim 66M$ trainable parameters. On unseen loci and distant variants, BioBLIP maintains performance gains over matched-text LLM baselines both on variant prioritization (70.2% vs. 64.9% top-1) and on target genes that are not the closest gene (60.5% vs. 43.7%). Our ablation studies in Section 4 probe the following key components of the model architecture: (1) multimodal tokens, (2) individual modalities, (3) the master Q-former model. Adding new modalities leads to a nonlinear effect on performance, and different data inputs contribute to different fields of variant annotation.

Here we evaluated BioBLIP on limited benchmarks on Mendelian traits, but most traits are complex, influenced by many genes, and not characterized well. Benchmarks should be constructed that would require models to infer from sequence and regulatory context to make correct predictions. Limitations of BioBLIP are that it encodes single-nucleotide variants only, and is limited in the number of variants it can rank at once. Interpretability, confidence scores, and uncertainty quantification are left to future work.

Our results provide strong evidence that multimodal representations learned from variant annotation through BioBLIP pretraining transfer to complex neighboring tasks with multiple entities. Corrupting the multimodal correspondence substantially destroys performance, which is strong evidence that the biological representations are actually being used. Overall, BioBLIP demonstrates the capability to perform multi-step inference by integrating information encoded in diverse representations to answer previously unseen biological questions.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods in software, to clean and reformat datasets, and to support qualitative analysis of model reasoning traces. We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, design research methodology or experiments, or interpret results. Formulating or proving mathematical claims, writing proofs, and translation are not applicable to this work. Additionally, we used generative AI tools to help format the paper and provide feedback on writing and experiments. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by multiple authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release the code publicly by the camera-ready version. The repository will include a model card describing the architecture, training data, intended use, limitations, and license. Section 2 and Appendix B describe the architecture, input representations, two-stage training procedure, and hyperparameters. Section 3.1 describes the OpenTargets training data, the data splits, and the held-out loci and neighborhood subsets. Sections 3.2 and 3.3 define the downstream evaluation sets and metrics, and Appendix E details how each baseline was trained and evaluated. All external datasets, models, and code are cited where they are introduced.

REFERENCES

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 10757–10765, 2024.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage the illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R Taylor, Tom Ward, Clare Bycroft, et al. Advancing regulatory variant effect prediction with alphagenome. *Nature*, 649(8099): 1206–1218, 2026.

Gonzalo Benegas, Carlos Albors, Alan J Aw, Chengzhong Ye, and Yun S Song. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 43(12):1960–1965, 2025a.

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

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*, pp. 1–13, 2026.

Annalisa Buniello, Daniel Suveges, Carlos Cruz-Castillo, Manuel Bernal Llinares, Helena Cornu, Irene Lopez, Kirill Tsukanov, Juan María Roldán-Romero, Chintan Mehta, Luca Fumis, Graham McNeill, James D Hayhurst, Ricardo Esteban Martinez Osorio, Ehsan Barkhordari, Javier Ferrer, Miguel Carmona, Prashant Uniyal, Maria J Falaguera, Polina Rusina, Ines Smit, Jeremy Schwartzentruber, Tobi Alegbe, Vivien W Ho, Daniel Considine, Xiangyu Ge, Szymon Szyszkowski, Yakov Tsepilov, Maya Ghoussaini, Ian Dunham, David G Hulcoop, Ellen M McDonagh, and David Ochoa. Open targets platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Research*, 53(D1):D1467–D1475, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1128. URL https://doi.org/10.1093/nar/gkae1128.

Yiqun Chen and James Zou. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *BioRxiv*, pp. 2023–10, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Bernardo P de Almeida, Guillaume Richard, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Chandana Rajesh, Marie Lopez, Alexandre Laterre, et al. A multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence*, 7(6): 928–941, 2025.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J Maddison, et al. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *arXiv preprint arXiv:2505.23579*, 2025.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Purav Gupta, Jack Naimer, Kevin Zhu, Arnav Shah, Shihao Ma, Abhinav Adduri, Talu Güloglu, Nuo Liu, Haotian Cui, Arihant Jain, Max de Castro, Amirfaham Fallahpour, Antonio Cembellin-Prieto, John S. Stiles, Filip Nemčko, Alexander A. Nevue, Hyungseok C. Moon, Lucas Sosnick, Olivia Markham, Haonan Duan, Michelle Y. Y. Lee, Andrea F. M. Salvador, Chris J. Maddison, Christoph A. Thaiss, Chiara Ricci-Tam, Brian S. Plosky, Dave P. Burke, Patrick D. Hsu, Hani Goodarzi, and Bo Wang. Bioreason-pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*, 2026. doi: 10.64898/2026.03.19.712954. URL `https://www.biorxiv.org/content/early/2026/03/20/2026.03.19.712954`.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024.

Manning Lab. Genelocator: Variant annotation by mapping chromosome positions to gene names. `https://github.com`, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

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.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In *European conference on computer vision*, pp. 126–142. Springer, 2024.

Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human genetics*, 141(10):1629–1647, 2022.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1102–1123, 2023.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Anna Hakobyan, Varun S Sharma, Thomas Krausgruber, Jörg Menche, Eleni Tomazou, and Christoph Bock. Joint embedding of transcriptomes and text enables interactive single-cell rna-seq data exploration via natural language. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, et al. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, pp. 1–11, 2025.

Yiqing Shen, Zan Chen, Michail Mamalakis, Yungeng Liu, Tianbin Li, Yanzhou Su, Junjun He, Pietro Liò, and Yu Guang Wang. Toursynbio: A multi-modal large model and agent framework to bridge text and protein sequences for protein engineering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2382–2389. IEEE, 2024.

Yoel Shoshan, Moshiko Raboh, Michal Ozery-Flato, Vadim Ratner, Alex Golts, Jeffrey K Weber, Ella Barkan, Simona Rabinovici-Cohen, Sagi Polaczek, Ido Amos, et al. Mammal–molecular aligned multi-modal architecture and language. *arXiv preprint arXiv:2410.22367*, 2024.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Ching-Huei Tsou, Michal Ozery-Flato, Ella Barkan, Diwakar Mahajan, and Ben Shapira. Bioverse: Representation alignment of biomedical modalities to llms for multi-modal reasoning. *arXiv preprint arXiv:2510.01428*, 2025.

Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*, 2025a.

Yujia Wang, Jihong Guan, Wengen Li, Shuigeng Zhou, and Xuhong Wang. Biobridge: Bridging proteins and language for enhanced biological reasoning with llms. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 70–75. IEEE, 2025b.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*, 2024.

Shashank Yadav, Dhvani Sandip Vora, Durai Sundar, and Jaspreet Kaur Dhanjal. Tcr-esm: Employing protein language embeddings to predict tcr-peptide-mhc binding. *Computational and Structural Biotechnology Journal*, 23:165–173, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Mingze Yin, Hanjing Zhou, Yiheng Zhu, Miao Lin, Yixuan Wu, Jialu Wu, Hongxia Xu, Chang-Yu Hsieh, Tingjun Hou, Jintai Chen, et al. Multi-modal clip-informed protein editing. *Health Data Science*, 4:0211, 2024.

A Extended Related Work

The encoder-projection-LLM paradigm (Liu et al., 2024) has seen wide adoption in computational biology. BioReason (Fallahpour et al., 2025) projects Evo2 DNA embeddings through a linear layer into an LLM which is trained with SFT and GPRO for variant effect prediction, and BioReason-Pro (Fallahpour et al., 2026) extends this pattern to ESM protein embeddings for protein function prediction. GPT-style models such as ProteinGPT (Xiao et al., 2024), Prot2Text (Abdine et al., 2024), BioVERSE (Tsou et al., 2025), and MAMMAL (Shoshan et al., 2024) similarly connect biological encoders to language models through lightweight linear projections and fine-tune the language model with next-token-prediction. These architectures share fundamental limitations. The linear projection layer provides no mechanism to actively distill task-relevant information from the biological embedding(s). Combined with supervised fine-tuning or LoRA on the LLM backbone, this creates an incentive for the model to rely on textual cues in the prompt and ignore the embeddings. In vision-language models, MIRAGE recently demonstrated that LLMs can attain high scores on multimodal benchmarks in X-ray Q&A even if they completely disregard the image inputs Asadi et al. (2026). More systematic ablations of embeddings should be conducted in these Llava-style biological LLMs. These models also do not evaluate zero-shot transfer to complex tasks beyond their training tasks.

