VANDAM: VIEWING A NUCLEOTIDE SEQUENCE WITH DNA MOLECULAR PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contemporary Genomic Foundation Models (GFMs) rely on a DNA-as-a-string paradigm that employs masked token prediction objectives for pretraining. However, this abstraction does not explicitly model the biochemical, structural, and physical properties essential to biological function. Many molecular properties can be estimated from sequence using established biophysical models, so their utility lies not in providing an independent modality, but in introducing priors that training objectives can explicitly exploit. We introduce VANDAM, a framework that extends the training of GFMs with DNA molecular priors. In self-supervised training, VANDAM predicts regional molecular properties from pooled representations. When functional labels are available and can reward retaining molecular priors, local features are additionally injected at the input. VANDAM consistently improves downstream performance across four architecture families and nine held-out genomic tasks by complementing token-based objectives. Probing experiments further demonstrate that the use of molecular priors generalizes to other unseen molecular properties. Code and trained model checkpoints will be made publicly available.

1 INTRODUCTION

Genomic foundation models (GFMs) learn transferable representations from large unlabeled genomic corpora and support tasks ranging from regulatory element annotation to variant effect prediction (Benegas et al., 2025c; Zhou et al., 2024; Boshar et al., 2025; Schiff et al., 2024; Brixi et al., 2026). Despite substantial diversity in architecture, scale, and context length, their training objectives remain largely inherited from natural language processing (NLP), where masked language modeling (MLM) and next-token prediction (NTP) train models to recover or predict token identity, treating nucleotides primarily as discrete symbols. In NLP, pre-training and many downstream applications share a common linguistic interface, as the model conditions on text to produce text. For generative tasks such as question answering, NTP directly trains the mechanism used to generate an answer. Genomic downstream tasks, by contrast, predict functional and phenotypic properties, such as regulatory activity, gene expression, and pathogenicity. Token prediction remains a useful general-purpose prior, but it does not directly require the representation to preserve the molecular properties relevant to these tasks.

DNA differs from natural language in an important respect: whereas linguistic semantics is inferred from usage, many molecular properties of DNA have been characterized and can be computed directly from sequence. For example, nearest-neighbor models provide duplex free-energy contributions (SantaLucia Jr, 1998), and helical shape parameters can be estimated using models calibrated against all-atom simulations (Zhou et al., 2013; Chiu et al., 2016; Li et al., 2024). We therefore argue that GFMs should represent DNA not merely as a string of symbols, but also as a molecule with predictable structural, energetic, and chemical properties.

Motivated by this distinction and by the successful use of structural and functional priors in protein models, we introduce VANDAM, a framework that complements the conventional DNA-as-a-string view with a DNA-as-a-molecule view. VANDAM incorporates molecular priors through two complementary mechanisms. Local Feature Injection (LFI) augments token embeddings with local molecular descriptors, while Region Property Prediction (RPP) adds an auxiliary objective that asks pooled representations to predict sequence-derived molecular properties of genomic regions. RPP is

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.

employed in self-supervised pre-training, while LFI is additionally used in multi-task training, where functional labels can reward the use of molecular descriptors. Together, these mechanisms provide input-level and representation-level routes for grounding GFMs in the physical properties of DNA.

Across four backbone families and nine held-out tasks, VANDAM improves the point estimate in 29 of 36 compute-matched pre-training comparisons and 33 of 36 multi-task comparisons; 25 and 23 comparisons, respectively, show statistical significance. Matched non-biological controls show that these gains depend on biological target content rather than auxiliary capacity alone, and RPP increases linear access to six of eight held-out qDNA descriptors on an independent confirmation split.

We make four contributions. First, we frame genomic representation learning around objective-appropriate molecular grounding, treating DNA as a molecule. Second, we introduce LFI and RPP, which respectively inject local descriptors when functional labels can reward their use and directly supervise regional molecular properties during self-supervised training. Third, we evaluate VANDAM through compute-matched experiments across four backbone families and nine held-out tasks, and use matched non-biological controls to distinguish biological content from auxiliary capacity and optimization. Finally, we show with frozen-representation probes that RPP increases access to an independent family of held-out molecular descriptors.

2 Related Work

Genomic foundation models. Genomic foundation models (GFMs) follow the paradigm established by large language models: pre-train a sequence model on unlabeled genomic corpora, then transfer it to downstream functional genomics tasks. The landscape comprises diverse architectures and tokenization schemes for genomic sequence modeling. DNABERT-2 (Zhou et al., 2024) keeps a flat transformer encoder but replaces the overlapping k-mer tokenization of its predecessor (Ji et al., 2021) with byte-pair encoding; NTv3 (Boshar et al., 2025) adopts a U-Net-style convolutional–transformer hybrid; Caduceus (Schiff et al., 2024) replaces attention with a bi-directional Mamba block and, in its parameter-sharing variant, enforces reverse-complement equivariance architecturally; and Evo2 (Brixi et al., 2026) scales a StripedHyena2 hybrid to 40B parameters trained on 9.3 trillion tokens. Despite this diversity, most prominent GFMs rely primarily on pre-training objectives inherited from natural language processing: masked recovery or autoregressive prediction of token identity. Whether a token denotes a single nucleotide or a byte-pair-encoded sequence fragment, the objective treats it as a discrete symbol and does not explicitly supervise the physical properties of the underlying DNA sequence.

This token-identity objective has recently been questioned, on the grounds that it captures local genomic syntax and motif structure while leaving broader functional context underdetermined. Several recent approaches therefore modify the training objective or the genomic representation: JEPA-DNA (Larey et al., 2026b) adds supervision of high-level functional embeddings of masked segments alongside token-level reconstruction; D3LM (Yang et al., 2026) reframes DNA modeling as discrete diffusion; UKBioBERT (Liu et al., 2026) continues masked-language modeling on sequences containing population-level genetic variation rather than reference sequence alone; and BioToken/BioFM (Medvedev et al., 2025) encodes variants and genomic annotations as biologically informed tokens. DNAGPT (Zhang et al., 2023) augments autoregressive pre-training with sequence-order classification and GC-content regression over sampled sequence regions. MutBERT (Long et al., 2025) represents each position as a population allele-frequency distribution; GPN-MSA (Benegas et al., 2025a) uses multi-species alignments as input, and PhyloGPN (Albors et al., 2025) incorporates phylogenetic alignments through its training objective while retaining single-sequence inference. These results suggest that the training signal and the representation of genomic context, not only capacity or architecture, affect the quality of learned representations.

Biophysical supervision in representation learning. A relevant precedent is UTR-LM (Chu et al., 2024), which augments masked-nucleotide modeling of 5′ untranslated regions with auxiliary targets computed from the same sequence by ViennaRNA: secondary structure at masked positions and sequence-level minimum free energy. These targets provide biophysical supervision without introducing an additional experimental or structural modality. In DNA modeling, MIXBend (Yang et al., 2024) aligns DNABERT sequence representations with physicochemical features during self-

supervised pre-training, but develops this approach specifically for DNA bendability prediction. Whereas UTR-LM focuses on RNA and MIXBend is tailored to DNA bendability, VANDAM develops a general framework for grounding existing GFMs in a broader set of sequence-derived DNA structural, energetic, and chemical properties under both self-supervised and functionally supervised training.

Biophysical features in supervised prediction. Biophysical features carry well-established predictive signal for DNA, but much of the existing evidence comes from task-specific supervised models. DNA-shape features such as minor-groove width, computed from pentamer lookup tables built from all-atom Monte Carlo simulations (Zhou et al., 2013; Chiu et al., 2016; Li et al., 2017), improve transcription-factor-binding-site prediction over position-weight matrices and sequence-only baselines (Zhou et al., 2015; Mathelier et al., 2016). More recent work replaces the lookup tables with learned predictors that account for extended flanking context (Li et al., 2024). DNABERT-CorePromMM (Papineni et al., 2025) augments DNABERT embeddings with nucleotide composition, CpG density, minor-groove width, electrostatic potential, and other DNA-intrinsic features for supervised core-promoter prediction. Together, these studies establish the utility of sequence-derived molecular descriptors, but primarily within task-specific supervised settings.

3 METHODS

3.1 FROM DNA AS A STRING TO DNA AS A MOLECULE

VANDAM complements the conventional DNA-as-a-string training paradigm with DNA molecular priors (Figure 1). These priors provide a biophysical inductive bias without requiring an additional experimental modality. In the following section, we first define the biological features used by VANDAM, then present Local Feature Injection and Region Property Prediction, and finally describe how these mechanisms are combined under different supervision regimes.

3.2 DNA MOLECULAR PRIORS

We define a mapping from each nucleotide in a DNA sequence to a set of molecular properties. We use the 13 structural features available in DNAshapeR (Chiu et al., 2016), whose original four-feature framework was introduced by Zhou et al. (2013) and expanded to 13 features by Li et al. (2017), plus minor-groove electrostatic potential (EP) from Chiu et al. (2017), for a total of 14 sequence-derived profiles. The features are organized into three groups:

- Minor-groove environment: Minor-groove width (MGW) and electrostatic potential (EP).
- Intra-base-pair geometry: Propeller twist (ProT), stretch, shear, buckle, opening, and stagger.
- Inter-base-pair geometry: Helical twist (HelT), roll, tilt, rise, shift, and slide.

All local properties are deterministic, model-derived molecular priors rather than direct measurements and do not capture cell-specific context, environmental conditions, or conformational variability. Appendix A details the complete local feature catalog.

3.3 LOCAL FEATURE INJECTION

Local feature injection supplies the backbone with biological priors at the input without modifying its internal blocks. Let $x = (x_1, \ldots, x_L)$ denote a DNA sequence of L nucleotides. Let ϕ be a feature map assigning each base-pair in x to a vector in $\mathbb{R}^F$ (§3.2). Because some descriptors depend on a local sequence window, $\phi(x)_i$ may depend on the sequence context surrounding x_i, rather than on x_i alone.

Let $\mathrm{Tok}(x) = (t_1, \ldots, t_M)$ denote the tokenization of x, where M need not equal L. For each sequence-bearing token t_j, let $S_j \subseteq \{1, \ldots, L\}$ denote the set of nucleotide positions spanned by that token. We align the nucleotide-level features with the token sequence by mean-pooling the feature

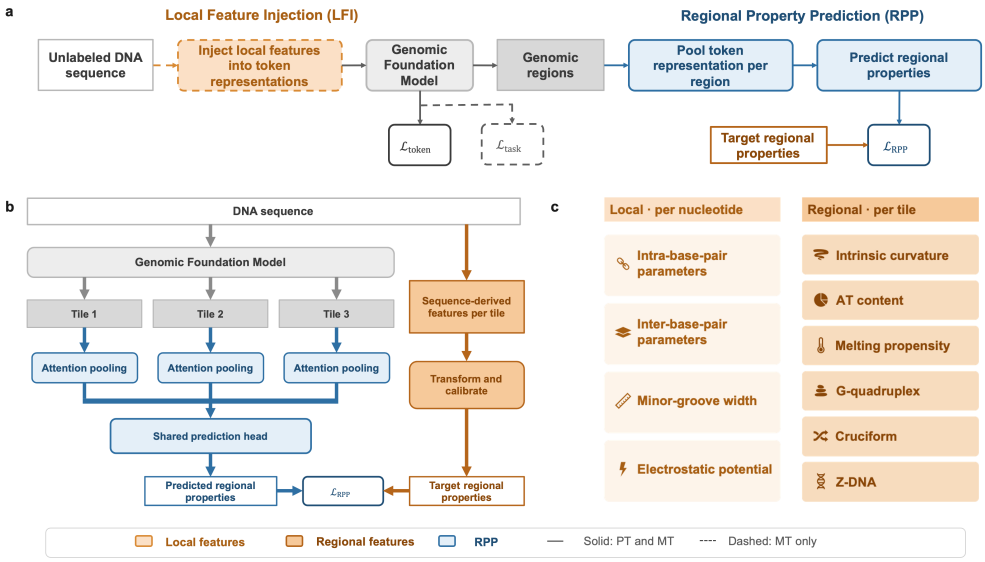


Figure 1: VANDAM architecture and molecular feature hierarchy. **a**, High-level training flow. Local Feature Injection (LFI; dashed) augments input token representations with local sequence-derived descriptors and is used only in VANDAM-MT, where functional-task supervision $\mathcal{L}_{\text{task}}$ can reward their use. Regional Property Prediction (RPP; blue) pools representations for genomic regions and predicts tile-level properties in both VANDAM-PT and VANDAM-MT, alongside the native token objective $\mathcal{L}_{\text{token}}$. **b**, RPP mechanism. Backbone representations are partitioned into tiles, attention-pooled within each tile, passed through a shared prediction head, and compared with transformed and calibrated sequence-derived regional targets using $\mathcal{L}_{\text{RPP}}$. **c**, Local features comprise per-nucleotide geometric and electrostatic descriptors; regional targets comprise tile-level measures of curvature, composition, duplex stability, and non-B-DNA structural propensity.

vectors over each token span:

$$\bar{\phi}(x)_j = \frac{1}{|S_j|} \sum_{i \in S_j} \phi(x)_i, \qquad \bar{\phi}(x)_j \in \mathbb{R}^F. \tag{1}$$

Under single-nucleotide tokenization, each sequence-bearing token spans one nucleotide, and therefore its token-aligned biological feature vector is equal to the feature vector of that nucleotide. Features are computed at nucleotide resolution before aggregation, preserving the local sequence context that may cross token boundaries.

