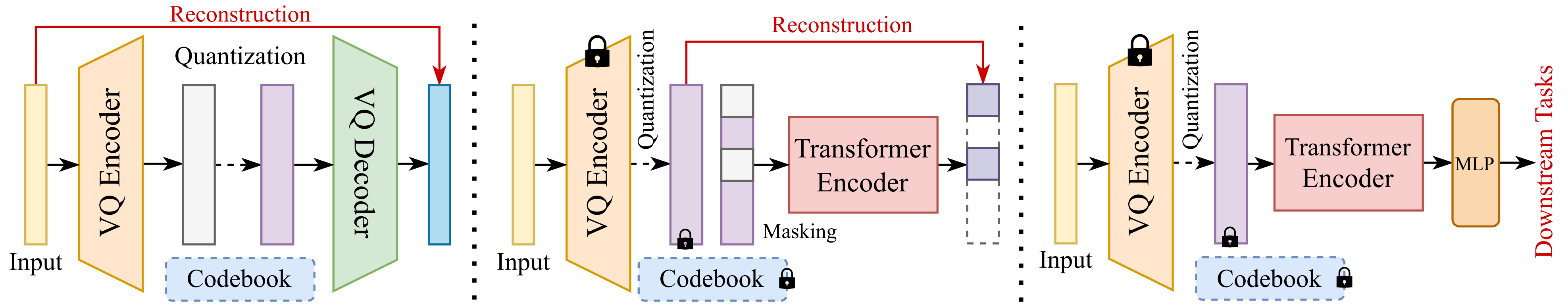
VQDNA: Unleashing the Power of Vector Quantization for Multi-Species Genomic Sequence Modeling

Siyuan Li[1,2,*], **Zedong Wang**[1,*], Zicheng Liu[1,2], Di Wu[1,2], Cheng Tan[1,2], Jiangbin Zheng[1,2], Yufei Huang[1,2], and Stan Z. Li[1,#]

[1] Westlake University, [2] Zhejiang University, [*] Equal contribution [#] Corresponding author

Summary of Contributions

- We introduce genome vocabulary learning framework that learns a genome tokenizer with a discriminative vocabulary for pattern-aware genome language.
- An HRQ tokenizer is designed to progressively enrich the limited genome vocabulary with a hierarchy of varying scales of codebooks in a coarse-to-fine manner.
- Extensive experiments across 32 datasets verify the exceptional generalizability of VQDNA, including an empirical case study on SARS-CoV-2 mutations.



(a) Stage-1: DNA tokenization (b) Stage-2: Masked Code Modeling (c) Stage-3: Fine-tuning with LoRA