MAMMAL integrates several biological modalities, such as small molecules, proteins, and cellular representations for regression and classification tasks, but has never been evaluated for complex reasoning tasks Shoshan et al. (2024). MAMMAL also does not support DNA embeddings. (de Almeida

et al., 2025) uses the encoder-projection-LLM paradigm to train a conversational DNA agent to answer various types of questions related to nucleic acid sequences. However, ChatNT also fine-tunes the DNA encoder jointly with the projection layer, which makes it difficult to assess whether the pretrained biological representations are meaningfully contributing to predictions or whether the system is simply learning task-specific classification features end-to-end.

A growing body of work applies CLIP-style contrastive alignment to pair biological embeddings with natural language representations. CellWhisperer Schaefer et al. (2024; 2025) aligns single-cell transcriptomic profiles with text descriptions, enabling natural language queries over gene expression space. In the protein domain, TourSynBio Shen et al. (2024) and ProteinCLIP Yin et al. (2024) align protein sequence embeddings with text to support retrieval and editing tasks, and BioBridge leverages such alignment for protein annotation Wang et al. (2025b). These approaches are effective for cross-modal retrieval: querying biological entities with natural language or predicting text annotations from biological inputs. However, they are architecturally limited in two respects. The shared embedding space supports matching and retrieval but not multi-step causal reasoning. Second, contrastive alignment becomes combinatorially cumbersome as the number of modalities grows, since each modality pair requires its own alignment objective; with k this scales as $O(k^2)$ pairwise losses, and the alignment process can distort the internal geometry of pretrained biological embeddings that already encode meaningful structure (e.g., protein homology or gene regulatory relationships).

An alternative to natively multimodal architectures is to convert biological entities into text and reason over them with standard LLMs. TxGemma (Wang et al., 2025a) fine-tunes Gemma on textualized therapeutic data (SMILES strings, protein sequences, disease names) for property prediction and conversational reasoning. BioT5 (Pei et al., 2023) and Galactica (Taylor et al., 2022) similarly operate on text-serialized biological inputs. While these models benefit from the full capabilities of the LLM, textualization discards rich geometric and functional representations learned by specialized biological foundation models. Textification of heterologous biological datasets is inherently lossy. Agentic systems such as GeneGPT (Jin et al., 2024) equip LLMs with tool-use capabilities to query biological databases at inference time. These systems can only surface information already catalogued in existing knowledge bases and cannot generalize to novel entities or relationships not present in the literature. Both approaches do not address the fundamental challenge of learning cross-modal biological representations that support generalizable downstream reasoning.

B Training Details

Modality-specific Q-Formers are 12-layer transformers with hidden dimension $d_q = 128$, 8 attention heads, a feed-forward width of $4d_q$, and cross-attention to the encoder every two layers. The AlphaGenome Q-former has 32 query vectors, while the GenePT Q-former and ESM Q-former are trained with 8 query vectors each. Varying number of query vectors was not found to provide a significant difference in downstream performance.

The master Q-former model is a 6-layer InstructBLIP-style Dai et al. (2023) Q-Former that fuses the Stage 1 modality-specific tokens with a natural-language instruction, and produces 32 refined query tokens for the LLM. It contains $N_m = 32$ learnable meta-query vectors of its own. At each forward pass the meta-query tokens and the tokenised instruction are concatenated into a single sequence; the query tokens cross-attend with both the instruction tokens and self-attend with each other (mimicking the InstructBLIP concatenation trick), and cross-attend to the layer-normalised Stage 1 visual tokens every two transformer layers.

The losses of the modality-specific Qformers are summed and optimised jointly with AdamW with learning rate lr $= 10^{-4}$ and weight decay 0.05, under a cosine schedule with a linear warm-up over 10% of total training steps. Stage 1 training proceeds for 15 epochs and requires approximately 4 hours/epoch with two Nvidia A100 GPUs. Stage two training uses AdamW (lr $= 10^{-4}$, weight decay 0.05) with a cosine schedule, 10% linear warm-up, gradient clipping at norm 1.0, and mixed-precision (BF16) AMP. The supervision signal is NTP loss on the answer tokens only; prompt tokens are masked with -100. Training proceeds for 60 epochs and requires approximately 2 hours/epoch with four Nvidia B200 GPUs.

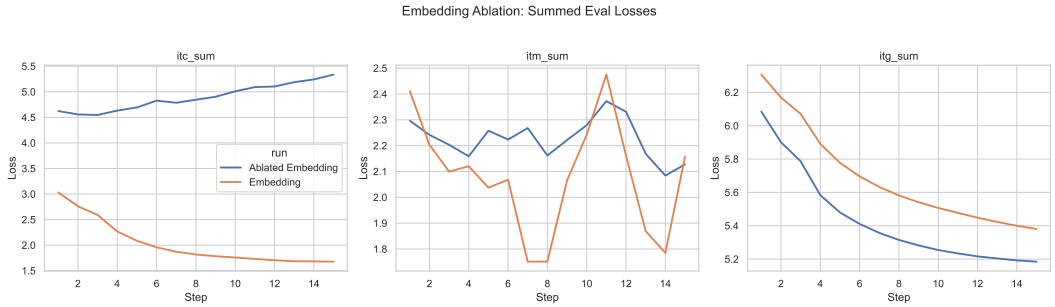


Figure 4: Stage 1 Evaluation Losses for BioBLIP with Embeddings permuted to incorrect variants (blue), versus Embeddings matched with the correct variants (orange).

C Embedding Ablations

We permute the variant text captions and embeddings within each batch, essentially assigning each variant a random caption. We observe in Figure 4 that the ITC loss increases instead of decreasing, as the Q-former is no longer able to provide high cosine similarity to variants matched with the wrong captions.

D Master Qformer Ablations

The Master Qformer ablation was tested with the BioBLIP architecture with Qwen3-4B as the backbone, trained for 15 epochs.

Table 6: **Ablation of the master qformer model in BioBLIP.**

Method	Chromosome	Type	Consequence	Nearest Genes
BioBLIP	48.37%	86.57%	62.09%	42.62%
BioBLIP: No MM	30.54%	85.77%	60.78%	7.63%
Difference	−17.83%	−0.80%	−1.31%	−34.99%

E Baseline Implementations

E.1 BioReason SFT Training Details

BioReason-VEP. BioReason-VEP was trained with SFT according to the original paper's VEP-Coding task for classifying coding variants Fallahpour et al. (2025). The input was paired reference and variant DNA embeddings and a textual query providing gene and chromosome context, and the task was to predict if a variant is benign or pathogenic, with its associated disease when pathogenic. Qwen/Qwen3-4B was used for the language model and Evo2-7B-base was used to produce DNA embeddings. BioReason-VEP was trained for 6,106 training steps using DeepSpeed stage 2 and batch size 1. The maximum text length was 1024 tokens and maximum DNA length was 2048 tokens. Evo2 embeddings were extracted from `blocks.28.mlp.l3`. Training was performed with learning rate 5×10^{-5}, weight decay 0.01, gradient accumulation over 8 steps, LoRA rank 32, LoRA alpha 64, and LoRA dropout 0.05.

BioReason-OT. BioReason-OT was trained with SFT on the OpenTargets variant captioning task. The input was paired reference and variant DNA embeddings together with a textual instruction asking the model to describe the variant as a structured JSON object containing chromosome, variant type, consequence, and nearest genes. The OpenTargets training set was the 250k SNV data augmented with DNA sequence context around the variant. Qwen/Qwen3-4B was used for the language model and Evo2-7B-base was used as the DNA encoder, with Evo2 embeddings extracted

from `blocks.28.mlp.l3`. BioReason-OT was trained for 6964 training steps using DeepSpeed stage 2 and batch size 2. The maximum text length was 512 tokens and maximum DNA length was 1024 tokens. Training was performed with learning rate 5×10^{-5}, weight decay 0.01, gradient accumulation over 8 steps, LoRA rank 32, LoRA alpha 64, and LoRA dropout 0.05.