We encode each token-aligned feature vector with a single linear layer:

$$h_j = W_f \, \bar{\phi}(x)_j + b_f, \qquad h_j, b_f \in \mathbb{R}^H, \quad W_f \in \mathbb{R}^{H \times F}. \tag{2}$$

Let $e(t_j) \in \mathbb{R}^D$ denote the backbone embedding of token t_j. We concatenate the feature embedding with the token embedding and project the result back to the backbone's input width:

$$m_j = W_m \, [e(t_j) \,\|\, h_j] + b_m, \qquad m_j \in \mathbb{R}^D, W_m \in \mathbb{R}^{D \times (D+H)}. \tag{3}$$

The mixed embeddings m_j replace the backbone's standard embedding outputs. Biological features are computed only from nucleotides visible under the native MLM or NTP objective, preventing leakage. We first train the injector against a frozen backbone and then train both jointly; Appendix B gives the masking rules, warm-up procedure, token-level alignment, and backbone-specific adaptations.

3.4 Region Property Prediction

A natural first approach to representation-level supervision is masked feature modeling (MFM), which predicts local molecular properties at masked positions. However, each target is determined by a

short sequence window, much of which remains visible after masking. MFM therefore does not need to integrate evidence beyond that window, making any downstream gain difficult to distinguish from generic local auxiliary regularization. Motivated by this limitation, RPP shifts supervision from local features to tile-level properties defined over genomic regions, encouraging the pooled representation to retain information distributed across a broader sequence context.

3.4.1 Regional Properties

For each genomic region, we compute six sequence-derived properties: intrinsic curvature, AT content, melting propensity, G-quadruplex propensity, cruciform propensity, and Z-DNA propensity. Together, these targets span regional geometry, composition, duplex stability, and non-B-DNA structural potential which quantifies sequence-based propensity. Appendix C gives the exact definition, unit, and biological motivation for each selected property.

3.4.2 Learning with Regional Properties

RPP directly rewards pooled representations for retaining molecular properties, rather than leaving their preservation to emerge from the native token objective. Given a DNA sequence, we partition it into fixed non-overlapping tiles $\mathcal{T}$ (regions). For each tile, the backbone's hidden states are pooled to a single vector, $z_j = \mathrm{pool}\big(\{H_i : i \in \mathcal{T}_j\}\big)$, where $\mathcal{T}_j$ contains the token indices of tile j and H_i is the backbone hidden state at position i.

A lightweight linear head predicts a vector of properties derived from the true sequence, $\hat{y}_j = W_{\mathrm{RPP}} z_j + b_{\mathrm{RPP}}$, where $W_{\mathrm{RPP}} \in \mathbb{R}^{N \times D}$, $b_{\mathrm{RPP}} \in \mathbb{R}^N$, and N is the number of features used. Let $\mathcal{V}$ be the set of valid tiles in a minibatch; the RPP loss is the mean squared error, averaged over valid tiles and target dimensions:

$$\mathcal{L}_{\mathrm{RPP}} = \frac{1}{N} \sum_{r=1}^{N} \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} (\hat{y}_{j,r} - y_{j,r})^2 \,. \tag{4}$$

The RPP loss is combined with the native token loss, $\mathcal{L}_{\mathrm{token}}$ (typically, MLM or NTP), using a linearly warmed coefficient,

$$\mathcal{L} = \mathcal{L}_{\mathrm{token}} + \lambda_{\mathrm{RPP}}(s)\mathcal{L}_{\mathrm{RPP}}, \qquad \lambda_{\mathrm{RPP}}(s) = \lambda_{\max} \min\Big(1, \frac{s}{1000}\Big) \tag{5}$$

where s is the optimizer step. Thus, λ_{RPP} increases linearly from zero during the first 1,000 steps and remains at its configured value thereafter. If $\mathcal{V} = \emptyset$, the RPP loss is omitted for that minibatch and optimization proceeds using only $\mathcal{L}_{\mathrm{token}}$. Both terms read from the same forward pass, so RPP adds no additional encoder evaluation. Appendix D gives target transformation and calibration details, precise tiling and validity rules, and backbone-specific BPE alignment details.

3.5 VANDAM Training Regimes

The appropriate mechanism for incorporating molecular priors depends on whether the training objective makes the supplied information consequential. We therefore consider two configurations. In the self-supervised pre-training phase, *VANDAM-PT* applies RPP without LFI. We omit LFI because its local descriptors are deterministic functions of context available to the native token objective and are not themselves supervised. In multi-task tuning, when functional task labels are available alongside the native token objective, *VANDAM-MT* combines LFI with RPP. Task gradients can reward use of the injected descriptors for functional prediction, while RPP directly supervises retention of regional properties.

4 Experiments

4.1 Setup

Models. We evaluate VANDAM on four backbones spanning four distinct architecture families to assess whether the effect generalizes beyond a single architectural choice: NTv3-8M (Boshar

et al., 2025), a U-Net convolutional–transformer hybrid; Caduceus-PS (Schiff et al., 2024), a bi-directional Mamba state-space model; DNABERT-2-117M (Zhou et al., 2024), a flat transformer; and HyenaDNA-small-32k (Nguyen et al., 2023), a long-range Hyena-operator model. We use the shorter family names Caduceus and HyenaDNA thereafter.

Table 1: Backbone models used in all experiments. Parameter counts are trainable backbone parameters in our implementation, excluding VANDAM and downstream heads.

Model	Params	Architecture	Tokenizer	Pre-training objective
NTv3-8M	8M	U-Net conv-transformer	Single-nucleotide	MLM
Caduceus-PS	7.73M	Bi-directional Mamba	Single-nucleotide	MLM
DNABERT-2	117M	Flat transformer	BPE	MLM
HyenaDNA	3.28M	Hyena operator	Single-nucleotide	NTP

Pre-training. We apply continual pre-training to all models on OpenGenome2 (Brixi et al., 2026), a sequence-only multi-species corpus of roughly 8.8 trillion nucleotides spanning bacteria, archaea, eukarya and phage. We compare VANDAM-PT with its released checkpoint and Baseline-PT, its compute-matched control using only the native token objective. Matching the data and compute budget isolates the effect of the injected biology from the effect of simply training longer. Appendix E gives the full training configuration.

Multi-task training. As a second regime, we perform mixed-objective multi-task continual training using eight contamination-controlled functional tasks disjoint from the evaluation suite (Appendix F). A task-batch fraction ρ_{task} controls the proportion of optimization steps assigned to labeled task batches; the remaining updates retain the backbone's native sequence objective. Both Baseline-MT and VANDAM-MT use the same eight task-specific supervised heads, one for each task in the multi-task pool; VANDAM-MT additionally includes the shared RPP head. VANDAM-MT applies LFI to both batch types and RPP only to sequence-only batches, where it augments the native MLM or NTP loss. Baseline-MT is its compute-matched control, using the same data, schedule, and task mixture without LFI or RPP. Appendix E gives the schedule, head warm-up, and compute budget.

Evaluation and statistical analysis. We evaluate through GFMBench-API (Larey et al., 2026a) on nine held-out tasks: four zero-shot variant-effect tasks, scored by cosine dissimilarity between pooled reference and alternate-sequence embeddings, and five tasks using frozen representations with linear probes. We use AUROC for variant-effect and classification tasks, and macro Pearson correlation for CAGE expression. Appendix G gives the task sources, sample sizes, metrics, and probe protocol.

For each task–backbone comparison, we evaluate the paired score difference $\Delta = \text{VANDAM} - \text{Baseline}$; for linear-probe tasks, this is averaged across five matched probe seeds. We construct 95% confidence intervals from 10,000 paired bootstrap replicates, resampling test units for all tasks and the five probe seeds for linear-probe tasks. Each replicate uses the same sampled test units for VANDAM and Baseline. Zero-shot tasks therefore resample test units only. We obtain one-sided p-values from 10,000 paired randomization replicates that swap VANDAM and Baseline predictions within each test unit; for linear-probe tasks, each swap is shared across seeds. We apply Benjamini–Hochberg (BH) adjustment separately within each training regime and direction. A VANDAM advantage requires a positive mean paired difference, a 95% confidence interval wholly above zero, and an adjusted positive-direction $q < 0.05$; we use the mirrored criterion for Baseline advantages. Appendix H provides the complete inferential procedure and comparison-level results.

4.2 Downstream Performance and Component Ablations

Tables 2 and 3 report the per-task comparisons for continual pre-training (Baseline-PT versus VANDAM-PT) and multi-task training (Baseline-MT versus VANDAM-MT), respectively. Scores for frozen linear-probe tasks are means across the five matched probe seeds. To probe the role of target locality, Appendix I contrasts RPP with MFM and a random features control, showing how downstream gains change when supervision moves from local features to regional properties. To understand the contribution of each component under the two training regimes, Table 4 compares LFI

Table 2: Pre-training results per task. Bold denotes the higher central score; superscript +/− denotes a BH-adjusted paired difference favoring VANDAM/Baseline.

	NTv3-8M		Caduceus		DNABERT-2		HyenaDNA	
Task	Baseline-PT	VANDAM-PT	Baseline-PT	VANDAM-PT	Baseline-PT	VANDAM-PT	Baseline-PT	VANDAM-PT
Clinical pathogenicity (ClinVar)	58.0	**61.7**$^{+}$	60.5	**61.9**$^{+}$	59.3	**65.9**$^{+}$	66.9	**75.3**$^{+}$
Disease variants (OMIM)	49.8	**59.0**$^{+}$	48.6	**55.0**$^{+}$	**55.7**$^{-}$	51.2	**52.2**	52.1
Mendelian variants (TraitGym)	49.7	**55.8**$^{+}$	47.2	**55.9**$^{+}$	**54.2**	53.7	53.9	**54.2**
BRCA1 variant effect	50.9	**51.6**	49.6	**54.0**$^{+}$	56.9	**57.6**	51.9	**55.3**$^{+}$
Transcription-factor binding	79.7	**80.8**$^{+}$	68.9	**75.2**$^{+}$	79.0	**80.3**$^{+}$	73.7	**80.0**$^{+}$
Regulatory variants (eQTL)	68.3	**70.2**$^{+}$	**67.0**	66.5	69.1	**71.9**$^{+}$	68.2	**69.3**$^{+}$
Gene-expression variants	61.8	**63.3**$^{+}$	**62.4**	61.8	65.2	**66.3**$^{+}$	63.3	**63.7**
Chromatin state (histone)	66.0	**68.1**$^{+}$	60.2	**65.7**$^{+}$	66.1	**69.8**$^{+}$	**66.2**$^{-}$	63.4
CAGE expression	24.2	**34.9**$^{+}$	**23.4**$^{-}$	22.2	34.1	**34.5**$^{+}$	23.5	**30.5**$^{+}$

Table 3: Multi-task training results per task. Bold denotes the higher central score; superscript +/− denotes a BH-adjusted paired difference favoring VANDAM/Baseline.

