UGM: Unified Multi-scale Genomic Event Modeling with Site-Level Joint Prediction

Jiayang Wu[1,2]* **Chenchen Qin**[2]* **Xu Yang**[2] **Yu Zhao**[2]
Bing He[2] **Jiale Zhou**[1,2] **Zhenchao Tang**[2] **Minghao Yang**[2]
Shouzhi Chen[2] **Yefeng Zheng**[1]† **Jianhua Yao**[2]†

[1]School of Engineering, Westlake University, Hangzhou, China
[2]AI for Life Sciences Lab, Tencent, Shenzhen, China
zhengyefeng@westlake.edu.cn jianhuayao@tencent.com

Abstract

Genomic foundation models learn powerful representations from DNA sequences, but reference sequence alone does not explicitly describe methylation states, reference-relative variants, or structural-variant boundaries. Incorporating these observations into pretraining is challenging because they differ in representation and spatial extent, while event-bearing positions are sparse relative to ordinary reference bases. To address these challenges, we introduce the **Unified Genomic Model (UGM)**, an encoder-only genomic foundation model that represents heterogeneous genomic observations as coordinate-aligned event-state tokens and learns to recover them from their surrounding context. UGM introduces three components: (i) a unified event tokenizer that maps reference bases, SNPs, CpG methylation states, short INDELs, and structural-variant boundaries into a shared vocabulary of 43 genomic event-state tokens; (ii) event-balanced pretraining, which increases training exposure to rare but biologically important events; and (iii) Site-Level Joint Prediction (SJP), a masked-prediction objective that recovers the resolved event state at each masked position from its surrounding sequence and event context. We evaluate UGM against specialized and generalist baselines across functional variant classification, regulatory benchmarks (NT and GUE), and structural-variant detection. It achieves the highest average performance across the functional variant classification tasks, demonstrating consistently strong results across diverse evaluation settings. UGM also achieves competitive splicing prediction on NT and outperforms DNA-only baselines on SV breakpoint detection, while regulatory results show task-dependent advantages. These results support coordinate-aligned event-state pretraining as an approach to extending genomic representation learning beyond reference sequence alone.

1 Introduction

Genomic foundation models learn powerful representations from DNA sequences, but reference sequence alone does not explicitly describe methylation states, reference-relative variants, or structural-variant (SV) boundaries. These observations provide complementary information about genomic state, from changes at individual nucleotides to modifications at specific sites and alterations spanning genomic intervals. For example, a SNP and a nearby methylated CpG describe distinct observations at neighboring genomic coordinates. Modeling their relationship requires access to both event identities and the surrounding sequence context. This motivates a central question: how can genomic

*Equal contribution.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

pretraining learn reference sequence together with the heterogeneous events associated with its coordinates?

Existing approaches incorporate genomic information through sequence pretraining, functional supervision, and event-specific representations. DNABERT-2 Zhou et al. [2024], Nucleotide Transformer Dalla-Torre et al. [2025], and HyenaDNA Nguyen et al. [2023] learn nucleotide-sequence representations, while Enformer Avsec et al. [2021], SPACE Yang et al. [2025b], and Borzoi Linder et al. [2025] learn functional signals from sequence. Methylation models Jeong et al. [2025], Ni et al. [2019], variant-effect predictors Benegas et al. [2025], Shearer et al. [2025], and SV callers Lin et al. [2022], Popic et al. [2023] further address particular event types. These approaches support effective prediction within their respective settings, but differ in how event identity is exposed to the model and incorporated into training. They leave open how heterogeneous events can become explicit prediction targets within a shared genomic pretraining framework.

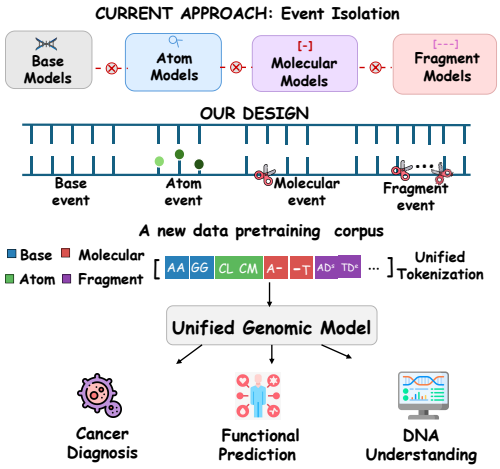


Figure 1: Overview of UGM-Bench.

Two challenges arise when extending pretraining beyond reference sequence. First, **representational heterogeneity**: substitutions describe reference-to-alternate changes, methylation provides continuous measurements at specific sites, and structural variants span intervals with distinct boundaries. A shared representation must preserve genomic locations and event identities while accommodating these differences in spatial extent. Second, **sparse training exposure**: most genomic positions retain ordinary reference states, whereas event-bearing positions are comparatively sparse and unevenly distributed. Merely including event records can therefore leave masked pretraining dominated by reference-base reconstruction, providing limited training exposure to rare event states.

We address these challenges by treating **coordinate-aligned event states as the units of genomic pretraining**. By making event identities explicit prediction targets, a shared encoder can learn to recover these states from their surrounding sequence and event context. This formulation motivates learning contextualized event representations that can transfer across tasks involving variation, methylation, and structural changes.

We introduce **UGM**, an encoder-only genomic foundation model built around three components. First, **unified event tokenization** maps reference bases, SNPs, CpG methylation states, short insertions and deletions (INDELs), and SV boundaries into a shared vocabulary of 43 genomic event-state tokens. Events are aligned to genomic coordinates, with sparse boundary tokens representing structural variants without dense token expansion. Second, **event-balanced pretraining** increases training exposure to rare event categories while preserving their surrounding genomic context. Third, **Site-Level Joint Prediction (SJP)** recovers the resolved event state at each masked coordinate through a shared prediction head. Together, these components connect a common event representation with a pretraining procedure that explicitly learns event identities in context.

We evaluate UGM against specialized and generalist baselines across functional variant classification, regulatory benchmarks (NT and GUE), and SV detection. UGM achieves the highest average performance across seven functional variant classification tasks, competitive splicing prediction on NT, and improved SV breakpoint detection over the DNA-only baseline. Its regulatory results show task-dependent advantages, indicating that the benefits of event-state pretraining vary across prediction settings. Component ablations examine the contributions of event-balanced pretraining and SJP, while attention visualizations provide a qualitative view of learned dependencies between nearby events. Our contributions are summarized as follows:

- **Coordinate-aligned event-state representation.** We represent reference bases and heterogeneous genomic events as native tokens in a shared vocabulary, making event identities explicit targets of genomic pretraining.

- **A shared framework for event-state pretraining.** We develop UGM with event-balanced sampling and Site-Level Joint Prediction to increase training exposure to sparse events and recover masked event states using a single encoder.

- **Evaluation across genomic prediction settings.** We evaluate UGM on functional variant, regulatory, and structural-variant tasks, and analyze component contributions and task-dependent performance to characterize the benefits and scope of event-state pretraining.

2 Related Work and Background

2.1 Genomic Foundation Models

DNA foundation models learn representations from large-scale genomic sequences and have become strong baselines for regulatory and variant-effect tasks. DNABERT-2 Zhou et al. [2024] uses byte-pair tokenization and efficient attention for multi-species genomes; Nucleotide Transformer Dalla-Torre et al. [2025] scales masked language modeling to billion-parameter models; HyenaDNA Nguyen et al. [2023] extends context length through long convolutions. Recent genome-scale and regulatory models such as Evo2 Brixi et al. [2026], Borzoi Linder et al. [2025], Enformer Avsec et al. [2021], SPACE Yang et al. [2025b], and AlphaGenome Avsec et al. [2026] further improve long-context sequence modeling, gene expression prediction, or chromatin-related track prediction. These models can be adapted to variant tasks through altered sequences, likelihood ratios, embeddings, or downstream heads. However, none consume multi-scale event vocabularies that include methylation, INDELs, and SV boundaries as first-class tokens.

2.2 Event-specific Genomic Models

Several model families specialize in particular genomic event types. Methylation models such as MethylBERT Jeong et al. [2025] and DeepSignal Ni et al. [2019] focus on read-level or signal-level epigenetic states. Evolutionary models such as GPN-MSA Benegas et al. [2025] use multiple sequence alignments to encode conservation signals for variant-effect prediction. INDEL-specialized models such as LOL-EVE Shearer et al. [2025] target short insertion/deletion effects, while SV tools such as SVision Lin et al. [2022], Cue Popic et al. [2023], and SVLearn Yang et al. [2025a] focus on structural-variant discovery or genotyping. These systems are valuable but are evaluated with different metrics, input formats, and task definitions, making cross-model comparison difficult. UGM unifies these event types within a single modeling framework.

2.3 Existing Genomic Benchmarks

Benchmarking resources have improved evaluation of genomic foundation models, but all share a common limitation, as they evaluate models primarily on DNA sequence tasks. Genomic Benchmarks Grešová et al. [2023] provides sequence-classification tasks across species and regulatory elements, but does not include methylation or variant states. BEND Marin et al. [2024] evaluates biologically meaningful DNA language model tasks, yet focuses on reference-sequence properties. GV-Rep Li et al. [2024] focuses on genetic variant representation learning, but tests only short variants without methylation or SVs. GenBench Liu et al. [2024] provides systematic evaluation for genomic foundation models, but still treats the genome as a nucleotide string. UGM addresses this gap by learning, and being evaluated on, a unified multi-scale event vocabulary. Across these lines of work, the central limitation is not model capacity alone but the representation of genomic state. Reference-only language models can incorporate variants by editing input sequences, and event-specific tools can achieve strong performance within their target modality. However, these approaches usually treat sequence, methylation, short variants, and structural variants as separate inputs or task-specific labels. UGM instead represents these signals as native states in a shared coordinate-aligned vocabulary. This design preserves the genomic locus as the unit of prediction and allows the model to learn dependencies among bases, epigenetic marks, and variant events within a single pretraining objective.

3 Method

UGM learns genomic representations by recovering coordinate-aligned event states from their surrounding context. Its pretraining framework combines three components: unified event tokenization specifies how heterogeneous observations are represented; event-balanced pretraining increases training exposure to sparse events; and Site-Level Joint Prediction (SJP) makes event states explicit prediction targets. A shared encoder processes the resulting token sequences and learns contextualized representations for downstream prediction. Figure 2 summarizes the framework.

