FROM K-MERS TO GENOMIC FOUNDATION MODELS: BENCHMARKING COX1 TAXONOMY UNDER EXTREME CLASS IMBALANCE

Luis Valenzuela, Sebastián Aguilera Madrid, Luis Martí & Nayat Shánchez-Pi
Inria Chile Research Center.
Avenida Apoquindo 2827, Las Condes, Santiago, Chile.
{luis.valenzuela,luis.marti,nayat.sanchez-pi}@inria.cl

ABSTRACT

Machine learning for genomics increasingly depends on pretrained genomic foundation models (gLMs) as reusable sequence encoders, yet adoption in biological discovery remains constrained by three linked challenges: tokenization mismatch with biological signal, domain shift between pretraining corpora and downstream assays, and extreme long-tail taxonomic labels that destabilize standard objectives. We study these issues in the ecologically central COI/COX1 gene through an alignment-free benchmark that converts nucleotide sequences into fixed-length embeddings (mean-pooled hidden states) and trains lightweight MLP classifiers for independent rank-wise prediction from Domain to Species. We evaluate two complementary regimes for scalable and interpretable genomic modeling: eKOI (15,947 sequences; protist-rich; 11,047 species) and MetaCOXI (5.6M metazoan sequences; 743,671 species). Across diverse gLM families (autoregressive decoders and masked-language encoders) and explicit compositional baselines (overlapping k-mer frequencies up to k=6), we find that effective motif length induced by tokenization is a dominant driver of fine-rank separability, while corpus alignment (eukaryote- vs. prokaryote-pretraining) materially affects transfer even under identical tokenization. Finally, imbalance-aware objectives (weighted cross-entropy and a hybrid weighted+contrastive loss) can stabilize rare-taxonomy performance but remain representation-dependent.

1 INTRODUCTION

Accurate taxonomic assignment is a cornerstone of modern biodiversity science and ecological monitoring, because it links sequence observations to species distributions, community composition, and downstream ecosystem functions (Wu et al., 2026). In aquatic ecosystems, this linkage is especially consequential for long-term assessments of climate-driven change: plankton and benthic communities regulate food webs and biogeochemical cycles, and sustained observations are essential to detect and attribute ecosystem shifts. Recent syntheses emphasize both the unique value of long-term plankton time series and the need to integrate emerging technologies, including imaging and molecular/omics approaches, alongside standard monitoring to obtain complementary, policy-relevant perspectives on pelagic biodiversity (Holland et al., 2025). In parallel, environmental DNA (eDNA) and metabarcoding provide minimally invasive, scalable routes to quantify community composition from sequencing data, but their impact depends critically on converting reads into reliable taxonomic labels (Wu et al., 2026).

At the same time, genomic language models (gLMs) have emerged as a foundation-model paradigm for DNA: pretrained on massive genomic corpora with self-supervised objectives, they learn contextual sequence representations that can be reused for downstream prediction without explicit alignment. This creates a compelling opportunity for biodiversity monitoring, where the computational bottleneck is increasingly shifting from sequencing throughput to robust and scalable interpretation of millions of barcode reads. Recent work has shown that tokenizer choice in genomic language models can materially affect downstream performance, especially in tasks that depend on fine-grained nucleotide- or motif-level signal, and that the most effective tokenization strategy may

depend on both the biological structure of the sequence and the target task (Lindsey et al., 2025; Eapen, 2025).Mitochondrial cytochrome c oxidase subunit I (COI/COX1; often referred to as COX) has become a *de facto* standard DNA barcode for eukaryotes due to its broad presence, strong discriminatory power across many metazoan lineages, and extensive reference coverage accumulated by the community. As a consequence, COX-based metabarcoding and eDNA surveys are increasingly used to translate short-read sequence data into ecological insights at multiple taxonomic ranks.

In this work, we study these barriers and frontiers through a unified, embedding-based alignment-free benchmark for COI/COX1 taxonomy across ranks from Domain to Species. We evaluate two complementary reference resources that jointly expand scale and phylogenetic breadth: **eKOI**, which extends COI coverage into protist diversity with rapidly increasing label cardinality toward fine ranks (González-Miguéns et al., 2025), and **MetaCOXI**, a compilation of metazoan COI sequences with extreme long-tail structure at genus and species (Balech et al., 2022).

Contributions. (i) We introduce a unified COI/COX1 benchmark spanning protists and metazoans (eKOI and MetaCOXI) with consistent rank-wise evaluation from Domain to Species. (ii) We provide a systematic comparison of pretrained gLM embeddings and k-mer baselines that reveals how tokenization granularity and pretraining-domain alignment govern transfer for mitochondrial barcodes. (iii) We assess imbalance-aware objectives (class reweighting and a hybrid weighted+contrastive loss) under extreme long-tail supervision, highlighting when they stabilize rare-taxonomy performance and when representation quality is the limiting factor. Together, these results clarify practical barriers for GenAI-in-genomics transfer and outline actionable frontiers for hierarchy-aware learning, and task-adaptive finetuning in biodiversity-scale deployments.

