arXiv:2611.29395v1 [cs.LG] 22 Nov 2026

BQDNA：用于基因组序列建模的多头二进制球面量化

王旭东[2,*]、程思维[1,*]胡浩哲[2,3,†]

[1]中国科学技术大学[2]数字孪生研究院，宁波东方理工大学[3]上海交通大学

摘要

从漫长且高度重复的脱氧核糖核酸中解码生物信息仍然具有挑战性。基因组语言模型（基因组语言模型）通常遵循传统的语言模型预训练-适配范式。然而，由于缺乏内在的语义分割和先验规则，k-mer 和字节对编码等手工设计的分词器并不理想。我们提出了用于脱氧核糖核酸标量量化的上下文感知 **BQVAE** ，以及支持多头编码簿的自回归基因组语言模型 **BQDNA**。综合实验表明：（1）脱氧核糖核酸空间中的向量量化具有独特的训练动态，包括训练语料与下游数据之间对齐的重要性。（2）BQVAE 的性能优于手工设计的分词器和基础 VQVAE。（3）BQDNA 具有卓越的跨物种泛化能力：仅通过 0.1B 的多物种 BQVAE 训练以及在 hg38 上进行 1 个 epoch 的 BQDNA 训练，即可达到最先进的性能，超越使用10∼100x 更多词元进行预训练的基因组语言模型。我们的研究结果凸显了学习式向量量化在基因组序列表示方面的潜力。代码可在 https://github.com/Sophie10001b/BQDNA 获取。

索引词— 基因组语言模型，DNA 分词

1. 引言

脱氧核糖核酸（DNA）由 4 种碱基 {A、C、G、T}组成，是编码生物体全部遗传信息的蓝图。鉴于 DNA 在生物语言中的基础性作用，人们已开发出众多启发式算法，用于 DNA 组装、注释以及与疾病和药物相关的高阶关联挖掘 [1, 2]。与此同时，基因组数据库和调控元件资源库（例如启动子、增强子和表观遗传标记） [3–5] 的持续扩展，推动了数据驱动的深度学习方法在基因组和分子分析中的广泛应用；这些方法已在多个下游任务中展现出优于传统启发式算法的性能 [6–9]。大规模自监督学习进一步推动了基因组语言模型预训练，即基于未标注的参考序列进行预训练，以服务于下游基因组应用 [10]。

随着大语言模型（LLMs）和基因组语言模型（gLMs）的快速发展，合适的分词仍然是连接原始基于字符的语料库与语言模型词元化输入的关键桥梁 [11, 12]。除了最简单的 k-mer 分词（$k \geqslant 1$，其中 $k = 1$ 采用基于单个核苷酸的分词方式），即将每 k 个相邻的核苷酸字符视为一个词元 [10, 13–16], ，近期的基因组语言模型已采用字节对编码（BPE），通过迭代合并高频子词



图 1。展示了在有无周围序列上下文的情况下，四头 BQVAE 分词的示例。

将其组合成更大的词元 [17, 18], ，并进一步探索可学习的分词层，将其作为更细粒度且更具泛化能力的预处理模块 [19, 20]。然而，传统的 k-mer 分词和基于统计的方法存在两个根本局限：（1）它们无法有效融入特定领域的生物学先验；（2）作为简单的预处理操作，它们难以被引导去捕捉下游任务所需的语义模式。尽管近期基于向量量化的方法（如 VQDNA）引入了残差量化（RQ） [21]，以增强序列表示，但在基因组语言模型预训练中，针对 VQVAE 训练所用语料库、序列下采样和多头码本的具体研究与细节仍然缺失，因此基于向量量化的基因组语言模型的潜力尚不明确。

基于标量量化（SQ）在图像离散化中的成功 [22], ，我们提出了一种可学习的 DNA 分词与预训练框架，包含三项贡献。(1) 我们开发了**BQVAE** ，采用多头二元球面量化（BSQ） [23]来扩展并丰富核苷酸序列的编码空间。图 1 展示了 BQVAE 如何根据周围的序列上下文，为同一核苷酸片段分配不同的多头码。(2) 我们构建了**BQDNA**，这是一种自回归基因组语言模型，可并行预测多头码本索引，从而基于 BQVAE 表征进行基因组预训练。(3) 我们通过广泛的实验验证了该框架的有效性和训练效率。与传统的手工设计的分词器和基础 VQVAE 相比，BQVAE 表现更优；同时，仅需使用来自多物种训练语料库的 0.1B 个碱基对，即可将跨物种先验迁移至下游基因组语言模型。在更多