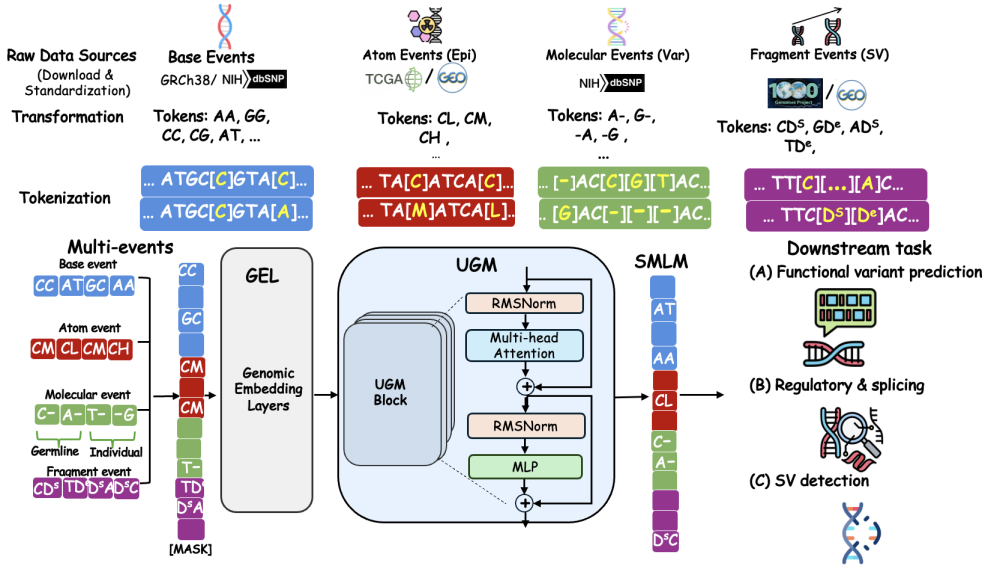


Figure 2: The UGM framework. Heterogeneous genomic events are standardized from raw sources, transformed into a unified vocabulary of 43 genomic event-state tokens, processed by genomic embedding layers and UGM blocks, and evaluated on functional variant prediction, regulatory and splicing tasks, and structural-variant detection.

3.1 Unified Event Tokenization

Genomic observations differ in their attributes and spatial extent, but can be organized using a common genomic coordinate system. UGM uses GRCh38 as the reference backbone and maps reference bases, substitutions, CpG methylation states, short INDELs, and structural-variant boundaries into a shared vocabulary $\mathcal{V}$ of 43 genomic event-state tokens. Special control tokens, such as `[MASK]` and, where applicable, `[PAD]`, are excluded from this count.

Coordinate alignment and state resolution. Preprocessing first locks individual-specific records to a single donor, normalizes all events to GRCh38, and collects candidate states at each coordinate. A deterministic coordinate-state resolver emits one token per represented coordinate: SV boundary and short-INDEL tokens take precedence at their breakpoints, SNP tokens replace the reference base when the allele changes, methylation tokens are retained only for compatible CpG cytosines without an incompatible variant at the same coordinate, and otherwise the reference-base token is used. Exact coordinate conflicts are resolved before tokenization, while neighboring observations, such as a SNP near a methylated CpG, remain available through their genomic context.

Let i denote a sequence index and (c_i, p_i) its chromosome and genomic coordinate. The resolved state is mapped to a token $x_i \in \mathcal{V}$, yielding a sequence

$$X = \big((x_i, c_i, p_i)\big)_{i=1}^{L}. \tag{1}$$

The vocabulary contains four event categories:

Base events (16 tokens). Reference bases and single-nucleotide substitutions are encoded as reference–alternate pairs over `A,C,G,T`. This yields 12 substitution tokens (`AG`, `AC`, `AT`, `GA`, `GC`, `GT`, `CA`, `CG`, `CT`, `TA`, `TG`, `TC`) plus four matched reference–alternate tokens (`AA`, `CC`, `GG`, `TT`), where identical reference and alternate bases denote unaltered genomic positions.

Atom events (3 tokens). CpG methylation states are encoded by discretizing β-values into low ($\beta < 0.2$), medium ($0.2 \leq \beta < 0.6$), and high ($\beta \geq 0.6$) bins, forming tokens `CL`, `CM`, and `CH`. This converts continuous methylation measurements into categorical states within the shared prediction space.

Molecular events (8 tokens). Short deletions are encoded as `A-`, `C-`, `G-`, `T-` and insertions as `-A`, `-C`, `-G`, `-T`, capturing both the reference base and the indel operation.

Fragment events (16 tokens). Large structural variants (≥ 50 bp) are represented through paired boundary tokens rather than enumerating every internal base. Deletion boundaries are encoded as $\texttt{AD}^s$, $\texttt{CD}^s$, $\texttt{GD}^s$, $\texttt{TD}^s$ (start) and $\texttt{AD}^e$, $\texttt{CD}^e$, $\texttt{GD}^e$, $\texttt{TD}^e$ (end); insertion boundaries use $\texttt{DA}^s$, $\texttt{DC}^s$, $\texttt{DG}^s$, $\texttt{DT}^s$ and their end counterparts. This sparse representation preserves breakpoint coordinates and flanking sequence while avoiding dense token expansion.

Together, these categories express heterogeneous observations within a common token space while retaining their coordinate alignment and encoded event attributes.

3.2 Event-Balanced Pretraining

A shared vocabulary makes heterogeneous events representable, but does not ensure that they receive sufficient training exposure. Ordinary reference positions substantially outnumber methylation, INDEL, and SV-boundary states. Sampling without accounting for this imbalance can therefore produce a masked-prediction task dominated by reference-base reconstruction. UGM uses event-balanced sequence construction to increase exposure to sparse event categories while retaining their surrounding genomic context. Each pretraining instance is a contiguous genomic segment of 262,144 positions sampled with target proportions: 40% base events, 25% atom events, 25% molecular events, and 10% fragment events. Individual-specific event streams are not mixed across donors, as personalized methylation and variant records are extracted from a single sample when available, while population-level variants are sampled from harmonized catalogs with coordinate normalization. Event balancing changes the composition of training examples; the prediction objective remains masked cross-entropy without explicit event-specific loss weighting. The complete multi-scale sampling and integration procedure is formalized in Algorithm 1.

Algorithm 1 UGM Multi-scale Sampling and Integration

Require: $L = 262,144$; Proportions $P = \{p_{base}, p_{atom}, p_{mol}, p_{frag}\}$; $X \leftarrow \emptyset$; $\mathcal{S} \leftarrow$ {base, atom, mol, frag}
Data Sources: $\mathcal{D}_{base}$: {Ref$_{GRCh38}$}; $\mathcal{D}_{atom}$: {TCGA, GEO}; $\mathcal{D}_{mol}$: {dbSNP, TCGA, 1000G}; $\mathcal{D}_{frag}$: {TCGA, 1000G}
Ensure: Integrated multi-scale sequence X
$f \leftarrow$ SelectDonor($\mathcal{D}_{base} \cup \mathcal{D}_{atom} \cup \mathcal{D}_{mol} \cup \mathcal{D}_{frag}$) (Lock donor based on event sources)
for each scale $s \in \mathcal{S}$ **do**
  $L_s \leftarrow L \times p_s$
  if $s =$ molecular **then**
    $E_{pop} \leftarrow$ Sample(dbSNP, $L_s/2$)
    $E_{indiv} \leftarrow$ Sample($\mathcal{D}_{mol} \setminus$ {dbSNP}, f, $L_s/2$)
    $E_s \leftarrow E_{pop} \cup E_{indiv}$
  else
    $E_s \leftarrow$ Sample($\mathcal{D}_s$, f, L_s)
  end if
  $C_s \leftarrow$ CandidateStates(E_s) (Map records to coordinate-level candidate states)
  $X \leftarrow X \cup C_s$
end for
$X \leftarrow$ ResolveCoordinateStates(X) (SV/INDEL $>$ SNP $>$ compatible CpG methylation $>$ reference)
SortByCoordinate(X)
return X

GRCh38 supplies the reference backbone. Methylation records are drawn from GEO and TCGA-derived resources, while dbSNP, 1000 Genomes, and donor-specific records provide short and structural variants. Selected records are aligned to genomic coordinates and processed by the state resolver described above. Corpus statistics and source details are provided in the appendix. Table 1 summarizes the event sources and biological levels.

Table 1: UGM event sources and biological levels.

Event Scale	Representative Source	Biological Level
Base	GRCh38	Reference
Atom	GEO / TCGA-derived records	Individual sample
Molecular	dbSNP / 1000G / donor variants	Population / individual
Fragment	dbSNP / 1000G / donor SV records	Population / individual

3.3 Site-Level Joint Prediction

SJP trains UGM to recover the resolved event-state token at each masked coordinate. Reference bases, substitutions, methylation states, INDEL operations, and SV boundaries share a prediction head and output vocabulary. Consequently, event identity and the attributes encoded by each token become explicit targets of pretraining. Given an event-state sequence X, a set of positions M is selected for masking to construct X_{masked}. At each selected position i, the model predicts the original token $x_i^* \in \mathcal{V}$ from the available sequence and event context:

$$p_\theta(x_i^* \mid X_{\text{masked}}), \quad x_i^* \in \mathcal{V}. \tag{2}$$

The target is the resolved state represented by the tokenizer. A substitution target identifies a reference–alternate pair, while a methylation target identifies a discretized CpG state. Nearby unmasked observations provide context for recovering either target. A shared linear head maps the contextual representation H_i to a distribution over the 43 event-state tokens:

$$p_\theta(x_i \mid X_{\text{masked}}) = \text{softmax}(H_i W + b). \tag{3}$$

Special control tokens are excluded from the prediction targets. UGM minimizes cross-entropy over the selected positions:

$$\mathcal{L}_{\text{SJP}} = -\frac{1}{|M|} \sum_{i \in M} \log p_\theta(x_i^* \mid X_{\text{masked}}). \tag{4}$$

This objective places heterogeneous event states within a common prediction task, allowing the encoder to learn statistical dependencies between event identities and their surrounding context.

3.4 UGM Architecture

UGM is an encoder-only Transformer that processes coordinate-aligned event-state sequences and special control tokens. Each input representation combines token, chromosome, and genomic-position embeddings:

$$E_i = E_{\text{token}}(x_i) + E_{\text{chr}}(c_i) + E_{\text{pos}}(p_i). \tag{5}$$

Token embeddings specify the encoded event state, while chromosome and position embeddings provide its genomic location. The resulting representations are processed by stacked Transformer blocks with RMSNorm Zhang and Sennrich [2019], rotary positional embeddings Su et al. [2024], and FlashAttention-2 Dao [2024]. We use sliding-window attention with default window size $w = 512$:

$$\text{Attn}(Q, K, V) = \text{FlashAttn}(Q, K, V, \mathcal{M}_{\text{win}}), \tag{6}$$

where $\mathcal{M}_{\text{win}}[i, j] = 1$ if $|i - j| \leq w$ and $-\infty$ otherwise.