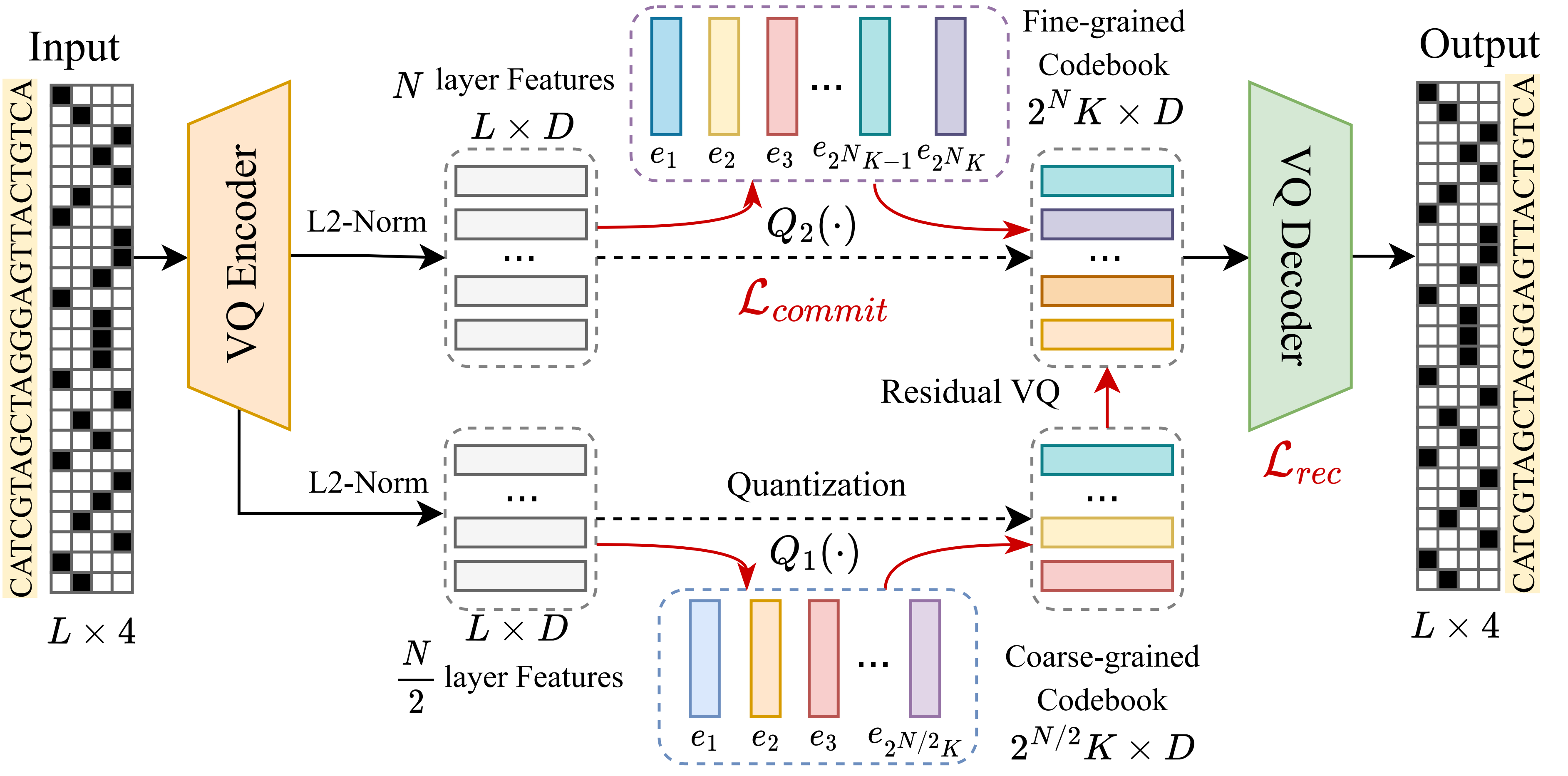
Data-efficient Pipeline for Genome

Three-stage pipeline: (a) genome vocabulary learning, (b) masked code modeling, (c) parameter-efficient fine-tuning.

- **Semantic Vocabulary:** Expanding 4 nucleotides to 4K learnable words (VQ dictionary). Using VQVAE as the baseline version of VQDNA, quantize embeddings by the code mapping function $Q(.,.)$ with a codebook with K words, $\mathcal{C}=\{(k,e(k))\}_{k\in[K]}$, where $e(k)\in\mathbb{R}^d$:

$$M_i=\mathcal{Q}(Z_i;\mathcal{C})=\text{argmin}_{k\in[K]}\|Z_i-e(k)\|_2$$

$$\mathcal{L}_{\text{VQ}}=\underbrace{\mathcal{L}_{CE}(X,\hat{X})}_{\mathcal{L}_{\text{rec}}}+\underbrace{\|\text{sg}[Z]-\hat{Z}\|_2^2}_{\mathcal{L}_{\text{code}}}+\beta\underbrace{\|Z-\text{sg}[\hat{Z}]\|_2^2}_{\mathcal{L}_{\text{commit}}}$$