	NTv3-8M		Caduceus		DNABERT-2		HyenaDNA	
Task	Baseline-MT	VANDAM-MT	Baseline-MT	VANDAM-MT	Baseline-MT	VANDAM-MT	Baseline-MT	VANDAM-MT
Clinical pathogenicity (ClinVar)	68.1	**70.2**$^{+}$	**54.4**$^{-}$	51.4	59.5	**69.7**$^{+}$	55.7	**71.2**$^{+}$
Disease variants (OMIM)	54.9	**57.4**	47.4	**53.1**$^{+}$	52.0	**66.4**$^{+}$	47.9	**51.0**
Mendelian variants (TraitGym)	51.5	**56.8**$^{+}$	47.8	**55.6**$^{+}$	55.0	**57.5**	51.4	**54.8**
BRCA1 variant effect	51.9	**54.6**	51.7	**52.5**	52.5	**56.2**$^{+}$	51.6	**53.4**
Transcription-factor binding	78.7	**79.8**$^{+}$	53.7	**66.4**$^{+}$	78.1	**82.6**$^{+}$	75.2	**81.0**$^{+}$
Regulatory variants (eQTL)	68.1	**69.3**$^{+}$	60.9	**63.0**$^{+}$	68.8	**71.8**$^{+}$	69.7	**70.0**
Gene-expression variants	62.6	**63.1**	57.8	**59.9**$^{+}$	68.2	**69.4**$^{+}$	**65.8**	64.9
Chromatin state (histone)	68.3	**69.3**$^{+}$	52.9	**55.4**$^{+}$	67.2	**72.6**$^{+}$	68.2	**69.2**$^{+}$
CAGE expression	31.7	**35.9**$^{+}$	13.7	**14.6**	**36.9**	36.6	20.5	**30.8**$^{+}$

and RPP in isolation and in combination on NTv3-8M. Appendix J further characterizes the learned LFI pathway, while Appendix K separately ablates the six RPP targets.

The evaluation suite covers clinical and Mendelian variant pathogenicity, experimentally measured BRCA1 effects, transcription-factor binding, regulatory and gene-expression variant effects, chromatin-state prediction, and CAGE expression. VANDAM has positive point-estimate improvements in 29 of 36 pre-training and 33 of 36 multi-task task–backbone comparisons, spanning all four backbones. After paired inference and within-regime BH adjustment, VANDAM shows statistically significant improvements on 25 pre-training and 23 multi-task comparisons. In contrast, the compute-matched baseline is significantly superior on three pre-training and one multi-task training comparison. This trend holds across convolutional, state-space, transformer, and Hyena architectures; single-nucleotide and BPE tokenization; and both masked-language and next-token prediction objectives. Together, the results support molecular grounding as a broadly useful auxiliary signal, both alone and when coupled with functional supervision. Intrinsic token-objective metrics (e.g. perplexity) vary by backbone and regime rather than increasing uniformly (Appendix L), indicating that the downstream gains are not explained solely by improved token-identity prediction.

4.3 Does Biological Content Matter?

We test whether VANDAM's gains arise from biological content rather than added capacity or optimization pathways. For LFI, we replace $\phi(x)_i$ with a random vector of the same dimension. For RPP, we use matched shuffled-table and algorithmic regional targets that remove the biological sequence-to-property mapping while preserving the regional prediction problem; Appendix M gives the complete control construction.

Figure 2a shows that, during pre-training, biological RPP exceeds the algorithmic and shuffled-table controls by 3.4 and 4.7 mean-AUROC points, respectively; thus, an auxiliary regional-prediction task is insufficient without biologically meaningful targets. Under multi-task training, randomizing LFI

Table 4: Component ablation across pre-training and multi-task training on NTv3-8M. Values are mean AUROC over the eight AUROC tasks.

Configuration	Pre-training	Multi-task
Original checkpoint	59.5	59.5
Baseline-PT / Baseline-MT	60.5	63.0
LFI only	60.5	64.2
RPP only	**63.8**	63.2
LFI + RPP	63.4	**65.1**
RPP without MLM loss, masking retained	60.2	60.2
RPP without MLM loss, no masking	59.7	60.1

reduces mean AUROC by 3.9 points (Figure 2b), while replacing biological RPP with shuffled-table or algorithmic targets reduces it by 1.1 and 1.6 points, respectively (Figure 2c).

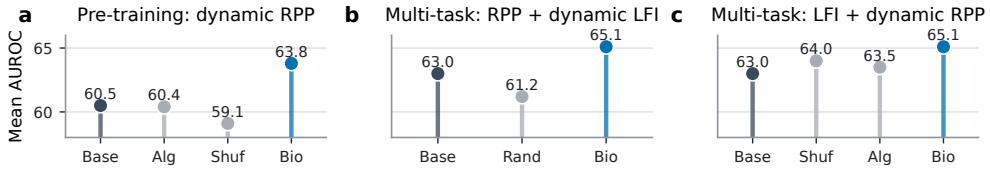


Figure 2: Biological-content controls on NTv3-8M (mean AUROC over eight tasks). **a**, Pre-training RPP-target controls. **b**, Multi-task LFI controls with biological RPP. **c**, Multi-task RPP-target controls with biological LFI.

4.4 RPP Transfers to Unseen Molecular Descriptors

The downstream evaluations above establish that VANDAM improves task performance, but do not show whether the use of molecular priors generalizes to other unseen molecular properties. To examine what changed in the representation, we turn to qDNA, which models sequence-dependent charge dynamics to compute model-derived electronic and photophysical observables of hole and electron localization, transfer, exciton decay, and electron–hole separation (Herb et al., 2025; Hawke et al., 2010). We use these quantities as held-out test-time molecular descriptors, which provide a physically distinct view of DNA from the properties used in RPP.

We fit ridge regression models using representations from Baseline-PT and VANDAM-PT to predict each one of the eight qDNA descriptors averaged over 300-bp tiles. Since qDNA descriptors may be linearly correlated with the six properties used as RPP targets, we further fit an OLS model on those properties to predict qDNA features. We then subtract its prediction from the ground-truth qDNA values and use the resulting residual as a new target for ridge probing. This ensures that the reported improvement (Table 5) is only attributed to the learned representations rather than to linear correlation between the two set of properties. Appendix N gives the complete target construction, probing, and inference protocol.

Table 5: qDNA transfer gain (ΔR^2, VANDAM-PT minus Baseline-PT). Bold denotes statistical significant improvement.

Descriptor	Original	Residual	Descriptor	Original	Residual
Hole central population	**0.060**	**0.054**	Electron central population	**0.083**	**0.057**
Hole maximum transfer	−0.041	−0.032	Electron maximum transfer	−0.055	−0.046
Hole delocalization (IPR)	**0.071**	**0.091**	Electron delocalization (IPR)	**0.067**	**0.051**
Exciton lifetime	**0.073**	**0.084**	Electron–hole separation	**0.069**	**0.076**

VANDAM-PT improves the predictability for six out of eight qDNA descriptors, both for original and residualized. A gain is considered statistically significant and reported as an improvement when the BH-adjusted $q \leq 0.05$ and $\Delta R^2 \geq 0.02$. Full results and 95% CIs are in Appendix Table 20.

These results show that VANDAM improves linear access to electronic and photophysical descriptors that were never used for training. Their persistence after residualization indicates that the transfer

is not explained solely by a linear recombination of the six supervised RPP targets. Thus, within NTv3-8M, VANDAM generalizes beyond its training properties to unseen molecular descriptors.

5 Discussion

Across four architecture families, including two tokenization schemes, and both MLM and NTP objectives, molecular grounding improves performance in most downstream task comparisons. This consistent performance supports an objective–task alignment interpretation: token recovery is a useful general-purpose pre-training signal, but it does not directly require representations to preserve molecular properties relevant to downstream functional and phenotypic prediction. The pre-training RPP-without-MLM ablation underperforms Baseline-PT (Table 4), further indicating that molecular grounding complements rather than replaces the token objective.

This complementarity is best understood through the role of the training objective. VANDAM does not introduce information unavailable from sequence: every local descriptor and regional property used in this work is a deterministic function of the input. The operative variable is instead what the objective obliges the representation to preserve. Making a property available is passive and can be ignored; attaching a loss to it, or attaching labels whose prediction it eases, makes its use consequential. This differs from settings in which a supplied structural prior contains information not directly present in the model's sequence input, whether that information is experimentally measured or obtained from another source. In the NTv3-8M component ablation, LFI alone matches Baseline-PT during continual pre-training (both 60.5 mean AUROC), but improves over Baseline-MT during multi-task training (64.2 versus 63.0; Table 4). This pattern is consistent with a simple asymmetry: the native token objective can be optimized from the same bounded local context from which an injected descriptor is computed, without requiring the representation to retain that descriptor. Functional labels, by contrast, can reward using a biologically meaningful featurization to predict targets not fixed by the local window.

Together, the biological-content controls and qDNA probing suggest that RPP does more than make its six supervised targets linearly accessible: it promotes representations that retain broader molecular structure. The persistence of transfer after residualization strengthens this interpretation.

Limitations and future work. The present study focuses on molecular properties defined over local genomic tiles and on comparatively small backbones with sequence contexts of at most 8,192 bp, enabling controlled compute-matched comparisons across architectures. We therefore do not establish how VANDAM behaves as the spatial scale of its molecular priors, model capacity, and sequence context grow together. Under MLM, zeroing descriptors whose pentamer intersects a mask makes approximately 47% of interior LFI vectors zero. Appendix O evaluates feature-only imputation on NTv3-8M; the effects are task-specific and were not tested across backbones. Extending these masking-policy experiments alongside supervision for long-range and three-dimensional genome organization is an important direction for future work.

6 Conclusion

VANDAM reframes genomic foundation-model training around DNA as a molecule as well as a sequence. By directly supervising regional molecular properties during self-supervised continual pre-training, and by incorporating local molecular descriptors when functional labels can reward their use, it provides objective-appropriate routes for grounding representations in sequence-derived biophysics. Across diverse architectures, tokenizers, and token-identity objectives, these interventions improve downstream genomic representations; RPP also increases access to held-out molecular descriptors. More broadly, molecular priors can help genomic foundation models capture sequence-dependent physical properties of DNA.

AI USE STATEMENT

We used generative AI tools to assist with research execution through drafting and debugging software code used in the experiments and for figure preparation; to identify potentially relevant literature; and to rephrase or polish selected sentences in the manuscript. We did not use generative AI for research ideation, to formulate research questions or hypotheses, design the study or scientific methods, select experiments or analyses, draft sections of the manuscript, interpret results, or generate scientific claims or conclusions. The authors inspected the original sources for all AI-suggested literature and independently verified all citations and related-work descriptions. All AI-assisted code was reviewed, tested, and validated by the authors, and all AI-assisted text was reviewed and revised for accuracy and faithful representation of the work. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a computational method for learning genomic representations from existing sequence resources and benchmark tasks. It does not involve the collection of new data from human participants or intervention with human subjects. The method is intended for research use and is not presented as a clinical decision-making system.

REPRODUCIBILITY STATEMENT

The Methods section specifies the VANDAM mechanisms and training regimes, while Appendix D documents the RPP targets, transformations, and backbone-specific alignment rules. Appendix E provides the training configurations and compute budgets; Appendix F describes the multi-task pool and overlap controls; Appendix G specifies the evaluation tasks, sources, sample sizes, metrics, and probe protocols; and Appendix H gives the statistical-inference procedure. The project code, training and evaluation configurations, and reproduction scripts will be released on GitHub, and the trained model checkpoints will be released on Hugging Face. Links to these resources will be added to the paper upon acceptance. Together with the cited public data sources, these materials are intended to enable reproduction of the reported results.

REFERENCES

C Albors, JC Li, G Benegas, C Ye, and YS Song. A phylogenetic approach to genomic language modeling in research in computational molecular biology, 2025.

Amina Bedrat, Laurent Lacroix, and Jean-Louis Mergny. Re-evaluation of g-quadruplex propensity with g4hunter. *Nucleic acids research*, 44(4):1746–1759, 2016.

