VEC2DNA: FEATURE-CONSISTENT GENOMIC EMBEDDING INVERSION ATTACKS AND TASK-ALIGNED DEFENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic foundation models are increasingly produced through mean-pooled embeddings, yet it remains comparatively underexplored when it comes to what such vectors leak about the underlying DNA from a privacy perspective. We introduce Vec2DNA, a black-box embedding inversion framework that constructs single-nucleotide-perturbed anchor–variant pairs and exploits their known correspondence through a feature-level consistency objective. Rather than treating these variants as conventional data augmentation, Vec2DNA uses their pairwise structure to preserve internal representations at unchanged target positions while localizing the effect of sequence perturbations. Across four genomic embedding models on hg38 20-mer dataset, Vec2DNA consistently outperforms alternative inversion methods in reconstruction quality. In particular, it achieves up to a 14.5% relative improvement in Levenshtein similarity over standard inversion baselines. Our analysis shows that the structured correspondence between anchor–variant pairs provides the key additional signal through feature-level consistency. Beyond sequence reconstruction, we further show that genomic embeddings retain sufficient fine-grained information to reliably distinguish single-nucleotide variants and recover allele identity well above chance. Furthermore, we propose Task-Aligned Genomic Quantization (TAG-Q), a utility-aware defense that selectively preserves task-relevant embedding directions while suppressing information useful for inversion. TAG-Q substantially reduces inversion performance while retaining near-clean downstream utility on DNABERT-2, and achieves a similarly favorable privacy–utility trade-off on GENA-LM. Together, our results highlight the privacy risks of genomic embeddings and demonstrate complementary attack and defense mechanisms for controlling fine-grained genomic information leakage.

1 INTRODUCTION

Modern machine-learning systems increasingly expose embeddings rather than raw data, with representations cached, shared across teams, shipped to vector databases, and served through Embedding-as-a-Service (EaaS) under the implicit assumption that a dense vector is an anonymous summary of its input. A growing line of work refutes this assumption. Prior work has challenged this assumption by showing that embeddings can retain substantial information about their underlying inputs. However, compared with image and text domains, the privacy implications of genomic embeddings remain far less explored. This gap is increasingly important as genomic foundation models (Ji et al., 2021) are widely used to encode DNA sequence into representations for a broad range of downstream tasks. Genomic data is particularly sensitive because an individual's genome is immutable, partially shared with biological relatives, and can support re-identification, as demonstrated by surname inference Gymrek et al. (2013) and broader studies of genetic privacy Erlich & Narayanan (2014). Moreover, single-nucleotide polymorphisms (SNPs) encode individual-level genetic variation The 1000 Genomes Project Consortium (2015). If such information can be recovered from released genomic embeddings, the privacy risk extends beyond approximate sequence reconstruction to fine-grained information about individual genetic variation.

Recent work has begun to investigate embedding inversion for genomic foundation models. Ouaari et al. (2026) showed that DNA sequences can be reconstructed from both token-level and mean-

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.

pooled representations, although reconstruction difficulty varies substantially across models and representation types. Mean pooling compresses these states into a single fixed-dimensional vector and removes direct access to their token-wise positional correspondence. Nevertheless, models such as Evo-2 Brixi et al. (2026) and Nucleotide Transformer v2 (NTv2) Dalla-Torre et al. (2025) remain highly recoverable from mean-pooled representations, particularly for short sequences. DNABERT-2 (Zhou et al., 2024) was considerably more resistant to inversion. Unlike nucleotide-level or largely fixed-length tokenization, DNABERT-2 employs byte-pair encoding (BPE) Sennrich et al. (2016), which produces content-dependent and variable-length token boundaries. An inversion model must therefore recover both token identity and its correspondence to nucleotide positions, introducing additional alignment ambiguity.

This observation raises a broader question about BPE-based genomic embeddings. The difficulty observed for DNABERT-2 does not necessarily imply that every BPE-based genomic model is equally resistant to inversion. Reconstruction difficulty may also depend on the representation geometry learned by the target embedding model and on how sequence information is retained after pooling. We therefore study fixed-dimensional, mean-pooled embeddings across multiple BPE-based genomic foundation models and examine whether the inversion difficulty previously observed for DNABERT-2 generalizes beyond a single model. We further investigate what fine-grained genomic information remains recoverable from these representations and whether such leakage can be reduced while preserving downstream utility.

These observations motivate the following research questions:

- **RQ1** How effectively can mean-pooled genomic embeddings be inverted, and can improve reconstruction across models with different levels of inversion difficulty?
- **RQ2** What privacy-sensitive genomic information is exposed beyond sequence-level reconstruction, particularly regarding allele identity and genomic locus information?
- **RQ3** Can genomic embedding inversion be mitigated while preserving downstream utility?

To address RQ1, we introduce VEC2DNA, a black-box embedding inversion framework that exploits local correspondence between nearby genomic sequences. Standard inversion training treats queried sequences as independent embedding and target pairs and therefore provides little explicit information about how a small sequence change is reflected in the target embedding space. Vec2DNA constructs paired sequences through controlled nucleotide substitutions and queries the frozen target embedding model for their representations. These anchor and variant pairs provide known local correspondence that can be exploited during inversion training.

The central component of VEC2DNA is a feature-level consistency objective. Nucleotide substitutions can alter BPE segmentation, so direct positional correspondence between an anchor and its variant is not always available. Vec2DNA therefore identifies target-token slots that remain unchanged between the paired sequences and encourages their internal inversion features to remain consistent. This supervision makes the relation between paired queries explicit and complements the standard reconstruction objective. Controlled ablations show that increasing the number of target-model queries alone already improves reconstruction, while feature-level consistency provides an additional gain beyond a matched-query control. This indicates that the effectiveness of Vec2DNA arises primarily from exploiting correspondence between paired sequences rather than from SNP augmentation alone.

Across four BPE-based genomic embedding models (DNABERT-2 (Zhou et al., 2024), DNABERT-S (Zhou et al., 2025), GENA-LM (Fishman et al., 2025), and GROVER (Sanabria et al., 2024)) on hg38 with $\ell = 20$, VEC2DNA improves Levenshtein similarity over learned inversion baselines, with a maximum relative gain of 14.5% on GENA-LM versus Encoder-Only (Table 1). Baseline recoverability still differs substantially across models, indicating that BPE tokenization alone does not determine inversion difficulty. Beyond sequence-level reconstruction, we show that the recovered information can distinguish single-nucleotide variants and identify allele states well above chance. We further analyze the extent to which accurate short-sequence reconstructions can narrow their genomic origin.

Despite recent work on genomic embedding inversion, post-hoc defenses for released mean-pooled genomic embeddings remain underexplored. A central challenge is reducing sequence recoverability while retaining utility for specified downstream tasks. Finally, we propose Task-Aligned Genomic Quantization (TAG-Q), a post-hoc defense that uses fitted linear probes to form a task-aligned

basis and allocates quantization bits according to task relevance and embedding variance. Across the evaluated encoders, TAG-Q reduces reconstruction by a frozen VEC2DNA inverter trained on undefended embeddings while retaining accuracy close to clean levels on the linear-probe tasks used for calibration.

Our contributions can be summarized as follows:

- We systematically characterize privacy leakage from mean-pooled genomic embeddings, showing that they expose not only sequence information but also fine-grained allele and locus information.
- We introduce VEC2DNA, a black-box inversion attack that exploits correspondence between anchor and variant sequences through feature-level consistency, yielding gains beyond simply increasing target-model queries.
- We propose Task-Aligned Genomic Quantization (TAG-Q), a task-aware post-hoc defense that reduces reconstruction by the frozen VEC2DNA inverter while retaining accuracy close to clean levels on the linear-probe tasks used for calibration.

2 RELATED WORK

2.1 EMBEDDING INVERSION ATTACK ON DIFFERENT MODALITIES

Embedding inversion attacks have been studied primarily in vision and text, and have recently begun to extend to biological sequences. Attack methods have evolved from directly optimizing inputs to match target representations toward learned decoders and iterative refinement at inference time.

In computer vision, Mahendran & Vedaldi (2015) inverted CNN representations by optimizing an input image to match layer activations, while Dosovitskiy & Brox (2016) replaced this per-image optimization with a learned feed-forward decoder. Zhmoginov & Sandler (2016) extended inversion to low-dimensional face embeddings, and Mai et al. (2019) showed that the reconstructed faces can compromise face verification and identification systems, connecting embedding inversion to biometric template security.

In the text domain, Song & Raghunathan (2020) recovered the set of words in an input sentence from sentence embeddings, and Morris et al. (2023) formulated inversion as controlled generation, iteratively refining candidates by re-embedding the generated text and comparing it with the target embedding. Chen et al. (2025) aligned the target and attacker-controlled embedding spaces using a few paired samples, enabling black-box inversion with limited supervision. In these methods, however, query access is used only to collect embedding–text pairs for training or to provide feedback at inference time; the relation between distinct queries is not itself a training signal.

Embedding inversion of genomic representations has been studied much less extensively. Ouaari et al. (2026) showed that DNA sequences can be reconstructed from genomic foundation model embeddings using learned decoders, with quality degrading for mean-pooled representations as sequence length increases and DNABERT-2 emerging as the most resistant target. Effective inversion of mean-pooled genomic representations therefore remains an open problem.

2.2 GENOMIC FOUNDATION MODEL

Genomic sequence models differ in both tokenization and the granularity of the representations they expose. Early work such as dna2vec learns context-independent embeddings of variable-length k-mers (Ng, 2017), whereas DNABERT produces contextual representations of overlapping k-mers through masked language modeling (Ji et al., 2021). DNABERT-2 replaces fixed k-mers with byte-pair encoding (BPE) and multi-species pretraining, the Nucleotide Transformer family predominantly uses non-overlapping 6-mers (Zhou et al., 2024; Dalla-Torre et al., 2025), and GROVER and GENA-LM explore learned DNA tokenization (Sanabria et al., 2024; Fishman et al., 2025). Longer-context models instead vary the architecture: HyenaDNA operates at single-nucleotide resolution with implicit convolutions, Caduceus introduces bidirectional state-space models, and Evo and Evo 2 use autoregressive hybrids (Nguyen et al., 2023; Schiff et al., 2024; Nguyen et al., 2024; Brixi et al., 2026), while GenSLMs is codon-tokenized for viral genome modeling (Zvyagin et al., 2023).