2 Related work

Classical COI-based taxonomic assignment in barcoding and metabarcoding pipelines often relies on sequence alignment against reference databases, typically via BLAST-style local alignment (Altschul et al., 1990). While effective, alignment-based search can be computationally expensive at biodiversity scale, sensitive to sequencing errors and indels, and difficult to deploy consistently when reference coverage is incomplete or when short, degraded fragments are analyzed. These limitations have motivated a growing body of alignment-free methods that bypass explicit alignment and instead learn discriminative signals directly from sequence composition. Early alignment-free COI classifiers have traditionally relied on explicit k-mer features. For instance, ALFIE represents COI barcode sequences using k-mer frequency vectors (rather than sequence alignment), and tested using 58,000 publicly available COI sequences. More recently, DEEPCOI proposed a large language model-driven framework tailored to animal metabarcoding, reporting fast taxonomic assignment with evaluation over a benchmark spanning eight animal phyla (Gwak & Rho, 2025).

More recently, gLMs have enabled representation learning directly from raw DNA, where contextual, fixed-length embeddings can be extracted and used as inputs to downstream classifiers (reviewed by (Veiner & Supek, 2026)). Prior efforts show that gLM embeddings can substantially improve alignment-free taxonomy prediction in bacteria (Leske et al., 2025), yet a comprehensive evaluation on COX datasets spanning metazoans and protists remains limited.

3 Proposed Methodology

We adopt a unified, alignment-free evaluation protocol that isolates representation quality by pairing frozen sequence encoders with a lightweight downstream classifier under a standardized preprocessing and training recipe across both datasets and all taxonomic ranks. Figure 1 summarizes the end-to-end pipeline used throughout this work.

3.1 Datasets

To evaluate comparative genomics signals in mitochondrial COI/COX1 for taxonomic inference, we analyze two complementary reference collections that span radically different regimes of scale and eukaryotic diversity. **eKOI** (González-Miguéns et al., 2025) comprises 15,947 curated COI sequences distributed across $\sim$76 eukaryotic phyla with a strong representation of protist lineages.

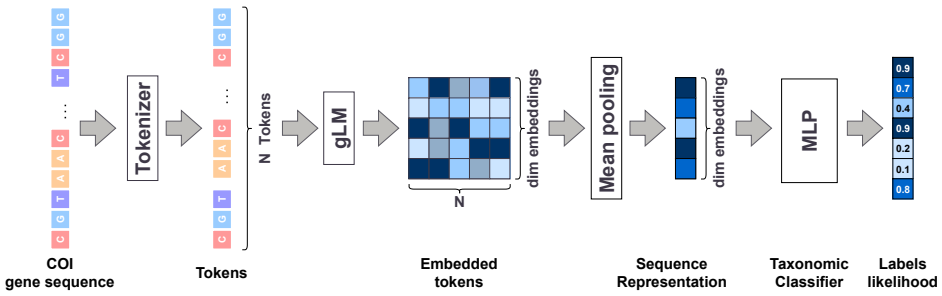

Figure 1: Embedding-based pipeline for COI/COX1 taxonomic classification. A COI sequence is tokenized and encoded by a pretrained genomic language model (gLM) to produce contextual token embeddings. Token-level representations are aggregated into a fixed-length sequence embedding via mean pooling over valid (non-padding) tokens, which is then provided to an MLP classifier to output taxonomic label likelihoods (Domain–Species).

The rapid growth in label cardinality toward genus (5,841) and species (11,047) creates an extreme long-tail setting in which many taxa are supported by few examples. In contrast, **MetaCOXI** (Balech et al., 2022) aggregates 5,608,848 metazoan COI sequences integrated from the European Nucleotide Archive and the Barcode of Life Data Systems, yielding near-complete coverage through higher ranks (e.g., 31 phyla; 100 classes) but exploding to 52,214 genera and 743,671 species, with a marked drop in genus completeness consistent with heterogeneous metadata across sources. For both datasets, we convert each nucleotide sequence into a fixed-length embedding and train supervised models to predict taxonomy from Domain to Species. Further dataset characterization is provided in Appendix A.

Sequence redundancy and split protocol. Because both eKOI and MetaCOXI exhibit extreme long-tail label distributions at fine taxonomic ranks, we prioritized preserving class coverage over applying aggressive redundancy reduction. In the classification benchmark reported here, we did not perform additional 100% identity deduplication within the downstream training pipeline, and train/validation partitions were not defined at the sequence-cluster level. Instead, after removing missing annotations and classes with fewer than two labeled examples for the target rank, we generated fixed 80/20 train/validation splits at the individual-sequence level using a seeded random partition. This design preserves rare taxa that would otherwise be lost under stricter deduplication or cluster-based partitioning, but it also means that the reported results should be interpreted primarily as sequence-level, in-distribution benchmark performance rather than as a strict homology-aware generalization test.

3.2 Genomic language models for COX-based taxonomy

We use pretrained genomic foundation models as frozen feature extractors for COI/COX1 sequences. For each sequence, we obtain contextual token-level hidden states and compute a fixed-length embedding by mean pooling over non-padding tokens, which is then used to train lightweight rank-specific classifiers. Mean pooling is also supported by recent large-scale benchmarking of DNA foundation models, which found that mean token embedding consistently outperformed summary-token and maximum pooling across 57 genomic datasets, including for DNABERT-2, Nucleotide Transformer v2, and HyenaDNA; the authors further argue that averaging non-padding token states provides a more comprehensive representation of sequence-wide signal and reduces architecture-dependent variation in downstream performance (Feng et al., 2025).