在超过 50 个下游数据集上的结果表明，BQDNA 的表现与基线方法相当或更优，而所需的预训练预算仅为其 2%。

*共同贡献。†通讯作者。

2. 使用 BQDNA 建模基因组

给定一个脱氧核糖核酸序列$\mathbf{S}$，BQVAE 首先将其映射为潜在特征$\mathbf{Z} = \mathcal{E}(\mathbf{S})$，并在 H 个头上独立量化这些特征，然后通过 $\mathcal{D}(\cdot)$ 进行重构。我们没有采用传统 VQVAE [24]，中使用的最近邻码本查找，而是采用二进制球面量化（BSQ） [23]:

$$\hat{\mathbf{z}}_i^h = \ell_2\text{-Norm}\left(\text{sign}\left(\ell_2\text{-Norm}(\mathbf{z}_i^h)\right)\right). \tag{1}$$

$d_e = 256$ 编码器特征被投影到 $D' = 48$ 量化维度，并拆分为 $H = 4$ 个头，每个头包含 $b = D'/H = 12$ 位。每个头定义 $K = 2^b = 4096$ 个编码，无需显式查找或优化码本。

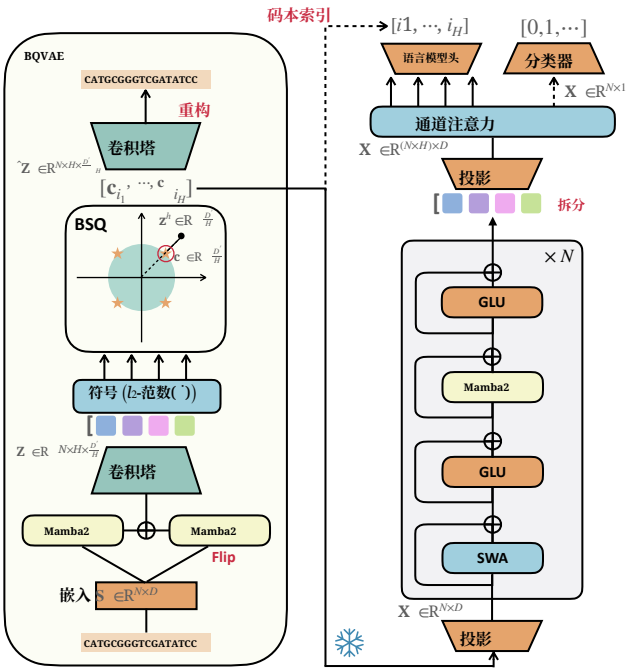


图 2。BQDNA 概览。在 BQDNA 的预训练阶段，BQVAE 会被冻结，将原始序列转换为码本索引，以进行因果语言建模。

与 VQDNA 和基于 VQVAE 的图像生成方法类似，BQDNA 采用两阶段预训练范式。在第一阶段，BQDNA 在功能元件数据集上训练通用 BQVAE，这些数据集遵循与下游任务类似的分布，将原始序列映射到丰富的离散潜在空间，以供后续基因组序列建模使用。在随后的阶段，冻结的 BQVAE 作为独立分词器，将预训练语料离散化为码本索引，语言模型（LM）直接在这些索引上执行因果语言建模（CLM）。图 2 展示了 BQDNA，以下段落将对此进行详细说明。

序列的 BQ 分词。图像分词器通常使用辅助训练目标 [25]。24 位 RGB 像素有 2^{24} 种可能状态，而一个 DNA 碱基有 2^2 种。使用 H 个有序头和一个大小为 K的共享码本$\mathbf{C}$，每个潜在位置最多有 K^H 种不同的编码元组。这种多头设计在保持每个头的预测词汇表精简（包含 K 个类别）的同时，扩展了编码组合空间。为了学习具备上下文感知能力的表示，并实现均衡的码本利用率，我们结合重建目标与基于熵的辅助损失，遵循 LFQ 和二进制球面量化（BSQ） [23, 26]:

$$\mathcal{L}_{BQ} = \mathcal{L}_{rec} + \beta\mathcal{L}_{entropy}$$
$$\mathcal{L}_{entropy} = \mathbb{E}(H(\hat{\mathcal{Q}}(\mathbf{c}|\mathbf{z}))) - H(\mathbb{E}(\hat{\mathcal{Q}}(\mathbf{c}|\mathbf{z})))$$
$$\hat{\mathcal{Q}}(\mathbf{c}|\mathbf{z}) = \text{softmax}(\mathbf{z}\mathbf{c}^\top/\tau)$$