For downstream use, token-level hidden states can be retained or aggregated into a fixed-dimensional sequence representation. Benchmarking five frozen DNA foundation models, Feng et al. (2025) found mean pooling generally more effective than summary-token or max pooling for sequence classification. Embedding quality is also task-dependent: DNABERT-S uses curriculum contrastive learning and manifold instance mixup to improve species-aware representations (Zhou et al., 2025), and Krupička et al. (2026) classified 10-kb windows for gene-fusion breakpoints with supervised models trained on frozen embeddings. These studies establish task-specific utility, but neither supervised prediction nor unsupervised clustering measures how much of an input sequence can be recovered from a released vector.

Recent work with DNABERT-2, NTv2, and Evo 2 report high reconstruction fidelity from per-token states and model- and length-dependent recovery from mean-pooled representations (Ouaari et al., 2026). Mean pooling removes the explicit ordering of token slots, although the contextual states being pooled can still encode sequence order. We therefore study sequence reconstruction from fixed-dimensional, mean-pooled embeddings across multiple BPE-based genomic foundation models, examining whether the inversion difficulty previously observed for DNABERT-2 generalizes beyond a single model.

3 Proposed Methodology

3.1 Threat Model and Problem Formulation

Let $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and let $x \in \Sigma^{\ell}$ be a genomic sequence, where ℓ is fixed within each experimental setting and known to the adversary. For a frozen BPE-based target model f, the native tokenizer τ partitions x into $m(x)$ sequence-bearing tokens. Each token spans at least one nucleotide, so $m(x) \leq \ell$, excluding added special tokens. The model produces contextual states $\boldsymbol{h}_t(x) \in \mathbb{R}^{d_e}$ and releases only the mean-pooled embedding

$$\boldsymbol{e}(x) = \frac{1}{|\mathcal{I}_{\mathrm{pool}}(x)|} \sum_{t \in \mathcal{I}_{\mathrm{pool}}(x)} \boldsymbol{h}_t(x) \in \mathbb{R}^{d_e}, \tag{1}$$

where $\mathcal{I}_{\mathrm{pool}}(x)$ is the nonempty set of encoder positions selected by the service's pooling rule.

The adversary is an authorized downstream recipient of $\boldsymbol{e}(x)$, with the underlying sequence x withheld. It is *honest-but-curious* (semi-honest): it follows the prescribed interfaces but performs additional local computations on its observations to infer private information (Goldreich, 2004; Lindell & Pinkas, 2009). Access to f is *black-box*: the attack uses neither model parameters nor contextual states, intermediate activations, logits, or gradients. The native tokenizer, vocabulary, and pooling rule are known; the private target's token sequence is not observed.

During offline preparation, the same adversary queries the service on public genomic sequences and synthetic variants, obtaining only their mean-pooled embeddings. Training targets for these known sequences are computed locally using τ. The total number of preparation evaluations, Q_{off}, includes both anchor and variant embeddings and must satisfy $Q_{\mathrm{off}} \leq B_{\mathrm{off}}$; requests also respect the service's rate limits. Once trained, the inversion model reconstructs released target embeddings locally without additional queries to f. Service-side query limits do not constrain this subsequent local computation, motivating protection of the released representations.

Let $\mathcal{V}$ be the native vocabulary, including PAD. The training target $\boldsymbol{y}(x) \in \mathcal{V}^L$ contains the sequence-bearing token IDs, right-padded to $L = \ell$, which accommodates $m(x) \leq \ell$. The trainable *inversion model* $g_\phi : \mathbb{R}^{d_e} \to (\Delta(\mathcal{V}))^L$ outputs one token distribution per slot, where $\Delta(\mathcal{V})$ denotes the probability simplex over $\mathcal{V}$. Slot-wise argmax followed by a fixed detokenization procedure, including special-token handling, yields $\hat{x}$. This decoding procedure does not enforce a nucleotide-length constraint, so $|\hat{x}|$ need not equal ℓ. The attack aims to recover x, evaluated using nucleotide-sequence reconstruction metrics.

Mean pooling is permutation-invariant over a fixed multiset of contextual states, but the complete sequence-to-embedding pipeline need not be: reordering the input can change both tokenization and the contextual states. The released vector therefore lacks an explicit token-slot axis while potentially retaining sequence-order information. Nucleotide substitutions can also alter BPE boundaries, token

counts, and subsequent slot indices. These properties motivate the paired substitutions and same-index ground-truth token-ID matching used by VEC2DNA for feature-level consistency.

3.2 VEC2DNA

Inversion Backbone. Following (Ouaari et al., 2026), we adopt a compact non-autoregressive encoder-only Transformer (Vaswani et al., 2017) with the same capacity as the strongest baseline, ensuring that performance gains arise from the proposed training scheme rather than increased model capacity. Given a mean-pooled embedding, we first standardize it using training-set statistics and project it into a length-L sequence of latent slots. After adding sinusoidal positional encodings, the slots are processed by N Transformer encoder layers and mapped to the target model's native vocabulary. We denote the resulting pre-head slot representations by $\boldsymbol{U} = [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_L]$, which are used by our consistency objective. Inference is performed in a single forward pass by taking the most likely token at each slot and detokenizing the resulting sequence.

Feature-Level Consistency on Anchor–Variant Pairs. For each training sequence x, we generate variants $x^{(k)} = \mathcal{A}_k(x)$ by substituting nucleotides at k distinct positions, yielding Hamming distance $d_H(x, x^{(k)}) = k$ for $k \in \mathcal{K} = \{1, 2, 3\}$. We obtain an embedding for each sequence from the frozen target model and its ground-truth token sequence from the model's native tokenizer. The anchor and each variant are processed separately by the same trainable inverter, providing paired features for consistency.

Because substitutions can change BPE segmentation, we select output slots by equality of the ground-truth target token IDs at the same slot index: $S(x, k) = \{j \in \{1, \ldots, L\} : y_j = y_j^{(k)}\}$. Let $\boldsymbol{U}_j$ and $\boldsymbol{U}_j^{(k)}$ denote the pre-head features of the anchor and its variant at slot j, respectively. We enforce feature-level consistency at the selected slots as:

$$\mathcal{L}_{\text{cons}} = \mathbb{E}_{(x,k)} \left[\frac{1}{d\,|S(x,k)|} \sum_{j \in S(x,k)} \left\| \boldsymbol{U}_j - \boldsymbol{U}_j^{(k)} \right\|_2^2 \right]. \tag{2}$$

This constraint encourages local sequence perturbations to affect only the representations associated with changed targets, while preserving features at unchanged positions. Unlike output-level matching, feature consistency provides supervision complementary to token-level reconstruction. Vec2DNA jointly optimizes the standard reconstruction loss and $\mathcal{L}_{\text{cons}}$; full optimization and training details are provided in the Appendix A.4.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We conduct all experiments on the human reference genome hg38 (GRCh38) Schneider et al. (2017), from which fixed-length DNA subsequences are extracted to construct the inversion dataset. Unless otherwise stated, we use sequences of length $\ell = 20$ as the primary evaluation setting, while additional experiments with longer sequences are provided to examine the effect of sequence length. The extracted sequences are divided into disjoint training, validation, and test splits, and only the training split is available for constructing the inversion model. Further details on sequence extraction and dataset splits are provided in Appendix A.9.

Embedding Models. We evaluate four representative genomic embedding models: DNABERT-2 Zhou et al. (2024), DNABERT-S Zhou et al. (2025), GENA-LM Fishman et al. (2025), and GROVER Sanabria et al. (2024). DNABERT-2 is a multi-species genomic foundation model that employs BPE tokenization, while DNABERT-S builds upon DNABERT-2 to learn species-aware DNA representations. GENA-LM is a family of Transformer-based genomic language models designed to process long DNA sequences, and GROVER is a DNA language model pretrained on the human genome with a vocabulary learned through BPE. For each model, we use its native tokenizer and obtain a single sequence-level representation by mean-pooling the token-level hidden states.

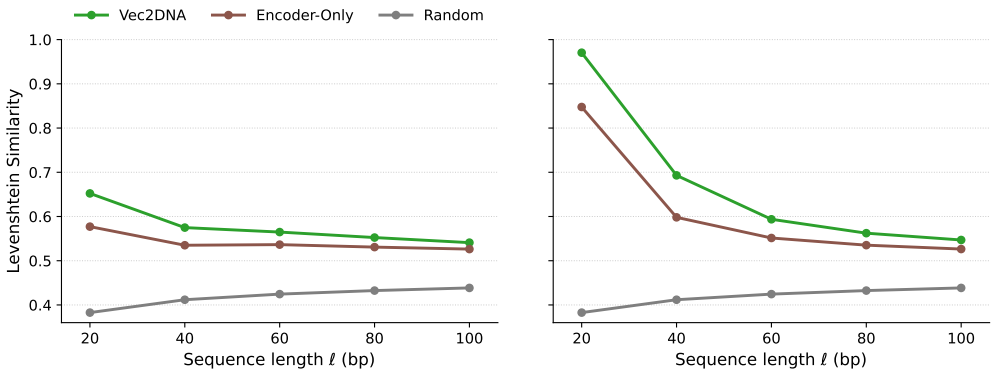

Figure 1: Levenshtein similarity as a function of sequence length ℓ on DNABERT-2 (Left) and GENA-LM (Right).