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.

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S Song. Genomic language models: opportunities and challenges. *Trends in Genetics*, 41(4):286–302, 2025c.

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

Václav Brázda, Rob C Laister, Eva B Jagelská, and Cheryl Arrowsmith. Cruciform structures are a common dna feature important for regulating biological processes. *BMC molecular biology*, 12(1): 33, 2011.

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.

Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. Dnashaper: an r/bioconductor package for dna shape prediction and feature encoding. *Bioinformatics*, 32(8): 1211–1213, 2016.

Tsu-Pei Chiu, Satyanarayan Rao, Richard S Mann, Barry Honig, and Remo Rohs. Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–dna binding. *Nucleic acids research*, 45(21):12565–12576, 2017.

Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. A 5' utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence*, 6(4):449–460, 2024.

Gregory M Findlay, Riza M Daza, Beth Martin, Melissa D Zhang, Anh P Leith, Molly Gasperini, Joseph D Janizek, Xingfan Huang, Lea M Starita, and Jay Shendure. Accurate classification of brca1 variants with saturation genome editing. *Nature*, 562(7726):217–222, 2018.

Laurence George Demosthenis Hawke, George Kalosakas, and Constantinos Simserides. Electronic parameters for charge transfer along dna. *The European Physical Journal E*, 32(3):291–305, 2010.

Dennis Herb, Marco Trenti, Marilena Mantela, Constantinos Simserides, Joachim Ankerhold, and Mirko Rossini. Quantumdna: A python package for analyzing quantum charge dynamics in dna and exploring its biological relevance. *Computer Physics Communications*, 313:109626, 2025.

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.

David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

Ariel Larey, Elay Dahan, Amit Bleiweiss, Raizy Kellerman, Guy Leib, Omri Nayshool, Dan Ofer, Tal Zinger, Dan Dominissini, Gideon Rechavi, et al. Gfmbench-api: A standardized interface for benchmarking genomic foundation models. *Biorxiv*, pp. 2026–02, 2026a.

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

Jinsen Li, Jared M Sagendorf, Tsu-Pei Chiu, Marco Pasi, Alberto Perez, and Remo Rohs. Expanding the repertoire of dna shape features for genome-scale studies of transcription factor binding. *Nucleic acids research*, 45(22):12877–12887, 2017.

Jinsen Li, Tsu-Pei Chiu, and Remo Rohs. Predicting dna structure using a deep learning method. *Nature communications*, 15(1):1243, 2024.

Tianyu Liu, Xiangyu Zhang, Jiecong Lin, Luca Pinello, Rex Ying, and Hongyu Zhao. Pre-training genomic language model with variants for better modeling functional genomics. *npj Artificial Intelligence*, 2(1):46, 2026.

Weicai Long, Houcheng Su, Jiaqi Xiong, and Yanlin Zhang. Mutbert: probabilistic genome representation improves genomics foundation models. *Bioinformatics*, 41(Supplement_1):i294–i303, 2025.

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

Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W Wasserman. Dna shape features improve transcription factor binding site predictions in vivo. *Cell systems*, 3(3): 278–286, 2016.

Aleksandr Medvedev, Karthik Viswanathan, Praveenkumar Kanithi, Kirill Vishniakov, Prateek Munjal, Clement Christophe, Marco AF Pimentel, Ronnie Rajan, and Shadab Khan. Biotoken and biofm–biologically-informed tokenization enables accurate and efficient genomic foundation models. *bioRxiv*, pp. 2025–03, 2025.

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

Alfred Nordheim and Alexander Rich. Negatively supercoiled simian virus 40 dna contains z-dna segments within transcriptional enhancer sequences. *Nature*, 303(5919):674–679, 1983.

Nikos Panayotatos and Robert D Wells. Cruciform structures in supercoiled dna. *Nature*, 289(5797), 1981.

Nimisha Papineni, Pratik Dutta, Max Chao, Orbin Acanto, Rekha Sathian, Pallavi Surana, and Ramana V Davuluri. Augmenting dnabert embeddings with multimodal dna features for improved regulatory sequence interpretation. In *MLCB*, pp. 294–303, 2025.

John SantaLucia Jr. A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4):1460–1465, 1998.

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

Evan Trop, Yair Schiff, Edgar Mariano Marroquin, Chia Hsiang Kao, Aaron Gokaslan, McKinley Polen, Mingyi Shao, Bernardo P de Almeida, Thomas Pierrot, Yang I Li, et al. The genomics long-range benchmark: advancing dna language models, 2024.

Andrew H-J Wang, Gary J Quigley, Francis J Kolpak, James L Crawford, Jacques H Van Boom, Gijs Van Der Marel, and Alexander Rich. Molecular structure of a left-handed double helical dna fragment at atomic resolution. *Nature*, 282(5740):680–686, 1979.

Minghao Yang, Shichen Zhang, Zhihang Zheng, Pengfei Zhang, Yan Liang, and Shaojun Tang. Employing bimodal representations to predict dna bendability within a self-supervised pre-trained framework. *Nucleic Acids Research*, 52(6):e33–e33, 2024.

Zhao Yang, Hengchang Liu, Chuan Cao, and Bing Su. D3lm: A discrete dna diffusion language model for bidirectional dna understanding and generation. *arXiv preprint arXiv:2603.01780*, 2026.

Daoan Zhang, Weitong Zhang, Bing He, Jianguo Zhang, Chenchen Qin, and Jianhua Yao. Dnagpt: a generalized pretrained tool for multiple dna sequence analysis tasks. *BioRxiv*, pp. 2023–07, 2023.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. Dnashape: a method for the high-throughput prediction of dna structural features on a genomic scale. *Nucleic acids research*, 41(W1):W56–W62, 2013.

Tianyin Zhou, Ning Shen, Lin Yang, Namiko Abe, John Horton, Richard S Mann, Harmen J Bussemaker, Raluca Gordân, and Remo Rohs. Quantitative modeling of transcription factor binding specificities using dna shape. *Proceedings of the National Academy of Sciences*, 112(15): 4654–4659, 2015.

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.

A Local Molecular Descriptor Definitions

This section defines the sequence-derived molecular descriptors used as local features in VANDAM. We specify the 14 DNAshapeR channels used by the local-feature-injection pathway. The mechanism used to incorporate these descriptors into the backbone is defined separately in §3.3. Table 6 summarizes the 14 DNAshapeR local descriptors.

Table 6: The 14 DNAshapeR channels used by VANDAM. Rotational quantities are measured in degrees and translational quantities in Å.

Feature	Full name	Units	Structural category	Biophysical interpretation
MGW	Minor Groove Width	Å	Groove geometry	Width of the minor groove at the center base; narrow grooves deepen the local negative electrostatic potential.
HelT	Helical Twist	degrees	Inter-bp rotation	Rotation about the helix axis between successive base-pair steps.
ProT	Propeller Twist	degrees	Intra-bp rotation	Out-of-plane twist of the two bases within a pair.
Roll	Roll	degrees	Inter-bp rotation	Rotation about the long axis of a base-pair step; a principal component of intrinsic bending.
EP	Electrostatic Potential	kT/e	Groove/electrostatic	Electrostatic potential in the minor groove at the center base.
Stretch	Stretch	Å	Intra-bp translation	Separation of the two bases along the base-pair long axis.
Tilt	Tilt	degrees	Inter-bp rotation	Rotation about the short axis of a base-pair step; the second bending component used with Roll.
Buckle	Buckle	degrees	Intra-bp rotation	Out-of-plane bending of the two bases toward the same groove.
Shear	Shear	Å	Intra-bp translation	Lateral displacement of the two bases along the short axis.
Opening	Opening	degrees	Intra-bp rotation	Opening of the base pair toward the major groove.
Rise	Rise	Å	Inter-bp translation	Distance along the helix axis between successive base pairs.
Shift	Shift	Å	Inter-bp translation	Displacement of the next base pair along the short axis.
Stagger	Stagger	Å	Intra-bp translation	Displacement of the two bases along the long axis relative to each other.
Slide	Slide	Å	Inter-bp translation	Displacement of the next base pair along the long axis.

B Local-Feature-Injection Alignment and Backbone-Specific Handling

This section specifies the visibility rules and backbone adaptations for the input-concatenation injector of §3.3.

Visibility and warm-up. Features are computed from the corrupted sequence available to the backbone, never from the uncorrupted target sequence. For MLM, a masked nucleotide is excluded both at its own position and from the feature windows of neighbouring positions. For NTP, feature windows exclude the target and all future nucleotides. Special tokens receive zero features, and no feature vectors are learned for them. The backbone is frozen while the feature projection W_f and local-feature-injection projection W_m are initialized against its fixed representation; all parameters are then trained jointly.

BPE token alignment. The features operate at nucleotide resolution, whereas a DNABERT-2 BPE token can span multiple bases. We compute features at base resolution and mean-pool them over the bases spanned by each token. Computing features before pooling preserves context that crosses BPE-token boundaries. Special tokens receive zero features.

Causal alignment. For causal HyenaDNA, the centered pentamer $(x_{i-2}, \ldots, x_{i+2})$ cannot be used at position i, because it includes future bases. We instead use the right-aligned window $(x_{i-4}, \ldots, x_i)$, which contains only bases x_j with $j \leq i$. Thus, the adaptation changes the feature's determining window, not the position at which it is injected.

Backbone-specific handling. Caduceus-PS represents forward and reverse-complement strands jointly, so VANDAM applies the same local-feature-injection projection W_m independently to each strand in its natural orientation before restoring the paired representation. This preserves reverse-complement equivariance.

C Regional Biological Feature Definitions

We define the six regional properties of §3.4 for a region $\mathcal{T} = (x_s, \ldots, x_{s+K-1})$ of length K. These properties are computed directly from the nucleotide sequence and summarize regional geometry, composition, duplex stability, and non-B-DNA structural propensity.

Two of the regional properties use additional sequence-derived quantities that are not included among the local molecular features supplied to LFI: the SantaLucia nearest-neighbor duplex free energy and the purine–pyrimidine classification of each nucleotide. These quantities are introduced below only as intermediate inputs to the melting-propensity and Z-DNA-propensity definitions; they are not separate RPP targets.

Nearest-neighbor duplex free energy. For a valid dinucleotide XY, let $\Delta G^{\circ}_{37}(X, Y)$ denote its sequence-dependent nearest-neighbor free-energy contribution at $37\,^{\circ}\mathrm{C}$. We use the SantaLucia values listed in Table 7 (SantaLucia Jr, 1998). Steps related by Watson–Crick reverse-complement symmetry share the same value.

Table 7: SantaLucia nearest-neighbor duplex free energies at $310.15\,\mathrm{K}$ ($37\,^{\circ}\mathrm{C}$) (SantaLucia Jr, 1998).

Step (5'–3')	Free energy (kcal/mol)	Step (5'–3')	Free energy (kcal/mol)
AA / TT	-1.00	CT / AG	-1.28
AT	-0.88	GA / TC	-1.30
TA	-0.58	CG	-2.17
CA / TG	-1.45	GC	-2.24
GT / AC	-1.44	GG / CC	-1.84

Purine–pyrimidine classification. For each valid nucleotide $b \in \{A, C, G, T\}$, define

$$\mathrm{class}(b) = \begin{cases} \text{purine}, & b \in \{A, G\}, \\ \text{pyrimidine}, & b \in \{C, T\}. \end{cases} \tag{6}$$

This classification is used below to identify alternating purine–pyrimidine tracts associated with Z-DNA formation.

Intrinsic curvature (dimensionless). Let ρ_k, τ_k, and θ_k be the raw DNAshapeR Roll, Tilt, and HelT values at position k, converted from degrees to radians. Starting from the identity frame $R_{-1} = I_3$, we propagate

$$Q_k = R_z(\theta_k)R_y(\rho_k)R_x(\tau_k), \qquad R_k = R_{k-1}Q_k, \qquad t_k = R_k e_3, \tag{7}$$

for every $k \in \{0, \ldots, K-1\}$, including positions whose boundary DNAshape values are zero. The curvature score is

$$c(\mathcal{T}) = \frac{1}{K}\left\|\sum_{k=0}^{K-1} t_k\right\|_2. \tag{8}$$

It is the end-to-end/contour-length ratio: values near one indicate aligned tangents, whereas lower values indicate stronger intrinsic curvature. The score captures whether sequence-dependent local bends reinforce coherently across the region (Zhou et al., 2013; Chiu et al., 2016).