E.2 AlphaGenome Prediction Details

AlphaGenome predictions were generated for reference and alternate alleles of each OpenTargets SNV. Each variant was evaluated in a fixed 131,072 bp window centered on the variant and using the hg38 reference genome. Outputs were saved at 128-bp resolution, giving 1024 genomic bins per variant. We included 128-bp resolution bins for ATAC-seq, DNase-seq, CAGE, PRO-cap, RNA-seq, TF ChIP-seq, and histone ChIP-seq tracks. For LLM baselines using AlphaGenome outputs, metadata from the AlphaGenome track metadata table was used to label each output track by assay, biosample, and, when applicable, transcription factor or histone mark Avsec et al. (2026).

E.3 LLM-Only Baseline Details

Text Only. The text-only baseline used Qwen/Qwen3-32B without DNA embeddings or AlphaGenome outputs. The model received only the variant identifier string and a JSON-captioning instruction asking for chromosome, coding status, consequence, and nearest genes. Evaluation used the OpenTargets 1k validation subset. The benchmark was run with maximum generation length 256 tokens.

Text Plus AlphaGenome Output Tracks. The AlphaGenome-augmented text baseline used the same Qwen/Qwen3-32B JSON-captioning setup, but the prompt additionally included raw AlphaGenome reference and alternate signal values. For each variant, the top three AlphaGenome tracks were selected by peak absolute log2 fold-change between alternate and reference predictions within a local window of ± 2 128-bp bins around the variant. The prompt included the raw reference and alternate values for these tracks, labeled with their AlphaGenome assay and biosample metadata, and the model was asked to return only the structured JSON answer. The model was evaluated with maximum generation length 256 tokens.

F Variant Annotation Extended Results

Table 7: **Variant annotation performance across evaluation subsets.**

Subset	n	Per-Field Accuracy (%)				Avg. Acc. (%)
		Chromosome	Variant Type	Consequence	Nearest Genes	
All	999	97.80	91.39	74.37	77.09	85.16
Held-out loci	942	97.66	91.30	74.52	77.05	85.13
Held-out neighborhoods	109	95.41	**97.25**	**81.65**	65.02	84.83

G Target Gene Prediction Results

Table 8: Concordance between BioBLIP and Qwen3-32B on non-closest gene identification at held-out loci ($n = 119$). BioBLIP correctly identifies the causal gene in 21 cases where Qwen3-32B fails, while the reverse occurs in only 1 case.

	Qwen3-32B Correct	Qwen3-32B Incorrect	Total
BioBLIP Correct	51	21	72 (60.5%)
BioBLIP Incorrect	1	46	47
Total	52 (43.7%)	67	119

Table 9: Error pattern breakdown for the 21 cases where BioBLIP identifies the correct causal gene and Qwen3-32B does not. All three patterns stem from Qwen3-32B's proximity anchoring heuristic: in 19 of the 21 cases it selects the gene that overlaps or lies closest to the variant.

Pattern	Cases	Description
Metabolite–enzyme specificity	11 (52.4%)	Qwen3-32B selects the overlapping or nearest gene. In 9 of 11 cases that gene has no established role in the metabolite's production, degradation, or transport; in the other two it is a neighboring enzyme of the same pathway (*DMGDH* for betaine) or a paralog (*IDO2*). BioBLIP selects the metabolic enzyme or transporter.
Multi-gene locus disambiguation	7 (33.3%)	At lipid, metabolic-disease, and pigmentation loci with several neighboring genes, Qwen3-32B anchors on a proximal gene. At the *SORT1* and *ASIP* loci it names the causal gene in its own trace but does not select it.
Canonical GWAS locus recognition	3 (14.3%)	At well-characterized loci where the causal gene lies 113–376 kb from the variant (8q24 *MYC*, 2q21 *LCT*, 7q11.23 *MLXIPL*), Qwen3-32B selects the gene the variant overlaps. BioBLIP's annotation lists the effector gene, and in two of the three cases omits the overlapping gene altogether.

G.1 Illustrative Reasoning Traces

We present six paired reasoning traces produced by BioBLIP and Qwen3-32B for non-closest causal gene identification at held-out loci. Each pair illustrates one of the three patterns in Table 9. For each case, we provide the variant and trait context, followed by the five-step trace from each model. Traces are reproduced verbatim apart from formatting; for BioBLIP, we show the variant type, consequence, and nearest-gene fields of its Step 1 annotation. Gene distances in the context paragraphs are measured from the variant. The correct prediction is marked with ✓ and the incorrect prediction with ×.

G.2 Pattern 1: Metabolite–Enzyme Specificity

Case A: Nonanoylcarnitine

Context. Variant rs3738934 (2_210009522_T_C); trait: nonanoylcarnitine; gold standard gene: *ACADL* (acyl-CoA dehydrogenase, long chain), a mitochondrial fatty acid β-oxidation enzyme. The variant lies within *RPE* (ribulose-5-phosphate-3-epimerase); *ACADL* is 177.6 kb away.

BioBLIP (*ACADL* ✓)

> *Step 1 — Variant Annotation.* Variant type: noncoding; consequence: intron variant; nearest genes: *ACADL*, *RPE*, *UNC80*, *KANSL1L*.
>
> *Step 2 — Gene-Level Effect.* The variant is located in the intronic region of *ACADL* (acyl-CoA dehydrogenase, long-chain), a gene encoding an enzyme involved in the mitochondrial fatty acid β-oxidation pathway. Intronic variants can affect splicing efficiency or gene expression levels. This variant may alter splicing or regulatory elements that control *ACADL* expression. Direction: Likely decreased expression or altered splicing of *ACADL*, leading to reduced enzymatic activity.
>
> *Step 3 — Cellular Process.* Impacted pathway: Mitochondrial fatty acid β-oxidation. Cellular consequence: Reduced activity of *ACADL* impairs the initial step of long-chain fatty acid oxidation, leading to the accumulation of long-chain acylcarnitines, including nonanoylcarnitine, in the cytosol due to incomplete mitochondrial entry or metabolism.
>
> *Step 4 — Tissue-Level Effect.* Tissue: Liver. Outcome: In hepatocytes, impaired fatty acid oxidation leads to the accumulation of nonanoylcarnitine and other long-chain acylcarnitines, which can disrupt mitochondrial function and energy metabolism. This may also contribute to hepatic steatosis or metabolic dysregulation.

Step 5 — Disease Mechanism. The accumulation of nonanoylcarnitine in the liver reflects a defect in long-chain fatty acid metabolism. This can lead to metabolic stress, mitochondrial dysfunction, and potentially contribute to metabolic disorders such as fatty liver disease or impaired energy homeostasis.
Causal gene: *ACADL*.

Qwen3-32B (*RPE* ×)

Step 1 — Variant Consequence. The variant 2_210009522_T_C is located on chromosome 2 at position 210,009,522. It overlaps with the gene *RPE* (Retinal Pigment Epithelium-Specific 65kDa Protein), which is a well-characterized gene involved in lipid metabolism and peroxisomal function. Given the variant's position within the gene, it is likely a missense variant that alters the amino acid sequence of the *RPE* protein. This type of variant can directly affect the structure and function of the encoded protein.
Step 2 — Gene-Level Effect. RPE encodes a peroxisomal membrane protein that is essential for the import of enzymes into peroxisomes, including those involved in fatty acid oxidation. A missense variant in *RPE* could impair its ability to facilitate the transport of these enzymes, thereby disrupting peroxisomal function. Direction: Loss of function — the variant likely reduces the efficiency or functionality of the *RPE* protein.
Step 3 — Cellular Process. Impacted pathway: Peroxisomal fatty acid oxidation — specifically, the breakdown of very long-chain fatty acids and branched-chain fatty acids such as nonanoic acid. Cellular consequence: Impaired peroxisomal function leads to the accumulation of fatty acid intermediates, including nonanoylcarnitine, which is a carnitine ester of nonanoic acid. This accumulation is a hallmark of peroxisomal dysfunction.
Step 4 — Tissue-Level Effect. Tissue: Liver and muscle — tissues with high metabolic activity and significant peroxisomal fatty acid oxidation. Outcome: Accumulation of nonanoylcarnitine in these tissues due to impaired peroxisomal import of enzymes, leading to metabolic stress and potential mitochondrial dysfunction.
Step 5 — Disease Mechanism. The accumulation of nonanoylcarnitine and other fatty acid intermediates due to impaired peroxisomal function disrupts normal lipid metabolism. This can lead to the observed trait of elevated nonanoylcarnitine levels, which is often associated with peroxisomal disorders or metabolic dysregulation. The trait reflects a downstream consequence of *RPE* dysfunction in peroxisomal enzyme trafficking.
Causal gene: *RPE*.