其中 $H(\cdot)$表示熵计算，例如，

$$-\hat{\mathcal{Q}}(\mathbf{c}|\mathbf{z}) \odot \log \hat{\mathcal{Q}}(\mathbf{c}|\mathbf{z}),$$

$\mathbb{E}(\cdot)$表示给定分布下的期望值，在此情况下类似于缩减求和；β 和 τ 分别是基于熵的损失和距离计算的超参数。 $\mathbb{E}(H(\hat{\mathcal{Q}}(\mathbf{c}|\mathbf{z})))$用于计算样本级熵，而 $\mathcal{L}_{entropy}$ 中的第二部分用于计算批次级多样性。

在我们的实验中，对于采用熵正则化的 EMA 更新的 VQVAE 和二进制球面量化变体，重建损失 $\mathcal{L}_{rec}$ 均在前 1k 步内收敛。受图像自编码器中重建质量与下游效用之间差异的启发 [27]，，我们优先考虑 BQVAE 训练语料的功能性内容，以支持后续的基因组建模。具体而言，我们从 Gener Tasks 基因分类数据集中筛选了长度各异的多物种序列 [10]，，而非使用包含大量重复序列的完整参考基因组。这样的选择使分词器训练聚焦于与下游基因组注释相关的序列模式。

在 $\mathcal{O}(N)$ 中预测多头码本索引。在预训练期间，按原始长度高效处理多头潜在向量颇具挑战。为将采用碱基层级多头码本检索进行长序列建模的复杂度限制在 $\mathcal{O}(N)$，BQDNA 采用混合架构，交替使用滑动窗口注意力（SWA） [28]（因果建模采用 1k-bps 窗口，下游双向分类采用 512-bps 窗口） [29], 以及线性注意力模块 [30–32]，；后者直接采用广泛使用的 Mamba2，并通过数据依赖的记忆衰减确保简洁高效 [33–38]。此外，借鉴图像生成中多头 VQVAE 的形式 [21]，，我们在投影头之前设置了一个通道级注意力层，以实现所有码本头之间的通道级特征交互 H，从而明确指导模型从每个融合 token$\mathbf{x}$ 预测 H 码本索引 $\hat{\mathbf{y}}$：

$$\mathbf{x} \in \mathbb{R}^D = [\mathbf{x}^1, \cdots, \mathbf{x}^H], \text{where } \mathbf{x}^H \in \mathbb{R}^{\frac{D}{H}}$$
$$\hat{\mathbf{y}} = [\hat{\mathbf{y}}^1, \cdots, \hat{\mathbf{y}}^H] = \text{Attn.}([\mathbf{x}^1, \cdots, \mathbf{x}^H])$$

由于我们多头 BQVAE 中每个头的潜在信息彼此独立，因此通道级注意力没有显式的位置编码。

训练细节。训练分为两个阶段。首先，我们在 Gene Task 基因分类数据集上训练一个含 170 万参数的 BQVAE，共训练 3 个 epoch（约 9k 步），其编码器特征维度为 $d_e = 256$，量化维度为 $D' = 48$，码本大小为 $K = 4096$，并使用 $H = 4$个注意力头。其双向编码器包含两个 Mamba2 模块，而对称的一维卷积塔则执行可选的下采样和上采样。随后，我们冻结 BQVAE，并使用因果语言建模对一个含 1.1 亿参数的 BQDNA 进行预训练。骨干网络的隐藏维度为 $D = 512$，采用 24 个全注意力层（注意力。）或 12 个 SWA 层和 12 个 Mamba2 层（混合）。所有变体均在人类 hg38 上预训练 1 个 epoch（约 24k 步）。标准变体使用 2^{10} bp 输入，而 8× 下采样变体则处理 2^{13} bp，以在相同训练预算下生成 2^{10} 个潜在标记。