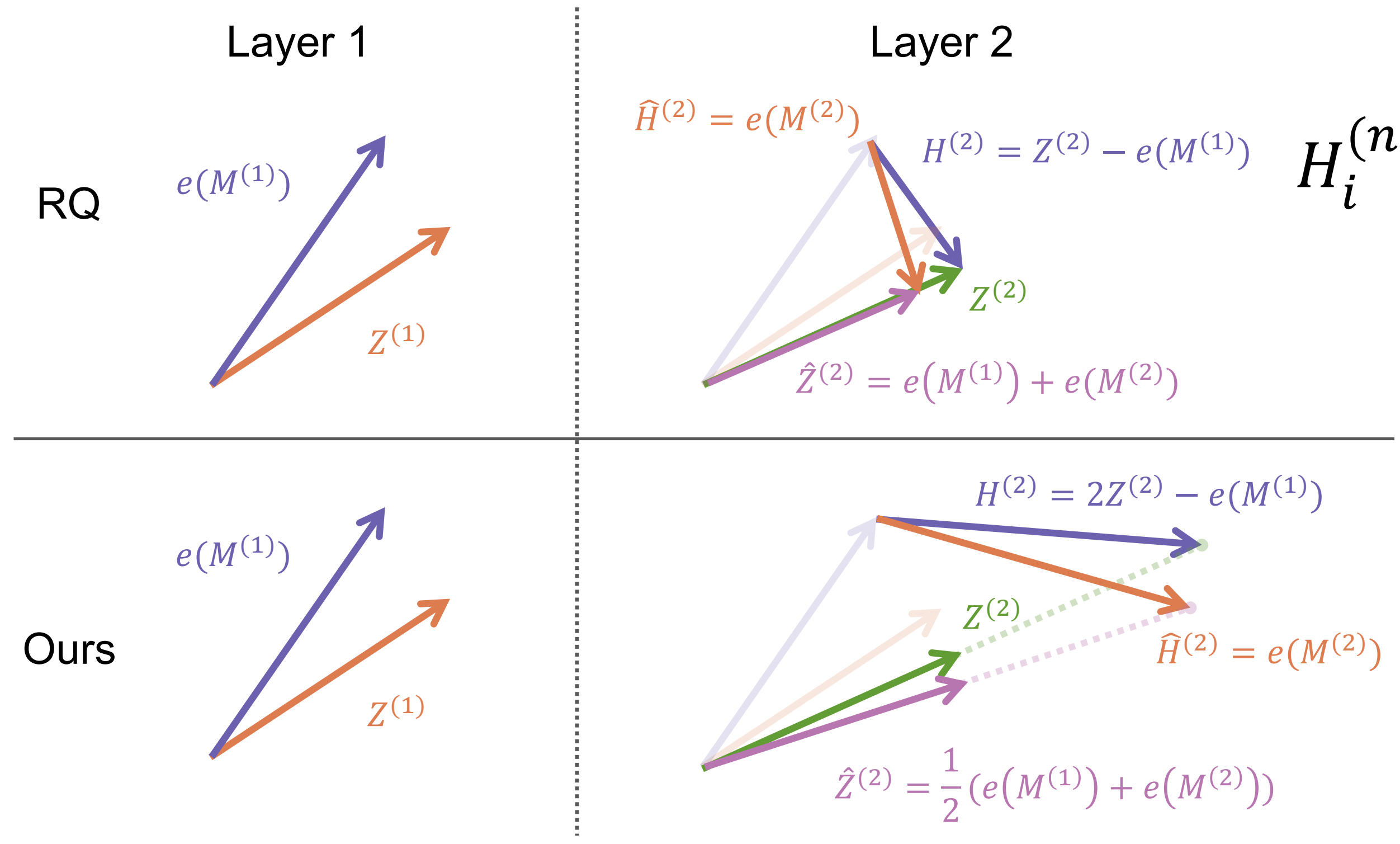


VQ-based Generation: Designing HRQ (hierarchical residual quantization) with two coarse-to-fine codebooks. The advanced hierachical codebooks merging by vector subtraction instead of addation proposed by RQ:

$$M_i^{(n)}=Q\left(H_i^{(n)};C^{(n)}\right),$$

$$M_i^{(n)}=\text{argmin}_{k\in[2^nK]}\left\|H_i^{(n)}-e\left(k^{(n)}\right)\right\|_2,\ M_i^{(n)}\in[2^nK]^L$$

$$\mathcal{L}_{\text{HRQ}}=\mathcal{L}_{CE}(X,\hat{X})+\beta\sum_{n=1}^{N}\left\|Z^{(n)}-\text{sg}[\hat{Z}^{(n)}]\right\|_2^2$$