Analysis. Qwen3-32B commits to *RPE* in Step 1 because the variant overlaps it, then describes the wrong protein: it names *RPE* as the "Retinal Pigment Epithelium-Specific 65kDa Protein" (the product of *RPE65*) and assigns it an invented role in peroxisomal enzyme import. *RPE* encodes ribulose-5-phosphate-3-epimerase of the pentose phosphate pathway and has no established role in acylcarnitine metabolism. BioBLIP selects *ACADL*, the β-oxidation enzyme, although its Step 2 places the variant in an *ACADL* intron, which is positionally incorrect (177.6 kb away), and it calls nonanoylcarnitine (C9) a long-chain acylcarnitine.

CASE B: GLUTARYLCARNITINE

Context. Variant rs13375749 (1_53230210_T_C); trait: glutarylcarnitine; gold standard gene: *CPT2* (carnitine palmitoyltransferase 2), the inner mitochondrial membrane enzyme of the carnitine shuttle that converts acylcarnitines back to acyl-CoAs. The variant lies within *MAGOH*; *CPT2* is 16.0 kb away.

BioBLIP (*CPT2* ✓)

Step 1 — Variant Annotation. Variant type: noncoding; consequence: intron variant; nearest genes: *CPT2*, *LRP8*, *MAGOH*, *CZIB*.

Step 2 — Gene-Level Effect. The variant is located in the intronic region of *CPT2* (Carnitine Palmitoyltransferase 2), a gene encoding a mitochondrial enzyme critical for the transport of long-chain fatty acids into the mitochondrial matrix for β-oxidation. Intronic variants can affect splicing efficiency or gene expression levels. Direction: Likely leads to altered splicing or reduced expression of *CPT2*, impairing its enzymatic activity.

Step 3 — Cellular Process. Impacted pathway: Fatty acid β-oxidation in mitochondria. Cellular consequence: Reduced *CPT2* activity leads to impaired mitochondrial fatty acid oxidation, resulting in the accumulation of long-chain acylcarnitines, including glutarylcarnitine.

Step 4 — Tissue-Level Effect. Tissue: Liver and skeletal muscle (primary sites of fatty acid oxidation). Outcome: Accumulation of long-chain acylcarnitines, including glutarylcarnitine, in the liver and bloodstream due to impaired mitochondrial β-oxidation.

Step 5 — Disease Mechanism. The accumulation of glutarylcarnitine and other long-chain acylcarnitines is a hallmark of mitochondrial fatty acid oxidation disorders. This can lead to metabolic stress, impaired energy production, and potentially contribute to hepatic steatosis or systemic metabolic dysregulation.

Causal gene: *CPT2*.

Qwen3-32B (*MAGOH* $\times$)

Step 1 — Variant Consequence. The variant 1_53230210_T_C is located on chromosome 1 at position 53,230,210. Based on the nearby genes, the closest gene is *MAGOH* (MAGO Homolog, meiosis-specific), which is overlapping with the variant. This suggests the variant may be located within the coding or regulatory region of *MAGOH*. Given the proximity and overlap, it is likely a missense or regulatory variant affecting *MAGOH* expression or function.

Step 2 — Gene-Level Effect. MAGOH is a component of the exon junction complex (EJC), which is involved in mRNA splicing, nonsense-mediated decay (NMD), and mRNA export. A variant in *MAGOH* could disrupt its ability to form the EJC or interfere with its interaction with other splicing factors. Direction: Likely a loss of function, as disruption of *MAGOH* could impair proper mRNA splicing and lead to the accumulation of aberrant transcripts.

Step 3 — Cellular Process. Impacted pathway: mRNA splicing and nonsense-mediated decay (NMD). Cellular consequence: Defective splicing could lead to the production of truncated or non-functional proteins, and impaired NMD could allow the accumulation of aberrant mRNAs, potentially leading to cellular stress or dysfunction.

Step 4 — Tissue-Level Effect. Tissue: Liver and muscle (as *MAGOH* is ubiquitously expressed but has particularly important roles in tissues with high metabolic activity). Outcome: Altered splicing and mRNA stability in liver and muscle cells could disrupt metabolic pathways, including fatty acid oxidation and amino acid metabolism, which are relevant to carnitine-related traits.

Step 5 — Disease Mechanism. Disruption of *MAGOH* function in liver and muscle could impair the regulation of genes involved in carnitine metabolism, leading to altered levels of glutarylcarnitine, a branched-chain acylcarnitine. This may reflect a disturbance in the metabolism of glutaric acid or related pathways, potentially contributing to metabolic dysregulation.

Causal gene: *MAGOH*.

Analysis. Both genes appear among the candidates; BioBLIP ranks *CPT2* first and ties it directly to the carnitine shuttle. Qwen3-32B selects *MAGOH* because it overlaps the variant and builds a long, indirect chain from exon junction complex function through mRNA splicing and nonsense-mediated decay to carnitine metabolism. *MAGOH* has no established role in acylcarnitine metabolism. Neither trace characterizes the metabolite correctly: glutarylcarnitine is a C5 dicarboxylic acylcarnitine, which BioBLIP calls long-chain and Qwen3-32B calls branched-chain. As in Case A, BioBLIP's Step 2 incorrectly places the variant in an intron of its chosen gene.

G.3 Pattern 2: Multi-Gene Locus Disambiguation

Case C: LDL Cholesterol (1p13.3 locus)

Context. Variant rs646776 (1_109275908_C_T), the canonical 1p13.3 LDL locus; trait: LDL cholesterol; gold standard gene: *SORT1* (sortilin), which controls hepatic VLDL/LDL secretion. The variant is 157 bp from *CELSR2* and 33.7 kb from *SORT1*. The rs12740374 LD block at this locus is one of the best-characterized noncoding causal mechanisms in cardiovascular genetics.

BioBLIP (*SORT1* ✓)

Step 1 — Variant Annotation. Variant type: noncoding; consequence: intron variant; nearest genes: *CELSR2*, *PSRC1*, *MYBPHL*, *SORT1*.

Step 2 — Gene-Level Effect. The variant is located in the intronic region of the *CELSR2* gene, which is closely linked to *SORT1*. *SORT1* encodes a protein involved in the sorting and trafficking of low-density lipoprotein receptor-related protein 1 (LRP1) and apolipoprotein B (ApoB), both of which are critical for lipid metabolism and LDL cholesterol homeostasis. This intronic variant may alter splicing or regulatory elements, leading to altered *SORT1* expression or function. Direction: Likely a gain of function or increased expression of *SORT1* due to altered regulatory elements.

Step 3 — Cellular Process. Impacted pathway: Lipoprotein metabolism and LDL receptor-mediated endocytosis. Cellular consequence: Increased *SORT1* expression enhances the retention of ApoB-containing lipoproteins in the liver, reducing their secretion into the bloodstream and thereby lowering circulating LDL cholesterol levels.

Step 4 — Tissue-Level Effect. Tissue: Liver. Outcome: Reduced secretion of LDL particles into the bloodstream due to enhanced retention of ApoB-containing lipoproteins in hepatocytes.

Step 5 — Disease Mechanism. The reduced secretion of LDL particles from the liver leads to lower circulating LDL cholesterol levels, which is associated with a decreased risk of atherosclerosis and cardiovascular disease.

Causal gene: *SORT1*.

Qwen3-32B (*CELSR2* ×)

Step 1 — Variant Consequence. The variant 1_109275908_C_T is located 157 bp away from the gene *CELSR2*, which is a strong candidate for being the affected gene. Given its proximity to the transcription start site of *CELSR2*, this variant is likely a regulatory variant, possibly affecting enhancer or promoter activity. Regulatory variants in this region can influence gene expression levels, particularly of *CELSR2*, which is known to be co-expressed with *PSRC1* and *SORT1* in lipid metabolism pathways. This variant is likely to alter the expression of *CELSR2* and possibly *PSRC1* or *SORT1* through cis-regulatory mechanisms.