Inversion Baselines. Following prior genomic embedding inversion work Ouaari et al. (2026), we compare Vec2DNA with five standard inversion baselines: *Random*, *Encoder-Only*, *Decoder-Only*, *ResNet*, and *NNL* (Nearest Neighbour Lookup). *Random* generates nucleotide sequences according to the empirical nucleotide distribution and serves as a chance-level reference. *Encoder-Only* and *Decoder-Only* employ Transformer-based architectures Vaswani et al. (2017) to reconstruct the original sequence from its embedding. *ResNet* adopts a one-dimensional residual convolutional architecture based on residual learning He et al. (2016), whereas *NNL* follows the nearest-neighbor principle Cover & Hart (1967) and returns the training sequence associated with the closest embedding. All baselines are evaluated using the same data splits and embedding extraction protocol as Vec2DNA.

Evaluation Metrics. Following prior genomic embedding inversion work (Ouaari et al., 2026), we report Levenshtein Similarity (LS) and Nucleotide Accuracy (NA). We additionally report Exact Reconstruction Rate (Exact) to quantify the fraction of sequences recovered without any nucleotide-level error. Levenshtein Similarity is derived from the classical Levenshtein edit distance (Levenshtein, 1965) and measures sequence-level reconstruction similarity as $\mathrm{LS}(\hat{x}, x) = 1 - \mathrm{Lev}(\hat{x}, x)/\max(|\hat{x}|, |x|)$, where higher values indicate more similar sequences. Nucleotide Accuracy measures the proportion of nucleotide positions correctly reconstructed with respect to the original sequence. Exact Reconstruction Rate is the fraction of test sequences for which the reconstructed nucleotide sequence exactly matches the ground truth, $\mathrm{Exact} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{x}_i = x_i]$. Higher values indicate better inversion performance for all metrics.

	DNABERT-2			DNABERT-S			GENA-LM			GROVER		
	LS	NA	Exact	LS	NA	Exact	LS	NA	Exact	LS	NA	Exact
Random	0.3827	0.2592	0.00	0.3827	0.2592	0.00	0.3827	0.2592	0.00	0.3826	0.2592	0.00
Encoder-Only	0.5771	0.3573	0.00	0.5716	0.3462	0.00	0.8476	0.6952	26.04	0.8300	0.6756	15.31
Decoder-Only	0.5756	0.3574	0.01	0.5707	0.3426	0.00	0.8618	0.7250	28.41	0.8314	0.6661	14.26
ResNet	0.5646	0.3474	0.00	0.5565	0.3370	0.01	0.8113	0.6434	14.57	0.8435	0.6850	16.97
NNL	0.4662	0.3347	0.00	0.4613	0.3291	0.00	0.5302	0.3848	0.00	0.5380	0.3801	0.00
Vec2DNA *(Ours)*	**0.6522**	**0.4227**	**0.22**	**0.6429**	**0.4060**	**0.15**	**0.9704**	**0.9365**	**77.11**	**0.9424**	**0.8871**	**56.93**

Table 1: Embedding inversion performance on hg38 dataset with different embedding models, where $\ell = 20$. **LS**, **NA** and **Exact** denote Levenshtein Similarity, Nucleotide Accuracy and Exact Reconstruction Rate (%). The best-performing method is highlighted in **bold**.

4.2 **RQ1:** Inversion Performance

Short-sequence inversion. Table 1 presents the embedding inversion results at $\ell = 20$ across the four genomic embedding models. Vec2DNA consistently achieves the best reconstruction performance across Levenshtein Similarity (LS), Nucleotide Accuracy (NA), and Exact Reconstruction Rate (Exact), outperforming all standard inversion baselines. Compared with the Encoder-Only baseline, which shares the same underlying inversion backbone, Vec2DNA improves LS from 0.5771 to 0.6522

on DNABERT-2, from 0.5716 to 0.6429 on DNABERT-S, from 0.8476 to 0.9704 on GENA-LM, and from 0.8300 to 0.9424 on GROVER. Exact recovery exhibits an even sharper model-dependent contrast: Vec2DNA exactly reconstructs 77.11% of GENA-LM and 56.93% of GROVER test sequences, compared with 26.04% and 15.31% for Encoder-Only, respectively, while exact recovery remains rare on DNABERT-2 (0.22%) and DNABERT-S (0.15%). This result shows that highly recoverable embedding spaces can expose complete sequences at substantial rates, whereas harder embedding spaces still retain partial sequence information captured by LS and NA.

A notable empirical finding is that baseline recoverability differs substantially even among BPE-based models. In particular, Encoder-Only already reaches an LS of 0.8476 on GENA-LM and 0.8300 on GROVER, compared with only 0.5771 on DNABERT-2. This contrast indicates that BPE tokenization alone is insufficient to explain inversion resistance. Instead, the amount and structure of sequence information retained in the mean-pooled embedding appear to be strongly model-dependent. While isolating the precise architectural or representation-level cause is beyond the scope of this study, the result shows that the difficulty previously observed for DNABERT-2 does not generalize uniformly across BPE-based genomic foundation models. Nevertheless, Vec2DNA provides consistent gains across both relatively difficult and highly recoverable embedding spaces, suggesting that the proposed training strategy is not tied to a particular level of baseline invertibility.

Effect of sequence length. Figure 1 further evaluates Vec2DNA on DNABERT-2 and GENA-LM as the sequence length increases from $\ell = 20$ to $\ell = 100$. Reconstruction performance decreases with increasing sequence length for both Vec2DNA and the Encoder-Only baseline, indicating that recovering longer sequences from a fixed-dimensional mean-pooled representation becomes progressively more challenging. Nevertheless, Vec2DNA consistently maintains higher Levenshtein similarity than the Encoder-Only baseline across all evaluated lengths and on both embedding models. The performance gap is largest in the short-sequence regime and gradually narrows as the sequence length increases, showing that the benefit of the proposed training strategy remains consistent while the overall inversion problem becomes increasingly difficult.

5 Analysis

5.1 Ablation Study

We analyze whether the improvement of Vec2DNA beyond increased query coverage is attributable to feature-level consistency on anchor–variant pairs. All comparisons use matched training and query budgets, while full ablations and negative controls are provided in Appendix A.5. As shown in Table 2, using unrelated hg38 sequences with the same query budget (A2) already yields a substantial improvement, confirming that broader query coverage itself is an important contributor to inversion performance. Replacing these independent queries with SNP variants alone (A5) does not improve over A2. In contrast, applying feature-level consistency to the same SNP pairs (A7) increases Levenshtein similarity by $+0.0105$ and $+0.0123$ over the matched-query control A2 on DNABERT-2 and GENA-LM, respectively. Thus, additional queries explain a large portion of the improvement over the base inverter, while the gain beyond query scaling arises from exploiting anchor–variant correspondence through feature-level consistency rather than from SNP augmentation alone.

ID	Training Signal	Cons.	DNABERT-2		GENA-LM	
			LS	NA	LS	NA
A2	Unrelated hg38	✗	.6400	.4179	.9581	.9119
A5	SNP pairs	✗	.6354	.4143	.9533	.9035
A7	Vec2DNA	✓	**.6505**	**.4196**	**.9704**	**.9373**

Table 2: Core ablation of Vec2DNA at $\ell = 20$. All variants use identical query and optimization budgets.

5.2 **RQ2:** Privacy Leakage Analysis

Allele-level disclosure. Sequence-level metrics such as Levenshtein similarity and nucleotide accuracy do not directly reveal whether an inversion model recovers the single-nucleotide differences that carry individual-level information (The 1000 Genomes Project Consortium, 2015). We therefore construct a Hamming-1 evaluation set by introducing one alternative nucleotide into each of the 15,000 DNABERT-2 test sequences, and evaluate pair discrimination, allele identification, and reference bias. The allele-level evaluations are candidate-conditioned: they test whether the embedding can

	Metric	Chance	Encoder-only	Vec2DNA
Reconstruction	Levenshtein similarity	0.383	0.5755	**0.6522**
	Nucleotide accuracy	0.259	0.3540	**0.4227**
Allele level	Pair discrimination AUC	0.500	0.9523	**0.9823**
	Two-alternative forced choice	0.500	0.9623	**0.9857**
	Four-way allele call, reference	0.250	0.7679	**0.8570**
	Four-way allele call, variant	0.250	0.7191	**0.8262**
	Four-way allele call, both	0.0625	0.5625	**0.7169**
Prior reliance	Reference bias ↓	0 (ideal)	0.1479	**0.0888**

Table 3: Allele-level disclosure on the Hamming-1 pair set (DNABERT-2, $\ell = 20$, $N = 15{,}000$ pairs). Reconstruction metrics evaluate embedding-only inversion, whereas allele-level metrics are candidate-conditioned and score known reference/variant or four-way allele hypotheses. Reference bias is the fraction of pairs on which the model prefers the reference allele although it is given the embedding of the variant, so lower is better.

distinguish allele hypotheses once the relevant reference context or variant position is known, whereas the reconstruction metrics evaluate the embedding-only inversion setting.

As shown in Table 3, Vec2DNA distinguishes reference–variant pairs with an AUC of 0.9823 and correctly identifies both alleles in 71.7% of the pairs, compared with 56.3% for the Encoder-Only baseline and a 6.25% chance rate. It also reduces reference bias from 0.1479 to 0.0888. These results show that the recovered information extends beyond approximate sequence similarity to fine-grained allele identity.

From sequence recovery to locus disclosure. Recovered genomic sequences can also disclose where an embedding originated in the reference genome, since sufficiently distinctive short sequences can be uniquely mapped to genomic locations (Karimzadeh et al., 2018). In our $\ell = 20$ test set, 81.55% of sequences occur uniquely in hg38, indicating that many sufficiently accurate reconstructions can be associated with a specific genomic locus. Combined with the allele-level results above, this implies that a released embedding can reveal both *where* a sequence originates and *which allele* is present at that locus. Detailed analysis of sequence-length identifiability, approximate locus matching, and representative reconstruction cases is provided in Appendix A.7.