Layer 1
Layer 2
RQ
$e(M^{(1)})$
$Z^{(1)}$
$\hat{H}^{(2)}=e(M^{(2)})$
$H^{(2)}=Z^{(2)}-e(M^{(1)})$
$Z^{(2)}$
$\hat{Z}^{(2)}=e(M^{(1)})+e(M^{(2)})$
Ours
$e(M^{(1)})$
$Z^{(1)}$
$H^{(2)}=2Z^{(2)}-e(M^{(1)})$
$Z^{(2)}$
$\hat{H}^{(2)}=e(M^{(2)})$
$\hat{Z}^{(2)}=\frac{1}{2}(e(M^{(1)})+e(M^{(2)}))$

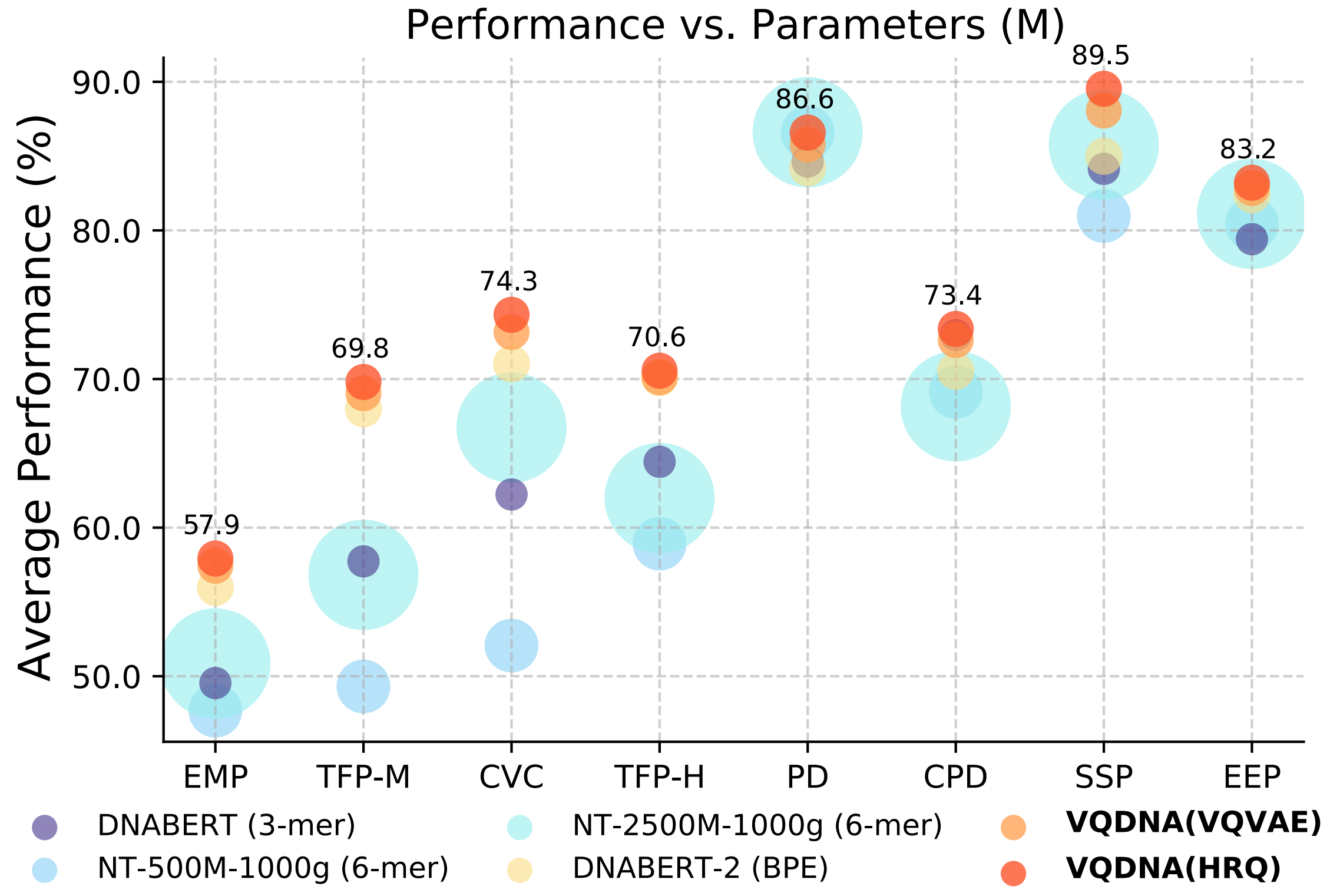
$$H_i^{(n)}=\begin{cases}2Z_i^{(n)}-e\left(M_i^{(n-1)}\right),\\ e(M_i^{(1)}), & n=1\end{cases}$$