Below we summarize the model families evaluated in this study, emphasizing: (i) architecture class (encoder vs. decoder), (ii) tokenization unit, (iii) parameter scale and representation (embedding) dimensionality, (iv) maximum supported context length, and (v) training corpus characteristics.

3.2.1 TRANSFORMER ENCODERS

The **Nucleotide Transformer (NT)** (Dalla-Torre et al., 2025; Boshar et al., 2025) suite represents a seminal collection of encoder-only genomic foundation models, with architectures ranging from 50 million to 2.5 billion parameters. Building upon the masked language modeling objective, the first two generations of NT were pre-trained on diverse datasets, including the human reference genome, the 1,000 Genomes Project corpus, and a multi-species alignment of 850 organisms. While **NT v1** and **v2** utilized *k*-mer tokenization to provide context windows of 6 kb and 12 kb respectively, the latest iteration, **NT v3**, introduces a generational shift in modeling capability. **NT v3** was pre-trained on an unprecedented 9Tbp (terabase pairs) from the OpenGenome2 corpus and subsequently post-trained on over 16,000 functional genomic tracks. By employing a U-Net-inspired architecture and single-base tokenization, **NT v3** facilitates high-resolution modeling of dependencies up to 1 Mb. In this study, we specifically utilized the `NT-(v1)-2.5B-multi-species`, the `NT-v2-500M-multi-species`, and the `NTv3-650M-post` models. For embedding extraction, mean pooling yields fixed-length vectors whose dimensionality equals the model hidden size (e.g., 2560 for NT v1 2.5B; 1024 for NT v2 500M). These encoders offer strong bidirectional context modeling, which may be advantageous for COX fragments with local mutations and variable region boundaries.

EnCodon (Naghipourfar et al., 2024) is a foundation model that represents protein-coding DNA at the codon level, aiming to capture how synonymous codon choices and local coding context shape higher-level protein landscape. Architecturally, it follows a RoFormer-inspired design with Rotary Positional Encoding (RoPE) and rotary self-attention, enabling efficient modeling of long-range dependencies within coding regions. **EnCodon** is pretrained with masked language modeling on an aggregated corpus of $\sim$60 million coding sequences spanning $>$5,000 species from the NCBI Genomes resource. Although the pretraining data are predominantly bacterial (98.7%), the model includes a dedicated eukaryotic adaptation stage intended to better reflect eukaryotic (including mammalian) codon-usage patterns. In this study, the `EnCodon-620M-euk` checkpoint (620M parameters) is used, supporting contexts up to 2,048 codons (6,144 nucleotides) and a hidden width of 2,048. Sequence-level representations are obtained by mean pooling token embeddings, producing 2,048-dimensional vectors. This codon tokenization provides a biologically grounded inductive bias for COI/COX1—a protein-coding gene—and may facilitate extraction of taxonomically informative signal beyond purely nucleotide-level tokenizations.

DNABERT-2 (Zhou et al., 2023) is an efficient, multi-species genomic foundation model that adopts a BERT-style Transformer encoder architecture (Devlin et al., 2019). Departing from traditional fixed-length *k*-mer tokenization, the model employs Byte Pair Encoding (BPE) to significantly enhance sample efficiency and reduce sequence redundancy. It was pre-trained via masked language modeling on a massive 32.49-billion nucleotide corpus, spanning the human reference genome and 135 diverse species. To overcome the input length constraints typical of BERT-style models, **DNABERT-2** integrates Attention with Linear Biases (ALiBi), allowing it to extrapolate to long-range sequences during inference. Furthermore, by incorporating Flash Attention and GEGLU activation functions, the model achieves performance comparable to state-of-the-art architectures while utilizing $21\times$ fewer parameters and requiring $92\times$ less GPU pre-training time. The 117M checkpoint corresponds to a BERT-base-sized encoder (12 layers; 768 hidden dimension) with a maximum input length of 512 tokens. We extract per-token hidden states from the final layer and apply mean pooling to obtain 768-dimensional sequence embeddings for downstream taxonomic classifiers.

3.2.2 TRANSFORMER DECODERS

GenomeOcean (Zhou et al., 2025) is a genomic foundation model trained with a next-token prediction objective on large-scale microbial metagenomic contig corpora, designed for efficient learning from assembled metagenomic sequences. It employs a BPE/SentencePiece-style tokenizer over DNA substrings (non-overlapping tokens) and uses rotary positional embeddings (RoPE), with large checkpoints leveraging architectural optimizations such as grouped-query attention and fast attention kernels. In our embedding pipeline, mean pooling yields a fixed-size representation whose dimensionality equals the model hidden size (e.g., 3072 for the 4B checkpoint; 1536 for the 500M checkpoint). GenomeOcean is trained with a finite maximum sequence length (token context win-

dow) and can concatenate contigs up to that limit, enabling scalable embedding extraction for COX fragments and longer COX contigs.

GENERator (Wu et al., 2025; Li et al., 2026) is a long-context decoder-only model optimized for DNA sequence modeling with next-token prediction, using a 6-mer tokenization scheme and a large context window (16,384 tokens; $\sim$98 kb at 6-mer resolution). **GENERator** was pre-trained on an extensive functional corpus comprising 386 billion nucleotides (Gbp) of eukaryotic DNA curated from the RefSeq database. This design supports expressive sequence modeling and flexible generative applications, including the synthesis of protein-coding genes and the design of cis-regulatory elements. In this work, we utilized the 3-billion parameter (3B) variants, specifically the `eukaryote-3b-base` and the v2 iterations for both eukaryotic and prokaryotic domains. These models support an expansive context length of 98,304 nucleotides (16,384 tokens), with each produced embedding vector featuring a hidden dimensionality of 3,072, providing long-range contextual representations that are potentially beneficial for noisy COX sequences and for capturing motifs beyond short k-mer statistics.