Step 2 — Gene-Level Effect. CELSR2 encodes a member of the cadherin-eptin-spondin-repeat-containing protein family and is involved in lipid metabolism, particularly in the regulation of low-density lipoprotein (LDL) cholesterol. It is part of a gene cluster with *PSRC1* and *SORT1*, which are also implicated in lipid homeostasis. The variant likely reduces the expression of *CELSR2*, which in turn affects the function of the *SORT1* gene, a known regulator of hepatic LDL receptor expression. Direction: Decreased expression of *CELSR2* and possibly *SORT1*.

Step 3 — Cellular Process. Impacted pathway: LDL metabolism and cholesterol homeostasis. Cellular consequence: Reduced expression of *CELSR2* and *SORT1* leads to impaired clearance of LDL cholesterol from the bloodstream. This results in increased circulating LDL levels due to reduced hepatic uptake of LDL particles.

Step 4 — Tissue-Level Effect. Tissue: Liver. Outcome: The liver is the primary site of LDL receptor expression and LDL clearance. Reduced *SORT1* activity in the liver leads to decreased LDL receptor expression and reduced LDL uptake, resulting in elevated plasma LDL cholesterol levels.

Step 5 — Disease Mechanism. The reduced expression of *CELSR2* and *SORT1* in the liver impairs the normal clearance of LDL cholesterol, leading to a chronic elevation in circulating LDL levels. This contributes to the development of dyslipidemia and increases the risk of atherosclerosis and cardiovascular disease.

Causal gene: *CELSR2*.

Analysis. BioBLIP's Step 1 lists *CELSR2*, *PSRC1*, and *SORT1* as candidate genes. In Step 2 it places the variant at *CELSR2* but identifies *SORT1* as the gene whose altered hepatic expression changes LDL secretion, the established mechanism at this locus. Qwen3-32B anchors on *CELSR2* because of its 157 bp distance and misdescribes it as a "cadherin-eptin-spondin-repeat" protein. It mentions *SORT1* throughout its trace and even attributes the LDL effect to reduced *SORT1* activity in Step 4, yet it still names *CELSR2* as causal.

CASE D: SUNTAN

Context. Variant 20_34077942_A_G; trait: suntan; gold standard gene: *ASIP* (Agouti Signaling Protein), an antagonist of MC1R. Increased ASIP expression suppresses eumelanin synthesis and reduces tanning capacity. The variant lies within *RALY*, 10.4 kb from *EIF2S2* and 116.6 kb from *ASIP*.

BioBLIP (*ASIP* ✓)

Step 1 — Variant Annotation. Variant type: noncoding; consequence: intron variant; nearest genes: *ASIP*, *RALY*, *EIF2S2*.

Step 2 — Gene-Level Effect. The variant is located in the intronic region of the *ASIP* (Agouti Signaling Protein) gene, which encodes a paracrine signaling protein that antagonizes the melanocortin 1 receptor (MC1R) in melanocytes. This interaction regulates the type of melanin produced (eumelanin vs. pheomelanin), which directly affects skin pigmentation. An intronic variant may alter splicing or regulatory elements, potentially affecting *ASIP* expression or function. Direction: Likely increased expression or altered splicing of *ASIP*, leading to enhanced antagonism of MC1R.

Step 3 — Cellular Process. Impacted pathway: Melanin synthesis pathway, specifically the MC1R signaling cascade. Cellular consequence: Increased *ASIP* activity reduces MC1R signaling, shifting melanin production from eumelanin (dark, photoprotective) to pheomelanin (light, less protective), resulting in lighter skin pigmentation and reduced tanning capacity.

Step 4 — Tissue-Level Effect. Tissue: Skin (melanocytes in the epidermis). Outcome: Reduced ability to produce eumelanin in response to UV exposure, leading to a diminished capacity to develop a suntan and increased susceptibility to UV-induced damage.

Step 5 — Disease Mechanism. The reduced MC1R signaling due to increased *ASIP* activity leads to a lighter baseline skin tone and a reduced capacity to tan upon sun exposure. This is a non-pathological trait but is associated with increased risk of UV-induced DNA damage and skin cancer in individuals with this genetic background.

Causal gene: *ASIP*.

Qwen3-32B (*RALY* ×)

Step 1 — Variant Consequence. The variant 20_34077942_A_G is located within the gene *RALY*, which is annotated as overlapping with the variant. This suggests the variant is likely a coding variant (e.g., missense or synonymous) within the *RALY* gene. Given the proximity to the gene and the overlapping annotation, it is most likely a missense variant that alters the amino acid sequence of the *RALY* protein. *RALY* is a RNA-binding protein involved in post-transcriptional regulation of gene expression, including splicing and mRNA stability.

> *Step 2 — Gene-Level Effect. RALY* encodes a RNA-binding protein that regulates the stability and splicing of mRNAs, particularly those involved in pigmentation and skin development. A missense variant in *RALY* could alter its RNA-binding affinity or splicing activity, leading to dysregulation of downstream targets. Direction: Likely a gain or loss of function, depending on the specific amino acid change. If the variant disrupts RNA-binding domains, it could result in loss of function.
>
> *Step 3 — Cellular Process.* Impacted pathway: RNA processing and splicing, particularly of genes involved in melanogenesis and skin pigmentation. Cellular consequence: Altered splicing or stability of mRNAs encoding pigmentation-related proteins (e.g., MC1R, *ASIP*, or others) could lead to changes in melanin production or distribution in melanocytes.
>
> *Step 4 — Tissue-Level Effect.* Tissue: Skin (melanocytes). Outcome: Altered melanin synthesis or distribution in melanocytes, potentially leading to increased or decreased pigmentation in response to UV exposure.
>
> *Step 5 — Disease Mechanism.* The altered function of *RALY* due to the missense variant may lead to dysregulation of pigmentation-related mRNAs, resulting in a more efficient or sustained tanning response to UV exposure. This could manifest as a trait of increased suntanability or reduced susceptibility to sunburn, depending on the downstream effects on melanin production and skin pigmentation.
>
> Causal gene: *RALY*.

Analysis. BioBLIP produces a specific mechanistic chain: ASIP $\dashv$ MC1R $\rightarrow$ eumelanin, with increased ASIP expression reducing tanning capacity. Its Step 2 places the variant in an *ASIP* intron, which is positionally incorrect: the variant lies in *RALY*. Qwen3-32B picks *RALY* because it overlaps the variant, assumes a missense change, and constructs a generic post-transcriptional regulation story. It lists ASIP and MC1R among the pigmentation mRNAs that RALY might regulate. It possesses the relevant biological knowledge but cannot use it to override the proximity anchor.

G.4 Pattern 3: Canonical GWAS Locus Recognition

Case E: Breast Carcinoma (8q24 locus)

Context. Variant 8_127359926_A_G; trait: breast carcinoma; gold standard gene: *MYC*. The 8q24 gene desert is one of the most-replicated cancer GWAS regions; noncoding variants here act through tissue-specific enhancers that loop to *MYC*. The variant lies within *POU5F1B*; *MYC* is 375.5 kb away.

BioBLIP (*MYC* ✓)

> *Step 1 — Variant Annotation.* Variant type: noncoding; consequence: intron variant; nearest genes: *POU5F1B*, *MYC*.
>
> *Step 2 — Gene-Level Effect.* The variant is located in the intronic region of *POU5F1B*, a gene encoding a transcription factor with regulatory roles in stem cell pluripotency and early development. It is also in close proximity to *MYC*, a well-known proto-oncogene that regulates cell proliferation and apoptosis. The intronic variant may alter splicing or regulatory elements, potentially affecting the expression or function of *POU5F1B* or *MYC*. Direction: Likely gain of regulatory activity or altered splicing, potentially increasing *MYC* expression or dysregulating *POU5F1B* function.
>
> *Step 3 — Cellular Process.* Impacted pathway: Wnt/β-catenin signaling and cell cycle regulation. *MYC* is a central player in these pathways, and dysregulation can lead to uncontrolled cell proliferation. Cellular consequence: Increased cell proliferation and reduced apoptosis in breast epithelial cells, promoting a pre-malignant or malignant phenotype.
>
> *Step 4 — Tissue-Level Effect.* Tissue: Breast tissue, particularly epithelial cells in the mammary gland. Outcome: Enhanced proliferation and reduced differentiation of epithelial cells, increasing the risk of tumor initiation and progression.
>
> *Step 5 — Disease Mechanism.* The dysregulation of *MYC* expression due to the variant contributes to the development of breast carcinoma by promoting

uncontrolled cell growth and inhibiting apoptosis. *POU5F1B* may also play a role in modulating stem cell behavior, and its disruption could further contribute to tumor heterogeneity and aggressiveness.