Code dim.	VQDNA		+HRQ	
	Rec.	Lin.	Rec.	Lin.
256	99.4	44.3	99.5	48.2
384	99.5	**44.8**	99.6	**48.9**
768	99.6	44.6	99.6	**48.9**
1024	99.8	44.7	99.7	48.8

Ablation of codebook dim.

Experiment Results

- Comparison experiments across 32 genome datasets.



Epigenetic Mark Prediction (EMP)
Transcription Factor Prediction (TFP)
TFP with Human or Mouse
Covid Variants Classification (CVC)
Promoter Detection (PD)
Core Promoter Detection (CPD)
Splice Site Prediction (SSP)
Editing Efficiency Prediction (EEP)
Species Classification (long-seq)

Method	Date	Tokenizer	# Params. (M)	FLOPs (G)	Train (B)	Average Rank
DNABERT	BioInfo'2021	3-mer	86	3.3	122	5
NT-500M-1000g	arXiv'2023	6-mer	480	3.2	50	6
NT-2500M-1000g	arXiv'2023	6-mer	2537	19.4	300	4
DNABERT-2	ICLR'2024	BPE	117	1.0	262	3
VQDNA	**Ours**	**VQVAE**	86+16	1.1+0.5	262	<u>2</u>
VQDNA	**Ours**	**HRQ**	86+17	1.1+0.6	262	**1**

Average performance ranking on GUE datasets.

Method	Tokenizer	Usage	Lin.	FT
DNABERT	6-mer	47	23.54	55.50
NT-2500M-1000g	6-mer (non)	47	23.54	66.73
HyenaDNA	one-hot	100	5.47	54.10
DNABERT-2	BPE (6-mer)	99	36.53	71.02
VQDNA	**VQVAE**	100	<u>44.76</u>	<u>73.16</u>
VQDNA	**HRQ**	100	**48.87**	**74.32**

Analysis of tokenization efficiency.

Code size	VQDNA		+HRQ	
	Rec.	Lin.	Rec.	Lin.
128	98.2	42.1	98.4	42.8
256	98.8	43.6	99.1	47.7
512	99.5	**44.8**	99.6	**48.9**
1024	**99.6**	44.5	**99.8**	48.2