HyenaDNA (Nguyen et al., 2023) is a long-range genomic foundation model designed for single-nucleotide resolution sequence processing, built upon a decoder-only architecture utilizing implicit convolutions. **HyenaDNA** scales sub-quadratically in sequence length (training up to $160\times$ faster than Transformer), enabling an expansive context window of up to 1 million tokens. The model utilizes a single-nucleotide tokenizer, treating each DNA base as an individual character to preserve the fine-grained resolution necessary for detecting single-nucleotide polymorphisms and long-range regulatory interactions. **HyenaDNA** was pre-trained on only a single human reference genome using a next-token prediction objective, facilitating full global context at every layer. In this work, we utilized the `HyenaDNA-Large` variant, which supports input sequences up to 1,000,000 nucleotides long. This configuration features 8 layers with a hidden dimensionality of 256, achieving state-of-the-art efficiency at the megabase scale.

Evo (Nguyen et al., 2024) is a foundation model designed for sequence modeling and design at the genome scale, built upon the **StripedHyena** architecture. This hybrid framework interleaves multi-head attention with data-controlled convolutional operators, facilitating near-linear scaling of compute and memory relative to sequence length. Comprising 7 billion parameters, **Evo** utilizes a byte-level, single-nucleotide tokenizer, which preserves high-resolution genetic information without the loss of granularity associated with subword or *k*-mer tokenization. The model was pre-trained using a next-token prediction objective on the OpenGenome dataset, an extensive corpus of 300 billion tokens derived from 2.7 million prokaryotic and phage genomes. In this work, we utilized the `evo-1-131k-base` variant, which was length-extended from the 8k-context base model to support an expansive context of 131,072 tokens. Each hidden state produced by the model has a dimensionality of 4,096.

3.3 k-MER-BASED APPROACH

As a lightweight alignment-free baseline, nucleotide sequences were embedded using normalized frequency vectors of overlapping *k*-mers. Using scikit-bio (Aton et al., 2025), we computed relative (length-normalized) *k*-mer frequencies with overlap and concatenated them into fixed-dimensional representations. The dimensionality is determined by the DNA alphabet size ($|\Sigma| = 4$ for {A,C,G,T}), yielding $|\Sigma|^k = 4^k$ possible *k*-mers at order k. In our experiments, this corresponds to (from monomers to hexamers).

$$\dim(\mathbf{x}_k) = 4^k, \quad k \in \{1, \ldots, 6\} \;\Rightarrow\; (4,\ 16,\ 64,\ 256,\ 1024,\ 4096).$$

This construction enumerates the full combinatorial space from monomers to hexamers, yielding a composition-based signature that is invariant to total sequence length while remaining sensitive to local motif distributions captured by higher-order *k*-mers. Missing *k*-mers are assigned zero frequency, ensuring consistent feature ordering across all sequences. In practice, cumulative feature sets were evaluated by progressively expanding the representation from $k \leq 1$, to $k \leq 2$, $k \leq 3$, and so on up to $k \leq 6$.

3.4 DOWNSTREAM CLASSIFIER ARCHITECTURE

To assess the predictive value of the precomputed embeddings, we used a lightweight *Multi-Layer Perceptron* (MLP) classifier as the downstream model. In this setup, each embedding vector is directly treated as a fixed-length feature representation. Let d denote the embedding dimension (e.g., $d = 4096$ for Evo-1 mean pooling or $d = 5460$ for the *6*-mer frequency baseline). The MLP maps $\mathbb{R}^d$ to a task-specific output space of size C, where C is the number of classes for the corresponding taxonomic rank. Concretely, the model applies a sequence of Linear $\rightarrow$ GELU transformations with dropout ($p = 0.3$) between layers, followed by a final linear classification layer that produces logits over the C classes. This architecture is intentionally simple to isolate the contribution of the embedding representations and to enable efficient training on millions of samples.

A separate MLP was trained for each prediction task and embedding type. For every task, the labeled subset was filtered to remove missing annotations and extremely rare classes, then split into training (80%) and validation (20%) partitions using a fixed random sequence-level split with a shared seed across runs. Optimization was performed with AdamW using a learning rate of 2×10^{-4} and weight decay of 1×10^{-2}. All models were trained for 200 epochs under mixed precision (FP16) to maximize throughput on GPU hardware. This standardized protocol provides a consistent and fair comparison across embedding sources and taxonomic targets.

3.5 ADDRESSING CLASS IMBALANCE WITH THREE COMPLEMENTARY LOSSES

Given the highly skewed label frequencies in our taxonomic targets, we compared three loss strategies that progressively increase the emphasis on minority classes and representation structure. *Standard Cross-Entropy (CE)* serves as a baseline objective that optimizes overall accuracy but is typically dominated by frequent classes under severe imbalance.

Weighted Cross-Entropy (WCE) explicitly rebalances the learning signal by up-weighting errors on rare classes via inverse-frequency class weights, improving sensitivity to minority taxa while keeping the same probabilistic classification objective.