6 Defending Against Inversion Attacks

Task-Aligned Genomic Quantization (TAG-Q). Simple post-hoc defenses such as additive Gaussian noise Morris et al. (2023) and uniform scalar quantization Seputis et al. (2025) face a privacy–utility trade-off: perturbations strong enough to suppress inversion may also remove information required by downstream tasks. To address this issue, we propose Task-Aligned Genomic Quantization (TAG-Q), a utility-aware defense that allocates representation capacity according to downstream relevance rather than treating all embedding dimensions equally. Given standardized embeddings, TAG-Q first fits linear probes for a small set of downstream tasks and uses their directions to construct a task-aligned orthonormal basis. The remaining subspace is represented by principal components of the residual embedding, separating directions directly used by the tasks from residual information that may facilitate inversion.

For each basis direction i, TAG-Q estimates its embedding variance λ_i and task relevance ρ_i, and distributes a global bit budget R through a relevance-weighted rate allocation. Directions that are both task-relevant and informative receive higher precision, whereas directions with low task relevance receive fewer bits or collapse to their mean. Thus, unlike uniform quantization, TAG-Q concentrates the released information on the subspace required for legitimate downstream use while suppressing fine-grained residual information exploitable by an inversion attacker. The rate R provides a single control over the resulting privacy–utility trade-off. Algorithm 1 summarizes the overall calibration and quantization procedure, while the complete formulation and derivation of the rate-allocation objective are provided in Appendix A.2.

Algorithm 1 Task-Aligned Genomic Quantization (TAG-Q)

Require: training embeddings $\{\boldsymbol{\phi}\}$, tasks $\{t\}$, target rate R, clipping factor c

1: estimate $\boldsymbol{\mu}, \boldsymbol{\sigma}$ and standardize embeddings
2: fit task probes $\{\boldsymbol{w}_t\}_{t=1}^{T}$
3: construct task-aligned basis $\boldsymbol{V} = [\boldsymbol{U}\ \widetilde{\boldsymbol{V}}]$
4: estimate variance λ_i and task relevance ρ_i
5: find θ such that $\sum_i b_i = R$
6: allocate $b_i \leftarrow \max\left(0, \frac{1}{2}\log_2 \frac{\rho_i \lambda_i}{\theta}\right)$
7: quantize $\boldsymbol{c} = \boldsymbol{V}^\top \boldsymbol{z}$ using the allocated rates
8: **return** $\boldsymbol{\phi}' = \boldsymbol{\sigma} \odot (\boldsymbol{V}\hat{\boldsymbol{c}}) + \boldsymbol{\mu}$

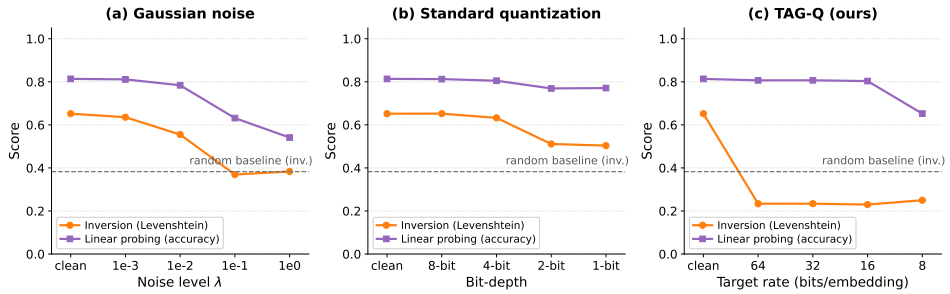

Figure 2: Privacy–utility trade-off on DNABERT-2 ($\ell = 20$). Each panel reports linear-probing accuracy and inversion Levenshtein similarity under increasing defense strength. TAG-Q (c) suppresses inversion below the random baseline while preserving accuracy close to the clean-embedding level, whereas Gaussian noise and Standard quantization do not achieve the same trade-off.

RQ3: Privacy–utility trade-off. Following the Vec2Text Morris et al. (2023) defense evaluation setup, we evaluate TAG-Q against Gaussian noise and Standard quantization using the pre-trained Vec2DNA attacker. For downstream utility, we perform linear probing with a Logistic Regression classifier for GC-content prediction. As shown in Figure 2, Gaussian noise reduces inversion only at the cost of substantial downstream degradation, whereas standard quantization preserves utility but leaves considerable sequence information recoverable. TAG-Q instead reduces inversion Levenshtein similarity on DNABERT-2 from approximately 0.65 to 0.23 while retaining linear-probing accuracy near the clean value of 0.81. A similar pattern is observed on GENA-LM, where TAG-Q substantially reduces inversion while preserving near-clean utility. Full results across all four genomic embedding models, together with defense sweeps and rate-sensitivity analysis, are provided in Appendix A.2.

7 Conclusion

We introduced Vec2DNA, a black-box attack on mean-pooled genomic embeddings. It encourages consistent features at each output position where the anchor and variant have matching ground-truth token IDs. Across four BPE-based encoders on hg38 reference 20-mers, it outperforms the evaluated baselines, with a 14.5% relative gain in Levenshtein similarity over Encoder-Only on GENA-LM. Ablations on two encoders show gains from additional queries. Under matched query and optimization budgets, substituted sequences used as independent examples do not outperform unrelated-query controls; adding consistency on the same pairs surpasses those controls. Baseline recoverability varies substantially despite shared BPE tokenization.

Candidate-conditioned analyses on DNABERT-2 show allele discrimination, while reference-mapping cases illustrate locus narrowing from accurate reconstructions of distinctive sequences. We further proposed TAG-Q, a task-aware post-hoc defense that reduces reconstruction by the frozen Vec2DNA inverter trained on undefended embeddings while retaining accuracy close to clean levels on calibrated downstream tasks. These results provide no formal privacy guarantee; robustness to adaptive attackers and utility on uncalibrated tasks remain open. Together, our results provide a model-specific empirical basis for assessing genomic leakage and its trade-off with downstream utility.

AI USE STATEMENT

AI assistants were used only for limited writing and language-editing support, as well as preparation of submission materials. All technical content, methodology, implementation, experimental results, and final verification were performed and checked by the authors.

ETHICS STATEMENT

This work studies privacy risks associated with released genomic embeddings and develops defenses intended to mitigate such risks. All experiments use the public hg38 reference assembly and synthetic sequence variants; no personal or identified genomic data are used. While embedding inversion techniques may have dual-use implications, our experiments are conducted solely in a controlled research setting to characterize potential information leakage and motivate stronger privacy protections. The methods described in this work should be used only on data and systems for which appropriate authorization has been obtained.

REFERENCES

Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, 2026.

Yiyi Chen, Qiongkai Xu, and Johannes Bjerva. Algen: Few-shot inversion attacks on textual embeddings via cross-model alignment and generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24330–24348, 2025.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

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

Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4829–4837, 2016.

Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2665–2679, 2023.

Yaniv Erlich and Arvind Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6):409–421, 2014. doi: 10.1038/nrg3723.

Haonan Feng, Lang Wu, Bingxin Zhao, Chad Huff, Jianjun Zhang, Jia Wu, Lifeng Lin, Peng Wei, and Chong Wu. Benchmarking DNA foundation models for genomic and genetic tasks. *Nature Communications*, 16:10780, 2025. doi: 10.1038/s41467-025-65823-8.

Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 2025.

Oded Goldreich. *Foundations of Cryptography: Volume 2, Basic Applications*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511721656.

Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013. doi: 10.1126/science.1229566.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021. doi: 10.1093/bioinformatics/btab083.

Shuaifan Jin, Xiaoyi Pang, Zhibo Wang, He Wang, Jiacheng Du, Jiahui Hu, and Kui Ren. Privacy-preserving LLM embedding transmission for end-cloud collaboration, 2026. URL https://arxiv.org/abs/2503.12896v2. Version 2; accepted manuscript for IEEE Transactions on Networking; first posted in 2025.

Mehran Karimzadeh, Carl Ernst, Anshul Kundaje, and Michael M. Hoffman. Umap and bismap: quantifying genome and methylome mappability. *Nucleic Acids Research*, 46(20):e120, 2018. doi: 10.1093/nar/gky677.

Radim Krupička, Mariana Komárková, Bohuslav Dvorský, Kateřina Kollinová, and Ondřej Klempíř. Benchmarking genomic foundation models for binary classification of gene fusion breakpoints from DNA sequences. *BioData Mining*, 19:41, 2026. doi: 10.1186/s13040-026-00553-1.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965. URL https://api.semanticscholar.org/CorpusID:60827152.

Yu Lin, Ruining Yang, Yunlong Mao, Qizhi Zhang, Jue Hong, Quanwei Cai, Ye Wu, Huiqi Liu, Zhiyu Chen, Bing Duan, and Sheng Zhong. ObfusLM: Privacy-preserving language model service against embedding inversion attacks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1160–1174, 2025. doi: 10.18653/v1/2025.acl-long.58.

Yehuda Lindell and Benny Pinkas. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality*, 1(1):59–98, 2009. doi: 10.29012/jpc.v1i1.566.

Tiantian Liu, Hongwei Yao, Feng Lin, Tong Wu, Zhan Qin, and Kui Ren. Eguard: Defending LLM embeddings against inversion attacks via text mutual information optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):35644–35652, 2026. doi: 10.1609/aaai.v40i42.40876.

Rui Lu, Yanting Liu, and Dan Wang. P4LLM: Protecting embedding privacy against model inversion attacks for EaaS LLM serving via token-wise partition and perturbation. *IEEE Transactions on Mobile Computing*, 2026. doi: 10.1109/TMC.2026.3667619.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5188–5196, 2015.

Guangcan Mai, Kai Cao, Pong C. Yuen, and Anil K. Jain. On the reconstruction of face images from deep face templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5):1188–1202, 2019. doi: 10.1109/TPAMI.2018.2827389.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.765/.

Patrick Ng. dna2vec: Consistent vector representations of variable-length k-mers, 2017. URL https://arxiv.org/abs/1701.06279.

Eric Nguyen, Michael Poli, Marjan Faizi, et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201, 2023. URL https://proceedings.nips.cc/paper_files/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract-Conference.html.

Eric Nguyen, Michael Poli, Matthew G. Durrant, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336.

Sofiane Ouaari, Jules Kreuer, and Nico Pfeifer. How private are DNA embeddings? inverting foundation model representations of genomic sequences. *arXiv preprint arXiv:2603.06950*, 2026.