表1。在 GUE 和 Nucleotide Transformer（修订版）基准测试上进行 3 次运行的马修斯相关系数（%）平均表现；排名第一和第二的表现分别以**粗体**和下划线标出。“–”表示模型在 3 次运行中均未能在该数据集上收敛。由于标有“*”的 DNABERT 和 NT-500M-1000G 无法在 GUE 和 NT 基准测试上进行微调，因此引入 DNABERT-2 报告中的基准测试结果作为比较基线。

模型	训练词元	GUE 基准测试							NT 基准测试				
		EMP	TFM	TFH	PD	CPD	SSP	**Avg.**	EMP	ED	PD	SSP	**Avg.**
DNABERT（6-mer）*	122B	49.10	56.43	64.17	81.70	71.81	84.07	60.69	–	–	–	–	–
DNABERT-2	262B	56.04	**73.68**	63.38	81.77	67.02	88.01	65.93	**56.79**	51.66	79.21	86.29	64.87
NT-500M-1000G*	50B	47.68	49.31	58.92	86.58	69.13	80.97	58.00	–	–	–	–	–
NTv2-100M	300B	51.79	65.68	61.23	88.87	73.19	87.84	63.95	55.12	55.35	81.32	96.75	66.45
NTv2-500M	300B	57.63	71.59	67.36	**91.39**	**74.74**	90.49	68.89	56.74	**56.53**	**84.01**	**97.06**	**67.98**
GENA-LM	275B	47.01	64.69	64.38	81.12	67.16	76.55	60.63	56.01	51.78	76.58	80.86	63.11
HyenaDNA	5B	47.69	65.40	64.28	78.53	68.76	77.23	60.90	52.70	51.24	79.41	74.36	60.60
Caduceus-Ps	52B	55.76	72.48	66.93	78.02	70.11	84.05	66.05	51.20	47.78	76.02	81.69	60.04
BQDNA													
混合	**3.2B**	**68.93**	72.44	67.56	84.32	70.20	89.79	**71.96**	54.93	49.08	80.17	96.76	65.46
不含通道注意力的混合	3.2B	66.24	71.33	67.76	83.67	70.06	**90.93**	70.75	54.81	50.42	78.40	96.79	65.26
含碱基的混合	3.2B	50.11	69.56	67.89	81.27	70.75	86.93	64.13	50.36	49.60	79.08	92.90	62.15
含 *VQVAE* 的混合	3.2B	52.55	69.86	64.83	81.33	69.77	90.11	64.53	50.34	51.79	80.58	96.26	63.19
注意力	3.2B	56.60	72.17	**67.93**	82.82	69.28	88.17	67.08	55.61	51.15	79.92	96.26	65.94
带下采样的注意力	3.2B	48.25	58.05	59.47	85.47	69.85	79.11	59.82	51.03	46.37	77.71	95.63	62.39

3. 实验

基线。我们将 BQDNA 与若干规模相近的预训练基因组语言模型进行比较，包括近期推出的 1 亿参数 DNABERT-2 [17], 、Nucleotide Transformer v2-100M（NTv2）[16], 和 GENA-LM [18]。DNABERT [13] 和 Nucleotide Transformer-500M-1000G [16], 均在人类基因组上进行过预训练，因此被纳入比较，以评估 BQ 与基础 k-mer 分词的样本效率。HyenaDNA [14], 、Caduceus-Ps [15], 以及不使用 BQVAE 的 BQDNA 作为基础分词基线。我们还纳入了当前最先进的 NTv2-500M, 以进行更有力的比较。由于没有官方的 VQDNA [19] 实现，我们加入了一个采用基础 VQVAE 分词器的 BQDNA 变体作为消融实验。

设置。所有预训练模型均使用 AdamW [39] 进行微调，峰值学习率为 5e-5，前 10% 的步数采用线性预热，且不进行衰减。我们将批量大小设为 32，训练 5k 个更新步数。为确保性能稳定，我们每 200 步评估一次模型，并将排名前 5 的检查点集成为最终结果；使用不同的随机种子重复此过程 3 次。