The final hidden states are passed to the shared SJP head during pretraining and used as contextualized representations for downstream fine-tuning.

4 Experiments

4.1 Experimental Setup

Evaluation organization. We structure the evaluation to directly test the design choices in UGM. The main text covers six experimental subsections, including (1) functional variant classification, which tests whether native event tokens improve SNP- and variant-level prediction; (2) regulatory benchmarks (NT and GUE), which assess transfer to chromatin, splicing, and TF-binding tasks; (3) SV detection, which evaluates fragment-scale boundary tokens on breakpoint identification; (4) component ablations, which isolate event-balanced pretraining and Site-Level Joint Prediction (SJP); (5) cross-event attention analysis, which visualizes learned multi-scale interactions; and (6) model scaling, which examines capacity effects within the UGM family. Additional results on Genomic Benchmarks and context-window ablations are provided in Appendix A.9. Detailed dataset descriptions for each subsection are provided in Appendix A.2.

Fine-tuning protocol. All tasks use GRCh38 coordinates and identical train/validation/test splits across models. For downstream classification, each backbone is paired with the same mean-pooled 768-dimensional MLP head and fine-tuned for 5 epochs with learning rate 1×10^{-5}. Unless otherwise stated, results are reported as mean $\pm$ standard deviation over five independent runs.

Model configuration. Cross-model comparisons use UGM (100M) with 12 layers, hidden size 768, 12 attention heads, and 113.3M parameters. Scaling experiments additionally evaluate UGM (500M; 24 layers, hidden size 1,024, 16 heads, 498.7M parameters) and UGM (1B; 32 layers, hidden size 1,280, 20 heads, 1.02B parameters). All variants are pretrained on the unified event corpus using 262,144-token segments and 15% masked positions, with FusedAdam, cosine scheduling, 2,000 warmup steps, batch size 8, and 8 NVIDIA A100 40GB GPUs.

4.2 Functional Variant Classification

Table 2 compares functional variant classification across seven tasks. We assemble datasets from ClinVar Landrum et al. [2020], NHGRI-EBI GWAS Catalog Sollis et al. [2023], Cell Passport Marin et al. [2024], MAVE Esposito et al. [2019], ENCODE histone modifications Consortium et al. [2012], and variant effect annotations Zhou and Troyanskaya [2015]. The baselines include Enformer, SPACE, DNABERT-2 Zhou et al. [2024], NT-V2 Dalla-Torre et al. [2025], and Evo 2 Brixi et al. [2026]. MSA-dependent models are excluded because aligned homologous context is not uniformly available across tasks. Evo 2 obtains the highest baseline mean and leads the Variant Effect task, while NT-V2 provides the strongest baseline on MAVE. UGM achieves the highest mean AUC and the best result on six of the seven tasks. Relative to the strongest baseline in each column, UGM improves performance by 0.16%–4.67% on these six tasks, with the largest gains on Histone-II and MAVE. Its mean score increases from Evo 2's 74.87 to 76.21, corresponding to a relative improvement of 1.79%. On Variant Effect, UGM remains below Evo 2, scoring 61.82 compared with 62.25. The results show uneven benefits across tasks rather than a uniform improvement.

MAVE provides a focused test of contextualized variant representations. UGM reaches 95.08, exceeding NT-V2 by 2.84 AUC points on the $\times 100$ scale. Neither MAVE assay scores nor nonsense/missense labels are included in pretraining. The event token specifies the reference-to-alternate change, whereas the downstream task classifies its consequence using the available sequence context. This distinction connects the evaluation to UGM's pretraining objective: recovering an event state and classifying its functional consequence are different prediction tasks.

Table 2: Functional variant classification results measured by AUC ($\times 100$). The best result in each column is shown in bold and the second-best is underlined. Imprv. reports the relative change over the strongest non-UGM baseline in each column, calculated as $(\mathrm{UGM} - \mathrm{baseline})/\mathrm{baseline} \times 100\%$.

Model	GWAS	ClinVar	Cell Passport	MAVE	Variant Effect	Histone-I	Histone-II	Mean
Enformer	47.01	94.70	<u>79.39</u>	62.82	60.34	74.51	64.40	69.02
SPACE	54.91	<u>94.71</u>	79.31	62.83	58.29	73.54	48.05	67.38
DNABERT-2	53.41	91.42	72.46	83.93	60.28	71.60	64.30	71.06
NT-V2	43.68	89.18	69.63	<u>92.24</u>	59.36	73.32	61.33	69.82
Evo 2	<u>55.83</u>	94.46	79.38	91.71	**62.25**	<u>74.64</u>	<u>65.80</u>	<u>74.87</u>
UGM	**57.42**	**95.21**	**79.52**	**95.08**	<u>61.82</u>	**75.57**	**68.87**	**76.21**
Imprv.	+2.85%	+0.53%	+0.16%	+3.08%	−0.69%	+1.25%	+4.67%	+1.79%

4.3 Regulatory Benchmarks

Table 3 compares the Chromatin, Regulatory, and Splicing groups of the NT benchmark Dalla-Torre et al. [2025], including LOL-EVE Shearer et al. [2025] as a specialized variant-effect baseline. SPACE obtains the highest scores on Chromatin and Regulatory, while NT-2.5B leads on Splicing. UGM ranks second on Splicing, reaching 0.947 compared with 0.964 for NT-2.5B and 0.938 for LOL-EVE. Its relative improvement over LOL-EVE is 0.96%, while its score is 1.76% below NT-2.5B. On Chromatin and Regulatory, UGM scores 0.563 and 0.705, respectively, remaining below SPACE. Thus the strongest transfer within NT occurs on Splicing rather than across all regulatory groups.

Table 4 reports results on the six GUE task groups Zhou et al. [2024]. SPACE leads on Promoter, TF-Human, and TF-Mouse, while NT-2.5B leads on Virus and Splice. UGM achieves the highest Epigenetic score of 0.659, a relative improvement of 1.23% over SPACE, and ranks second on Splice. Compared with LOL-EVE, UGM improves Epigenetic, TF-Mouse, Virus, and Splice, but remains lower on Promoter and TF-Human. The comparison shows selective gains across regulatory settings rather than a general advantage over specialized models.

Across NT and GUE, UGM's most favorable results occur on Splicing and Epigenetic tasks. This pattern supports event-state pretraining as a useful representation choice for some downstream settings, while the lower scores on other groups delineate its current scope.

Table 3: NT benchmark results (AUC, mean $\pm$ std).

Model	Chromatin	Regulatory	Splicing
DNABERT-2	0.557 ± 0.072	0.660 ± 0.134	0.851 ± 0.010
NT-2.5B	0.549 ± 0.068	0.698 ± 0.170	$\mathbf{0.964 \pm 0.005}$
Enformer	0.558 ± 0.056	0.698 ± 0.087	0.756 ± 0.017
SPACE	$\mathbf{0.590 \pm 0.054}$	$\mathbf{0.718 \pm 0.096}$	0.917 ± 0.018
LOL-EVE	0.583 ± 0.0388	0.691 ± 0.216	0.938 ± 0.003
UGM	0.563 ± 0.071	0.705 ± 0.142	0.947 ± 0.004

Table 4: GUE benchmark results (AUC or task-group score, mean $\pm$ std).

Model	Epigenetic	Promoter	TF-Human	TF-Mouse	Virus	Splice
DNABERT-2	0.635 ± 0.001	0.773 ± 0.001	0.702 ± 0.007	0.680 ± 0.001	0.710 ± 0.001	0.850 ± 0.002
NT-2.5B	0.629 ± 0.001	0.799 ± 0.003	0.633 ± 0.007	0.670 ± 0.006	$\mathbf{0.730 \pm 0.001}$	$\mathbf{0.893 \pm 0.002}$
Enformer	0.536 ± 0.025	0.703 ± 0.055	0.737 ± 0.006	0.693 ± 0.047	0.613 ± 0.002	0.815 ± 0.005
SPACE	0.651 ± 0.035	$\mathbf{0.801 \pm 0.022}$	$\mathbf{0.741 \pm 0.050}$	$\mathbf{0.757 \pm 0.051}$	0.703 ± 0.029	0.875 ± 0.025
LOL-EVE	0.635 ± 0.027	0.767 ± 0.018	0.731 ± 0.038	0.696 ± 0.022	0.663 ± 0.034	0.853 ± 0.015
UGM	$\mathbf{0.659 \pm 0.030}$	0.763 ± 0.028	0.714 ± 0.011	0.708 ± 0.025	0.671 ± 0.017	$\underline{0.883 \pm 0.015}$

4.4 SV Detection

Table 5 evaluates breakpoint detection for deletion, duplication, and inversion. SVision leads on deletion and inversion, while UGM achieves the highest duplication AUC. Compared with the DNA-only baseline, UGM increases AUC from 0.624 to 0.856 on deletion, from 0.598 to 0.867 on duplication, and from 0.551 to 0.778 on inversion. These correspond to relative improvements of 37.18%, 44.98%, and 41.20%, respectively. UGM also exceeds Cue across all three categories. Against SVision, UGM improves duplication by 2.60% in relative terms, while remaining below it on deletion and inversion. The results establish competitive breakpoint detection using the shared event-state encoder, with performance varying across SV categories.

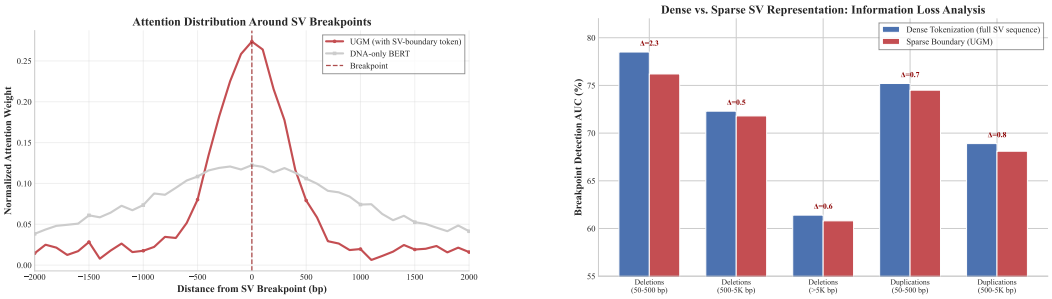
Figure 3b evaluates the trade-off introduced by sparse boundary tokenization. Dense representations retain the internal sequence of an SV, whereas the sparse representation encodes its boundaries and uses the surrounding genomic context for prediction. Comparing the two therefore examines how much breakpoint-detection performance is retained when the event is represented through its endpoints. Across the evaluated event categories and size ranges, sparse boundary tokens achieve performance close to the dense representation, although the remaining gaps show that the two representations are not interchangeable. These results support boundary-based encoding for breakpoint detection, where locating the event endpoints is the prediction target; they do not establish that the same representation preserves all information needed for other SV-related tasks. Figure 3a provides a complementary view of the learned representations. UGM exhibits more concentrated attention near breakpoints than the DNA-only baseline, indicating that boundary neighborhoods receive greater attention in this example. Together, the two analyses connect the representation choice to its downstream behavior: the dense–sparse comparison evaluates predictive performance, while the attention visualization illustrates how the model distributes attention around event boundaries.