Melissa Sanabria, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923, 2024.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43632–43648. PMLR, 2024. URL https://proceedings.mlr.press/v235/schiff24a.html.

Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, 27(5):849–864, 2017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1715–1725, 2016.

Dominykas Seputis, Yongkang Li, Karsten Langerak, and Serghei Mihailov. Rethinking the privacy of text embeddings: A reproducibility study of "text embeddings reveal (almost) as much as text". In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 822–831, 2025.

Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 377–390, 2020. doi: 10.1145/3372297.3417270.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68–74, 2015. doi: 10.1038/nature15393.

Yu-Che Tsai, Hsiang Hsiao, Kuan-Yu Chen, and Shou-De Lin. Concept-aware privacy mechanisms for defending embedding inversion attacks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/01025a4e79355bb37a10ba39605944b5-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yubo Wang, Shujie Cui, James Bailey, Hongzhi Yin, Wenyu Liang, Min Tang, Shiyue Qin, and Weiqing Wang. Denoising-aware inversion: Revealing privacy risks in noise-protected text embeddings, 2026. URL https://arxiv.org/abs/2608.18610. Preprint.

Andrey Zhmoginov and Mark Sandler. Inverting face embeddings with convolutional neural networks. *arXiv preprint arXiv:1606.04189*, 2016.

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuanjing Huang. TextObfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5459–5473, 2023. doi: 10.18653/v1/2023.findings-acl.337.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *International Conference on Learning Representations*, volume 2024, pp. 41642–41665, 2024.

Zhihan Zhou, Weimin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V Davuluri, Zhong Wang, and Han Liu. Dnabert-s: pioneering species differentiation with species-aware dna embeddings. *Bioinformatics*, 41(Supplement_1):i255–i264, 2025.

Shengyao Zhuang, Bevan Koopman, Xiaoran Chu, and Guido Zuccon. Understanding and mitigating the threat of vec2text to dense retrieval systems. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, pp. 259–268, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707247. doi: 10.1145/3673791.3698414. URL https://doi.org/10.1145/3673791.3698414.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, et al. GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications*, 37(6):683–705, 2023. doi: 10.1177/10943420231201154.

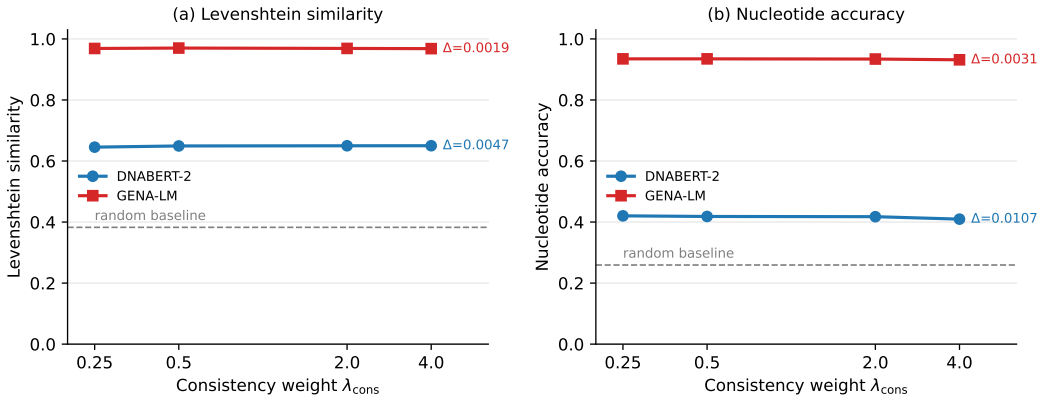

Figure 3: Robustness of the attack to the consistency weight λ_{cons} ($\ell = 20$, seed 42). Across a sixteenfold range of λ_{cons} (log-scaled axis), both (a) Levenshtein similarity and (b) nucleotide accuracy stay essentially flat and well above the random baseline for both victim models. The annotated Δ is the total spread of each curve: the test Levenshtein similarity moves by only 0.0047 for DNABERT-2 and 0.0019 for GENA-LM.

A Appendix

A.1 Hyperparameter Analysis

To assess how sensitive the attack is to this choice, we sweep $\lambda_{\text{cons}} \in \{0.25, 0.5, 2.0, 4.0\}$, a sixteenfold range, on two embedding models at sequence length $\ell = 20$ with a fixed seed. The two models bracket the difficulty spectrum seen in our main results: DNABERT-2, which is comparatively hard to invert (Levenshtein similarity near 0.65), and GENA-LM, which is easy to invert (similarity near 0.97). Figure 3 plots test Levenshtein similarity and test nucleotide accuracy against λ_{cons}, with the empirical random baseline shown for reference.

Robustness to λ_{cons}. The attack is highly insensitive to this hyperparameter. As λ_{cons} ranges over the full sixteenfold interval, the test Levenshtein similarity stays within a band of 0.0047 for DNABERT-2 (from 0.6454 to 0.6501) and 0.0019 for GENA-LM (from 0.9682 to 0.9701), that is, below half a percentage point in both regimes (Figure 3a). Test nucleotide accuracy is similarly stable, changing by 0.0107 and 0.0031 respectively (Figure 3b). Every curve remains far above the random baseline (0.383 Levenshtein similarity, 0.259 nucleotide accuracy) at all four settings. Variations of this size are within run to run noise, so reproducing our results does not require careful tuning of λ_{cons}.

Takeaway. Where a trend is discernible it is weak and consistent. For DNABERT-2, a larger λ_{cons} slightly raises Levenshtein similarity while slightly lowering nucleotide accuracy, so stronger consistency trades a small amount of per-position precision for marginally better alignment; both effects are minor. For GENA-LM the curves are essentially flat with a shallow optimum near $\lambda_{\text{cons}} = 0.5$. Since this near-horizontal behavior holds for both an easy and a hard victim model, it reflects a property of the method rather than of a particular embedding space. Consequently the consistency weight can be fixed to a single default for all victim models rather than tuned per model, and any value in $[0.5, 4.0]$ yields essentially the same reconstruction quality.

A.2 TAG-Q: Detailed Formulation and Additional Analysis

Detailed Formulation. Let $\boldsymbol{\phi} \in \mathbb{R}^d$ denote an embedding produced by a frozen genomic foundation model. Using training-set statistics $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, we first standardize the embedding as $\boldsymbol{z} = (\boldsymbol{\phi} - \boldsymbol{\mu})/\boldsymbol{\sigma}$. For each downstream task $t = 1, \ldots, T$, we fit a linear probe $\boldsymbol{w}_t$ on $\boldsymbol{z}$.

TAG-Q constructs a task-aligned basis by placing the probe directions first,

$$\boldsymbol{U} = \operatorname{orth}([\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T]), \qquad \boldsymbol{V} = [\boldsymbol{U}\ \widetilde{\boldsymbol{V}}], \tag{3}$$

	R^2 ↑	Pearson ↑	MAE ↓	Acc ↑	AUC ↑
DNABERT-2	0.6230	0.7894	0.0686	0.8135	0.8975
+ Gaussian Noise	0.1299	0.3610	0.1067	0.6314	0.6678
+ Quantization	0.6070	0.7792	0.0700	0.8049	0.8898
+ **TAG-Q (*Ours*)**	0.6134	0.7833	0.0695	0.8069	0.8918
DNABERT-S	0.6827	0.8263	0.0632	0.8371	0.9166
+ Gaussian Noise	0.3277	0.5725	0.0926	0.7012	0.7737
+ Quantization	0.6591	0.8119	0.0654	0.8229	0.9076
+ **TAG-Q (*Ours*)**	0.6744	0.8213	0.0641	0.8214	0.9079
GENA-LM	0.9673	0.9835	0.0195	0.9673	0.9955
+ Gaussian Noise	0.0818	0.2862	0.1096	0.6056	0.6366
+ Quantization	0.9640	0.9818	0.0207	0.9633	0.9947
+ **TAG-Q (*Ours*)**	0.9661	0.9829	0.0199	0.9644	0.9948
GROVER	0.9502	0.9748	0.0249	0.9499	0.9900
+ Gaussian Noise	0.6540	0.8087	0.0668	0.8164	0.9005
+ Quantization	0.9453	0.9723	0.0261	0.9471	0.9892
+ **TAG-Q (*Ours*)**	0.9488	0.9741	0.0253	0.9439	0.9890

Table 4: Linear probing performance comparison of different DNA embedding models under embedding perturbation and transformation methods. For Gaussian Noise, we apply the basic defense method from Vec2Text (Morris et al., 2023) with a noise level of 10^{-1}. The standard quantization defense applies 4-bit scalar quantization. ↑: higher is better; ↓: lower is better.

	DNABERT-2		DNABERT-S		GENA-LM		GROVER	
	LS ↓	NA ↓	LS ↓	NA ↓	LS ↓	NA ↓	LS ↓	NA ↓
VEC2DNA (w/o defense methods)	0.6522	0.4227	0.6429	0.4060	0.9704	0.9365	0.9424	0.8871
+ Gaussian Noise	0.3844	0.2028	0.4034	0.2377	0.4245	0.2231	0.4055	0.2464
+ Quantization	0.6325	0.4037	0.6142	0.3799	0.9650	0.9255	0.9316	0.8662
+ **TAG-Q (*Ours*)**	0.2338	0.1278	0.3481	0.2189	0.4910	0.3463	0.2586	0.1618

Table 5: Embedding inversion performance of pre-trained VEC2DNA with different defense methods on hg38 dataset, where $L = 20$. **LS** and **NA** denote Levenshtein Similarity and Nucleotide Accuracy, respectively. The noise level of Gaussian Noise method is 10^{-1}. The standard quantization defense applies 4-bit scalar quantization. ↓: lower inversion (less leakage) is better.

where $\widetilde{\boldsymbol{V}}$ contains principal components of the residual $\boldsymbol{z} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{z}$. This construction concentrates task-relevant information into a small number of directions.