GUE 基准测试。我们在基因组理解评估（GUE）[17], 的 28 个数据集中选取 27 个对 BQDNA 进行评估，未纳入新冠变异株分类（CVC），因为部分方法超出了我们的 GPU 内存限制。这 6 项任务包括 1 个剪接位点预测（SSP）数据集、6 个启动子和核心启动子检测（PD 和 CPD）数据集、10 个人类和小鼠转录因子预测（TFH 和 TFM）数据集，以及 10 个酵母表观遗传标记预测（EMP）数据集。我们报告马修斯相关系数（MCC）。所选数据集涵盖人类、小鼠和酵母序列，可用于评估跨物种泛化能力。

如表1所示，在评估的模型中，混合型 BQDNA 模型在 GUE 上取得了最高的平均马修斯相关系数。BQDNA 在预训练期间使用的词元数量最少。在采用相同混合骨干网络的情况下，BQVAE 将平均马修斯相关系数从基础分词时的 64.13 和基础 VQVAE 时的 64.53 提高到了 71.96，凸显了其学习得到的词元表示的优势。此外，它的表现也超过了在 GUE 上最强的外部基线 NTv2-500M，后者的平均马修斯相关系数为 68.89。通道注意力进一步将平均马修斯相关系数从 70.75 提高到了 71.96。这一优势在 EMP 上尤为明显：在相同混合骨干网络下，BQVAE 的马修斯相关系数达到 68.93，而基础分词和 VQVAE 分别为 50.11 和 52.55。BQDNA 在 EMP 上稳健的跨物种泛化能力进一步彰显了其学习高效多物种序列量化的强大能力，这表明，经过充分训练的 BQVAE 分词器可以使单物种预训练基因组语言模型很好地泛化到下游多物种任务。8×下采样变体在 GUE 上也超过了采用 6-mer 分词的 NT-500M-1000G，在 EMP 和 TFM 上尤为突出。这些结果表明，BQVAE 优于人工设计的分段方法。

NT（修订版）基准。Nucleotide Transformer [16]引入的第二个基准是评估基因组语言模型的另一项重要参考，主要关注人类基因组中的功能元件和调控元件。NT 基准测试包含 4 项任务和 18 个数据集：表观遗传标记预测包含 10 个数据集，增强子检测（ED）包含 2 个数据集，启动子检测包含 3 个数据集，剪接位点预测包含 3 个数据集，评估指标为马修斯相关系数。我们从训练集中提取 10 号染色体（chr10）上的所有序列，构建留出验证集。如表1右侧所示，BQDNA 的两种变体在预训练词元数量最少的情况下，仍在 NT 基准测试中保持了有竞争力的表现。BQVAE 带来的提升也延伸到了该基准：在使用相同混合骨干网络的情况下，平均马修斯相关系数从基础分词时的 62.15、基础 VQVAE 时的 63.19 提升至 65.46。两个基准上持续取得的提升表明，BQVAE 不仅适用于单一评估套件，也是一种有效的分词器。与全分辨率注意力变体相比，8×下采样变体在 NT 上的马修斯相关系数降幅小于在 GUE 上的降幅，这表明对词元化分辨率的敏感性取决于任务。

冻结嵌入调控分类。为评估表示的可区分性，我们使用下游标签 [41], 在每个冻结 gLM 的均值池化嵌入上训练监督式 XGBoost [40]分类器。我们使用马修斯相关系数作为评估指标，在包含 10 个数据集的 3 项任务上评估所有模型：2 个用于检测人类甲基化修饰的数据集（即 EMP 任务）、4 个用于检测不同细胞系中人类启动子的数据集（PDH），以及 4 个用于多物种启动子检测的数据集（PDM）。如表2所示，在结果完整的模型中，混合 BQDNA 在 EMP 和多物种PDM上的平均马修斯相关系数最高，在全部

表2。在调控分类任务上，使用冻结的基因语言模型嵌入进行3次运行后得到的平均马修斯相关系数（%）；将每个预训练模型的最终隐藏状态输入XGBoost模型，以获得预测结果。“–”表示模型无法处理长度不同的序列和2千碱基对上下文。