Causal gene: *MYC*.

Qwen3-32B (*POU5F1B* ×)

Step 1 — Variant Consequence. The variant 8_127359926_A_G is located on chromosome 8 at position 127,359,926. Given the nearby genes, the most proximal gene is *POU5F1B*, which is described as overlapping with the variant. This suggests the variant may lie within the gene body or regulatory region of *POU5F1B*. *POU5F1B* is a transcription factor gene, and variants within or near it are likely to affect gene expression or function. The variant may be a regulatory variant (e.g., enhancer or promoter) or an intronic/splice variant that alters splicing or transcriptional regulation of *POU5F1B*.

Step 2 — Gene-Level Effect. POU5F1B is a member of the POU (Pit-Oct-Unc) family of transcription factors and is known to regulate genes involved in cell proliferation and differentiation. It is a pseudogene of POU5F1 (OCT4), which is a key regulator of pluripotency and early embryonic development. However, *POU5F1B* has retained some regulatory function and is expressed in certain tissues. A regulatory variant in *POU5F1B* could alter its expression levels or disrupt its ability to regulate downstream genes. Direction: The variant is likely to cause altered expression of *POU5F1B*, potentially increasing its expression or disrupting its regulatory function.

Step 3 — Cellular Process. Impacted pathway: Transcriptional regulation of genes involved in cell cycle control and proliferation. *POU5F1B* may regulate genes that are also regulated by POU5F1 (OCT4), such as those involved in stem cell maintenance and differentiation. Cellular consequence: Altered *POU5F1B* expression may lead to dysregulation of cell cycle genes, promoting uncontrolled cell proliferation or reduced differentiation, both of which are hallmarks of cancer.

Step 4 — Tissue-Level Effect. Tissue: Breast tissue, particularly mammary epithelial cells, where *POU5F1B* may be expressed or where its regulatory effects are most relevant. Outcome: In breast tissue, dysregulated *POU5F1B* expression may lead to increased proliferation of epithelial cells, reduced apoptosis, or loss of differentiation, contributing to the formation of breast carcinoma.

Step 5 — Disease Mechanism. The altered expression or function of *POU5F1B* due to the 8_127359926_A_G variant may disrupt normal transcriptional regulation in breast epithelial cells. This could lead to the activation of oncogenic pathways or the suppression of tumor suppressor genes, promoting malignant transformation and tumor progression in the breast.

Causal gene: *POU5F1B*.

Analysis. Both models agree that the variant falls within *POU5F1B*. BioBLIP lists only *POU5F1B* and *MYC* as candidates and attributes the risk to *MYC* dysregulation, consistent with the known regulatory architecture at 8q24, although it describes *MYC* as being "in close proximity" to a variant 375.5 kb away. Qwen3-32B commits to *POU5F1B* in Step 1 and, despite noting that it is a pseudogene of *POU5F1*, builds a coherent but incorrect transcription-factor mechanism. It never mentions *MYC*. This suggests that Qwen3-32B lacks reliable knowledge of long-range regulatory loci and defaults to a proximity heuristic.

CASE F: DIET MEASUREMENT (2Q21 LACTASE LOCUS)

Context. Variant 2_135950412_T_C; trait: diet measurement; gold standard gene: *LCT* (lactase). The lactase-persistence haplotype on 2q21 spans roughly a megabase as a result of recent positive selection; its functional variant (rs4988235, −13910C>T) lies in an *MCM6* intron and acts as an enhancer of *LCT*. The variant here lies within *DARS*, 74.0 kb from *MCM6* and 113.2 kb from *LCT*.

BioBLIP (*LCT* ✓)

Step 1 — Variant Annotation. Variant type: noncoding; consequence: intron variant; nearest genes: *LCT*, *MCM6*.

Step 2 — Gene-Level Effect. The variant is located in the intronic region of the *LCT* gene, which encodes lactase, an enzyme responsible for the hydrolysis of lactose into glucose and galactose in the small intestine. Intronic variants can affect splicing efficiency or regulatory elements that influence gene expression. Direction: Likely affects *LCT* expression or splicing, potentially leading to reduced lactase activity.

Step 3 — Cellular Process. Impacted pathway: Carbohydrate digestion and absorption. Cellular consequence: Reduced lactase activity in intestinal epithelial cells leads to impaired lactose digestion, resulting in lactose intolerance symptoms in some individuals.

Step 4 — Tissue-Level Effect. Tissue: Small intestine. Outcome: Decreased lactase activity in the brush border of the small intestine reduces the ability to digest lactose, potentially altering dietary lactose tolerance and influencing dietary patterns.

Step 5 — Disease Mechanism. Individuals with reduced lactase activity may avoid dairy products or consume less lactose, which can influence overall diet composition and nutrient intake (e.g., calcium, vitamin D). This dietary adaptation can be reflected in diet measurement studies, particularly in populations with high prevalence of lactose intolerance.

Causal gene: *LCT*.

Qwen3-32B (*DARS* $\times$)

Step 1 — Variant Consequence. The variant 2_135950412_T_C is located on chromosome 2 at position 135,950,412. It overlaps with the gene *DARS* (Dihydroxyacetone synthase), which is directly annotated as the gene containing the variant. *DARS* is an enzyme involved in the biosynthesis of isoprenoids, a class of molecules essential for various cellular functions, including membrane integrity and signaling. This variant is likely a missense or regulatory variant within the *DARS* gene, potentially altering its expression or enzymatic activity.

Step 2 — Gene-Level Effect. DARS catalyzes the formation of dihydroxyacetone phosphate, a precursor in the mevalonate pathway, which is critical for cholesterol and isoprenoid synthesis. A variant in *DARS* could disrupt this enzymatic activity or alter gene expression. Direction: Depending on the nature of the variant, it could result in reduced enzymatic activity or altered expression levels of *DARS*, leading to decreased isoprenoid production.

Step 3 — Cellular Process. Impacted pathway: Mevalonate pathway (isoprenoid biosynthesis). Cellular consequence: Reduced isoprenoid synthesis can impair post-translational modification of proteins (e.g., prenylation), which is essential for the function of signaling proteins such as small GTPases (e.g., Ras, Rho). This can disrupt cellular signaling and metabolic homeostasis.

Step 4 — Tissue-Level Effect. Tissue: Adipose tissue and liver, which are central to lipid metabolism and energy homeostasis. Outcome: Altered isoprenoid metabolism in these tissues may affect lipid storage, insulin sensitivity, and energy expenditure, potentially influencing dietary intake and metabolic responses to food.

Step 5 — Disease Mechanism. Variants in *DARS* that reduce isoprenoid synthesis may alter metabolic signaling pathways in adipose and liver tissues, leading to changes in energy balance and nutrient utilization. These changes could influence dietary preferences or metabolic responses to diet, contributing to the observed association with diet measurement traits such as caloric intake or macronutrient distribution.

Causal gene: *DARS*.

Analysis. BioBLIP's annotation lists only *LCT* and *MCM6*, omitting *DARS*, and its Step 2 incorrectly places the variant in an *LCT* intron. Its gene choice and mechanism nonetheless match the known biology: lactase expression in the small intestine shapes dairy intake. Qwen3-32B anchors on

DARS, misidentifies it as "dihydroxyacetone synthase" (*DARS* encodes aspartyl-tRNA synthetase), and constructs a mevalonate-pathway explanation for dietary behavior. It never mentions *LCT*, although lactase persistence is one of the best-known diet-related variants in the human genome.