4.5 Component Ablations

Table 6 evaluates event-balanced pretraining and SJP by removing them individually and jointly while retaining the unified event tokenizer. Full UGM achieves the highest score on all five tasks.

Table 5: SV breakpoint detection AUC on held-out structural variants.

Model	Deletion	Duplication	Inversion
DNA-BERT	0.624±0.018	0.598±0.021	0.551±0.024
Cue	0.812±0.012	0.789±0.014	0.734±0.016
SVision	**0.891±0.008**	0.845±0.011	**0.812±0.013**
UGM	0.856±0.010	**0.867±0.009**	0.778±0.015



(a) SV attention.

(b) Sparse tokenization.

Figure 3: Structural-variant analyses. UGM concentrates attention near SV breakpoints, and sparse boundary tokens preserve breakpoint-detection performance with small AUC loss.

Removing event-balanced pretraining lowers the five-task mean from 79.89 to 77.87, a relative reduction of 2.53%. Histone-II and SV detection show larger absolute decreases than GWAS, MAVE, and NT Splicing, consistent with the motivation to increase exposure to sparse event categories. Removing SJP lowers the mean to 73.97, with the largest decrease on MAVE, from 95.08 to 75.38. This result supports the contribution of the shared event-state prediction objective to downstream transfer. Removing both components yields a mean of 75.79. The joint removal performs better than removing SJP alone, indicating that the component effects depend on the training configuration and are not additive. The complete configuration improves the five-task mean by 2.02–5.92 points over the ablated variants. The contrast is particularly clear on MAVE: removing SJP lowers performance more than removing both components, whereas event-balanced pretraining with SJP retained remains close to full UGM. This pattern suggests an interaction between the sampling distribution and prediction objective. It motivates evaluating the two choices together rather than assuming that either component contributes a fixed gain independently of the other.

Table 6: Component ablations of event-balanced pretraining and Site-Level Joint Prediction (SJP). All metrics are reported on the $\times 100$ scale; SV Mean averages deletion, duplication, and inversion. The best result in each column is shown in bold.

Configuration	Event-Balanced	SJP	GWAS	MAVE	Histone-II	SV Mean	NT Splicing	Five-Task Mean
w/o Both	No	No	54.48	91.72	63.95	76.88	91.90	75.79
w/o Event-Balanced	No	Yes	55.96	93.74	66.21	79.84	93.60	77.87
w/o SJP	Yes	No	55.31	75.38	65.74	80.21	93.20	73.97
UGM	Yes	Yes	**57.42**	**95.08**	**68.87**	**83.37**	**94.70**	**79.89**

4.6 Cross-Event Attention Analysis

Figure 4 provides a qualitative comparison of attention patterns at SNP–methylation co-occurrence sites. We examine loci where a biallelic SNP lies within 10 bp of a CpG site with medium-to-high methylation ($\beta > 0.5$, $n = 2,847$ genome-wide). In UGM, attention is concentrated around the SNP position and the nearby methylated CpG position, whereas the DNA-only baseline exhibits a more diffuse pattern at the same loci. These qualitative patterns provide evidence that UGM captures dependencies between nearby SNP and methylation events.

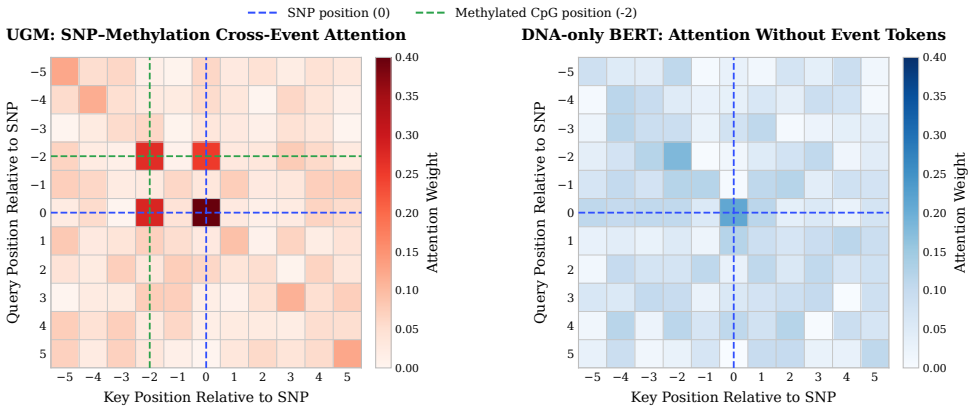


Figure 4: Attention visualization at SNP–methylation co-occurrence sites. UGM exhibits localized attention around the SNP and nearby methylated CpG positions, whereas DNA-only BERT shows a more diffuse attention pattern.

4.7 Model Scaling

Figure 5 compares UGM at 100M, 500M, and 1B parameters on seven representative tasks. The models use the same pretraining corpus and hyperparameters, with changes in depth, width, and attention heads. Increasing model size improves performance across the evaluated tasks, but the magnitude varies. SV Boundary and GWAS show the largest gains: increasing capacity from 100M to 1B raises their scores by 9.0 and 6.4 points, respectively, on the $\times 100$ scale. ClinVar and MAVE show smaller changes. This task-dependent scaling behavior is consistent with additional capacity benefiting some event-related prediction settings more than others. The curves characterize scaling within the UGM family; the cross-model comparisons above use the 100M configuration.

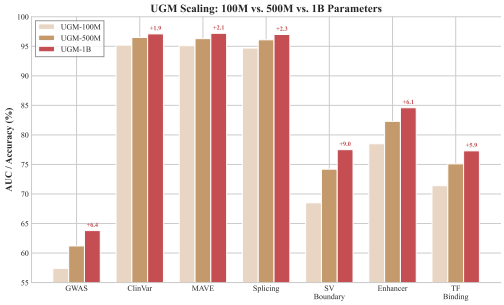


Figure 5: UGM scaling across 100M, 500M, and 1B parameters. Scores increase with model size across the evaluated tasks, with larger gains on SV Boundary and GWAS than on ClinVar and MAVE.

5 Conclusion

We presented **UGM**, a genomic foundation model that treats coordinate-aligned event states as the units of pretraining. Its shared vocabulary, event-balanced sampling, and Site-Level Joint Prediction enable learning heterogeneous event states from their surrounding context. UGM achieves the highest average performance on the evaluated functional variant classification tasks and improves SV breakpoint detection over the DNA-only baseline, with task-dependent results on regulatory benchmarks. Component ablations support the complete training configuration, while the sparse-representation analysis supports boundary tokens for breakpoint detection. These findings support extending genomic pretraining beyond reference sequence to include heterogeneous event states. Beyond genomics, this formulation motivates exploring event-state pretraining for other scientific data in which heterogeneous observations share spatial or temporal coordinates Li et al. [2026].

Acknowledgments

This work was supported by the Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (2024R01007) and the "Pioneer" and "Leading Goose" Research and Development Program of Zhejiang (2025C02077).

References

1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021.

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

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, 2025. doi: 10.1038/s41587-024-02511-w.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali Taghibakhshi, Anton Vorontsov, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112):1349–1361, 2026. doi: 10.1038/s41586-026-10176-5. URL `https://doi.org/10.1038/s41586-026-10176-5`.

ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. doi: 10.1038/nature11247.

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

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024.

Daniel Esposito, Jochen Weile, Jay Shendure, Lea M Starita, Anthony T Papenfuss, Frederick P Roth, Douglas M Fowler, and Alan F Rubin. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, 20(1):223, 2019.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic Benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24:25, 2023.

Yunhee Jeong, Clarissa Gerhäuser, Guido Sauter, Thorsten Schlomm, Karl Rohr, and Pavlo Lutsik. MethylBERT enables read-level DNA methylation pattern identification and tumour deconvolution using a transformer-based model. *Nature Communications*, 16(1):788, 2025.

Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, et al. ClinVar: improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844, 2020.

Mingyang Li, Yurou Liu, Jieping Ye, Bing Su, Ji-Rong Wen, and Zheng Wang. Speaking the language of science: Toward a general-purpose generative foundation model for the natural sciences. *arXiv preprint arXiv:2606.16905*, 2026.

Zehui Li, Vallijah Subasri, Guy-Bart Stan, Yiren Zhao, and Bo Wang. GV-Rep: A large-scale dataset for genetic variant representation learning. *NeurIPS*, 37:134701–134720, 2024.

Jiadong Lin, Songbo Wang, Peter A Audano, Deyu Meng, Jacob I Flores, Walter Kosters, Xiaofei Yang, Peng Jia, Tobias Marschall, Christine R Beck, et al. Svision: a deep learning approach to resolve complex structural variants. *Nature Methods*, 19(10):1230–1233, 2022.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics*, 57(4):949–961, 2025. doi: 10.1038/s41588-024-02053-6.

Zicheng Liu, Jiahui Li, Siyuan Li, Zelin Zang, Cheng Tan, Yufei Huang, Yajing Bai, and Stan Z Li. Genbench: A benchmarking suite for systematic evaluation of genomic foundation models. *arXiv preprint arXiv:2406.01627*, 2024.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA language models on biologically meaningful tasks. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=uKB4cFNQFg`.

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. *NeurIPS*, 36:43177–43201, 2023.

Peng Ni, Neng Huang, Zhi Zhang, De-Peng Wang, Fan Liang, Yu Miao, Chuan-Le Xiao, Feng Luo, and Jianxin Wang. Deepsignal: detecting dna methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics*, 35(22):4586–4595, 2019.

Victoria Popic, Chris Rohlicek, Fabio Cunial, Iman Hajirasouliha, Dmitry Meleshko, Kiran Garimella, and Anant Maheshwari. Cue: a deep-learning framework for structural variant discovery and genotyping. *Nature Methods*, 20(4):559–568, 2023.

Courtney Shearer, Rose Orenbuch, Felix Teufel, Christian Steinmetz, Daniel Ritter, Erik Xie, Artem Gazizov, Aviv Spinner, Mafalda Dias, Jonathan Frazer, Pascal Notin, and Debora Marks. A genomic language model for zero-shot prediction of promoter indel effects. In David A. Knowles and Peter K. Koo, editors, *Proceedings of the 20th Machine Learning in Computational Biology Meeting*, volume 311 of *Proceedings of Machine Learning Research*, pages 95–127. PMLR, 2025. URL `https://proceedings.mlr.press/v311/shearer25a.html`.

Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Dieudonne van der Meer, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E Francies, and Mathew J Garnett. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Research*, 47(D1): D923–D929, 2019.

Qimeng Yang, Jianfeng Sun, Xinyu Wang, Jiong Wang, Quanzhong Liu, Jinlong Ru, Xin Zhang, Sizhe Wang, Ran Hao, Peipei Bian, et al. SVLearn: a dual-reference machine learning approach enables accurate cross-species genotyping of structural variants. *Nature Communications*, 16(1): 2406, 2025a.

Zhao Yang, Jiwei Zhu, and Bing Su. SPACE: Your genomic profile predictor is a powerful DNA foundation model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 71648–71667. PMLR, 2025b. URL `https://proceedings.mlr.press/v267/yang25ay.html`.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *NeurIPS*, 32, 2019.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 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*, 2024. URL https://openreview.net/forum?id=oMLQB4EZE1.

A Appendix

A.1 UGM Pretraining Corpus Construction Details

This section provides a detailed description of the data preparation pipeline underlying the UGM pretraining corpus. These details complement the brief summary in the methodology section.

(i) Data Download and Processing

The UGM corpus is assembled from four classes of genomic data obtained from public databases and curated extensions, all referenced to the GRCh38 coordinate system. All VCF and MAF files undergo normalization with bcftools and coordinate standardization to promote consistent positioning before sampling and tokenization.

- **Reference Genome**: The GRCh38 primary assembly serves as the coordinate backbone, encompassing 3.2 billion base pairs across 25 chromosomes. This provides the stable genomic context for positioning all multi-scale events.
- **Atom Events (Epigenetic Marks)**: Methylation data are obtained from GEO and TCGA-derived records to represent atom-level events. Methylation levels are quantified as CpG β-values and discretized into low, medium, and high levels. These data represent sample-specific epigenetic states without a germline equivalent.
- **Molecular and Fragment Catalogs (dbSNP)**: dbSNP (build 155) serves as the comprehensive population-level catalog of known genetic variations. This dataset provides a vast pool of molecular-level short variants (INDELs) and fragment-level structural variants ($\geq$ 50 bp) discovered across diverse human populations. Only high-quality biallelic variants with a PASS filter are retained to serve as a background variant diversity source.
- **Individual Events (Personalized Variants)**: Somatic and germline variants from individual donor records are integrated where available. TCGA-derived tumor records and 1000 Genomes germline variants contribute molecular events and fragment events. Only variants with at least five supporting reads are retained, and hypermutated outliers are excluded to maintain biological integrity during training.

(ii) Rich Token Distributions for Unified Genomic Prediction

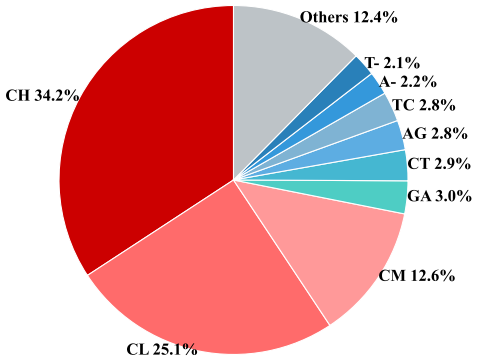


Figure 6: Token distribution across multi-scale genomic event sources in UGM pretraining corpus.

We define a vocabulary of 43 genomic event-state tokens encompassing four genomic event scales. Special control tokens used for sequence processing are excluded from this count. To clarify how raw records from diverse sources are transformed into unified training instances, Figure 9 presents representative examples for each event type, and Figure 6 shows the major token distributions in the UGM corpus:

- **Bases and SNPs**: These events are sourced from the reference genome and dbSNP. Each token encodes a single-nucleotide variant as a reference base followed by an alternate base, allowing the model to distinguish all six substitution classes while preserving strand orientation (`AG`, `AC`, `AT`, `GA`, `GC`, `GT`, `CA`, `CG`, `CT`, `TA`, `TG`, `TC`).

Table 7: The unified vocabulary of 43 genomic event-state tokens (fraction in %). Special control tokens are excluded. Fragment (Structural) tokens use a sparse boundary representation, resulting in efficient token usage despite high event coverage.

ID	Variant Type	Methylation	Short	Structural	Total Count	Fraction
Base Events (Reference & SNPs)						
0	AA	0	35,004,206	0	35,004,206	0.7%
1	AG	0	137,688,182	0	137,688,182	2.8%
2	AC	0	38,633,354	0	38,633,354	0.8%
3	AT	0	32,709,700	0	32,709,700	0.7%
4	GA	0	152,038,354	0	152,038,354	3.0%
5	GG	0	26,032,565	0	26,032,565	0.5%
6	GC	0	38,021,818	0	38,021,818	0.8%
7	GT	0	38,498,902	0	38,498,902	0.8%
8	CA	0	44,792,231	0	44,792,231	0.9%
9	CG	0	40,530,520	0	40,530,520	0.8%
10	CC	0	46,546,777	0	46,546,777	0.9%
11	CT	0	142,929,349	0	142,929,349	2.9%
12	TA	0	35,647,313	0	35,647,313	0.7%
13	TG	0	35,875,140	0	35,875,140	0.7%
14	TC	0	137,722,682	0	137,722,682	2.8%
15	TT	0	34,178,126	0	34,178,126	0.7%
Atom Events (Methylation)						
24	CL	1,255,511,363	N/A	N/A	1,255,511,363	25.1%
25	CM	628,212,572	N/A	N/A	628,212,572	12.6%
26	CH	1,710,442,981	N/A	N/A	1,710,442,981	34.2%
Molecular Events (Short Indels $\leq$ 50bp)						
16	A-	0	108,197,663	N/A	108,197,663	2.2%
17	G-	0	44,499,084	N/A	44,499,084	0.9%
18	C-	0	44,784,533	N/A	44,784,533	0.9%
19	T-	0	107,333,575	N/A	107,333,575	2.1%
20	-A	0	64,634,169	N/A	64,634,169	1.3%
21	-G	0	24,632,730	N/A	24,632,730	0.5%
22	-C	0	24,779,684	N/A	24,779,684	0.5%
23	-T	0	63,772,789	N/A	63,772,789	1.3%
Fragment Events (Structural Variants > 50bp, Sparse Boundary)						
27-30	AD^s, CD^s, GD^s, TD^s	0	0	2,841,593	2,841,593	<0.1%
31-34	AD^e, CD^e, GD^e, TD^e	0	0	2,839,104	2,839,104	<0.1%
35-38	DA^s, DC^s, DG^s, DT^s	0	0	2,158,472	2,158,472	<0.1%
39-42	DA^e, DC^e, DG^e, DT^e	0	0	2,157,911	2,157,911	<0.1%
Total		**3.6 B**	**1.5 B**	**≈0.01 B**	**5.0 B**	**100%**

- **Methylation**: These events are sourced from TCGA-derived records and GEO. CpG methylation levels are encoded as CH, CM, and CL. Methylation β-values (0–1) are discretized into low ($\beta < 0.2$, L), medium ($0.2 \leq \beta < 0.6$, M), and high ($\beta \geq 0.6$, H) bins, allowing joint learning of methylation and genetic signals in one token space.
- **Short variants**: These events are sourced from dbSNP, the 1000 Genomes Project, and TCGA-derived records. Short insertions (-A, -C, -G, -T) and deletions (A-, C-, G-, T-) are encoded as reference-base-plus-dash for deletions and dash-plus-inserted-base for insertions.
- **Structural variants**: These events are sourced from dbSNP, the 1000 Genomes Project, and TCGA-derived records. Large structural variants are represented by paired breakpoint tokens, where the first letter denotes the nucleotide adjacent to the breakpoint, superscript s marks the start coordinate, and superscript e marks the end coordinate. Deletion signatures are AD^s, CD^s, GD^s, TD^s and their corresponding end partners AD^e, CD^e, GD^e, TD^e; insertion signatures are DA^s, DC^s, DG^s, DT^s and their end partners DA^e, DC^e, DG^e, DT^e.

(iii). Sampling strategy and sequence construction

To construct training sequences of 262,144 tokens, UGM employs a multi-scale sampling strategy that integrates diverse genomic event types based on their specific scale characteristics. The comprehensive logic for this multi-scale integration is formalized in Algorithm 1, which helps ensure that while events are sampled according to pre-defined proportions, they remain unified under a consistent coordinate system and a single-donor isolation protocol. The distinct sampling logic for each event type is described below:

- **Base events (40%)**: Pure genomic sequences extracted directly from the GRCh38 reference assembly without any modifications.
- **Atom events (25%)**: Individual-specific epigenetic sequences where CpG sites from methylation sequences (TCGA/GEO) are annotated with discretized methylation levels (`CL`, `CM`, `CH`). Methylation represents epigenetic modifications rather than genetic variants. Methylation data are sampled within each sequence sample to maintain individual-specific epigenetic signatures.
- **Molecular events (25%)**: Short genomic variants, including SNPs and small INDELs ($<$ 50 bp), which provide site-specific genetic diversity across the genome. These events are sampled from two hierarchical sources: (1) *Germline variants* from dbSNP are sampled at the population level to capture common alleles and genetic diversity patterns. (2) *Individual variants* from TCGA and 1000 Genomes are sampled under a single-donor isolation protocol to preserve the specific linkage disequilibrium and co-occurrence patterns unique to an individual.
- **Fragment events (10%)**: Large-scale structural variations ($\geq$ 50 bp) such as long INDELs. To optimize computational efficiency while preserving structural information, UGM employs a sparse boundary-replacement strategy, tokenizing only the start and end boundaries (e.g., $\texttt{AD}^s$, $\texttt{GD}^e$) of structural events. We identified individual-specific structural variants from TCGA or 1000 Genomes by sampling fragment events from single donors.

After candidate events are sampled, records are merged by genomic coordinate rather than concatenated as independent modality channels. For each coordinate, UGM applies the same deterministic resolver used in the main method, where structural-variant boundary and short-INDEL states take precedence at altered bases, SNP states replace the reference state when the allele changes, CpG methylation is retained only when the resolved base remains a compatible cytosine, and unresolved positions fall back to the reference token. This rule prevents duplicated tokens at the same locus while preserving cross-event dependencies through nearby coordinates within the same donor-specific sequence. **Algorithmic details.** The complete multi-scale sampling and integration procedure is formalized in Algorithm 1 (see §3.2).

(i). Multi-Scale Genomic Embedding.