AT content (dimensionless fraction). Let N_b denote the number of occurrences of base b, and let $N_{\text{valid}} = N_A + N_C + N_G + N_T$. The implemented score is

$$a(\mathcal{T}) = 1 - \frac{N_C + N_G}{\max(N_{\text{valid}}, 1)}. \tag{9}$$

Non-ACGT positions contribute to neither the numerator nor the valid-base count. This property captures AT-rich regional composition, a component of bendable and low-stability origin-like sequence contexts.

Melting propensity (kcal/mol). Let $d_0 = 0$ and, for $i \geq 1$, define

$$d_i = \begin{cases} \Delta G^\circ_{37}(x_{s+i-1}, x_{s+i}), & x_{s+i-1}, x_{s+i} \in \{A, C, G, T\}, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Thus, $(d_0, \ldots, d_{K-1})$ is a position-aligned track of nearest-neighbor free-energy contributions, with a leading zero because the first nucleotide has no preceding nucleotide within the region. Using the SantaLucia values in Table 7, the regional score is

$$m(\mathcal{T}) = \frac{1}{K-1}\sum_{i=0}^{K-1} d_i = \frac{1}{K-1}\sum_{i=1}^{K-1} d_i. \tag{11}$$

More negative values indicate a more stable duplex, whereas less negative values indicate easier strand separation (SantaLucia Jr, 1998).

G-quadruplex propensity (dimensionless). We follow the standard G4Hunter definition, which caps run scores at 4 and uses a default 25-nt window (Bedrat et al., 2016). For each position i, let ℓ_i be the length of the maximal contiguous run of the same base containing i. Define the per-base G4Hunter track

$$g_i = \begin{cases} \min(\ell_i, 4), & x_{s+i} = G, \\ -\min(\ell_i, 4), & x_{s+i} = C, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where A, T, and non-ACGT positions receive zero and terminate G/C runs. For $K > 0$, set $w = \min(25, K)$ and compute

$$q(\mathcal{T}) = \max_{0 \leq j \leq K-w} \left| \frac{1}{w} \sum_{i=j}^{j+w-1} g_i \right|; \qquad q(\emptyset) = 0. \tag{13}$$

Thus, regions shorter than 25 nucleotides are evaluated using one window spanning the complete region. High values identify clustered G- or C-runs with G-quadruplex potential on either strand (Bedrat et al., 2016).

Cruciform propensity (bp). Let $\overline{x}$ denote the Watson–Crick complement of base x. We define the one-arm stem length

$$\ell_{\mathrm{stem}}(\mathcal{T}) = \max\left\{\ell : \begin{array}{l} \exists\, i < j, \quad 3 \le j - i - 1 \le 12, \\ x_{s+i-r} = \overline{x_{s+j+r}} \quad \text{for all } r = 0, \ldots, \ell - 1 \end{array}\right\}. \tag{14}$$

The indices are restricted to $\mathcal{T}$, and the score has a default value of zero when no admissible stem exists. Both bases in every pair must be valid ACGT; an invalid base fails the complement match and terminates the stem. The score measures one arm rather than the total 2ℓ paired bases. Long inverted-repeat stems can extrude as cruciform or hairpin structures under supercoiling and are associated with recombination and genomic instability (Panayotatos & Wells, 1981; Brázda et al., 2011).

Z-DNA propensity (bp). For $i = 0, \ldots, K-2$, define

$$u_i = \begin{cases} 1, & x_{s+i}, x_{s+i+1} \in \{A, C, G, T\} \text{ and } \mathrm{class}(x_{s+i}) \neq \mathrm{class}(x_{s+i+1}), \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

If $r(u)$ is the length of the longest consecutive run of ones in $(u_0, \ldots, u_{K-2})$, then

$$z_{\mathrm{DNA}}(\mathcal{T}) = r(u) + 1. \tag{16}$$

An invalid base makes both adjacent alternation indicators zero and therefore terminates a tract; a valid homopolymer has score one. The score therefore measures the length of an alternating purine–pyrimidine tract, a sequence class associated with left-handed Z-DNA formation, particularly under negative supercoiling (Wang et al., 1979; Nordheim & Rich, 1983). It is a coarse sequence proxy and does not model sequence-dependent Z-DNA stability.

D RPP Implementation and Backbone Alignment

The main text defines the shared prediction head, RPP loss, and loss schedule. Here we give the remaining implementation details. For each base-defined tile, the regional-property target vector is formed using target-specific transformations and calibration statistics estimated offline from training tiles.

Tiling and pooling. Targets are computed only from bases within the crop; no flanking sequence outside the crop is consulted. Tiles with insufficient valid A/C/G/T content are excluded from the RPP loss. For a tile $\mathcal{T}_j$ of length K, let $N_{\mathrm{valid}}(\mathcal{T}_j)$ be its number of A/C/G/T bases. We define the set of valid tiles as

$$\mathcal{V} = \left\{ j : \frac{N_{\mathrm{valid}}(\mathcal{T}_j)}{K} \ge 0.5 \right\}. \tag{17}$$

Tiles with fewer than 50% valid A/C/G/T bases are excluded from the RPP loss.

For a valid tile, its hidden states are aggregated by scalar-attention pooling,

$$\alpha_i = \frac{\exp(w^\top H_i)}{\sum_{k \in \mathcal{T}_t} \exp(w^\top H_k)}, \qquad z_t = \sum_{i \in \mathcal{T}_t} \alpha_i H_i, \tag{18}$$

where the attention weights are normalized independently within each tile. The pooling scorer is shared across target dimensions.

BPE alignment. For DNABERT-2, tile boundaries and targets remain defined in nucleotide coordinates rather than BPE-token counts. We reconstruct the retained nucleotide sequence by decoding BPE tokens and assign each nonempty token to one tile according to the midpoint of its decoded nucleotide span. A token crossing a tile boundary is assigned wholly to the tile containing its midpoint; it is neither split nor duplicated. Special, padding, and non-nucleotide tokens are excluded from pooling. Consequently, all backbones are supervised against the same type of base-defined regional targets while their native token states are pooled according to their respective tokenizations.

Backbone-specific hyperparameters. Table 8 reports the tile length and $\lambda_{\max}$ used for the reported VANDAM-PT configuration of each backbone. DNABERT-2 uses a 150 bp tile rather than the 300 bp used elsewhere, reflecting its shorter native context length (Table 1). $\lambda_{\max}$ is set per backbone to hold the ratio between the native token loss and $\mathcal{L}_{\mathrm{RPP}}$ approximately constant, rather than fixed to a single value across backbones.

Table 8: RPP tile length and loss-weighting coefficient used for the reported VANDAM-PT configuration of each backbone.

Backbone	Tile length (bp)	$\lambda_{\max}$
NTv3-8M	300	0.5
Caduceus	300	0.5
DNABERT-2	150	1.0
HyenaDNA	300	0.1

E Training Configuration

This section reports the training setup used across backbones. Both regimes continue training from each backbone's released checkpoint. Within each regime, the VANDAM arms use the settings below unchanged and differ only in the biological components under study, so the comparisons in Section 4.1 hold data, optimizer, schedule and number of updates fixed.

Checkpoint sources. We initialize from *InstaDeepAI/NTv3_8M_pre* (NTv3-8M), *kuleshov-group/caduceus-ps_seqlen-131k_d_model-256_n_layer-16* (Caduceus-PS), *zhihan1996/DNABERT-2-117M* (DNABERT-2-117M), and *LongSafari/hyenadna-small-32k-seqlen-hf* (HyenaDNA-small-32k). We use the shorter family names Caduceus and HyenaDNA elsewhere in the paper.

Table 9: Training configuration used across backbones.

Parameter	Value
Initialization	Corresponding released backbone checkpoint (all parameters trainable)
Optimizer steps	54,000
Global batch size	128 sequences
Peak learning rate	6×10^{-4}
Schedule	cosine, 600 warmup steps, decayed to zero at the final step
Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 0.01, gradient clipping 1.0
Precision	bf16

Table 10: Per-backbone compute, used both in Baseline and VANDAM.

Backbone	Nucleotides / tokens seen	Total compute
NTv3-8M	38.221B nucleotides	1.764 EFLOP
Caduceus-PS	56.623B nucleotides	2.623 EFLOP
DNABERT-2-117M	21.234B source nucleotides; 3.539B BPE tokens	2.419 EFLOP
HyenaDNA-small-32k	56.623B nucleotides	1.114 EFLOP

Data and augmentation. Training uses OpenGenome2. NTv3 uses a mixed-length curriculum: sequence lengths of 1,024, 2,048, 4,096 and 8,192 bp are sampled with probabilities 0.1, 0.15, 0.25 and 0.5, giving a weighted mean of 5,530 bp per example. Caduceus and HyenaDNA instead use a fixed 8,192 bp context. DNABERT-2 uses a fixed 3,072-bp crop, tokenized to a fixed 512-BPE-token input including [CLS] and [SEP]; longer tokenizations are truncated and shorter ones padded. Evaluation uses the full 8,192 bp context for NTv3, Caduceus, and HyenaDNA. DNABERT-2 is evaluated on 2,500-bp crops. Two augmentations are applied during training only: a random crop of the target length from each record, and reverse-complementation with probability 0.5. Evaluation is forward-strand.

Native token objectives. NTv3, Caduceus, and DNABERT-2 use standard random per-token masking at a rate of 15%, following the usual 80/10/10 split between mask, random, and unchanged tokens. HyenaDNA retains its native autoregressive next-token prediction objective.

Multi-task schedule. Every tenth optimizer step is a task-only update. We select one task from the pool in Appendix F in deterministic round-robin order and optimize only its supervised task loss; the native sequence objective and, for VANDAM-MT, the RPP losses are not evaluated on these steps. All remaining updates are sequence-only: Baseline-MT optimizes the native sequence objective, whereas VANDAM-MT optimizes the native sequence objective together with the shared RPP losses, with unit weight for each active loss term. Both Baseline-MT and VANDAM-MT include a task-specific supervised head for each supervised pool task; VANDAM-MT additionally includes the shared RPP head, which is updated only on sequence-only batches.

For the first 1,000 steps, the backbone is frozen during task-only updates so that the task heads calibrate against a stable representation; it remains trainable during sequence-only updates. Task batches use the full 8,192 bp context, except for DNABERT-2, which uses its fixed 3,072-bp, 512-BPE-token training configuration. For the two zero-shot variant tasks, we do not use a supervised task head; instead, we score variants using the sequence log-likelihood ratio between the reference and alternate alleles.

F Multi-Task Training Pool and Overlap Control

The multi-task pool contains eight supervised tasks in three families: single-sequence classification, paired variant-effect classification, and zero-shot SNV scoring.

Table 11: The eight-task multi-task training pool. The training tasks are task-disjoint from the evaluation suite; record-level filtering further removes training examples that overlap an evaluation example under the policy below.

Family	Tasks
Single-sequence classification	promoter detection, splice-site detection, enhancer annotation
Variant-effect classification (paired)	common-vs-rare, meQTL, sQTL
Zero-shot SNV scoring	BEND VEP disease, BEND VEP expression

All reported multi-task results use a single strict retained training pool. We compared every candidate training example with every available split of each of the nine evaluation tasks. The nine tasks comprise 16 evaluation task/split partitions, yielding 128 training–evaluation-split comparisons. For coordinate-resolvable records, coordinates were harmonized to GRCh38, and GRCh37/hg19 records were lifted to GRCh38 before comparison. For sequence-only tasks, we assigned a genomic locus only when an exact, full-length alignment was unique. Training examples with ambiguous, unmapped, or unliftable loci were excluded rather than assigned arbitrary coordinates.