Finally, the *Hybrid* objective couples WCE with a *Balanced Contrastive Loss (BCL)*, jointly optimizing for correct predictions and a more discriminative feature geometry: samples from the same class are encouraged to cluster, while samples from different classes are separated by a margin. This metric-learning component helps mitigate class imbalance by improving inter-class separability and intra-class compactness, particularly for underrepresented taxa that would otherwise be poorly modeled by purely frequency-driven gradients. See Appendix B for details.

4 RESULTS AND DISCUSSION

Our experiments follow a unified embedding–plus–classifier evaluation setting applied to both eKOI and MetaCOXI. For each pretrained representation (and each k-mer approach), we first extract a fixed-length sequence embedding via mean pooling over valid token states, and then train a lightweight MLP classifier independently at each taxonomic rank (Domain–Species) using the training protocol described in Section 3. We report downstream performance for both datasets in Figures 2–3, using Macro-F1 to emphasize robustness under long-tailed label distributions, and comparing standard cross-entropy (CE) against imbalance-aware objectives (weighted cross-entropy and the hybrid WCE+BCL).

4.1 EKOI: DIVERSITY-DRIVEN DIFFICULTY, TOKENIZATION, AND OBJECTIVE SENSITIVITY

Despite its modest size (15,947 sequences), eKOI remains a demanding fine-grained benchmark because label diversity explodes from Domain (2 labels) to Species (11,047 labels), yielding a severe long-tail where Macro-F1 is particularly stringent. Across model families, two effects emerge clearly.

First, the explicit k-mer series provides the cleanest ablation: under CE, extending the receptive field from $k = 1$ to $k =$ 1–6 yields a dramatic Species gain (macro-F1 $\approx 0.05 \rightarrow 0.76$) and consistent improvements across intermediate ranks (Fig. 2), mirroring the increased diagonal concentration observed in the phylum-level confusion-matrix panel for k-mer baselines (see Fig. 5 in

appendix). This confirms that COI taxonomic signal is not captured by low-order composition alone but requires longer motifs and codon-adjacent dependencies. This interpretation is also aligned with recent studies showing that tokenizer design can be a major determinant of genomic language model performance: optimized k-mer strategies can in some settings outweigh model scale, guided tokenization can preserve biologically meaningful subsequences for classification, and adaptive segmentation can improve robustness by matching token granularity to local sequence structure (Suzuki et al., 2025; Mahangade et al., 2026; Kim et al., 2026).

We observe that representation quality is strongly shaped by both the effective motif length induced by tokenization and the pretraining paradigm. In particular, models with multi-scale tokenization, such as the autoregressive decoder GenomeOcean and the encoder DNABERT-2 (BPE over variable-length k-mers, k=1 . . . 12), achieve consistently strong performance across both coarse and fine taxonomic ranks (e.g., Phylum ≈ 0.90 and Species ≈ 0.71; Fig. 2). In contrast, fixed 6-mer tokenization exhibits a clearer dependence on pretraining corpus: GENERator-v2 trained on eukaryotic sequences generalizes better to protist-rich eKOI than its prokaryotic counterpart, especially at finer ranks, consistent with domain-shift effects between bacterial/archaeal genomic statistics and eukaryotic mitochondrial barcodes (see Fig. 6 in appendix). Encoder-style masked language models such as Nucleotide Transformer and EnCodon remain highly competitive, with NT-v2 achieving the strongest Species performance within the MLM family (Fig. 2), indicating that 6-mer masked pretraining can yield discriminative embeddings for COI even in the presence of extreme label fragmentation (see Fig. 7 in appendix).

Second, weighted objectives often stabilize performance in the long-tail regime, but the magnitude of the benefit is highly model-dependent (Fig. 2). For several learned embeddings (e.g., GenomeOcean and the Nucleotide Transformer), performance is comparatively robust across CE vs. weighted losses, with consistently strong mid-to-fine rank Macro-F1 (Fig. 2). In contrast, some representations are markedly objective-sensitive: for example, Evo1 shows very low Species Macro-F1 under CE yet becomes highly competitive under class reweighting, suggesting that its embedding geometry can support fine discrimination but is harder to optimize without explicit imbalance handling (Fig. 2; Appx., Fig. 7).

4.2 METACOXI: WEB-SCALE SUPERVISION, EXTREME LONG-TAIL, AND ROBUSTNESS TO DOMAIN SHIFT.

MetaCOXI scales the task to 5.6M metazoan sequences with near-complete coverage at higher ranks (31 phyla; 100 classes) but an explosive growth to 52,214 genera and 743,671 species, creating an ultra long-tailed regime where per-class sample density collapses at fine resolution and Macro-F1 is therefore most penalized at Species. In this setting, the same previous pattern emerges: methods whose tokenization captures longer, biologically meaningful context—multi-scale BPE (GenomeOcean) or longer fixed units (6-mers; codons)—tend to dominate mid-to-fine ranks, consistent with COI being a compact coding marker where discriminative substitutions are structured in motif-like and codon-adjacent dependencies (Fig. 3). The explicit k-mer results provide an especially clean ablation: k-mers 1–6 substantially outperform 1–3 across Phylum–Genus (Fig. 3), and this improvement is reflected in the phylum-level confusion matrices, where increasing k progressively concentrates mass along the diagonal (Appx., Fig. 8). Among learned models, NTP decoders pretrained on genomic corpora achieve strong performance across ranks, while MLM encoders (Nucleotide Transformer, DNABERT-2) remain competitive, reflecting their ability to produce linearly separable embeddings for downstream classification (Fig. 3; Appx., Figs. 9–10). Crucially, corpus alignment strongly affects generalization: eukaryote-pretrained representations outperform prokaryote-pretrained ones on metazoan mitochondrial COI, underscoring that pretraining on mismatched genomic domains can degrade taxonomic resolution even when tokenization is identical (e.g., 6-mer) (Fig. 3; Appx., Fig. 9).