Let $\boldsymbol{c} = \boldsymbol{V}^\top \boldsymbol{z}$, with $\lambda_i = \mathrm{Var}(c_i)$, and define the task relevance of each direction as $\rho_i = \sum_t (\boldsymbol{V}^\top \boldsymbol{w}_t)_i^2$. TAG-Q allocates a total rate R by minimizing the relevance-weighted distortion

$$D = \sum_{i=1}^{d} \rho_i \, \mathbb{E}\left[(\hat{c}_i - c_i)^2\right]. \tag{4}$$

Using the Gaussian rate–distortion approximation $\mathbb{E}[(\hat{c}_i - c_i)^2] \approx \lambda_i 2^{-2b_i}$, the resulting reverse-water-filling allocation is

$$b_i = \max\left(0, \, \frac{1}{2}\log_2 \frac{\rho_i \lambda_i}{\theta}\right), \tag{5}$$

where the water level θ is selected such that $\sum_i b_i = R$. Consequently, directions that are both high-variance and task-relevant receive more bits, whereas directions with low relevance receive little or no capacity.

Each retained coefficient is uniformly quantized after clipping,

$$\hat{c}_i = \Delta_i \left\lfloor \frac{\mathrm{clip}(c_i, \pm c\sqrt{\lambda_i})}{\Delta_i} \right\rceil, \qquad \Delta_i = \frac{2c\sqrt{\lambda_i}}{2^{b_i} - 1}. \tag{6}$$

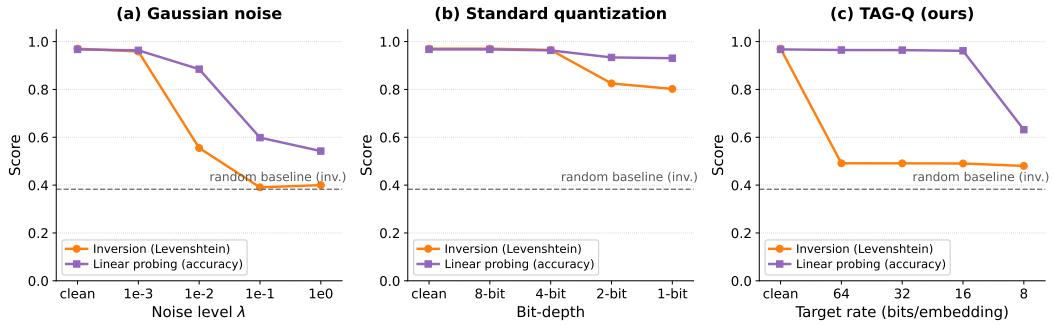

Figure 4: GENA-LM ($\ell = 20$), same protocol as Figure 2. The clean GENA-LM embedding is almost perfectly invertible (0.970), yet TAG-Q (c) still halves inversion at near-clean utility, a far better operating point than either baseline.

The defended embedding is then mapped back to the original coordinate system as $\phi' = \boldsymbol{\sigma} \odot (\boldsymbol{V}\hat{\boldsymbol{c}}) + \boldsymbol{\mu}$. All calibration statistics are estimated only from the training split.

Additional Analysis. We further examine the three defenses (Gaussian noise, standard quantization, and TAG-Q) by sweeping the hyperparameter of each and reporting, at every setting, both the utility retained (linear-probing accuracy) and the privacy achieved (inversion Levenshtein similarity of the frozen Vec2DNA attacker). All runs use GENA-LM and DNABERT-2 victim embeddings at $\ell = 20$; the empirical random baseline for inversion is 0.383. Figures 2 and 4 show the results.

TAG-Q resolves the privacy and utility trade-off. The two baselines occupy opposite failure modes. Gaussian noise couples the two objectives: the inversion and probing curves fall together, so the only way to push inversion toward the random baseline also destroys utility. On DNABERT-2 the attacker reaches the random baseline only at $\lambda = 0.1$, where probing accuracy has already collapsed from 0.81 to 0.63, and on GENA-LM reaching 0.39 inversion costs a drop from 0.97 to 0.60 accuracy. Standard quantization has the reverse problem: it preserves utility well (accuracy stays at 0.77 on DNABERT-2 and 0.93 on GENA-LM even at one bit) but is unable to lower inversion, which floors at 0.50 and 0.80 respectively, both far above the random baseline. TAG-Q breaks this trade-off. On DNABERT-2 it reduces inversion to 0.23, which is below the random baseline of 0.383, while keeping probing accuracy at 0.81, essentially the clean value; no setting of either baseline reaches this region. On GENA-LM, whose clean embedding is almost perfectly invertible (0.970), TAG-Q lowers inversion to 0.49 at 0.96 accuracy: it does not fall below the random baseline for this hardest model, but it still attains a far lower inversion at a far higher utility than either baseline can. In short, TAG-Q spends its bit budget only on the task-relevant directions, so it removes the information the attacker needs while retaining the information the downstream task uses.

Robustness of the hyperparameter choice across embedding models. TAG-Q also removes the need to tune its hyperparameter per model. Over the range $R \in \{16, 32, 64\}$, a fourfold change in the target rate, both curves are essentially flat: on DNABERT-2 accuracy stays within $[0.803, 0.807]$ and inversion within $[0.230, 0.234]$, and on GENA-LM accuracy stays within $[0.962, 0.965]$ and inversion within $[0.490, 0.491]$. Utility degrades only at the extreme rate $R = 8$, where the budget is too small to quantize even the few kept directions. Crucially, the *same* rate transfers across two very different embedding models: DNABERT-2 and GENA-LM differ sharply in scale and in clean invertibility (0.652 versus 0.970), yet a single setting such as $R = 32$ places both at their favorable operating point. This is a direct consequence of TAG-Q calibrating its statistics (μ, σ, the task-aligned basis, and the per-direction variances) from each model's own training embeddings, so one rate induces a comparable effect everywhere. The baselines lack this property: Gaussian noise applies a fixed absolute perturbation whose relative strength depends on the embedding scale, so the same λ acts very differently across models, and the quantization floor is likewise model dependent. TAG-Q therefore offers a single, transferable operating point rather than a per-model search.

A.3 Implementation Details

Inversion network. The inversion model is a compact non-autoregressive encoder. A linear projection expands the 768-dimensional pooled embedding into ℓ latent slot representations of width $d = 128$. Sinusoidal positional encodings are then added, and the resulting slots are jointly refined by a Transformer encoder with $N = 6$ layers, 8 attention heads, $d_{\text{ff}} = 1024$, GELU activations, and dropout of 0.1. A linear prediction head maps each slot to the native vocabulary of the target embedding model, allowing all positions to be predicted in a single forward pass without autoregressive decoding. The resulting parameter count depends primarily on the target vocabulary size and sequence length ℓ, ranging from 4.03M parameters for GROVER ($|\mathcal{V}| = 610$) to 8.08M for GENA-LM ($|\mathcal{V}| = 32{,}000$) at $\ell = 20$, and reaching 12.36M for DNABERT-2 at $\ell = 100$.

Embedding extraction and targets. All target embedding models are kept frozen throughout the experiments and are used only for embedding extraction; we do not fine-tune them or access their gradients. For each sequence, we take the final hidden layer, exclude special tokens, and mean-pool the remaining sequence-bearing token representations into a single 768-dimensional vector. The resulting embeddings are cached in HDF5 together with checksums so that all methods operate on exactly the same inputs. Embeddings are extracted with a batch size of 256. Reconstruction targets are defined using the target model's own token identifiers and are right-padded to ℓ slots, since the inversion model predicts directly over the native tokenizer vocabulary rather than over individual nucleotides. Input embeddings are standardized using statistics computed exclusively from the training split. All experiments use the publicly released checkpoints of the corresponding target models.

Optimization. Unless otherwise stated, we use $\mathcal{K} = \{1, 2, 3\}$ with one variant generated at each perturbation distance, consistency weight $\lambda = 1$, hidden width $d = 128$, $N = 6$ Transformer layers, 8 attention heads, $d_{\text{ff}} = 1024$, dropout of 0.1, and AdamW with a learning rate of 3×10^{-4} and weight decay of 10^{-2}. We use a reconstruction batch size of 512, a pair batch size of 256, and train for 50 epochs. The training objective combines token-level cross-entropy over the target model's vocabulary with the feature-level consistency loss weighted by λ. The reconstruction and consistency terms are computed from independently sampled batches within each optimization step. We select the checkpoint with the highest Levenshtein similarity on the validation split. At inference time, the predicted token at each slot is obtained by argmax, and the resulting token sequence is detokenized back into a nucleotide sequence using the target model's native tokenizer.

Reproducibility and compute. Unless otherwise stated, experiments use a fixed random seed of 42 for Python, NumPy, and PyTorch. The ablation experiments additionally report the mean and standard deviation over three random seeds. All experiments in this work, including embedding extraction, inversion-model training, and defense evaluation, are conducted on a single NVIDIA B200 GPU using PyTorch 2.11, CUDA 12.8, and Transformers 5.12.1. Because target embeddings are extracted once and subsequently cached, training the inversion model is comparatively lightweight. The inversion network contains fewer than 10M parameters in all $\ell = 20$ settings, making the one-time embedding extraction from the target model the dominant computational cost rather than inversion-model optimization.

A.4 Training Algorithm

We provide detailed training procedure of our Vec2DNA framework in presented in Algorithm 2.

A.5 Additional Ablation Studies

Table 6 provides the complete ablation of Vec2DNA, including optimization, data-volume, perturbation, and distance controls. All augmented settings use the same training-set size, number of gradient updates, and target-model query budget.