The strict filter excluded a training example if it shared with any evaluation example: (i) a normalized allele, (ii) a genomic position irrespective of alternate allele, (iii) an overlapping labelled anchor interval, (iv) an identical model-input sequence, treating reverse complements as equal, (v) complete containment of the shorter labelled or model-input sequence, (vi) overlap between the two loci after expanding each locus by 4,096 bp on each flank, or (vii) overlap between the corresponding model-input windows. Accordingly, for coordinate-resolvable records, retained multi-task training examples are locus/window-disjoint from evaluation examples under these declared rules. Table 12 reports the resulting training pool.

G Evaluation Suite

G.1 Tasks

Table 13 summarizes the nine evaluation tasks.

Table 12: Record-level filtering of the eight-task multi-task training pool. Counts are examples before filtering; the retained pool was used for all reported multi-task results.

Training task	Visible	Excluded	Retained
GUE promoter	47,356	40,183 (84.85%)	7,173
GUE splice site	36,496	18,006 (49.34%)	18,490
LRB enhancer	99,995	28,987 (28.99%)	71,008
Common-vs-rare	193,908	50,620 (26.11%)	143,288
meQTL	73,601	25,108 (34.11%)	48,493
sQTL	1,000,741	385,758 (38.55%)	614,983
BEND VEP disease	295,495	154,346 (52.23%)	141,149
BEND VEP expression	105,263	38,825 (36.88%)	66,438
Total	**1,852,855**	**741,833 (40.04%)**	**1,111,022**

Table 13: The nine-task evaluation suite. Len denotes the task-level sequence window. NTv3, Caduceus, and HyenaDNA use the listed context; DNABERT-2 uses 2,500-bp crops.

Task	Access	Prediction	Metric	Test n	Len (bp)
BRCA1 variant effect	zero-shot	cosine dissimilarity	AUROC	3,893	8,192
Clinical pathogenicity (ClinVar)	zero-shot	cosine dissimilarity	AUROC	258,944	8,192
Disease variants (OMIM)	zero-shot	cosine dissimilarity	AUROC	200,406	8,192
Mendelian variants (TraitGym)	zero-shot	cosine dissimilarity	AUROC	3,380	8,192
Transcription-factor binding	linear probe	probability	AUROC	5,000	100
Regulatory variants (eQTL)	linear probe	probability	AUROC	8,862	8,192
Gene-expression variants	linear probe	probability	AUROC	9,147	1,023
Chromatin state (histone)	linear probe	20 outputs	macro AUROC	227,456	8,192
CAGE expression	linear probe	regression	macro Pearson r	1,922	8,192

Zero-shot tasks. BRCA1 uses saturation genome-editing data, distinguishing functionally impaired from functional single-nucleotide variants (Findlay et al., 2018). ClinVar uses the VEP-eval benchmark and asks for pathogenic versus benign classification (Lu et al., 2025). The OMIM task originates from LRB and asks for pathogenic non-coding regulatory variants against common background variation. Our GFMBench evaluation retains all 406 positives and uses 200,000 background variants, giving 200,406 examples (Trop et al., 2024; Larey et al., 2026a). TraitGym asks for causal non-coding variants for Mendelian traits against matched controls (Benegas et al., 2025b).

Supervised tasks. Transcription-factor binding is binary classification on short sequences, the only task in the suite running at 100 bp (Zhou et al., 2024). The eQTL task predicts whether a variant is causal for nearby gene expression, conditioned on tissue (Trop et al., 2024). Gene-expression variants uses the expression-effect subset of the BioFM Variant Benchmark, derived from DeepSEA regulatory annotations (Medvedev et al., 2025; Zhou & Troyanskaya, 2015). Histone marks follows LRB's 20-track multi-label task over annotated 200 bp bins, derived from the DeepSEA chromatin-feature dataset (Trop et al., 2024; Zhou & Troyanskaya, 2015). CAGE follows LRB's binned-expression task across 50 FANTOM5 tracks in log space, constructed from Basenji data (Trop et al., 2024; Kelley et al., 2018).

G.2 Prediction Protocols

Each task is evaluated by first producing a score or prediction and then applying the appropriate evaluation metric. We describe these two steps separately below.

Zero-shot variant scoring. Reference and alternate sequences are encoded independently using each backbone's native tokenizer. Their final contextualized token states are pooled by an attention-masked mean to obtain sequence representations h_{ref} and h_{alt}. We score each variant by cosine dissimilarity, $s_i = 1 - \cos(h_{\mathrm{alt}}^{(i)}, h_{\mathrm{ref}}^{(i)})$, so that a larger representational perturbation ranks as more likely to be pathogenic. Nothing is fitted on the evaluation task. Although each sequence pair differs at only one nucleotide, the resulting representation change is not necessarily confined to one hidden state. In single-nucleotide models, the perturbation can propagate across contextualized positions

through convolution, attention, or state-space mixing; in DNABERT-2, it may additionally alter the local BPE segmentation. The number and magnitude of affected hidden states are therefore backbone-dependent. Because the reference and alternate sequences share nearly all of their genomic context and are summarized by global mean pooling, however, their sequence-level representations are generally similar.

Linear probing. For supervised tasks, the backbone is frozen and a single linear head is trained for three epochs with AdamW at learning rate 3×10^{-5}, weight decay 0.01, and batch size 16. For binary single-sequence tasks, the head maps a pooled representation to a positive-class probability. For the two paired variant-effect tasks, it consumes $[\, h_{\mathrm{alt}} \,\|\, h_{\mathrm{ref}} \,]$, allowing an asymmetric comparison. The histone-marks head emits twenty independent sigmoid outputs trained with binary cross-entropy, and the CAGE head maps bin-pooled states to predicted $\log 1p$ counts using MSE.

G.3 Evaluation Metrics

AUROC. AUROC is used for the four zero-shot variant tasks and the binary linear-probe tasks. For the former, it evaluates whether cosine dissimilarity ranks pathogenic above benign variants within each backbone; absolute cosine distances are not compared across backbones. For the latter, it is computed from the predicted positive-class probability.

Macro AUROC. For histone marks, AUROC is computed independently for each track and then averaged, excluding tracks with only one class present in the test set. Thus, each track contributes equally to the reported score.

Macro Pearson correlation. For CAGE, all (sequence, bin) pairs are flattened separately for each track, Pearson r is computed per track, and the mean over the fifty tracks is reported. Chance is zero rather than 0.5, which is why this task is never folded into the AUROC mean.

H Paired Statistical Inference for Main Results

This appendix gives the inferential procedure used for Tables 2 and 3. For each task–backbone comparison, we estimate the paired score difference $\Delta = \mathrm{VANDAM} - \mathrm{Baseline}$, reported in the 0–100 units used by the main tables.

For the four zero-shot tasks and the two variant-effect probe tasks (eQTL and gene-expression variants), a test unit is one labeled reference–alternate variant pair, including the tissue covariate for eQTL. For TF binding, it is one labeled test sequence. For histone marks, it is one source sequence with all 20 track labels; for CAGE, it is one source region with all contiguous output bins across the 50 tracks. For all nine tasks, we construct 95% confidence intervals using 10,000 paired bootstrap replicates that resample these test units with replacement, using the same selected units for VANDAM and Baseline. For the five frozen-probe tasks, each replicate additionally resamples the five matched probe seeds with replacement and recomputes the mean paired difference across the selected seeds. For the four zero-shot AUROC tasks, no probe is trained, so bootstrap replicates resample test units only. We use 10,000 paired randomization replicates for one-sided tests, swapping VANDAM and Baseline predictions within each test unit; for frozen-probe tasks, each swap is shared across seeds. Bootstrap resampling and paired randomization swaps preserve these test units intact. Benjamini–Hochberg adjustment is applied separately to the 36 comparisons in each regime and direction. A decision requires the corresponding adjusted one-sided $q < 0.05$ and a 95% paired interval wholly in that direction.

Tables 14 and 15 report the complete results. The q column is the adjusted value for the displayed supported direction; it is omitted when neither direction meets the decision rule.

Table 14: Paired inference for pre-training comparisons. Δ is VANDAM minus Baseline in points; brackets give the paired 95% confidence interval. The q column reports the BH-adjusted one-sided q-value in the direction of the observed Δ. "Supported" denotes the direction meeting the paired-interval and BH-adjusted $q < 0.05$ criteria.

Backbone	**Task**	**Δ [95% CI]**	q	**Supported**
NTv3-8M	ClinVar	$3.6\ [3.3, 4.0]$	< 0.001	VANDAM
NTv3-8M	OMIM	$9.3\ [5.4, 13.1]$	< 0.001	VANDAM
NTv3-8M	TraitGym	$6.1\ [1.4, 10.7]$	0.008	VANDAM
NTv3-8M	BRCA1	$0.6\ [-2.1, 3.3]$	0.425	–
NTv3-8M	TF binding	$1.1\ [0.4, 1.7]$	0.001	VANDAM
NTv3-8M	eQTL	$1.9\ [1.4, 2.3]$	< 0.001	VANDAM
NTv3-8M	Gene expression	$1.5\ [0.9, 2.1]$	< 0.001	VANDAM
NTv3-8M	Histone	$2.1\ [1.9, 2.3]$	< 0.001	VANDAM
NTv3-8M	CAGE	$10.7\ [9.5, 12.0]$	< 0.001	VANDAM
Caduceus	ClinVar	$1.5\ [1.2, 1.8]$	< 0.001	VANDAM
Caduceus	OMIM	$6.3\ [2.3, 10.4]$	0.002	VANDAM
Caduceus	TraitGym	$8.8\ [4.1, 13.4]$	< 0.001	VANDAM
Caduceus	BRCA1	$4.5\ [1.5, 7.4]$	0.003	VANDAM
Caduceus	TF binding	$6.3\ [4.7, 7.9]$	< 0.001	VANDAM
Caduceus	eQTL	$-0.5\ [-0.7, -0.3]$	0.653	–
Caduceus	Gene expression	$-0.6\ [-1.1, -0.0]$	0.282	–
Caduceus	Histone	$5.4\ [5.3, 5.6]$	< 0.001	VANDAM
Caduceus	CAGE	$-1.2\ [-1.5, -0.9]$	0.023	Baseline
DNABERT-2	ClinVar	$6.6\ [6.3, 6.9]$	< 0.001	VANDAM
DNABERT-2	OMIM	$-4.4\ [-7.5, -1.4]$	0.027	Baseline
DNABERT-2	TraitGym	$-0.5\ [-4.0, 3.1]$	1.000	–
DNABERT-2	BRCA1	$0.7\ [-1.7, 3.2]$	0.371	–
DNABERT-2	TF binding	$1.3\ [0.4, 2.2]$	0.003	VANDAM
DNABERT-2	eQTL	$2.8\ [2.1, 3.4]$	< 0.001	VANDAM
DNABERT-2	Gene expression	$1.1\ [0.2, 2.0]$	0.018	VANDAM
DNABERT-2	Histone	$3.8\ [3.6, 4.0]$	< 0.001	VANDAM
DNABERT-2	CAGE	$0.3\ [0.0, 0.6]$	0.044	VANDAM
HyenaDNA	ClinVar	$8.4\ [8.2, 8.6]$	< 0.001	VANDAM
HyenaDNA	OMIM	$-0.1\ [-3.2, 3.1]$	1.000	–
HyenaDNA	TraitGym	$0.2\ [-3.3, 3.8]$	0.557	–
HyenaDNA	BRCA1	$3.4\ [0.9, 6.0]$	0.007	VANDAM
HyenaDNA	TF binding	$6.4\ [5.3, 7.4]$	< 0.001	VANDAM
HyenaDNA	eQTL	$1.1\ [0.7, 1.5]$	< 0.001	VANDAM
HyenaDNA	Gene expression	$0.4\ [-0.2, 1.0]$	0.133	–
HyenaDNA	Histone	$-2.8\ [-3.1, -2.5]$	0.004	Baseline
HyenaDNA	CAGE	$7.0\ [6.2, 7.7]$	< 0.001	VANDAM

Table 15: Paired inference for multi-task comparisons. Δ is VANDAM minus Baseline in points; brackets give the paired 95% confidence interval. The q column reports the BH-adjusted one-sided q-value in the direction of the observed Δ. "Supported" denotes the direction meeting the paired-interval and BH-adjusted $q < 0.05$ criteria.