For each genomic position $i \in \{1, \ldots, L\}$, UGM constructs unified genomic tokens across four hierarchical scales. For base events, $v_i \in \Sigma_{\text{var}} = \{\texttt{AA},\texttt{CC},\texttt{GG},\texttt{TT},\texttt{AC},\texttt{AG},\texttt{AT},\texttt{CA},\texttt{CG},\texttt{CT},\texttt{GA},\texttt{GC},\texttt{GT},\texttt{TA},\texttt{TC},\texttt{TG}\}$ represents matched bases (homozygous) or substituted bases (heterozygous), where the first character denotes the reference base and the second denotes the observed allele. For atom events (methylation modifications), $m_i \in \Sigma_{\text{met}} = \{\texttt{CH},\texttt{CM},\texttt{CL}\}$ encodes discretized methylation levels where `CH`, `CM`, and `CL` represent high, medium, and low methylation states at CpG sites. For molecular events (short variants with INDELs $<$ 50 bp), deletions are denoted as $v_i \in \{\texttt{A-},\texttt{G-},\texttt{C-},\texttt{T-}\}$ where the reference base is followed by a dash, while insertions are represented as $v_i \in \{\texttt{-A},\texttt{-G},\texttt{-C},\texttt{-T}\}$ where a dash precedes the inserted base. For fragment events (large structural variants $\geq$ 50 bp), we employ a sparse boundary-replacement strategy where each structural variant is represented by two boundary tokens marking its start and end positions. For deletions, the start boundaries are encoded as $v_i \in \{\texttt{AD}^s, \texttt{CD}^s, \texttt{GD}^s, \texttt{TD}^s\}$ and end boundaries as $v_i \in \{\texttt{AD}^e, \texttt{CD}^e, \texttt{GD}^e, \texttt{TD}^e\}$, while for insertions, start boundaries are $v_i \in \{\texttt{DA}^s, \texttt{DC}^s, \texttt{DG}^s, \texttt{DT}^s\}$ and end boundaries are $v_i \in \{\texttt{DA}^e, \texttt{DC}^e, \texttt{DG}^e, \texttt{DT}^e\}$, where the superscripts s and e indicate the start and end coordinates of the structural variant span, and the base letter represents the reference nucleotide at that boundary position. The unified token $x_i \in \mathcal{V}$ is formed by concatenating these components: $x_i = r_i \| v_i$ for DNA sequences and variations, or $x_i = r_i \| m_i$ for methylation modifications, yielding $|\mathcal{V}| = 43$ possible token types spanning all four genomic scales.

$$E_i = E_{\text{token}}(x_i) + E_{\text{chr}}(c_i) + E_{\text{pos}}(p_i), \tag{7}$$

where $E_{\text{token}}(x_i) \in \mathbb{R}^d$ maps the unified token x_i to a d-dimensional space, $E_{\text{chr}}(c_i) \in \mathbb{R}^d$ encodes chromosomal information, and $E_{\text{pos}}(p_i) \in \mathbb{R}^d$ provides positional encoding. For molecular events, we further incorporate an inheritance-type embedding $E_i = E_i + E_{\text{type}}(t_i)$ where $t_i \in \{\text{germline, individual}\}$ to distinguish between population-level and individual-specific inheritance patterns.

(ii). UGM blocks.

UGM uses a stack of N identical UGM blocks with pre-normalization and residual connections. At layer $\ell \in \{1, \ldots, N\}$ the input is $H^{(\ell-1)} \in \mathbb{R}^{L\times d}$. The embedded sequence $E = (E_1, \ldots, E_L)$ is processed through N UGM blocks for joint representation learning:

$$H^{(0)} = E, \tag{8}$$

$$H^{(\ell)} = \text{UGMBlock}^{(\ell)}(H^{(\ell-1)}), \quad \ell = 1, \ldots, N, \tag{9}$$

where each block uses pre-normalization with RMSNorm and rotary positional encoding (RoPE) for relative position modeling. We first apply RMS normalization,

$$\tilde{H}^{(\ell-1)} = \text{RMSNorm}^{(\ell)}\big(H^{(\ell-1)}\big), \tag{10}$$

followed by multi-head self-attention. Let h be the number of heads and $d_h = d/h$ the head dimension. We use projection matrices

$$W_Q^{(\ell)}, W_K^{(\ell)}, W_V^{(\ell)} \in \mathbb{R}^{d\times d}, \quad W_O^{(\ell)} \in \mathbb{R}^{d\times d}. \tag{11}$$

The joint query, key and value sequences are

$$Q^{(\ell)} = \tilde{H}^{(\ell-1)}W_Q^{(\ell)}, \quad K^{(\ell)} = \tilde{H}^{(\ell-1)}W_K^{(\ell)}, \quad V^{(\ell)} = \tilde{H}^{(\ell-1)}W_V^{(\ell)}, \tag{12}$$

which we view as h heads by reshaping the last dimension. For head $j \in \{1, \ldots, h\}$ we write $Q_j^{(\ell)}, K_j^{(\ell)}, V_j^{(\ell)} \in \mathbb{R}^{n\times d_h}$. Before computing attention scores, UGM applies rotary positional embeddings (RoPE) to $(Q_j^{(\ell)}, K_j^{(\ell)})$ using precomputed complex rotations $\Omega_{1:n}$; we denote this operation abstractly as

$$(\hat{Q}_j^{(\ell)}, \hat{K}_j^{(\ell)}) = \text{RoPE}\big(Q_j^{(\ell)}, K_j^{(\ell)}; \Omega_{1:n}\big). \tag{13}$$

Given an additive attention mask $M \in \mathbb{R}^{n\times n}$ (with large negative values at padded or disallowed positions and 0 elsewhere), the attention weights and outputs for head j are

$$A_j^{(\ell)} = \text{softmax}\Big(\frac{\hat{Q}_j^{(\ell)}(\hat{K}_j^{(\ell)})^\top}{\sqrt{d_h}} + M\Big) \in \mathbb{R}^{n\times n}, \tag{14}$$

$$O_j^{(\ell)} = A_j^{(\ell)}V_j^{(\ell)} \in \mathbb{R}^{n\times d_h}. \tag{15}$$

Concatenating all heads and projecting back to dimension d yields

$$O^{(\ell)} = \text{Concat}_{j=1}^{h} O_j^{(\ell)}W_O^{(\ell)} \in \mathbb{R}^{n\times d}. \tag{16}$$

With a drop-path operator $\text{DropPath}^{(\ell)}$ and a second RMS normalization, the residual self-attention update is

$$H^{(\ell-1/2)} = H^{(\ell-1)} + \text{DropPath}^{(\ell)}\big(O^{(\ell)}\big). \tag{17}$$

The position-wise feed-forward network (FFN) then applies a SwiGLU-style nonlinearity:

$$U^{(\ell)} = \text{RMSNorm}_2^{(\ell)}\big(H^{(\ell-1/2)}\big), \tag{18}$$

$$F^{(\ell)} = \phi\big(U^{(\ell)}W_1^{(\ell)} + b_1^{(\ell)}\big)W_2^{(\ell)} + b_2^{(\ell)}, \tag{19}$$

$$H^{(\ell)} = H^{(\ell-1/2)} + \text{DropPath}^{(\ell)}\big(F^{(\ell)}\big), \tag{20}$$

where $W_1^{(\ell)} \in \mathbb{R}^{d\times d_{\text{ff}}}$, $W_2^{(\ell)} \in \mathbb{R}^{d_{\text{ff}}\times d}$, $b_1^{(\ell)}, b_2^{(\ell)}$ are learned parameters and ϕ denotes the SwiGLU activation. Stacking L such blocks produces final hidden states $H^{(L)} \in \mathbb{R}^{n\times d}$.

(iii) Site-Level Joint Prediction objective and loss.

UGM employs a site-level masked language modeling objective over the 43 genomic event-state tokens in $\mathcal{V}$ for Site-Level Joint Prediction. Before prediction, UGM applies a final RMS normalization,

$$\tilde{H}^{(N)} = \mathrm{RMSNorm}^{\mathrm{final}}\big(H^{(N)}\big). \tag{21}$$

Let $|\mathcal{V}| = 43$ be the size of the genomic event-state vocabulary and let $W \in \mathbb{R}^{d\times 43}$, $b \in \mathbb{R}^{43}$ be the parameters of the prediction head. Special control tokens are excluded from $\mathcal{V}$ and are not prediction targets. The unified prediction head computes:

$$p_\theta(x_i \mid X_{\mathrm{masked}}) = \mathrm{softmax}(H_i W + b), \tag{22}$$

Let $S \subseteq \{1, \ldots, L\}$ be the set of masked positions and let $x_i^\star$ be the original (unmasked) token at position i. The cross-entropy loss enables simultaneous learning across all genomic event scales:

$$\mathcal{L} = -\frac{1}{|S|}\sum_{i\in S} \log p_\theta(x_i^* \mid X_{\mathrm{masked}}), \tag{23}$$

This unified objective allows the model to jointly predict DNA sequence states, modifications, and variants at each genomic site, capturing interactions across genomic event scales.

A.2 Functional Variant Classification Datasets

This section provides detailed statistics and dataset descriptions for the functional variant classification subsection in the main text. The evaluation follows the same train/validation/test splits and downstream head used in Table 2. Following the protocols of existing benchmarks BEND Marin et al. [2024] and GV-Rep Li et al. [2024], the datasets are assembled from established public genomic databases. Table 8 lists the resulting dataset statistics. All tasks use standard train/validation/test splits with a fixed random seed (42). For each dataset, 70% of samples are used for training, 15% for validation, and 15% for testing. All models in the main comparison (Table 2) are evaluated on the same splits with the same downstream MLP classification head.

- **Cell Passport van der Meer et al. [2019]:** Derived from the Wellcome Sanger Institute's Cell Model Passports database, containing somatic variations from cancer cell lines. We processed VCF files from multiple cell lines to extract variants and classify genomic sequences based on their location within genes (intronic vs. exonic regions). This task evaluates the model's ability to recognize functional gene elements and understand splicing patterns. Through balanced sampling, we obtained 4,915 samples per class.
- **ClinVar Landrum et al. [2020]:** Sourced from NCBI's ClinVar database (version 2024-04-16), a public archive of relationships between human variations and phenotypes. We extracted variants with clinical significance annotations, focusing on the binary classification of variants as benign or pathogenic. After filtering uncertain classifications and balanced sampling, we obtained 2,500 samples per class for robust clinical variant interpretation evaluation.
- **GWAS Sollis et al. [2023]:** Based on the NHGRI-EBI GWAS Catalog, focusing on breast cancer association studies. We extracted single nucleotide polymorphisms (SNPs) with different levels of statistical significance and classified them as either highly significant (p $<$ 1e-8) or genome-wide significant (1e-8 $\leq$ p $<$ 1e-5). The dataset contains 293 and 262 samples respectively, reflecting the natural distribution of significant variants in association studies.
- **MAVE Esposito et al. [2019]:** Constructed from the Multiplexed Assays of Variant Effects (MAVE) database, which contains experimental measurements of variant functional effects. We processed functional scores to classify variants based on their protein-coding consequences, distinguishing between nonsense and missense variations. The dataset is nearly balanced with 2,511 nonsense and 2,489 missense variants.
- **Histone-I Consortium et al. [2012]:** *Histone-I:* Derived from ChIP-seq data in BED format containing histone modification annotations. We selected the two most abundant single histone modification types (H3K4me3 and H2AFZ) and performed balanced sampling to create a binary classification task. This epigenetic task requires models to learn sequence patterns associated with different chromatin states, with 2,500 samples per modification type.