Optimization and data-volume controls. Matching the number of optimization steps on the original data (A0′) leaves DNABERT-2 unchanged, while producing a modest improvement on GENA-LM. Replicating the same base data four times (A1) provides no further gain, indicating that

Algorithm 2 Training Procedure of VEC2DNA

Require: corpus $\mathcal{D}$, frozen embedding model f, distances $\mathcal{K}$, weight λ

construct anchor–variant pairs $\mathcal{P} = \{(x, \mathcal{A}_k(x)) : x \in \mathcal{D},\ k \in \mathcal{K}\}$ and cache their embeddings and targets
for epoch $= 1$ **to** E **do**
  for each training batch **do**
    compute reconstruction loss $\mathcal{L}_{\text{rec}}$
    compute shared slots $\mathcal{S}(x, k) = \{j : y_j = y_j^{(k)}\}$
    compute feature-level consistency loss $\mathcal{L}_{\text{cons}}$ ▷ Eq. equation 2
    update $\phi \leftarrow \text{AdamW}(\phi, \nabla_\phi[\mathcal{L}_{\text{rec}} + \lambda\mathcal{L}_{\text{cons}}])$
  end for
end for
return $\phi^\star$

ID	Training Signal	Cons.	N	DNABERT-2		GENA-LM	
				Lev.	Acc.	Lev.	Acc.
References and negative controls							
A0	Base data only	✗	70k	$0.5765_{\pm.0002}$	.3545	$0.8476_{\pm.0014}$	.6964
A0′	Base data, step matched	✗	70k	$0.5765_{\pm.0002}$	.3545	$0.8589_{\pm.0016}$	.7259
A1	Base data replicated $\times 4$	✗	280k	$0.5756_{\pm.0007}$	.3558	$0.8585_{\pm.0020}$	.7264
A3	Gaussian jitter in embedding space	✗	280k	$0.5596_{\pm.0003}$	.3414	$0.7946_{\pm.0012}$	.6127
Matched query budget, reconstruction loss only							
A4	SNP variants, $k = 1, \times 3$	✗	280k	$0.6293_{\pm.0002}$	.4118	$0.9460_{\pm.0002}$	.8881
A5	SNP variants, $k \in \{1, 2, 3\}$	✗	280k	$0.6354_{\pm.0015}$	.4143	$0.9533_{\pm.0005}$	.9035
A2	**Unrelated hg38 sequences**	✗	280k	$0.6400_{\pm.0005}$	.4179	$0.9581_{\pm.0006}$	.9119
With feature level consistency							
A6	SNP variants, $k = 1, \times 3$	✓	280k	$0.6482_{\pm.0010}$	.4173	$0.9685_{\pm.0004}$	.9314
A7	**Ours**, $k \in \{1, 2, 3\}$	✓	280k	$\mathbf{0.6505}_{\pm.0017}$	**.4196**	$\mathbf{0.9704}_{\pm.0005}$	**.9373**

Table 6: Ablation studies of Vec2DNA training procedure at $l = 20$ on two embedding models that differ in tokenization and positional encoding. Every augmented variant is held to the same training set size ($N = 280$k) and the same number of gradient updates (27,350), and the augmented sequences are identical across the two encoders, so the only quantity that varies is the training signal itself. **A2** is the matched query control: it spends exactly as many victim queries as our method, but on unrelated genomic sequences, and therefore admits no anchor to variant correspondence.

nominal dataset size alone does not explain the improvement. Similarly, Gaussian perturbation in embedding space (A3) degrades reconstruction on both models, showing that generic embedding noise does not reproduce the benefit of querying additional genomic sequences.

Query budget versus paired structure. The matched-query control A2 uses unrelated hg38 sequences and already provides a substantial improvement over the base setting. SNP variants trained only with reconstruction loss (A5) perform slightly below A2, indicating that SNP perturbations are not intrinsically more informative than additional independent queries. Their primary role in Vec2DNA is instead to create anchor–variant pairs with known correspondence, which enables the feature-level consistency objective.

Role of perturbation distance. Using multiple SNP distances provides a comparatively small additional benefit. Without consistency, moving from single-distance variants (A4) to $k \in \{1, 2, 3\}$ (A5) improves Levenshtein similarity by approximately 0.006–0.007. Once feature-level consistency is applied, however, the corresponding gain from A6 to A7 becomes only 0.0023 on DNABERT-2 and 0.0019 on GENA-LM. Thus, multi-distance perturbation plays a supporting role, whereas feature-level consistency accounts for the substantive improvement of Vec2DNA.

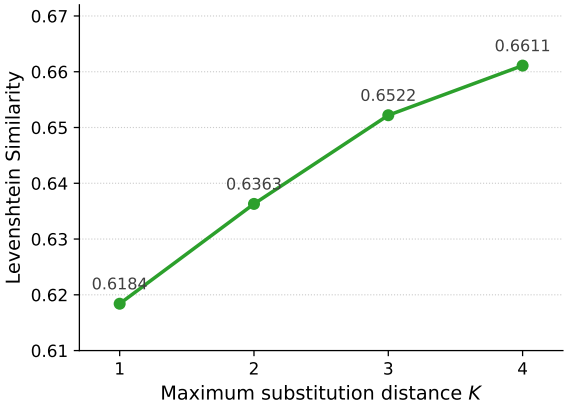

Figure 5: Levenshtein similarity of VEC2DNA on DNABERT-2 at $\ell = 20$ as the maximum substitution distance K increases, with one variant per distance and a fixed budget of 50 epochs. Larger K also increases the number of queries and gradient updates, so the trend should be read as the combined effect rather than as the effect of K alone.

A.6 QUERY-BUDGET SCALING AND PRIVACY IMPLICATIONS

Scaling with controlled sequence variants. We further examine how the inversion attack scales when additional anchor–variant queries are available. Figure 5 reports the performance of Vec2DNA on DNABERT-2 at $\ell = 20$ while increasing the maximum substitution distance from $K = 1$ to $K = 4$, with one synthetic variant generated at each distance. Levenshtein similarity increases consistently from 0.6184 at $K = 1$ to 0.6363, 0.6522, and 0.6611 at $K = 2$, $K = 3$, and $K = 4$, respectively. Thus, the inversion attack continues to improve as a larger set of controlled local perturbations is queried from the target embedding model.

Importantly, this scaling does not require collecting additional genomic source data. All variants are synthesized locally from the same set of training anchors by introducing controlled nucleotide substitutions. Increasing K therefore expands the attack data through additional queries to the target embedding model rather than through access to additional genomes or genomic loci. This observation is particularly relevant to embedding-service settings: an adversary that already possesses a collection of public or otherwise authorized sequences can construct multiple nearby variants of each sequence and query their released embeddings. Consequently, a permissive offline query budget can substantially enlarge the supervision available to an inversion attack even when the underlying genomic source dataset remains fixed.

Relation to the matched-query ablation. This result should be interpreted together with the matched-query controls in Appendix A.5. Those experiments show that increasing the number of target-model queries itself provides a substantial benefit, while feature-level consistency yields an additional gain beyond an equal-query control using unrelated hg38 sequences. The role of Figure 5 is therefore not to attribute the improvement solely to larger substitution distance. Rather, it illustrates a complementary privacy concern: controlled variants provide a practical mechanism through which an attacker can convert additional query access into progressively stronger inversion performance without requiring additional genomic source sequences.

There is an important confounding factor in this scaling experiment. Because the number of generated variants increases with K while the number of epochs is fixed, larger K also increases both the number of target-model queries and the number of optimization updates. The monotonic trend in Figure 5 should therefore be interpreted as the combined effect of additional query-derived supervision and additional training, rather than as a causal estimate of substitution distance alone. Nevertheless, from a privacy perspective, this combined scaling is itself informative because both resources become available naturally when an embedding service permits a larger offline query budget.

ℓ	12	16	17	18	20	24
Expected occurrences $2G/4^\ell$	3.7×10^2	1.44	0.36	0.090	5.6×10^{-3}	2.2×10^{-5}
Genome-unique fraction (hg38)	·	·	·	·	**81.55%**	·

Table 7: Identifiability as a function of sequence length. A specific ℓ-mer is expected to occur $2G/4^\ell$ times in a genome of length $G = 3.1 \times 10^9$ searched on both strands, which falls below one at $\ell \approx 16.3$. The second row reports the measured fraction of the 15,000 test ℓ-mers that occur exactly once in hg38 on both strands; entries shown as · were not measured in this work.

ℓ	20	40	60	80	100
Levenshtein similarity	0.5765	0.5350	0.5364	0.5308	0.5263
Random baseline	0.3827	0.4118	0.4245	0.4325	0.4385
Margin over random	**0.1938**	0.1232	0.1119	0.0983	0.0878

Table 8: Encoder-only inversion of DNABERT-2 embeddings as a function of sequence length. The margin over the random baseline is the informative quantity, since the baseline itself rises with ℓ.

A.7 Further Analysis of Genomic Privacy Leakage

Allele-level evaluation protocol. For each of the 15,000 test sequences x, we construct a Hamming-1 variant x' by replacing one randomly selected nucleotide with an alternative base and obtain its embedding from the same frozen encoder. Let $s(e, y)$ denote the sequence log-likelihood assigned by the inversion model to target y given embedding e. We use the score difference between the reference and variant sequences to evaluate pair discrimination, and additionally report a four-way allele call over the four possible nucleotides at the perturbed position. Reference bias measures how often the model incorrectly favors the reference allele when given the variant embedding.

These allele-level evaluations are candidate-conditioned analyses rather than unconstrained sequence reconstruction. For pair discrimination and the two-alternative forced-choice task, the reference and Hamming-1 variant sequences are treated as known candidate hypotheses, and the inversion model selects between them using their sequence likelihoods. For the four-way allele call, the genomic context and perturbed position are fixed, and the four possible nucleotides at that position are scored as candidate alleles. These experiments therefore measure whether a released embedding contains sufficient information to resolve allele identity when the relevant candidate set or locus is known. They should be distinguished from the embedding-only reconstruction setting, in which Vec2DNA receives only the released embedding and reconstructs the sequence without such candidate information.

Why $\ell = 20$ is a privacy-relevant operating point. A reconstructed sequence creates a meaningful privacy risk only if it is both invertible and sufficiently identifying. Under a uniform background, a specific ℓ-mer is expected to occur approximately $2G/4^\ell$ times in a genome of length G when both strands are considered. For GRCh38, this expectation falls below one near $\ell \approx 16.3$. Because the human genome contains substantial repetitive sequence, we verify uniqueness directly: 12,233 of the 15,000 test 20-mers (81.55%) occur exactly once in hg38.