H VARIANT PRIORITIZATION RESULTS

H.1 ILLUSTRATIVE REASONING TRACES

The traces below come from BioBLIP on the TraitGym Mendelian traits benchmark. Candidate variant IDs are anonymized as `var_001`–`var_010` before ranking, so the ranking model sees only BioBLIP's annotation of each candidate (predicted chromosome, consequence, and nearest genes), shown in the table above each trace. The true variant IDs are listed for reference only, and ⋆ marks the causal variant. Annotations are reproduced as generated, including errors.

Trace 1 Trait: POLYDACTYLY, PREAXIAL II Match group: `dELS_1`

Result: `var_001` (7_156791413_A_C) ✓ Rank 1. BioBLIP annotates the causal variant correctly as a chr7 intron variant in *LMBR1*, the gene that contains the ZRS limb enhancer of *SHH*.

		BioBLIP annotation		
ID	**Variant**	**Chr.**	**Consequence**	**Nearest genes**
`var_001` ⋆	7_156791413_A_C	chr7	intron variant	*LMBR1*
`var_002`	7_152538209_T_C	chr7	intron variant	*AK092532.1*, *KMT2C*, *XRCC5*
`var_003`	7_47710497_G_A	chr7	intron variant	*PKD1L1*
`var_004`	7_101919211_G_T	chr7	intron variant	*CUX1*
`var_005`	7_135335371_G_A	chr22	upstream gene variant	*C20orf71*
`var_006`	7_17030613_T_C	chr7	intron variant	*AHR*, *AC074145.1*, *AC074145.2*
`var_007`	7_157519986_C_T	chr7	intron variant	*PTRPN*
`var_008`	7_76728900_G_T	chr7	intron variant	*AC004088.1*, *POMZP3*, *FGL2*
`var_009`	7_148032157_A_T	chr7	intron variant	*CNTNAP2*
`var_010`	7_47710453_C_T	chr7	intron variant	*PKD1L1*

BioBLIP (*LMBR1* ✓)

> Okay, let's tackle this problem. The trait in question is Preaxial Polydactyly Type II, which is a condition where extra fingers or toes are present on the thumb or big toe side. I remember that this condition is often associated with mutations in certain genes, particularly those involved in limb development.
>
> First, I need to look at the candidate variants provided. All of them are noncoding intron variants on chromosome 7 except for var_005, which is on chr22. The key here is to identify which gene is most likely involved in preaxial polydactyly.
>
> I recall that the LMBR1 gene is part of the ZRS (Zone of Polarizing Activity Regulatory Sequence) region, which is a long-range enhancer for the Sonic Hedgehog (SHH) gene. SHH is crucial for limb development, and mutations in the ZRS region are known to cause various limb malformations, including preaxial polydactyly. So var_001, which is near LMBR1, is a strong candidate.
>
> Looking at the other variants, var_003 and var_010 are both near PKD1L1. I'm not as familiar with PKD1L1's role in limb development, so maybe they're less likely. The other genes like CUX1, AHR, CNTNAP2, etc., don't immediately ring a bell in the context of polydactyly.
>
> Since var_001 is the only one near LMBR1, which is linked to SHH regulation and known to be involved in this type of polydactyly, it's probably the causal variant. The rest are either on different chromosomes or near genes not typically associated with this condition. Even though some are on chr7, their proximity to other genes doesn't seem relevant here.
>
> So the ranking would put var_001 first, followed by others based on their gene relevance. But since most are not linked to known polydactyly genes, they would be lower in the ranking.
>
> Ranked Candidates: 1. `var_001`, 2. `var_003`, 3. `var_010`, 4. `var_002`, 5. `var_004`, 6. `var_006`, 7. `var_007`, 8. `var_008`, 9. `var_009`, 10. `var_005`

Trace 2 Trait: IMMUNODYSREGULATION, POLYENDOCRINOPATHY, AND ENTEROPATHY, X-LINKED (IPEX) Match group: `3_prime_UTR_variant_27`

Result: `var_004` (X_49250456_T_C) × Rank 10. The causal variant is a 3′ UTR variant at the *FOXP3*/*CCDC22* locus on chrX, but BioBLIP annotates it as a chr11 missense variant near *POC5*, so *FOXP3* never appears among the candidate annotations.

		BioBLIP annotation		
ID	**Variant**	**Chr.**	**Consequence**	**Nearest genes**
var_001	X_136214250_C_T	chrX	intron variant	*MAP7D3*, *FAM120A*, *FAM120A-AS1*
var_002	X_41696942_A_T	chrX	missense variant	*GPX7*, *GPR149*
var_003	X_20125252_G_T	chrX	3 prime UTR variant	*EIF1AX*
var_004 ⋆	X_49250456_T_C	chr11	missense variant	*POC5*
var_005	X_111689561_G_C	chrX	splice region variant	*ALG13*
var_006	X_48608798_C_T	chr1	3 prime UTR variant	*WDR37*
var_007	X_155057576_C_G	chr20	intron variant	*FTCD2C1*, *FTCD2C2*, *FTCD2C3*
var_008	X_143626437_T_C	chrX	3 prime UTR variant	*SLITRK4*
var_009	X_145825876_A_G	chrX	3 prime UTR variant	*SLITRK2*
var_010	X_111519499_T_C	chrX	intron variant	*STRN2*

BioBLIP (*ALG13* ×)

Okay, let's tackle this problem. The trait in question is Immunodysregulation, Polyendocrinopathy, and Enteropathy, X-linked (IPEX). I know that IPEX is a rare genetic disorder typically caused by mutations in the FOXP3 gene, which is located on the X chromosome. FOXP3 is crucial for the development and function of regulatory T cells, and its dysfunction leads to the autoimmune symptoms seen in IPEX.

Looking at the candidate variants provided, all are on the X chromosome except var_004 (chr11) and var_006 (chr1), so those two can be ruled out immediately since IPEX is X-linked. That leaves us with variants on chrX.

Now, let's go through the remaining variants. The key here is to identify which gene is associated with IPEX. As mentioned, FOXP3 is the primary gene involved. However, none of the variants here are in FOXP3. Wait, maybe I need to check the nearest genes listed for each variant. Let me look again.

Var_001 is near MAP7D3, FAM120A, FAM120A-AS1. Not familiar with these, but not FOXP3.

Var_002 is near GPX7, GPR149. Again, not FOXP3.

Var_003 is near EIF1AX. EIF1AX is involved in translation initiation, but not linked to IPEX as far as I know.

Var_005 is near ALG13. ALG13 is part of the oligosaccharyltransferase complex, involved in N-linked glycosylation. Not FOXP3.

Var_008 is near SLITRK4, with an amino acid change R148. SLITRK4 is a receptor for SLIT proteins, involved in axon guidance. Not FOXP3.

Var_009 is near SLITRK2. Similar to above, not FOXP3.

Var_010 is near STRN2. STRN2 is a striatin family member, involved in signaling complexes. Not FOXP3.

Wait, none of these variants are in FOXP3. That's confusing. Maybe the question is testing if I know that the correct gene isn't listed here, but perhaps there's a mistake? Or maybe I'm missing something. Let me check again.

Wait, the user says exactly one is the true causal variant. But if none of the variants are in FOXP3, then perhaps there's an error in the question. Alternatively, maybe the correct gene is not FOXP3? But IPEX is almost always caused by FOXP3 mutations. Let me confirm. Yes, IPEX is caused by mutations in FOXP3. So if none of the variants here are in FOXP3, then perhaps the correct answer is not listed. But the user says exactly one is the true causal variant. Hmm.

Wait, maybe I made a mistake in checking the nearest genes. Let me go through each variant again carefully.

Var_005 is a splice region variant in ALG13. ALG13 is on X chromosome. But IPEX is FOXP3. So unless there's another gene involved, which I'm not aware of. Maybe the question is a trick, but the user says exactly one is correct. Alternatively, maybe the correct variant is in a gene that's not FOXP3 but is related. For example, maybe a gene involved in the same pathway. But I don't recall any other genes being the primary cause for IPEX.

Wait, perhaps the user provided a typo in the gene names? Let me check the nearest genes again. For example, maybe one of the genes is a pseudogene or a related gene. Let me check each variant's nearest genes again.

Var_003 is near EIF1AX. EIF1AX is on X chromosome. But not related to IPEX.