模型	EMP	PDH	PDM	Avg.
DNABERT-2	26.09	83.95	51.44	59.38
NTv2-100M	35.50	83.61	55.45	62.72
NTv2-500M	37.17	84.67	56.77	64.01
GENA-LM	40.57	–	59.14	–
HyenaDNA	30.36	84.05	57.78	62.80
Caduceus-Ps	39.31	**86.28**	60.77	66.68
BQDNA				
混合	**41.47**	82.86	**63.74**	**66.94**
注意力	40.06	82.74	63.30	66.43

十个评估数据集上的总体平均值也最高。其在多物种PDM上的马修斯相关系数为 63.74，高于所列最强基线 Caduceus-Ps 的 60.77；其总体平均值为 66.94，而该基线为 66.68。这些结果凸显了 BQVAE 能够为仅在 hg38 上预训练的 gLM 骨干网络提供跨物种表示。这些冻结嵌入结果补充了微调基准，表明 BQDNA 的表示无需针对下游标签调整 gLM 本身，也能支持监督分类。

消融研究。 由于下游骨干网络是从头开始训练的，因此该比较考察了预训练分词器如何影响下游任务中的后续学习，而不依赖于预训练语言模型骨干网络中的知识。

在多个数据集上的基准测试表明，与朴素的碱基分词相比，具备上下文感知能力的 BQ-VAE 仅使用有限的多物种训练语料库，就能向 gLM 主干网络迁移丰富的跨物种知识。为了系统考察训练语料库如何影响 BQVAE 的跨物种迁移能力，我们引入了 3 种 BQDNA 变体，它们使用不同的数据集（Gener 任务、多物种数据集和仅人类参考基因组）训练 BQVAE，并在 GUE EMP 任务上从头开始训练。如图 3(a) 所示的验证曲线所示，与仅使用 hg38 的版本相比，采用多物种语料库的 BQVAE 能通过其分词模式提供信息量显著更高的先验知识，从而在跨物种任务上取得显著的性能提升。这一发现不仅证实了 BQVAE 对训练数据分布的敏感性，也表明，尽管 BQVAE 只需少量 DNA 空间训练数据即可实现无损重建，但谨慎筛选语料库

化可显著提升下游泛化能力。

除了对训练语料进行消融实验外，码本大小也是 BQVAE 的另一项关键超参数。为评估这一权衡，我们报告了从头开始训练的 H3K14ac 测试集性能、H3K14ac 训练期间按步骤平均的码本利用率，以及采用冻结嵌入的 BQVAE 在 NT 基准测试中 4 种不同功能元件（即 NT 生物类型数据集）上的监督式多分类性能，并比较了不同码本大小的影响。图 3(b) 显示，仅增加码本大小并不能持续提升下游马修斯相关系数，因此需要在码本大小和头数之间取得平衡。多头配置实现了更高的码本利用率和更强的功能元件区分能力，说明多个头有助于丰富词元表示。因此，BQDNA 采用 $K = 4096$

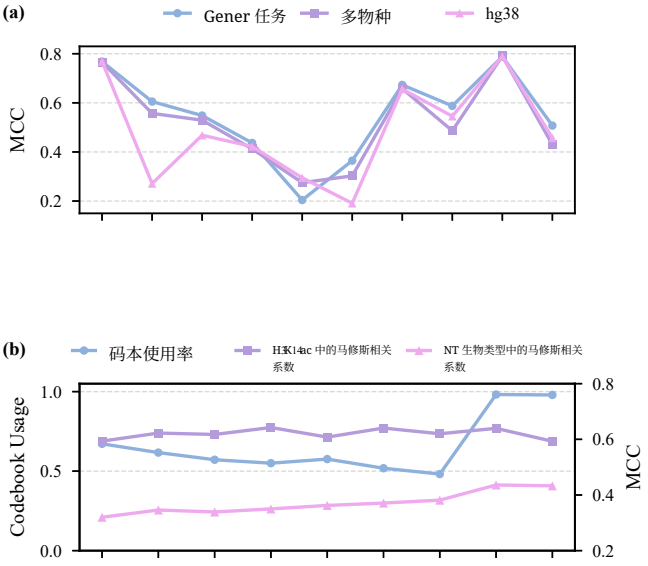


图 3。BQVAE 的综合消融研究。(a) 使用不同语料训练的 BQVAE 在 GUE EMP 任务上的马修斯相关系数。所有混合模型均从头开始训练，而使用仅含人类或多物种参考基因组的 BQVAE 变体训练了 10k 步。(b) 码本大小和头数 ($K \times H$) 对 H3K14ac 马修斯相关系数、码本使用率以及在冻结嵌入上使用监督式 XGBoost 测得的 NT 生物类型马修斯相关系数的影响。