5 CONCLUSION

We present a unified benchmark for alignment-free COI/COX1 taxonomic classification spanning protist-rich eKOI and metazoan MetaCOXI, and compare pretrained genomic foundation model embeddings against explicit k-mer baselines under a consistent frozen-embedding + MLP protocol. Across ranks from Domain to Species, results show that representation choices—especially the

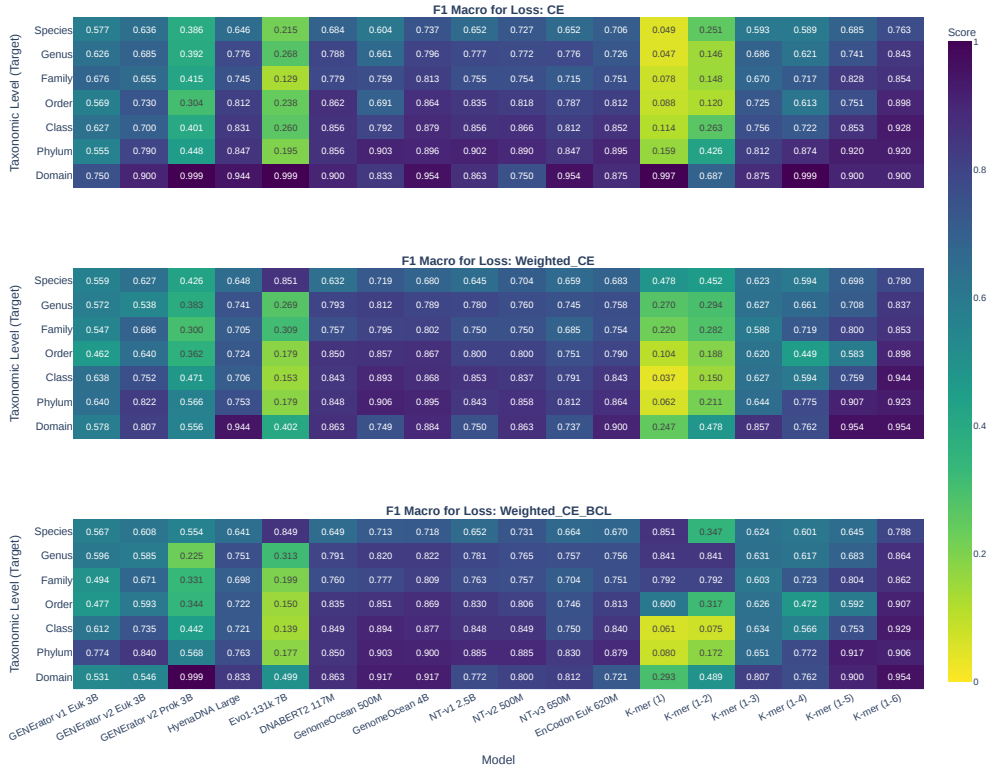

Figure 2: **Taxonomy prediction on eKOI.** Macro-F1 scores for MLP classifiers trained on eKOI embeddings to predict taxonomic labels at multiple ranks (rows). Columns correspond to embedding models. We report three training objectives: CE (top), Weighted_CE (middle), and Weighted_CE_BCL (bottom), enabling a controlled comparison of standard versus imbalance-aware losses under the strongly long-tailed label distributions at fine taxonomic resolution.

effective motif/context length implied by tokenization—strongly determine taxonomic separability under severe long-tail label distributions.

Three practical takeaways emerge for other researchers. (i) Longer context is decisive: the k-mer ablation demonstrates that expanding from k=1 to k=1–6 yields consistent gains from Phylum through Genus and large improvements at fine ranks, indicating that COI signal is captured by motif- and codon-adjacent dependencies rather than low-order composition alone. (ii) Tokenization and pretraining interact: multi-scale BPE models and codon/k-mer MLM encoders produce consistently separable embeddings, while fixed-unit models are more sensitive to the training corpus. (iii) Corpus alignment matters: eukaryote-pretrained representations generalize more reliably to mitochondrial COI than prokaryote-pretrained counterparts under similar tokenization, highlighting domain shift as a concrete deployment risk.

REFERENCES

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

Matthew Aton, Daniel McDonald, Jorge Cañardo Alastuey, Raeed Azom, et al. Scikit-bio: a fundamental python library for biological omic data analysis. *Nature Methods*, 2025. doi: 10.1038/s41592-025-02981-z.