Backbone	**Task**	**Δ [95% CI]**	q	**Supported**
NTv3-8M	ClinVar	$2.1\ [1.9, 2.4]$	< 0.001	VANDAM
NTv3-8M	OMIM	$2.5\ [-1.0, 6.0]$	0.102	–
NTv3-8M	TraitGym	$5.3\ [0.9, 9.8]$	0.015	VANDAM
NTv3-8M	BRCA1	$2.7\ [-0.2, 5.7]$	0.051	–
NTv3-8M	TF binding	$1.1\ [0.2, 2.0]$	0.010	VANDAM
NTv3-8M	eQTL	$1.2\ [0.7, 1.6]$	< 0.001	VANDAM
NTv3-8M	Gene expression	$0.5\ [-0.3, 1.2]$	0.124	–
NTv3-8M	Histone	$1.0\ [0.9, 1.1]$	< 0.001	VANDAM
NTv3-8M	CAGE	$4.2\ [3.5, 5.0]$	< 0.001	VANDAM
Caduceus	ClinVar	$-3.0\ [-3.3, -2.6]$	< 0.001	Baseline
Caduceus	OMIM	$5.7\ [1.8, 9.7]$	0.004	VANDAM
Caduceus	TraitGym	$7.9\ [3.1, 12.6]$	0.001	VANDAM
Caduceus	BRCA1	$0.7\ [-2.3, 3.8]$	0.343	–
Caduceus	TF binding	$12.7\ [11.6, 13.8]$	< 0.001	VANDAM
Caduceus	eQTL	$2.1\ [1.4, 2.9]$	< 0.001	VANDAM
Caduceus	Gene expression	$2.0\ [0.8, 3.3]$	0.003	VANDAM
Caduceus	Histone	$2.5\ [2.2, 2.9]$	< 0.001	VANDAM
Caduceus	CAGE	$0.9\ [0.3, 1.5]$	0.163	–
DNABERT-2	ClinVar	$10.2\ [9.9, 10.4]$	< 0.001	VANDAM
DNABERT-2	OMIM	$14.4\ [11.2, 17.7]$	< 0.001	VANDAM
DNABERT-2	TraitGym	$2.5\ [-1.3, 6.3]$	0.121	–
DNABERT-2	BRCA1	$3.8\ [1.1, 6.4]$	0.005	VANDAM
DNABERT-2	TF binding	$4.5\ [3.7, 5.3]$	< 0.001	VANDAM
DNABERT-2	eQTL	$3.0\ [2.4, 3.5]$	< 0.001	VANDAM
DNABERT-2	Gene expression	$1.2\ [0.3, 2.1]$	0.007	VANDAM
DNABERT-2	Histone	$5.4\ [5.2, 5.6]$	< 0.001	VANDAM
DNABERT-2	CAGE	$-0.2\ [-0.8, 0.4]$	1.000	–
HyenaDNA	ClinVar	$15.5\ [15.2, 15.8]$	< 0.001	VANDAM
HyenaDNA	OMIM	$3.1\ [-0.5, 6.6]$	0.064	–
HyenaDNA	TraitGym	$3.4\ [-0.9, 7.7]$	0.084	–
HyenaDNA	BRCA1	$1.9\ [-1.1, 4.9]$	0.129	–
HyenaDNA	TF binding	$5.8\ [4.8, 6.8]$	< 0.001	VANDAM
HyenaDNA	eQTL	$0.3\ [-0.1, 0.6]$	0.083	–
HyenaDNA	Gene expression	$-0.9\ [-1.7, -0.1]$	0.286	–
HyenaDNA	Histone	$1.0\ [0.9, 1.2]$	< 0.001	VANDAM
HyenaDNA	CAGE	$10.4\ [9.2, 11.5]$	< 0.001	VANDAM

I Masked Feature Modeling and the Locality Control

MFM predicts the local feature vector $\phi(x)_i$ from the backbone hidden state at each masked index $i \in \mathcal{M}$ using a thin head g_{MFM}:

$$\mathcal{L}_{\mathrm{MFM}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|g_{\mathrm{MFM}}(H_i) - \phi(x)_i\|_2^2 . \tag{19}$$

For MFM (bio), $\phi(x)_i$ is the 15-dimensional local feature vector obtained by concatenating the 14 DNAshapeR descriptors listed in Table 6 with the SantaLucia nearest-neighbor duplex free-energy descriptor defined in Appendix C. Because $\phi(x)_i$ is determined by a $2R + 1$ base window and much of that window remains visible when its central nucleotide is masked, the objective can be optimized without integrating evidence beyond a short local context. We therefore compare it with a random 5-mer control, whose target shape and prediction head are matched but whose values carry no biological content. This is the appropriate MFM control because it isolates the biological content of a *local* target while preserving the positional prediction problem. For RPP, random local targets would not test the regional-integration hypothesis; we instead use algorithmic and shuffled-table regional controls that retain regional prediction while removing the biological sequence-to-property mapping.

Table 16 compares MFM, our local feature-reconstruction objective, with RPP on NTv3-8M. MFM improves mean AUROC by 1.2 points over the random 5-mer control, but the gain is concentrated in the ClinVar task. The control is comparable or better on OMIM, gene-expression variants, and CAGE. Biological RPP reaches 63.8 mean AUROC, exceeding the algorithmic and shuffled-table regional controls by 3.4 and 4.7 points, respectively. Its largest improvements over both controls occur on ClinVar and TraitGym. This pattern is consistent with the locality argument of §3.4: local reconstruction can provide auxiliary regularization, whereas regional supervision is more effective on tasks that benefit from representations integrating evidence across a broader context.

Table 16: Local feature reconstruction versus regional property prediction on NTv3-8M. MFM (bio) uses the composite biological feature provider and MFM (rand) a random 5-mer provider. RPP (alg) predicts generic regional sequence statistics, whereas RPP (shuf) predicts properties recomputed from a shuffled pentamer shape table. Values are AUROC except for CAGE expression, which uses macro Pearson r; mean AUROC averages the eight AUROC tasks. Bold marks the best value in each row.

Task	MFM (bio)	MFM (rand)	RPP (alg)	RPP (shuf)	RPP (bio)
Clinical pathogenicity (ClinVar)	**67.2**	60.7	54.0	53.8	61.7
Disease variants (OMIM)	52.6	53.5	51.9	47.0	**59.0**
Mendelian variants (TraitGym)	49.0	48.0	48.9	47.2	**55.8**
BRCA1 variant effect	51.5	49.8	50.2	48.9	**51.6**
Transcription-factor binding	**81.2**	81.1	78.9	78.7	80.8
Regulatory variants (eQTL)	69.1	68.4	69.0	67.2	**70.2**
Gene-expression variants	62.6	63.1	63.1	63.0	**63.3**
Chromatin state (histone)	67.5	66.7	67.3	67.1	**68.1**
CAGE expression	33.1	33.2	32.8	29.2	**34.9**
Mean AUROC	62.6	61.4	60.4	59.1	**63.8**

J Learned Utilization of the DNAshape Feature Pathway

We inspect the trained local-feature-injection pathway of the NTv3-8M VANDAM-MT checkpoint. This analysis uses 256 OpenGenome2 test sequences (2,097,152 valid A/C/G/T positions).

Let $e_i = \mathbf{e}(x_i) \in \mathbb{R}^D$ be the token embedding and let $f_i = \phi(x)_i$ be the z-scored 14-channel DNAshape vector used in this analysis. The feature projection of §3.3 is $h_i = W_f f_i + b_f \in \mathbb{R}^H$. Writing the learned mixing projection as $W_m = [W_m^{\mathrm{seq}} \mid W_m^{\mathrm{feat}}]$, the branch contributions are $z_i^{\mathrm{seq}} = W_m^{\mathrm{seq}} e_i$ and $z_i^{\mathrm{feat}} = W_m^{\mathrm{feat}} h_i$. To measure position-varying rather than constant contributions, we center each branch over valid positions and report

$$\log_{10} \frac{\left\| z_i^{\mathrm{feat}} - \overline{z}^{\mathrm{feat}} \right\|_2}{\left\| z_i^{\mathrm{seq}} - \overline{z}^{\mathrm{seq}} \right\|_2 + \epsilon}. \tag{20}$$

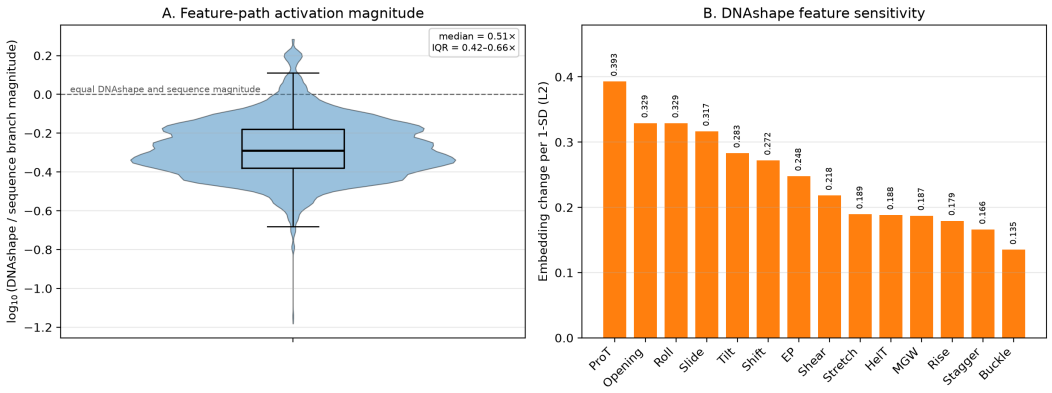


Figure 3: Learned utilization of the DNAshape local-feature-injection pathway. **A:** Distribution across valid positions of the centered DNAshape-branch contribution relative to the centered sequence-embedding contribution after the learned mixing projection. **B:** Mixed-embedding change induced by a one-standard-deviation perturbation of each feature, computed from the composed feature-encoder and mixer map. These analyses describe learned activation and parameter scales, not causal downstream-task importance.

Here $\epsilon = 10^{-8}$ is a numerical stabilizer. We also compose the feature encoder and feature block of the mixing projection, $W_{\text{eff}} = W_m^{\text{feat}} W_f$. For channel f, changing that coordinate by one empirical standard deviation σ_f changes the linear mixed embedding by $\sigma_f W_{\text{eff}}[:, f]$ when the remaining coordinates are fixed; we report the magnitude of this mixed-embedding change induced by a one-standard-deviation feature perturbation, $\sigma_f \|W_{\text{eff}}[:, f]\|_2$.

Figure 3 shows that the DNAshape branch is not suppressed: its centered contribution has median magnitude $0.51\times$ that of the sequence branch (IQR 0.42–$0.66\times$). The composed map also reweights channels non-uniformly: ProT induces the largest mixed-embedding change under a one-standard-deviation feature perturbation (0.393), followed by Opening and Roll (0.329 each), whereas Buckle is smallest (0.135). These are descriptive properties of a single trained checkpoint; they show that the injector retains and selectively transforms DNAshape inputs, but do not establish causal importance for an individual feature or downstream task.

K RPP Target Ablation

We ablate the six RPP targets used by VANDAM-PT on NTv3-8M: intrinsic curvature, AT content, melting propensity, G-quadruplex propensity, cruciform propensity, and Z-DNA propensity. All arms use the same 54,000-step continual-pre-training protocol, native MLM objective, 300-bp tiles, scalar-attention pooling, and total RPP coefficient ($\lambda_{\max} = 0.5$). Each ablation removes the indicated target or target group while retaining the remaining targets. Table 17 shows that the effects are task-dependent: no target removal lowers every score.

Table 17: NTv3-8M VANDAM-PT ablation of RPP targets. "− folding" removes G-quadruplex, cruciform, and Z-DNA propensities; "− origin" removes AT content, melting propensity, and intrinsic curvature; "− curvature" removes intrinsic curvature. For supervised tasks, all columns use the same fixed probe seed, so differences isolate the effect of removing each RPP target. Zero-shot scores are seed-independent.