每个头的编码数，包含 $H = 4$个头。

词元分词行为分析。为分析 BQVAE 的分词行为，我们考察了用于训练 BQVAE 的 Gene Task 基因分类数据集。具体而言，我们计算了所有样本中每个码本词元的平均使用频率。作为基线，我们采用 DNABERT-2 中的 BPE 分词器，这是一种基于频率的人工设计方法。对比结果显示，BPE 的词元使用率呈现显著的长尾分布，而 BQVAE 的词元激活分布均衡，仅有极少数低概率离群值。这种分配方式有助于在后续的 gLM 主干网络中实现更全面的词元表示学习，凸显了 BQVAE 在词元使用和进一步扩展码本方面的优势。

4. 结论

本文探讨了多头 BSQ 在上下文感知 DNA 分词中的潜力，并提出了 BQDNA。这是一种基于 Mamba2 和对称卷积塔的双向 BQVAE 量化器，采用 gLM 框架，支持在多头码本上进行自回归预测。对50+个数据集开展的大量实验表明，只需使用规模适当且最小的训练语料进行训练，BQVAE 就能为 BQDNA 的骨干网络提供宝贵的跨物种先验；这使 BQDNA 仅需在 hg38 上进行一个周期的预训练，且训练词元数量仅为基线方法的 2 %，便在下游任务中取得了出色表现。这些基准测试结果进一步凸显了 BQVAE 作为可学习 DNA 分词方法的有效性，以及与手工设计策略和基础 VQVAE 相比显著的样本效率。