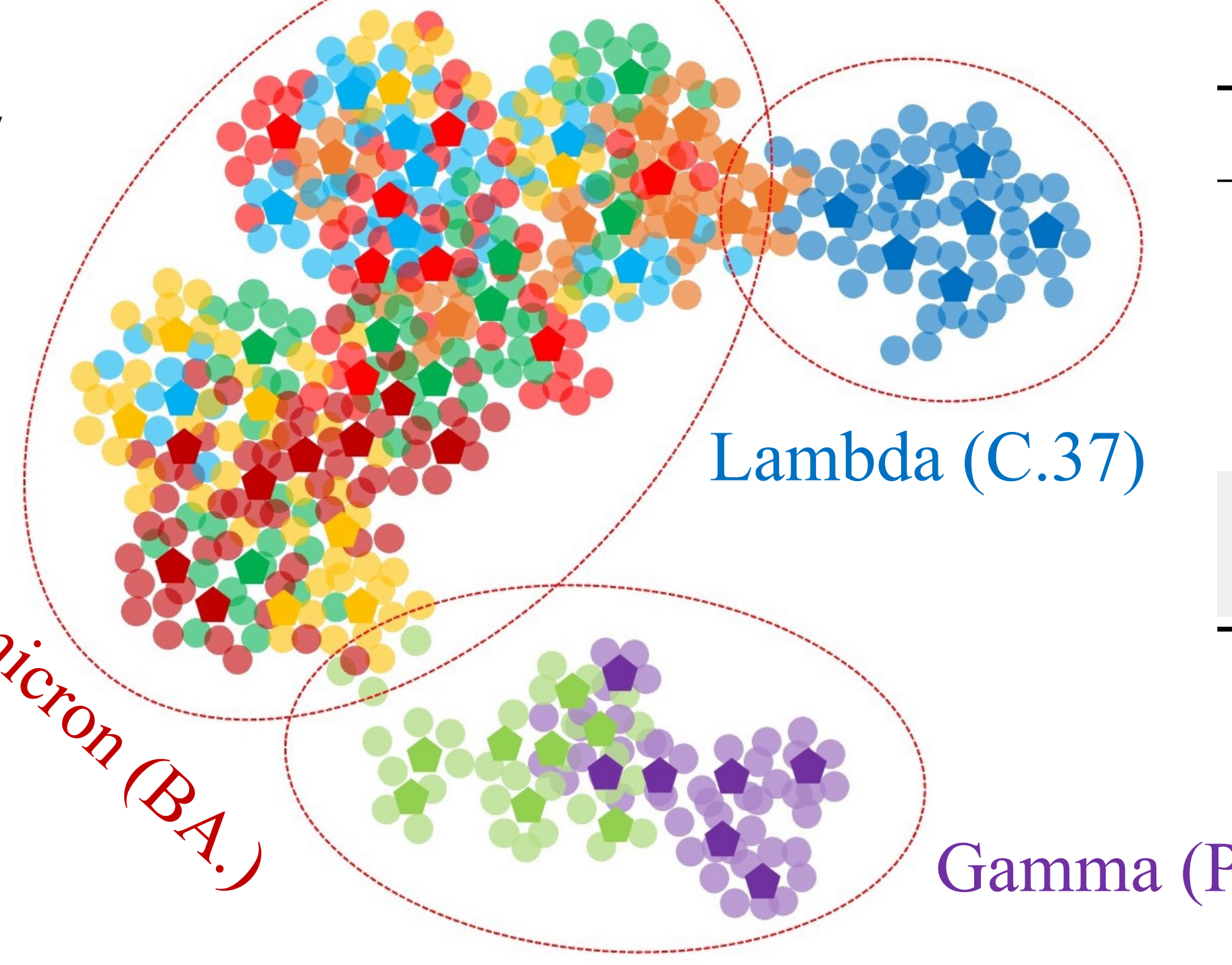
Code dim.	VQDNA	
	Rec.	Lin.
256	99.4	44.3
384	99.5	**44.8**
768	99.6	44.6
1024	99.8	44.7

Method	1k	20k	32k	250k	450k
HyenaDNA	61.13	87.42	93.42	97.90	99.40
DNABERT	39.61	76.21	91.93	N/A	N/A
DNABERT-2	61.04	86.83	99.28	N/A	N/A
VQDNA (HRQ)	**61.57**	**88.05**	**99.46**	N/A	N/A

Species classification (1k to 1M tokens)

- Case study: SARS-CoV-2

1 Alpha (B.1.1.7)
2 Beta (B.1.351)
3 Delta (B.1.617.2)
4 Eta (B.1.525)
5 Gamma (P.1)
6 Lota (B.1.526)
7 Kappa (B.1.617.1)
8 Lambda (C.37)
9 Zeta (P.2)



Method	CVC
DNABERT	62.23
NT-500M-1000g	52.06
NT-2500M-1000g	66.73
DNABERT-2	71.02
VQDNA	73.16
VQDNA (HRQ)	**74.32**

Virus classification