Contrastive Deep Learning for Variant Detection in Wastewater Genomic Sequencing

Adele Chinda* · Richmond Azumah* · Hemanth Demakethepalli Venkateswara
*Equal contribution
Georgia State University, Atlanta, GA

1. MOTIVATION & BACKGROUND

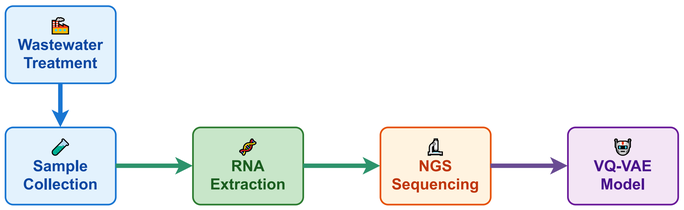
WHY WASTEWATER SURVEILLANCE?

- Non-invasive community-wide viral monitoring
- Detects asymptomatic & pre-symptomatic cases
- Cost-effective alternative to clinical testing
- Early warning system for variant emergence

CHALLENGES:

✗ Highly fragmented reads (100-300 bp)
✗ High sequencing noise & quality variation
✗ Low viral RNA concentration
✗ Multiple co-circulating strains
✗ Traditional pipelines require reference genomes

Wastewater Surveillance Workflow



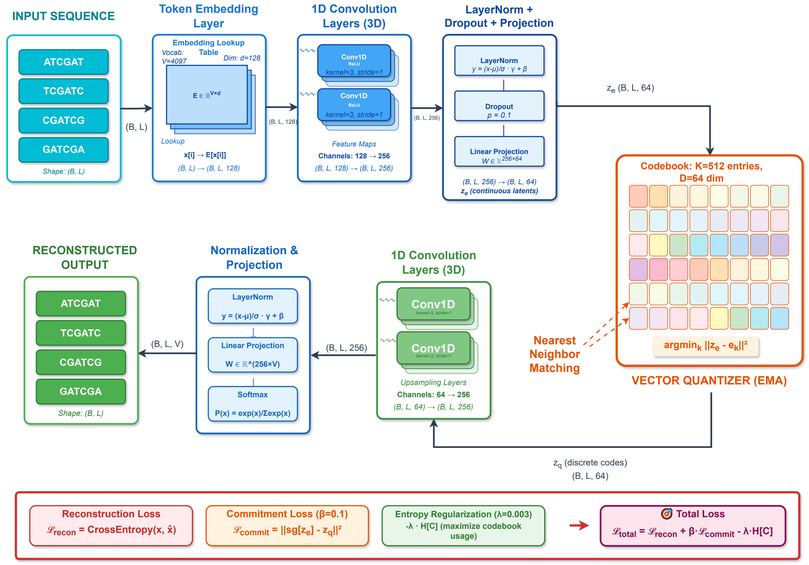
2. DATASET & PREPROCESSING

DATA SOURCE:
• SARS-CoV-2 wastewater sequencing reads
• FASTQ format, variable length (36-300 bp)
• Total sequences: ~100,000 reads

PREPROCESSING PIPELINE:
1. Quality Control (FastQC)
2. Adapter Removal (Trimmomatic)
- Leading/trailing quality: 3
- Sliding window: 4:15
- Min length: 36 bp
3. K-mer Tokenization (k=6)
- Vocabulary size: 4,097 tokens
- Canonical k-mer mapping
- Pad/truncate to 150 tokens

3. METHOD: VQ-VAE ARCHITECTURE

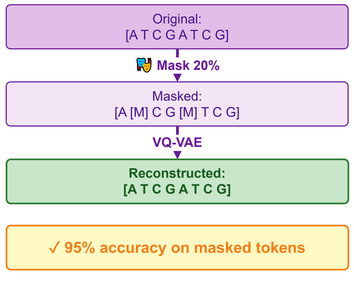
VECTOR-QUANTIZED VARIATIONAL AUTOENCODER



4. EXTENSIONS

A. MASKED VQ-VAE

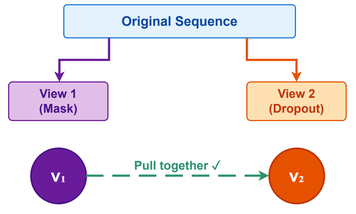
- Randomly mask 20% of input tokens
- Model learns to reconstruct masked regions
- Improves robustness to missing data
- Similar to BERT for genomic sequences



✓ 95% accuracy on masked tokens

B. CONTRASTIVE LEARNING

- Fine-tune encoder with InfoNCE loss
- Generate augmented views (mask + dropout)
- Pull similar sequences together
- Push different sequences apart
- Enhances clustering separability



Silhouette: 0.31 → 0.42 (+35%)

5. RESULTS

QUANTITATIVE RESULTS

Reconstruction Metrics:

- Mean Token Accuracy: **99.52%**
- Exact Sequence Match: **56.33%**
- Codebook Utilization: **19.73%**

Clustering Quality (k=10):

VQ-VAE Contrastive

- Silhouette: 0.31 **0.42**
- Davies-Bouldin: 1.68 **1.34**
- Calinski-H: 1248 **1876**

KEY FINDINGS

✓ VQ-VAE achieves 99.5% reconstruction accuracy
✓ Discrete codebook captures genomic patterns
✓ Entropy regularization prevents collapse
✓ Contrastive learning improves clustering 35%
✓ Reference-free, scalable approach

🌍 IMPACT: Democratizes genomic surveillance for public health monitoring worldwide

6. FUTURE WORK

→ Hierarchical VQ-VAE for multi-scale patterns
→ Integration with phylogenetic analysis
→ Validation on diverse pathogen datasets
→ Real-time surveillance deployment
→ Temporal dynamics modeling

ADVANTAGES OVER TRADITIONAL METHODS:
• No reference genome required
• Computationally efficient (~minutes vs hours)
• Learns meaningful representations
• Robust to sequencing noise

7. REFERENCES

[1] van den Oord et al. (2017). Neural Discrete Representation Learning. NeurIPS.
[2] Crits-Christoph et al. (2021). Genome Sequencing of Sewage. mBio.
[3] Chen et al. (2020). A Simple Framework for Contrastive Learning. ICML.
[4] Abdel-Aziz et al. (2024). VQ-DNA: Discrete Latent Representations. arXiv.

Code: github.com/arrdel/genomic_sequence_detection

CONTRASTIVE LOSS FORMULATION (InfoNCE)

$$\mathcal{L}_{contrast} = -\mathbb{E} \log \frac{\exp(sim(v_i, v_{i'})/\tau)}{\Sigma_{k=1}^{2B} \exp(sim(v_i, v_k)/\tau)}$$

where:

- v_i, $v_{i'}$ = positive pair (same sequence, different views)
- v_k = all samples in batch (2B total)
- $sim(u,v) = u^T v / (||u|| \, ||v||)$ = cosine similarity
- τ = temperature parameter (controls sharpness)

RESULTS & CLUSTERING METRICS

📈 CLUSTERING IMPROVEMENT
Silhouette Score: 0.31 → 0.42
(+35% improvement)

🎯 SimCLR-style Contrastive Learning

💪 Enhances Discriminative Power for Clustering

Alignment ↓
(lower is better)
Positive pairs close together

Uniformity ↑
(higher is better)
Points evenly distributed

Training Stability
✓ Converges in ~30 epochs
✓ No mode collapse
✓ Robust to hyperparams

Acknowledgments: Georgia State University Department of Computer Science

Contact: achinda1@student.gsu.edu | razumah1@student.gsu.edu