Table 8: Dataset statistics of Genomic Functional Classification.

Dataset	Label Class	# Samples
Cell Passport	Intronic	4915
	Exonic	4915
ClinVar	Benign	2500
	Pathogenic	2500
GWAS	Highly Significant	293
	Genome-wide Significant	262
MAVE	Nonsense	2511
	Missense	2489
Histone-I	H3K4me3	2500
	H2AFZ	2500
Histone-II	Repressive	2500
	Activating	2500
Variant Effect	Non-functional	2500
	Functional	2500

Table 9: Dataset statistics of Genomic Benchmarks.

Dataset	Positive	Negative	Total
Human Enhancers Cohn	13,896	13,895	40,791
Human Enhancers Ensembl	77,421	77,421	154,842
Human vs. Worm	50,000	50,000	100,000
Coding vs. Intergenic	50,000	50,000	100,000
OCR Ensembl	87,378	87,378	174,756
Non-TATA Promoters	16,423	19,708	36,131
Mouse Enhancers	605	605	1,210
Drosophila Enhancers	3,457	3,457	6,914

- **Histone-II Consortium et al. [2012]:** Built from the same histone modification data source, but with functional annotation grouping modifications as repressive or activating based on their known biological roles. This task tests understanding of epigenetic regulation mechanisms and their impact on gene expression, with 2,500 samples per functional class.
- **Variant Effect Zhou and Troyanskaya [2015]:** Processed from variant effect prediction data in BED format, specifically using expression-related variant effects containing consequence annotations for genomic variants. We classified variants as functional or non-functional based on their predicted impact on gene expression, using consequence severity to determine labels. The balanced dataset contains 2,500 samples per class, providing a robust testbed for expression-related variant interpretation.

A.3 Genomic Benchmark Datasets

We evaluate UGM on genomic benchmark datasets that cover diverse classification tasks across species and genomic elements Grešová et al. [2023]. Table 9 summarizes the dataset statistics for these benchmarks. The datasets range from small-scale tasks such as Mouse Enhancers (1,210 samples) to large-scale tasks such as Human Enhancers Ensembl (154,842 samples) and OCR Ensembl (174,756 samples). The benchmarks include enhancer identification in humans, cross-species sequence classification (Human vs. Worm), coding-region detection (Coding vs. Intergenic), open-chromatin prediction (OCR Ensembl), promoter classification (Non-TATA Promoters), and enhancer prediction in model organisms (Mouse and Drosophila). Most datasets maintain balanced positive and negative classes, except for Non-TATA Promoters, which has a slight class imbalance (16,423 positive vs. 19,708 negative samples). This collection supports evaluation of genomic sequence understanding across biological contexts and scales.

A.4 Regulatory Benchmark Datasets

This section describes the datasets used in the Regulatory Benchmarks subsection of the main text. We follow the official NT benchmark Dalla-Torre et al. [2025] and GUE benchmark protocols under the same fine-tuning setting used for all cross-model comparisons. **NT benchmark.** The NT benchmark contributes three task groups: Chromatin, Regulatory, and Splicing. The Chromatin group measures sequence-to-chromatin-state prediction, the Regulatory group covers promoter- and enhancer-related sequence classification, and the Splicing group evaluates recognition of splice-related sequence patterns. We preserve the benchmark-standard train/validation/test partitions for each task and report the category-level average AUC in Table 3. **GUE benchmark.** The GUE benchmark contributes six reported groups: Epigenetic, Promoter, TF-Human, TF-Mouse, Virus, and Splice. These groups cover epigenetic-state prediction, promoter recognition, transcription-factor binding in human and mouse, viral-sequence classification, and splice-related sequence understanding. We use the same published benchmark splits for all models and report task-group scores aggregated under the same evaluation protocol in Table 4.

A.5 SV Detection Datasets

This section describes the data used for SV detection. We construct the benchmark from curated structural-variant call sets, including deletion, duplication, and inversion events drawn from the 1000 Genomes Project 1000 Genomes Project Consortium et al. [2015] and TCGA-derived records. Positive examples are breakpoint-centered genomic windows around validated SV breakpoints, and negative examples are non-SV windows sampled from matched genomic regions to control for chromosome- and context-level biases so that models must identify breakpoint structure rather than rely on coarse genomic priors. The resulting dataset covers three SV categories with the following sample counts: deletion (4,852 positive / 4,852 negative), duplication (2,134 positive / 2,134 negative), and inversion (1,567 positive / 1,567 negative). All methods are evaluated on the same train/validation/test splits, and UGM uses the same sparse boundary-token representation in both pretraining and downstream prediction. The three-category evaluation protocol is summarized in Table 5.

A.6 Cross-Event Attention Analysis Data

This section documents the loci used in the qualitative cross-event attention analysis. Candidate loci are identified by intersecting biallelic SNP sites from dbSNP with CpG methylation records and retaining cases where a SNP lies within 10 bp of a CpG site with medium-to-high methylation ($\beta > 0.5$). The figure in the main text averages attention weights across all heads in the final UGM block at these co-occurrence sites. This analysis is qualitative rather than a standalone benchmark, and it is intended to illustrate the type of cross-event dependency learned by the unified tokenizer.

A.7 Context-Length Evaluation Tasks

This section describes the tasks used for the context-length study. The local-signal panel uses TF Binding, Promoter, Splice Site, and Chromatin tasks, which are dominated by motif-scale or neighborhood-scale context. The long-range panel uses enhancer-promoter interaction, gene-expression prediction, CTCF binding, and chromatin accessibility, where distal regulatory dependencies are more important. For every task, only the context window varies across runs; the data splits, downstream head, and optimization settings are kept fixed, allowing Table 12 and Table 11 to isolate the effect of context length.

A.8 Model Scaling Evaluation Tasks

This section describes the representative tasks used in the model scaling analysis. We evaluate internal UGM variants on a fixed subset spanning variant prediction, regulatory prediction, and structural-event prediction, including GWAS, splicing, enhancer-related tasks, transcription-factor binding, and SV-boundary detection. The dataset splits and fine-tuning protocol are held constant across UGM model sizes, so performance differences reflect parameter scaling rather than changes in data or optimization.

Table 10: Results on Genomic Benchmarks (accuracy, mean ± std).

Model	Mouse	Demo		Drosophila
	Enhancers	**Coding vs. Intergenic**	**Human vs. Worm**	**Enhancers**
Enformer	0.835 ± 0.012	0.913 ± 0.001	0.958 ± 0.001	0.613 ± 0.005
SPACE	$\mathbf{0.905 \pm 0.010}$	0.922 ± 0.001	0.967 ± 0.004	0.721 ± 0.016
UGM	0.791 ± 0.019	$\mathbf{0.952 \pm 0.001}$	$\mathbf{0.973 \pm 0.002}$	$\mathbf{0.727 \pm 0.025}$

Model	Human				
	Enhancers Cohn	**Enhancers Ensembl**	**Regulatory**	**OCR Ensembl**	**Non-TATA Promoters**
Enformer	0.723 ± 0.001	0.844 ± 0.001	0.903 ± 0.001	$\mathbf{0.876 \pm 0.001}$	0.878 ± 0.002
SPACE	$\mathbf{0.769 \pm 0.006}$	0.919 ± 0.014	$\mathbf{0.944 \pm 0.002}$	0.854 ± 0.001	0.940 ± 0.002
UGM	0.745 ± 0.008	$\mathbf{0.923 \pm 0.005}$	0.918 ± 0.005	0.813 ± 0.008	$\mathbf{0.959 \pm 0.005}$

Table 11: Window-size ablation on long-range tasks. Metrics are AUC except gene expression (R^2). Δ reports relative change from $w = 512$ to $w = 1024$.

Task	**Input Len**	$w = 128$	$w = 256$	$w = 512$	$w = 1024$	**Δ(1024 vs 512)**
Enhancer-Promoter Interaction	8K bp	0.653 ± 0.018	0.685 ± 0.015	0.718 ± 0.012	0.794 ± 0.008	$+10.58\%$
Gene Expression (TPM)	16K bp	0.389 ± 0.021	0.421 ± 0.018	0.467 ± 0.014	0.512 ± 0.011	$+9.64\%$
CTCF Binding	8K bp	0.692 ± 0.012	0.714 ± 0.010	0.735 ± 0.008	0.741 ± 0.007	$+0.82\%$
Chromatin Accessibility (ATAC)	4K bp	0.771 ± 0.009	0.779 ± 0.008	0.783 ± 0.007	0.789 ± 0.006	$+0.77\%$

A.9 Supplementary Experimental Results

Genomic Benchmarks. Table 10 evaluates whether unified event pretraining transfers to the Genomic Benchmarks suite Grešová et al. [2023], a collection of standard sequence classification tasks that do not explicitly expose multi-scale event labels. UGM is strongest on Coding vs. Intergenic, Human vs. Worm, Drosophila Enhancers, Human Enhancers Ensembl, and Non-TATA Promoters. These gains suggest that event-aware pretraining can improve sequence classification even when the downstream input is represented as ordinary genomic sequence, likely because coding boundaries, splice-related context, and promoter grammar are correlated with variant and epigenetic states in the pretraining corpus. The weaker cases are also informative. SPACE remains stronger on Mouse Enhancers, Human Enhancers Cohn, and Human Regulatory, and Enformer is strongest on OCR Ensembl. Thus the benchmark does not show a blanket advantage for UGM. Instead, it separates two regimes, where unified event pretraining helps tasks linked to event-state and coding/regulatory boundaries, while specialized regulatory-track or cross-species objectives remain advantageous for motif-heavy and chromatin-track prediction.

A.10 Context Length

The context-length analysis varies the sliding-window size $w \in \{128, 256, 512, 1024\}$. On the local-signal tasks in Table 12, increasing w from 128 to 1024 yields relative AUC improvements of 0.64%–0.85%. Table 11 reports larger changes on enhancer-promoter interaction and gene-expression prediction: increasing w from 512 to 1024 improves AUC from 0.718 to 0.794 and R^2 from 0.467 to 0.512, corresponding to relative improvements of 10.58% and 9.64%, respectively. CTCF binding and chromatin accessibility show relative AUC improvements of 0.82% and 0.77% over the same range. These observations show task-dependent sensitivity to the attention window and motivate choosing context length according to the downstream setting.