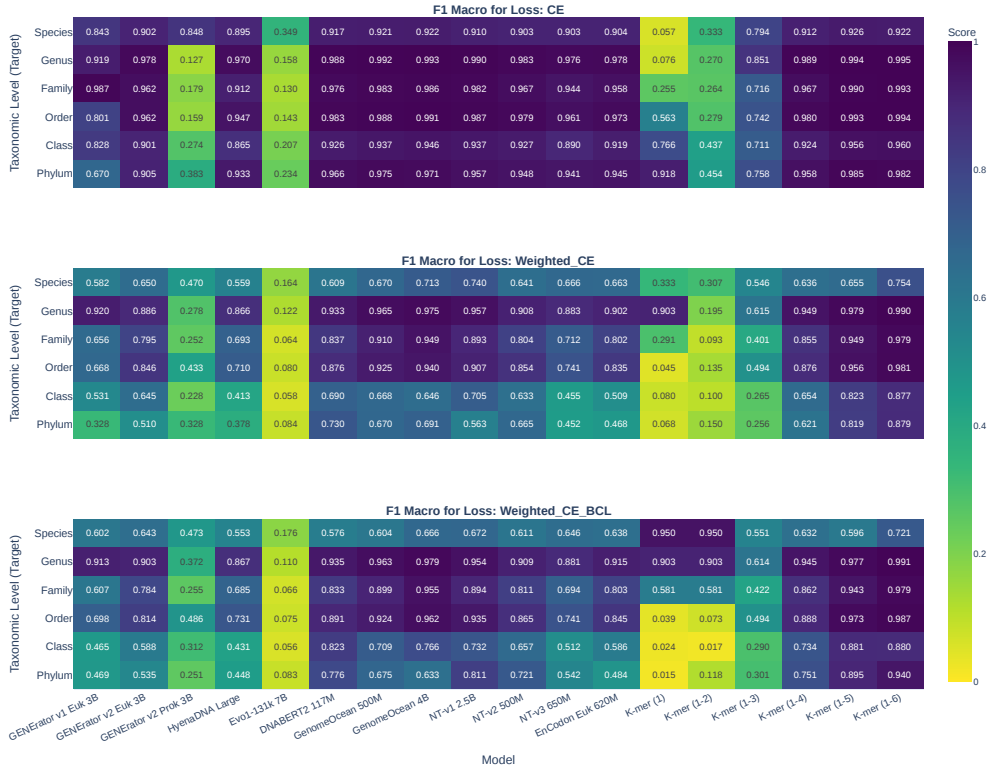

Figure 3: **Taxonomy prediction on MetaCOXI.** Macro-F1 scores for MLP classifiers trained on MetaCOXI embeddings across taxonomic ranks (rows) and embedding models (columns). As in Figure 2, we compare CE (top), Weighted_CE (middle), and Weighted_CE_BCL (bottom). This benchmark emphasizes scalability to web-scale COI barcoding data and evaluates robustness to extreme class imbalance at genus and species levels.

Bachir Balech, Anna Sandionigi, Marinella Marzano, Graziano Pesole, and Monica Santamaria. Metacoxi: an integrated collection of metazoan mitochondrial cytochrome oxidase subunit-i dna sequences. *Database*, 2022:1–6, 2022. doi: 10.1093/database/baab084. URL `https://doi.org/10.1093/database/baab084`.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K. Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, Javier Mendoza-Revilla, Fatimah Al-Ani, Evan Seitz, Yair Schiff, Yohan Bornachot, Ariana Hernandez, Marie Lopez, Alexandre Laterre, Karim Beguir, Peter Koo, Volodymyr Kuleshov, Alexander Stark, Bernardo P. de Almeida, and Thomas Pierrot. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

Bell Raj Eapen. Genomic tokenizer: Toward a biology-driven tokenization in transformer models for dna sequences. *bioRxiv*, 2025. doi: 10.1101/2025.04.02.646836. Preprint.

Han Feng, Lei Wu, Bin Zhao, et al. Benchmarking DNA foundation models for genomic and genetic tasks. *Nature Communications*, 16:10780, 2025. doi: 10.1038/s41467-025-65823-8.

Rubén González-Miguéns, Àlex Gàlvez-Morante, Margarita Skamnelou, Meritxell Antó, Elena Casacuberta, Daniel J. Richter, Enrique Lara, Daniel Vaulot, Javier del Campo, and Iñaki Ruiz-Trillo. A novel taxonomic database for eukaryotic mitochondrial cytochrome oxidase subunit i gene (ekoi), with a focus on protists diversity. *Database*, 2025:1–10, 2025. doi: 10.1093/database/baaf057. URL https://doi.org/10.1093/database/baaf057.

Ho-Jin Gwak and Mina Rho. Deepcoi: a large language model-driven framework for fast and accurate taxonomic assignment in animal metabarcoding. *Genome Biology*, 26(1):393, November 2025. doi: 10.1186/s13059-025-03861-7. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-025-03861-7.

Matthew M. Holland, Luis Felipe Artigas, Angus Atkinson, Mike Best, Eileen Bresnan, Michelle Devlin, Dafne Eerkes-Medrano, Marie Johansen, David G. Johns, Margarita Machairopoulou, Sophie Pitois, James Scott, Jos Schilder, Rowena Stern, Karen Tait, Callum Whyte, Claire Widdicombe, and Abigail McQuatters-Gollop. Mind the gap – the need to integrate novel plankton methods alongside ongoing long-term monitoring. *Ocean & Coastal Management*, pp. 107542, 2025. doi: 10.1016/j.ocecoaman.2025.107542.

Taewon Kim, Jihwan Shin, Hyomin Kim, Youngmok Jung, Jonghoon Lee, Won-Chul Lee, Sungsoo Ahn, and Insu Han. Dnachunker: Learnable tokenization for dna language models. *arXiv preprint arXiv:2601.03019*, 2026. Under review.