Invertibility exhibits the opposite trend. As sequence length increases, the fixed-dimensional mean-pooled representation must encode increasingly more nucleotides, and the margin between inversion performance and the random baseline decreases substantially. As shown in Table 8, this margin drops from 0.1938 at $\ell = 20$ to 0.0878 at $\ell = 100$. Thus, $\ell = 20$ represents a regime in which genomic sequences are already frequently identifiable while remaining comparatively amenable to inversion.

Approximate reconstruction and locus disclosure. Exact reconstruction is not required for locus disclosure. Vec2DNA exactly reconstructs 32 of the 15,000 test sequences, while 110 and 370 sequences achieve Levenshtein similarity of at least 0.95 and 0.90, respectively. Table 9 illustrates representative cases. For non-repetitive sequences, even a small number of nucleotide errors can leave only a small set of candidate loci when matched against the public reference genome, allowing approximate reconstructions to substantially narrow the genomic origin.

	Sequence	Cplx.	Allele calls	Genomic loci	$r \leq 1$	$r \leq 2$	$r \leq 3$
A	GCTTCCCTGAGATGTAATAT	18	20/20	1 (chr19:28,174,221)	0	23	279
B	TTACGGATCCCACAGAAAAG	17	20/20	1 (chr12:28,242,481)	1	18	1,975
C	AAAAATAATAAGCAAAATAA	8	20/20	2	71	1,691	29,630

Table 9: Case study on the $\ell = 20$ test set. Complexity is the number of distinct 3-mers, with a maximum of 18. Genomic loci are exact matches in hg38 on both strands. The last three columns give the number of genomic sequences inside a Hamming ball of radius r around the true sequence, which is the candidate set an adversary retains after a reconstruction with r errors.

Limits of locus identification. Locus disclosure is weaker for repetitive or low-complexity sequences. For example, the repetitive sequence in Table 9 maps to multiple exact or near-exact genomic locations, preventing unique identification. Consistent with this limitation, only 21 of the 32 exactly reconstructed sequences are genome-unique. Our locus-level privacy claim therefore applies primarily to non-repetitive regions of the genome rather than uniformly to all genomic sequences.

A.8 Related Work: Defense Against Inversion Attack

Various defense mechanisms have also been proposed to mitigate embedding inversion attacks. Song & Raghunathan (2020) evaluated adversarial training to learn representations from which sensitive attributes of the original input are difficult to infer. Morris et al. (2023) showed that adding Gaussian noise to embeddings can reduce text reconstruction while partially retaining downstream utility. Zhuang et al. (2024) proposed a linear transformation of embeddings that suppresses inversion while preserving retrieval rankings, and Seputis et al. (2025) evaluated scalar quantization of embeddings as a lightweight defense that requires no retraining, reporting that reduced bit-depth preserves downstream performance. However, these post-hoc defenses do not explicitly distinguish between information required for downstream tasks and information exploited by inversion models, so that privacy and utility degrade together. Moreover, these defenses have been studied primarily for text embeddings, and, to the best of our knowledge, no defense has yet been evaluated against inversion attacks on genomic embeddings.

Learned defenses explicitly address privacy–utility trade-offs. Eguard combines a projection network with global information-suppression, local semantic-contrast, and downstream-task objectives Liu et al. (2026). SPARSE learns concept-sensitive dimensions and applies sensitivity-guided elliptical noise, with guarantees defined relative to a Mahalanobis metric Tsai et al. (2026). EntroGuard learns entropy-driven perturbations while constraining embedding distortion to retain retrieval utility Jin et al. (2026). Consequently, existing defenses cannot generally be characterized as ignoring task-relevant information.

Related protections for client–server inference operate on token-level or intermediate representations: DP-Forward perturbs forward-pass embedding matrices Du et al. (2023), TextObfuscator learns task-aware prototype clusters Zhou et al. (2023), ObfusLM combines vocabulary permutation with clustered embedding synthesis Lin et al. (2025), and P4LLM partitions inference and perturbs transmitted hidden states Lu et al. (2026). Their deployment assumptions differ from post-hoc protection of a released sequence-level vector. Moreover, alignment-based inversion Chen et al. (2025) and the recent denoising-aware DAEI preprint Wang et al. (2026) demonstrate limitations of particular transformation and noise defenses, motivating defense-aware evaluation rather than conclusions based solely on a fixed inverter.

A.9 Dataset Splits and Experimental Setup Details

Sequence extraction. All experiments use the human reference genome hg38 (GRCh38). We scan the primary assembly and retain the 22 autosomes together with chrX, chrY and chrM, discarding unplaced scaffolds, unlocalized contigs and alternate haplotypes so that no region of the genome is represented more than once. From each retained chromosome we take *non-overlapping* windows of length ℓ, advancing the read head by exactly ℓ nucleotides, so that no two extracted sequences share a nucleotide position and no sequence is a shifted copy of another. Windows containing the ambiguity symbol N are discarded, which removes centromeric and telomeric gaps as well as unresolved regions.

Finally we deduplicate globally across the whole genome, so that a sequence occurring at several loci contributes a single candidate. Table 10 reports the resulting pool at each sequence length. The pool shrinks as ℓ grows because the number of non-overlapping windows is inversely proportional to ℓ, while the duplicate rate falls from 14.1% at $\ell = 20$ to 1.1% at $\ell = 100$ as longer sequences become increasingly unique within the genome.

ℓ	Non-overlapping windows	Unique sequences	Duplicate rate	Sampled fraction
20	146,881,897	126,110,599	14.1%	0.079%
40	73,440,948	70,076,787	4.6%	0.143%
60	48,960,632	47,869,494	2.2%	0.209%
80	36,720,474	36,148,534	1.6%	0.277%
100	29,375,683	29,048,944	1.1%	0.344%

Table 10: Candidate pool of non-ambiguous, non-overlapping, globally unique hg38 sequences at each length ℓ. The sampled fraction is the share of the pool covered by the 100,000 sequences that form our dataset.

Sampling and splits. From the pool at each length we draw 100,000 sequences uniformly at random without replacement under a fixed seed of 42, and partition them into training, validation and test splits with ratios $0.70/0.15/0.15$, giving 70,000, 15,000 and 15,000 sequences respectively.

Splits are performed at the sequence-window level rather than by chromosome or contiguous genomic blocks. Therefore, although no extracted windows overlap in genomic coordinates and exact sequence duplicates are excluded across splits, adjacent non-overlapping windows may belong to different splits.

Because the pool is globally deduplicated and the sample is drawn without replacement, the three splits are disjoint by construction, which we verify explicitly: the pairwise intersections of the three split files are empty at every length. The splits are materialised once per length and stored on disk, and every subsequent experiment reads those same files. All four victim encoders, all five inversion baselines, every ablation arm and every defense configuration therefore operate on byte-identical sequences, so that a difference between two reported numbers can be attributed to the method rather than to the data. Standardisation statistics for the inversion model inputs are computed on the training split alone, model selection uses Levenshtein similarity on the validation split, and the test split is consumed only when final numbers are reported.

Augmentation is confined to the training split. The multi-distance variants used by Vec2DNA are generated exclusively from the 70,000 training sequences. With $K = \{1, 2, 3\}$ and one variant per distance, this yields 210,000 variants and a total of 280,000 training examples, each of which costs one query to the victim encoder. We confirm that the 210,000 variants are mutually distinct and that none of them coincides with any training, validation or test sequence, so the perturbation operator cannot reintroduce an evaluation sequence into training. The matched query budget control that replaces variants with unrelated hg38 sequences is constructed at the same size of 210,000 and satisfies the same disjointness condition, which is what allows the two augmentation sources to be compared at equal cost. Neither the validation nor the test split contributes to training, to augmentation, or to any normalisation statistic.

Scope of the experiments. The pool, the sample and the splits are rebuilt independently at each length, so the datasets for different ℓ are not nested and results across lengths should be read as separate experiments rather than as a single dataset viewed at different resolutions. We adopt $\ell = 20$ as the primary setting and run the main comparison across four victim encoders, the ablation study, the allele-level analysis and all defense evaluations at that length. The length study additionally covers $\ell \in \{40, 60, 80, 100\}$ on DNABERT-2 and GENA-LM, which are the two encoders that differ most in tokenizer vocabulary and positional encoding. Defenses are applied post hoc to the stored embeddings, leaving the sequences and the splits untouched, and are evaluated against an inversion model trained on undefended embeddings and kept frozen, which corresponds to a non-adaptive adversary. The privacy analyses, covering genome uniqueness, Hamming-distance-one discrimination, allele calling and locus disclosure, are all carried out on the 15,000 test sequences at $\ell = 20$. Finally, we note

that hg38 is a reference assembly rather than the genome of an identified individual, so no personal genomic data is used anywhere in this work; our privacy claims concern what an inversion attack would recover were the same pipeline applied to embeddings of individual sequences.

A.10 LIMITATIONS

Our study has several limitations that delineate its current scope and motivate concrete avenues for future inquiry.

Adaptive attacks and task scope. TAG-Q demonstrates an empirical reconstruction–utility trade-off for calibrated linear probes and a fixed VEC2DNA inversion model, assuming that the defender can calibrate the defense for the intended downstream tasks before embedding release. Our evaluation does not cover attackers retrained on defended embeddings or tasks outside the calibration set, and these empirical results do not establish a formal privacy guarantee. Evaluating defense-aware adaptation and utility transfer to unseen tasks is an important direction for future work.

Short sequence regime. All controlled experiments use $l = 20$. This setting serves as a controlled probe to isolate token-order disruption under mean pooling and BPE instability, where even standard baselines struggle. In practice, genomic embeddings routinely operate on 10^2–10^3 bp. Larger l sharply increases the mean-pooling compression ratio, which should make inversion harder, but it also provides richer biological context that an autoregressive prior could exploit. Systematically scaling VEC2DNA to longer regimes is an essential next step.