A.11 Baseline Models

We compare UGM against representative genomic foundation models, categorized into sequence-only pretraining models and event-specific architectures. These baselines include leading DNA language models that focus on evolutionary conservation and sequence patterns, as well as specialized models designed for epigenetic or variant-level tasks.

DNA-sequence foundation models:

Table 12: Window-size ablation on local signal tasks (AUC). Δ reports relative change from $w = 128$ to $w = 1024$.

Task	w=128	w=256	w=512	w=1024	Δ
TF Binding	0.827	0.830	0.831	0.834	+0.85%
Promoter	0.871	0.874	0.875	0.878	+0.80%
Splice Site	0.943	0.946	0.947	0.949	+0.64%
Chromatin	0.854	0.857	0.858	0.861	+0.82%

- **DNABERT-2 Zhou et al. [2024]**: DNABERT-2 is a BERT-based DNA language model pretrained on multi-species genomes using byte-pair tokenization. It replaces fixed k-mer vocabulary with a learned subword tokenizer, reducing vocabulary size and improving handling of rare or variable-length genomic motifs. The model is trained with masked language modeling and is widely used as a baseline for sequence-level regulatory and variant-effect tasks.
- **Nucleotide Transformer (NT-2.5B) Dalla-Torre et al. [2025]**: The Nucleotide Transformer is a large-scale genomic language model built on the Transformer architecture and pretrained on human and multi-species reference genomes. The 2.5B parameter variant provides strong baselines for chromatin, splicing, promoter, and viral task groups, leveraging extensive pretraining data and large model capacity to capture regulatory sequence patterns.
- **Enformer Avsec et al. [2021]**: Enformer processes 200 kbp single-sequence inputs through a seven-layer CNN that downsamples the sequence to 128 bp resolution. It then applies an 11-layer Transformer for global self-attention, expanding the receptive field from 20 kbp to 100 kbp. The model uses multi-task learning to predict multiple human epigenomic and gene-expression tracks.
- **SPACE Yang et al. [2025b]**: SPACE uses supervised genomic track prediction as a pre-training objective and employs a sparse mixture-of-experts (MoE) architecture to learn both species-specific and cross-species regulatory features. The encoder maps sequences to species-aware hidden representations through species embeddings and Top-K routing. The decoder groups outputs by track type and uses dual-gated expert aggregation to capture cross-track dependencies.

Structural-variant models:

- **Cue Popic et al. [2023]**: Cue is a deep-learning framework for structural variant discovery and genotyping that treats SV detection as a breakpoint-classification problem over read-depth and split-read signals. It learns to distinguish true structural variants from sequencing artifacts by modeling local coverage patterns and discordant read alignments around candidate breakpoints.
- **SVision Lin et al. [2022]**: SVision is a specialized structural-variant detection model that integrates multi-source sequencing evidence, including split reads, discordant pairs, and read-depth signals, to identify deletion, duplication, and inversion events. It is optimized specifically for SV calling and serves as a strong specialized baseline for breakpoint-level detection tasks.

Epigenetic model:

- **MethylBERT Jeong et al. [2025]**: MethylBERT adapts the BERT architecture for read-level methylation modeling. Its inputs include 3-mer DNA sequences with corresponding CpG methylation states (0/1/2 labels). It uses masked language modeling for pretraining and is fine-tuned to output posterior probabilities for downstream tissue-state prediction. The model integrates maximum likelihood estimation with a Fisher information matrix for tumor purity estimation.

A.12 UGM Attention Matrix Analysis

This section provides an attention matrix visualization for exploratory cross-position dependency analysis. The heatmap shows attention weights between positions, where darker red indicates stronger

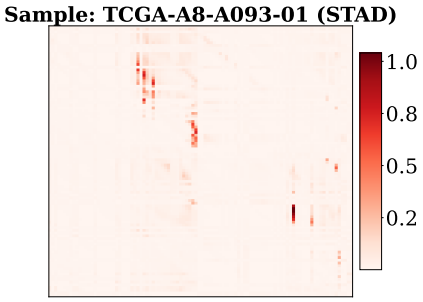


Figure 7: UGM reference-model attention matrix for exploratory cross-position dependency analysis.

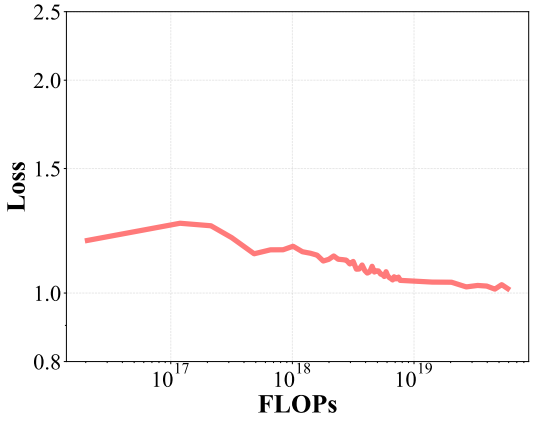


Figure 8: UGM-100M training loss versus pretraining FLOPs.

attention. Off-diagonal weights illustrate attention between nonadjacent positions within the displayed window. The visualization uses an illustrative sample with a 100-position window for qualitative inspection; it does not establish a biological regulatory mechanism or genomic distances between the displayed positions.

A.13 Scaling Analysis

Figure 8 plots the reported training loss of UGM-100M against floating-point operations (FLOPs) during pretraining. The curve provides a descriptive view of optimization as compute accumulates. The displayed loss decreases with increasing compute. No fitted power-law parameters or goodness-of-fit statistics are reported, so this curve is not used to establish a quantitative scaling law or extrapolate performance to larger models.

A.14 Limitations

Several limitations remain. The parameter scaling study covers three model sizes and a representative subset of downstream tasks; larger models and broader task coverage are needed to establish more precise genomic scaling laws. UGM uses a standard encoder backbone, leaving room for architectures designed specifically for long-range genomic dependencies. The current corpus is also primarily human-centric, which may limit transfer to species-specific regulatory tasks. Future work will expand cross-species event coverage, incorporate additional regulatory modalities, and investigate architectures that better capture distal genomic interactions.

Sample 1: Base Events (Reference & SNPs)

```
Source: GRCh38 Reference Genome + dbSNP (Germline)
* Base events include Reference backbone and single nucleotide polymorphisms.

- chr1:1,000,004: Reference G -> Variant A (SNP) -> Token: GA

ref  G G T G G A G C G
var  G G T G A A G C G  <-- SNP (G->A) is a Base Event
```

Sample 2: Atom Events (Single TCGA Methylation Sample)

```
Source: TCGA-BRCA-0001 (Individual Tumor Sample)
* All methylation states below are from this specific patient

- chr14:93347430 : High-methylation (beta=0.98 -> H)
- chr14:93347435 : Medium-methylation (beta=0.58 -> M)
- chr14:93347440 : Low-methylation (beta=0.03 -> L)

ref  C   C   C
var  H   M   L  <-- Patient-specific Epigenetic State
```

Sample 3: Molecular Events (Single 1000G Individual)

```
Source: 1000G Individual HG00096 (Whole Genome Sequencing)
* Variants sampled strictly from HG00096.
* Focuses on small INDELs (<50bp). SNPs are handled in Base Events.

- chr1:10390: 2-bp deletion (Specific to HG00096)

ref  A A C C C ...
var  C - - C C ...  <-- Small Deletion (Molecular Event)
```

Sample 4: Fragment Events (Single Structural Variant Source)

```
Source: 1000G Individual HG00154
* Large structural deletion (>50bp) using sparse boundaries.
* Intermediate positions are NOT tokenized (left empty).

- Start: chr2:15678 (Ref: A) -> Token: AD^s
- End:   chr2:16678 (Ref: G) -> Token: GD^e

ref  A    T  C  G  ... [1000bp gap] ...  A  T  C  G
var  AD^s                                GD^e
```

Figure 9: **Real samples.** Each training instance is constructed from a **single data source** to prevent locus conflicts. **Sample 1** illustrates Base Events (Reference and SNPs). **Sample 2** shows Atom Events (Methylation). **Sample 3** captures Molecular Events (INDELs), while **Sample 4** depicts Fragment Events (Structural Variants) using sparse boundaries.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction present UGM as a genomic foundation model based on coordinate-aligned event states, event-balanced pretraining, and Site-Level Joint Prediction. The experiments report task-dependent strengths and limitations.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The appendix includes a Limitations section discussing the human-centered corpus, the scope of the model-scaling study, and architectural limitations for long-range dependencies. The main text discusses the task-specific scope of sparse SV boundary encoding.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: The paper introduces a genomic pretraining framework and reports empirical evaluations; it does not present theoretical results requiring proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No].

   Justification: The paper reports data sources, event vocabulary, architecture, and major training settings, but does not yet provide every implementation detail needed for full reproduction, including multi-base INDEL encoding and the exact sampling construction.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The paper does not provide a public code and data release with complete reproduction instructions. Some genomic resources require controlled access.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The experimental setup specifies evaluation metrics, the reported run convention, fine-tuning settings, model configurations, and major pretraining hyperparameters.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

   Justification: Several tables report mean $\pm$ standard deviation, but functional variant classification and component ablations do not report variability. The paper therefore does not consistently provide uncertainty estimates for all main comparisons.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: The setup specifies GPU type, memory, and batch size for UGM pretraining, but execution times and complete per-experiment compute costs are not reported.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses existing public or access-controlled genomic resources. Access permissions and data-use restrictions must be respected when reproducing the experiments.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No].

    Justification: The current manuscript does not include a dedicated discussion of both positive societal impacts and potential negative impacts such as genomic privacy, population bias, or clinical misuse.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes].

    Justification: The paper describes the use of controlled-access genomic resources. Any release of derived data or models must follow the applicable data-use restrictions.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No].

    Justification: Major data and model sources are credited, but the manuscript does not provide a complete asset-by-asset account of licenses and terms of use.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No].

    Justification: The manuscript describes UGM's vocabulary, pretraining procedure, and evaluations, but does not provide a complete accompanying release and documentation for code, model weights, or derived datasets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not conduct crowdsourcing or recruit human subjects. It uses existing public or controlled-access genomic resources governed by their original access policies.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The work does not involve new human-subject recruitment or intervention. Existing controlled-access resources are used only under their corresponding access agreements, and restricted raw data are not redistributed.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs are not used as an important, original, or non-standard component of the model design, training, or evaluation methodology.