5. 遵守伦理规范

本研究使用了公开可用的基因组参考序列和基准数据集。未采集任何新的人类或动物样本。本研究无需额外的伦理批准。

6. 参考文献

[1] C. Perez-Iratxeta、M. Wjst 等，“G2D：一种用于挖掘疾病相关基因的工具”，*BMC Genetics*，2005年8月。[2] E. J. Culp、N. Waglechner 等，“通过进化引导发现抑制肽聚糖重塑的抗生素”，*Nature*，2020年2月。[3] P. Meylan、R. Dreos 等，“EPD in 2020：增强数据可视化并扩展至非编码RNA启动子”，*Nucleic acids research*，2020年。[4] W. Meuleman、A. Muratov 等，“人类DNase I超敏感位点的索引与生物学谱系”，*Nature*，2020年8月。[5] D. Marakulina、I. E. Vorontsov 等，“EpiFactors 2022：扩展和完善人类表观遗传因子及复合物的精选数据库”，*Nucleic Acids Research*，2023年。[6] B. P. de Almeida、F. Reiter 等，“DeepSTARR根据DNA序列预测增强子活性，并支持从头设计合成增强子”，*Nature Genetics*，2022年5月。[7] K. Jaganathan、S. Kyriazopoulou Panagiotopoulou 等，“利用深度学习从初级序列预测剪接”，*Cell*，2019年1月。[8] J. Zhou 和 O. G. Troyanskaya，“利用基于深度学习的序列模型预测非编码变异的影响”，*Na- ture Methods*，2015年10月。[9] H. Hu、Y. Jiang 等，“结合分子图与基于序列的数据增强，提升无模板反应预测”，载于*CIKM*，2023年。[10] W. Wu、Q. Li 等，“GENERator：长上下文生成式基因组基础模型”，arXiv预印本 arXiv:2502.07272，2025 年 2 月。[11] G. Dagan 、 G. Synnaeve 等，“充分发挥分词器在预训练和领域适应中的作用”，载于*ICML*，2024年6月。[12] C. Tao、Q. Liu 等，“词汇表缩放定律：更大的模型值得配备更大的词汇表”，载于*NeurIPS*，2024年11月。[13] Y. Ji、Z. Zhou 等，“DNABERT：用于基因组DNA语言的预训练双向变换器编码器表示模型”，*Bioinformatics*，2021年8月。[14] E. Nguyen、M. Poli 等，“HyenaDNA：单核苷酸分辨率下的长程基因组序列建模”，载于*NeurIPS*，2023年11月。[15] Y. Schiff、C. H. Kao 等，“Caduceus：双向等变长程DNA序列建模”，载于*ICML*，2024年6月。[16] H. Dalla-Torre、L. Gonzalez 等，“Nucleotide Transformer：构建并评估用于人类基因组学的稳健基础模型”，*NatureMethods*，2025年2月。[17] Z. Zhou、Y. Ji 等，“Dnabert-2：面向多物种基因组的高效基础模型与基准测试”，载于国际学习表征会议，2024年。[18] V. Fishman、Y. Kuratov 等，“GENA-LM：面向长序列的开源基础DNA语言模型系列”，*Nucleic Acids Research*，2025年1月。[19] S. Li、Z. Wang 等，“VQDNA：释放向量量化的强大能力，适用于多物种基因组序列建模”载于ICML，2024年6月。[20] L. Qiao、P. Ye 等，“模型决定如何分词：利用MxDNA进行自适应DNA序列分词”，载于NeurIPS，2024年11月。[21] D. Lee、C. Kim 等，“利用残差量化进行自回归图像生成”，载于CVPR，2022年6月。[22] F. Mentzer、D. Minnen 等，“有限标量量化：让VQ-VAE化繁为简”，载于国际学习表征会议，2024年。[23] Y. Zhao、Y. Xiong 等，“利用二进制球面量化进行图像与视频分词”，载于国际学习表征会议，2025年。[24] A. van den Oord、O. Vinyals 等，“神经离散表示学习”，载于NeurIPS，2017年。[25] P. Esser、R. Rombach 等，“驾驭Transformer进行高分辨率图像合成”，载于CVPR，2021年6月。[26] L. Yu、J. Lezama 等，“语言模型胜过扩散模型——分词器是视觉生成的关键”，载于国际学习表征会议，2024年。[27] I. Skorokhodov、S. Girish 等，“提升自编码器的可扩散性”，载于ICML，2025年6月。[28] I. Beltagy、M. E. Peters 等，“Longformer：长文档Transformer”，arXiv预印本 arXiv:2004 05150，2020年。.[29]T. Dao，“FlashAttention-2：更快的注意力机制，以及更好的并行性和工作划分”，arXiv预印本 arXiv:2307.08691，2023年7月。[30] L. Ren、Y. Liu 等，“Samba：用于高效无限上下文语言建模的简单混合状态空间模型”，载于国际学习表征会议，2025年。[31] S. Yang、B. Wang 等，“利用Delta规则沿序列长度并行化线性Transformer”，载于NeurIPS，2024年11月。[32] K. A. Wang、J. Shi 等，“测试时回归：利用联想记忆设计序列模型的统一框架”，arXiv预印本 arXiv:2501.12352，2025年1月。[33] T. Dao 和 A. Gu，“Transformer即状态空间模型：通过结构化状态空间对偶实现通用模型与高效算法”，载于ICML，2024年6月。[34] H. Wu、J. Tong 等，“从数据到模型：多模态大语言模型压缩生命周期综述”，TechRxiv，2026年。[35] H. Hu、H. Wu 等，“超越浮点运算次数：以通用矩阵乘法为中心的分类体系下的大语言模型剪枝实际推理加速基准测试”，arXiv预印本 arXiv:2606.09080，2026年。[36] H. Hu、H. Wu 等，“WIDE：通过令牌级动态宽度剪枝提升自适应大语言模型推理”，arXiv预印本 arXiv:2607.28418，2026年。[37] W. Liu、H. Wu 等，“ViCA：利用仅视觉交叉注意力实现高效多模态大语言模型”，arXiv预印本 arXiv:2602 07574，2026年。.[38] H. Wu、X. Wang 等，“UTPTrack：迈向简洁统一的视觉跟踪令牌剪枝”，载于CVPR，2026年6月。[39] I. Loshchilov 和 F. Hutter，“解耦权重衰减正则化”，arXiv预印本 arXiv:1711.05101，2017年。[40] T. Chen 和 C. Guestrin，“XGBoost：一种可扩展的树提升系统”，载于第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集，2016年。[41] H. Feng、L. Wu 等，“基因组序列分类的DNA基础模型基准测试”，bioRxiv预印本 2024.08.16.608288，2024年8月。