Var_008 is near SLITRK4. SLITRK4 is on X. Not related.

Var_009 is near SLITRK2. Also not related.

Var_002 is near GPX7 and GPR149. GPX7 is a glutathione peroxidase, but again, not related.

Var_005 is near ALG13. ALG13 is part of the oligosaccharyltransferase complex. Not related to IPEX.

Hmm. This is perplexing. Since the user insists that one of these is the correct answer, perhaps I need to consider other factors. Maybe the variant type? For example, splice region variants can be very impactful. Var_005 is a splice region variant in ALG13. But again, not FOXP3.

Alternatively, maybe the correct variant is in a gene that's not listed here. Wait, maybe the user made a mistake in the gene names. Let me check if any of the genes listed are actually FOXP3. For example, maybe a typo in the gene name. For example, "FOXP3" might be misspelled as something else. Let me check each variant's nearest genes again.

Var_001: MAP7D3, FAM120A, FAM120A-AS1. No.

Var_002: GPX7, GPR149, GPX7. No.

Var_003: EIF1AX. No.

Var_005: ALG13. No.

Var_008: SLITRK4. No.

Var_009: SLITRK2. No.

Var_010: STRN2. No.

So none of the variants are in FOXP3. Therefore, perhaps the correct answer is not listed here, but the user says exactly one is. Maybe I'm missing something. Wait, maybe the variant is in a non-coding region but affects FOXP3 through a regulatory element? For example, a long-range enhancer. But the nearest genes listed for the non-coding variants are not FOXP3. For example, var_001 is near MAP7D3, etc. So unless there's a regulatory element in those introns that affects FOXP3, which is possible but not indicated here.

Alternatively, maybe the correct variant is in a gene that's not FOXP3 but is involved in a similar pathway. For example, maybe a gene involved in T cell function. But I don't know of any other genes causing IPEX.

Wait, perhaps the user is using a different nomenclature. Let me check the gene names again. For example, maybe "ALG13" is part of a complex that interacts with FOXP3? Not that I'm aware of.

Alternatively, maybe the correct variant is in a gene that's not listed here. But the user says exactly one is correct. This is confusing. Given that, perhaps the intended answer is var_005 (ALG13) because it's a splice region variant, which is more likely to be pathogenic. But I'm not sure. Alternatively, maybe the user intended to include FOXP3 in one of the variants but made a typo. For example, maybe var_005's nearest gene is FOXP3 but was written as ALG13. If that's the case, then var_005 would be the answer. But given the information provided, I have to work with what's there.

Given that, perhaps the intended answer is var_005, even though it's ALG13. Maybe the user made a mistake in the gene name. Alternatively, maybe the correct variant is in a gene that's not FOXP3. But I don't know of any other gene causing IPEX. Therefore, perhaps the correct answer is not listed here, but the user says exactly one is. Given that, I'll have to make an educated guess based on the information provided.

Since IPEX is X-linked and caused by FOXP3 mutations, and none of the variants here are in FOXP3, perhaps the correct answer is not present. However, the user insists that one is correct. Maybe the correct variant is in a gene that's not listed here. Alternatively, maybe the user intended to include FOXP3 in one of the variants but made a typo. For example, maybe var_005's nearest gene is FOXP3 but was written as ALG13. If that's the case, then var_005 would be the answer. But given the information provided, I have to work with what's there.

Alternatively, maybe the correct variant is in a gene that's not FOXP3 but is involved in a similar pathway. For example, maybe a gene involved in T cell development. But I don't know of any. Given that, perhaps the intended answer is var_005, as a splice region variant, which is more likely to be pathogenic. So I'll go with that, even though it's not FOXP3. Alternatively, maybe the correct answer is var_008 or var_009, but those are in SLITRK genes, which are not related.

Alternatively, maybe the correct variant is in a gene that's not listed here. But the user says exactly one is correct. Given that, I'll have to choose the most plausible one based on the information given. Since IPEX is X-linked and caused by FOXP3, and none of the variants are in FOXP3, perhaps the correct answer is not listed. But the user says exactly one is. Therefore, perhaps the intended answer is var_005, even though it's ALG13. Maybe the user made a mistake in the gene name. So I'll proceed with that.

Ranked Candidates: 1. `var_005`, 2. `var_008`, 3. `var_009`, 4. `var_002`, 5. `var_003`, 6. `var_010`, 7. `var_001`, 8. `var_007`, 9. `var_006`, 10. `var_004`

Trace 3 Trait: BETA-THALASSEMIA Match group: `3_prime_UTR_variant_12`

Result: `var_001` (11_5225566_T_G) ✓ Rank 1. Two candidates are annotated in the β-globin cluster: `var_001` at *HBB* and `var_009` near *HBD*.

		BioBLIP annotation		
ID	**Variant**	**Chr.**	**Consequence**	**Nearest genes**
`var_001` ⋆	11_5225566_T_G	chr11	missense variant	*HBB*
`var_002`	11_5737924_A_C	chr11	intron variant	*TRIM8*, *TRIM46*, *TRIM5*
`var_003`	11_67631410_A_G	chr17	intron variant	*TBX10*
`var_004`	11_31651675_T_C	chr11	intron variant	*ELP4*
`var_005`	11_61510226_A_G	chr1	intron variant	*LRRC8B*
`var_006`	11_114398544_C_T	chr1	intron variant	*C11orf49*
`var_007`	11_82732375_C_T	chr10	non coding transcript exon variant	*FAM48A*
`var_008`	11_68032459_T_C	chr2	missense variant	*COX6D*
`var_009`	11_5248354_T_A	chr11	non coding transcript exon variant	*HBD*, *AC104995.1*, *AL031984.1*, *AL031985.1*
`var_010`	11_64239012_C_T	chr11	upstream gene variant	*EGFL8*, *C11orf49*, *FAM120A*

BioBLIP (*HBB* ✓)

Okay, let's tackle this problem. The goal is to determine which of the 10 candidate variants is the true causal variant for Beta-Thalassemia. Let's start by recalling what Beta-Thalassemia is. It's a genetic disorder related to the production of hemoglobin, specifically the beta-globin chain. The primary gene involved here is the HBB gene, which is on chromosome 11. So, any variant in or near HBB is a strong candidate.

Looking at the variants provided, let's go through them one by one.

Var_001 is on chr11, a coding variant (missense) in the HBB gene. That's a direct hit. Missense mutations in HBB are classic causes of Beta-Thalassemia. This is a very strong candidate.

Var_002 is also on chr11 but is a noncoding intron variant near TRIM8, TRIM46, TRIM5. These genes aren't related to hemoglobin, so probably not causal.

Var_003 is on chr17 near TBX10. Not relevant to HBB.

Var_004 is on chr11 near ELP4. ELP4 isn't related to beta-globin. Unlikely.

Var_005 on chr1 near LRRC8B. Not related.

Var_006 on chr1 near C11orf49. Again, not HBB-related.

Var_007 on chr10 near FAM48A. No connection here.

Var_008 on chr2 in COX6D. Coding missense, but COX6D is part of the cytochrome c oxidase complex, not hemoglobin. Not relevant.

Var_009 on chr11, noncoding, near HBD and others. HBD is the delta-globin gene, which is part of the beta-globin cluster. Variants in this region could affect HBB expression. However, the consequence is a non-coding transcript exon variant. This might be a regulatory variant affecting HBB. Possible, but not as strong as a direct HBB mutation.

Var_010 on chr11, upstream variant near EGFL8, C11orf49, FAM120A. Upstream variants can affect gene regulation. However, these genes aren't related to HBB. So probably not causal.

So, the top candidate is var_001 because it's a missense in HBB. Var_009 is next because it's near HBD, which is part of the beta-globin cluster. The rest are either on different chromosomes, non-HBB related genes, or non-coding variants in non-relevant regions. Even though var_009 is non-coding, its proximity to HBD (which is part of the same gene family) makes it a possible regulatory variant. The others are less likely.

Ranked Candidates: 1. `var_001`, 2. `var_009`, 3. `var_002`, 4. `var_004`, 5. `var_010`, 6. `var_006`, 7. `var_005`, 8. `var_007`, 9. `var_003`, 10. `var_008`