Task	Full RPP	− folding	− origin	− curvature
Transcription-factor binding	80.91	80.44	80.50	80.72
Regulatory variants (eQTL)	70.20	68.54	68.56	68.72
Gene-expression variants	63.44	63.77	63.09	63.33
BRCA1 variant effect	51.55	52.49	53.02	48.87
Disease variants (OMIM)	59.00	49.72	53.28	49.51
Mendelian variants (TraitGym)	55.78	56.83	56.67	52.16
Clinical pathogenicity (ClinVar)	61.66	60.27	62.68	65.95
Chromatin state (histone)	68.04	66.94	66.91	67.10
CAGE expression	34.80	34.12	32.90	33.68

The effects are task-dependent: no target removal lowers every score. Nevertheless, removing any target group decreases eQTL, histone, and CAGE performance, while the largest declines occur on OMIM when folding or curvature supervision is removed (59.00 to 49.72 and 49.51, respectively). These results indicate that the regional properties provide complementary signals rather than identifying a single target that is uniformly responsible for the downstream gains.

L Intrinsic Token-Objective Evaluation

We evaluate each checkpoint on the held-out OpenGenome2 test shards using its backbone's native token objective. These metrics complement the downstream evaluations: they ask whether VANDAM changes native-token performance, rather than whether its representations support the held-out functional tasks.

NTv3-8M, Caduceus, and HyenaDNA are evaluated on 3,733 deterministic non-overlapping 8,192-bp windows; DNABERT-2 is evaluated on 10,378 2,500-bp windows. For MLM backbones, we use standard random 15% masking with a fixed seed. Loss and accuracy are computed over masked tokens. For causal HyenaDNA, they are computed over all scored next-token positions. We report mean cross-entropy loss and top-1 accuracy. Tables 18 and 19 show that VANDAM's downstream gains do not require uniform improvement in native-token performance.

Table 18: Intrinsic test metrics after continual pre-training. Each cell gives Baseline-PT → VANDAM-PT.

Backbone	Objective	Loss ↓	Accuracy ↑
NTv3-8M	MLM	$1.098 \rightarrow 1.110$	$50.6\% \rightarrow 50.2\%$
Caduceus	MLM	$1.095 \rightarrow 1.109$	$51.1\% \rightarrow 50.2\%$
DNABERT-2	MLM (BPE)	$5.060 \rightarrow 5.058$	$16.4\% \rightarrow 16.4\%$
HyenaDNA	Causal NTP	$1.265 \rightarrow 1.250$	$40.2\% \rightarrow 41.3\%$

Table 19: Intrinsic test metrics after multi-task training. Each cell gives Baseline-MT → VANDAM-MT.

Backbone	Objective	Loss ↓	Accuracy ↑
NTv3-8M	MLM	$1.136 \rightarrow 1.118$	$48.6\% \rightarrow 49.7\%$
Caduceus	MLM	$1.207 \rightarrow 1.195$	$44.0\% \rightarrow 44.8\%$
DNABERT-2	MLM (BPE)	$5.007 \rightarrow 5.673$	$16.8\% \rightarrow 14.3\%$
HyenaDNA	Causal NTP	$1.267 \rightarrow 1.252$	$40.1\% \rightarrow 41.1\%$

M BIOLOGICAL-CONTENT CONTROLS

This section specifies the matched controls used in Figure 2. Each control uses the same backbone, data, and training budget as its biological comparison; the intervention is the content of the injected features or RPP targets.

Local-feature-injection control. For the LFI (rand) control, the biological local descriptor vector $\phi(x)_i$ is replaced by a random vector of the same dimension at every position. This leaves the injector's parameter count and gradient path unchanged while removing the sequence-derived molecular signal.

RPP target controls. The shuffled-table control applies one fixed permutation to the rows of the pentamer shape table and recomputes the same regional properties through the same estimators. It therefore preserves the targets' marginal distributions, cross-feature correlations, and order sensitivity, while destroying the pentamer-to-shape correspondence. The algorithmic control instead predicts three generic regional sequence statistics with no molecular interpretation: regional compression ratio, third-order entropy rate, and minimizer density. Two of these depend on base arrangement rather than composition alone, preserving the regional integration requirement of RPP while removing the biological sequence-to-property mapping.

Multi-task configurations. LFI (rand) + RPP retains biological RPP while randomizing the injected descriptors. LFI (bio) + RPP (shuf) and LFI (bio) + RPP (alg) retain biological injected descriptors while replacing only the RPP targets. Thus, each multi-task comparison isolates the content of one biological pathway while holding the remaining architecture and training setup fixed.

N QDNA REPRESENTATION PROBING

This analysis tests whether the RPP objective changes the linear accessibility of a molecular descriptor family absent from its supervision. We compare three frozen NTv3-8M checkpoints: the original released checkpoint (Baseline), Baseline-PT, and VANDAM-PT.

Targets and representations. qDNA combines a tight-binding description of sequence-dependent electronic couplings with open-quantum-system dynamics to characterize how charge and excitation behave in DNA (Herb et al., 2025), using electronic parameters from Hawke et al. (2010). From qDNA simulations, we derive eight molecular descriptors of hole and electron localization and transfer, exciton decay, and electron–hole separation. They are distinct from the geometry, melting, and non-B-structure properties used to train VANDAM-PT, and no qDNA descriptor enters the RPP loss. At each nucleotide position, the centered heptamer indexes the corresponding qDNA values. For descriptor j, the qDNA target for a 300-bp tile t is

$$y_{t,j} = \frac{1}{|V_t|} \sum_{i \in V_t} Q_j(x_{i-3}, \ldots, x_{i+3}), \tag{21}$$

where V_t contains positions with complete canonical centered heptamers.

Residual targets. To isolate qDNA variation not linearly predicted by the six regional properties used as RPP targets (Section 3.4), we form a six-dimensional nuisance vector z_t from intrinsic curvature, AT content, melting propensity, G-quadruplex propensity, cruciform propensity, and Z-DNA propensity. We standardize these variables using development data only and, separately for each qDNA descriptor, define

$$(\hat{\alpha}_j, \hat{\beta}_j) = \arg\min_{\alpha,\beta} \sum_{t \in \mathcal{D}_{\mathrm{dev}}} \left(y_{t,j} - \alpha - \beta^\top z_t\right)^2, \qquad y^{\mathrm{res}}_{t,j} = y_{t,j} - \hat{\alpha}_j - \hat{\beta}_j^\top z_t. \tag{22}$$

The fitted standardization and OLS model are applied unchanged to discovery and confirmation data. Thus, residualization removes only the component explained by this linear six-property basis; it does not imply independence from all sequence properties.

The descriptor set comprises central population, maximum transfer, and inverse participation ratio (IPR) separately for holes and electrons, together with exciton lifetime and mean electron–hole separation. Central population measures the time-averaged population retained at the initial base; maximum transfer is the largest time-averaged transfer away from it; and IPR measures state delocalization. Exciton lifetime is the time to decay to $1/e$, while electron–hole separation is their time-averaged distance.

We sample 400 OpenGenome2 validation sequences of length 8,192 and assign entire source sequences to development (280), discovery (60), or confirmation (60), so tiles from a source cannot cross splits. Each sequence yields 27 non-overlapping 300-bp tiles, giving 7,560 development and 1,620 tiles in each held-out phase before any invalid-tile removal. Every complete sequence is encoded by each frozen backbone, and the final hidden states within a tile are mean-pooled to form its representation.

Probes and inference. For every checkpoint–target pair, we fit a ridge-regression probe on development tiles. The penalty is selected from $\{0.1, 1, 10, 100, 1000, 10000\}$ using five-fold grouped cross-validation, with folds defined by source sequence and representation standardization fitted within each training fold. The selected probe is refitted on the full development split and evaluated without refitting on discovery and confirmation.

The primary effect is $\Delta R^2 = R^2_{\text{VANDAM-PT}} - R^2_{\text{Baseline-PT}}$. Confidence intervals use 2,000 paired bootstrap replicates over held-out source sequences, retaining all tiles from each sampled source. One-sided p-values use 5,000 source-cluster permutations that swap the two checkpoints' fixed predictions within a random subset of sources. Benjamini–Hochberg correction is applied separately over the eight-descriptor family for the original and residual-target analyses. A target enters confirmation only when its discovery $q \leq 0.05$ and $\Delta R^2 \geq 0.02$; final selection requires the same two criteria on the independent confirmation split.

Complete results. Table 20 gives confirmation performance for all eight reported qDNA descriptors and both target definitions. The same six descriptors meet both selection stages in the original and residual-target analyses. Their residual-target effects range from 0.051 to 0.091 absolute R^2 (mean 0.069), while both maximum-transfer channels remain exceptions.

Table 20: Frozen linear probing of qDNA descriptors at 300 bp on the independent confirmation split. Checkpoint R^2 values use the original targets. Both ΔR^2 columns compare VANDAM-PT with Baseline-PT; intervals are paired source-sequence-bootstrap 95% confidence intervals.

Descriptor	Baseline R^2	Baseline-PT R^2	VANDAM-PT R^2	Original ΔR^2 [95% CI]	Residual ΔR^2 [95% CI]
Hole central population	0.717	0.709	0.769	0.060 [0.023, 0.104]	0.054 [0.027, 0.083]
Hole maximum transfer	0.300	0.310	0.268	−0.041 [−0.113, 0.006]	−0.032 [−0.078, 0.007]
Hole delocalization (IPR)	0.766	0.742	0.812	0.071 [0.040, 0.106]	0.091 [0.063, 0.122]
Electron central population	0.671	0.663	0.746	0.083 [0.050, 0.122]	0.057 [0.027, 0.089]
Electron maximum transfer	0.343	0.353	0.298	−0.055 [−0.186, 0.025]	−0.046 [−0.194, 0.041]
Electron delocalization (IPR)	0.772	0.771	0.838	0.067 [0.047, 0.090]	0.051 [0.021, 0.082]
Exciton lifetime	0.742	0.721	0.794	0.073 [0.053, 0.095]	0.084 [0.053, 0.121]
Electron–hole separation	0.786	0.758	0.826	0.069 [0.049, 0.091]	0.076 [0.052, 0.109]

O Sensitivity to Masked-Base Handling in LFI

The reported LFI policy sets a feature vector to zero whenever its determining pentamer contains an explicit $\langle\text{mask}\rangle$ token, preventing the feature stream from revealing a masked nucleotide. Under the 15% MLM masking rate and standard 80/10/10 replacement policy, a nucleotide is replaced by $\langle\text{mask}\rangle$ with probability $0.15 \times 0.8 = 0.12$. An interior pentamer therefore contains at least one mask with probability $1 - (1 - 0.12)^5 \approx 47.2\%$, making the injected feature stream sparse during MLM training.

We evaluated two leakage-safe alternatives on NTv3-8M VANDAM-MT. Both construct a separate completed sequence used only to compute DNAshape features; the backbone still receives the original corrupted sequence. *Uniform* replaces each masked base with A, C, G, or T with equal probability. *Conditional* samples it from the OpenGenome2 pentamer distribution conditioned on the visible

pentamer context, with pseudocount smoothing. Replacements are resampled during training and applied simultaneously, so adjacent masks introduce no ordering effect. A single replacement can restore features for up to five overlapping pentamers.

Table 21: Sensitivity of NTv3-8M VANDAM-MT to masked-base handling. Scores are AUROC except CAGE expression, which uses macro Pearson r.

Handling	ClinVar	OMIM	TraitGym	BRCA1	TF	eQTL	Gene expression	Histone	CAGE
Zeroing (VANDAM)	70.2	57.4	**56.8**	54.6	**79.8**	69.3	63.1	69.3	**35.9**
Uniform	68.6	**58.1**	56.1	59.4	78.9	69.8	**70.7**	71.0	33.3
Conditional	**70.8**	55.9	54.3	**61.2**	78.6	**70.1**	70.0	**71.4**	33.5

Effects vary by task: imputation improves BRCA1, gene-expression variants, and histone prediction, while both alternatives reduce TF-binding and CAGE performance. Conditional sampling does not consistently outperform uniform sampling. Because this sensitivity analysis is limited to NTv3-8M, we retain zeroing as the reported multi-backbone policy.