Mike Leske, Jamie A. FitzGerald, Keith Coughlan, Francesca Bottacini, Haithem Afli, and Bruno Gabriel Nascimento Andrade. Alignment-free bacterial taxonomy classification with genomic language models. *bioRxiv*, June 2025. doi: 10.1101/2025.06.27.662019. URL https://doi.org/10.1101/2025.06.27.662019. Preprint.

Qiuyi Li, Zhihao Zhan, Shikun Feng, Yiheng Zhu, Yuan He, Wei Wu, Zhenghang Shi, Shengjie Wang, Zongyong Hu, Zhao Yang, Jiaoyang Li, Jian Tang, Haiguang Liu, and Tao Qin. Functional in-context learning in genomic language models with nucleotide-level supervision and genome compression. *bioRxiv*, 2026. doi: 10.64898/2026.01.27.702015. URL https://www.biorxiv.org/content/early/2026/01/29/2026.01.27.702015.

LeAnn M. Lindsey, Nicole L. Pershing, Anisa Habib, Keith Dufault-Thompson, W. Zac Stephens, Anne J. Blaschke, Xiaofang Jiang, and Hari Sundar. The impact of tokenizer selection in genomic language models. *Bioinformatics*, 41(9):btaf456, 2025. doi: 10.1093/bioinformatics/btaf456.

Vedant Mahangade, Matthew Mollerus, Keith A. Crandall, and Ali Rahnavard. Guided tokenization and domain knowledge enhance genomic language models' performance. *bioRxiv*, 2026. doi: 10.64898/2026.02.16.706213. Preprint.

Mohsen Naghipourfar, Siyu Chen, Mathew Howard, Christian Macdonald, Ali Saberi, Timo Hagen, Mohammad Mofrad, Willow Coyote-Maestas, and Hani Goodarzi. A suite of foundation models captures the contextual interplay between codons. October 2024. doi: 10.1101/2024.10.10.617568. URL http://dx.doi.org/10.1101/2024.10.10.617568.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. 2023.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL https://www.science.org/doi/abs/10.1126/science.ado9336.

Table 1: Taxonomic summary of the eKOI database, featuring 15,947 COI sequences across 76 phyla with a focus on protists. Lower taxonomic levels display increased label diversity but reduced classification completeness.

TAXONOMIC LEVEL	UNIQUE LABELS	NON-NULL RECORDS
Domain	2	15,947
Phylum	76	15,931
Class	205	15,610
Order	730	15,263
Family	2,517	15,036
Genus	5,841	14,620
Species	11,047	14,618

Shosuke Suzuki, Kazumasa Horie, Toshiyuki Amagasa, and Naoya Fukuda. Genomic language models with k-mer tokenization strategies for plant genome annotation and regulatory element strength prediction. *Plant Molecular Biology*, 115:100, 2025. doi: 10.1007/s11103-025-01604-7.

Marcell Veiner and Fran Supek. The DNA dialect: a comprehensive guide to pretrained genomic language models. *Mol. Syst. Biol.*, January 2026.

Shuwen Wu, Yun Wang, Haiyan Qin, Zeyu Zhang, Shijun Liu, Yunjie Ruan, Guangsuo Chen, Xia Yuan, and Hangjun Zhang. Environmental dna (edna) technology in biodiversity and ecosystem health research: Advances and prospects. *Ecology and Evolution*, 16(1):e72891, 2026. doi: 10.1002/ece3.72891.

Wei Wu, Qiuyi Li, Mingyang Li, Kun Fu, Fuli Feng, Jieping Ye, Hui Xiong, and Zheng Wang. Generator: A long-context generative genomic foundation model, 2025. URL `https://arxiv.org/abs/2502.07272`.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome, 2023.

Zhihan Zhou, Robert Riley, Satria Kautsar, Weimin Wu, Rob Egan, Steven Hofmeyr, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Fengchen Liu, et al. Genomeocean: An efficient genome foundation model trained on large-scale metagenomic assemblies. *bioRxiv*, pp. 2025–01, 2025.

A Characterization of eKOI and MetaCOXI datasets

We summarize the taxonomic composition and sequence-length distributions (bp) of eKOI and MetaCOXI to contextualize downstream preprocessing choices (e.g., truncation/padding) and potential length-related effects in model training and evaluation. Tables 1 and 2 report rank-wise label cardinality and annotation completeness, showing that eKOI is smaller but highly diverse (protist-rich) with rapidly increasing label fragmentation toward genus/species, whereas MetaCOXI provides web-scale supervision with far larger label spaces and reduced fine-rank coverage. Table 3 further indicates that MetaCOXI exhibits a tighter length distribution while eKOI is more variable and generally longer. Complementarily, Figure 4 offers a qualitative diagnostic of eKOI diversity in embedding space: 3D UMAP projections colored by phylum reveal method-dependent clustering and varying degrees of separation/overlap among broad eukaryotic lineages, illustrating how different representations organize the same protist-rich barcode collection.

B Loss functions and imbalance handling

To train the downstream MLP classifiers under the extreme long-tailed label distributions of eKOI and MetaCOXI, we compare three objectives: standard cross-entropy (CE), weighted cross-entropy (WCE), and a hybrid objective that combines WCE with a balanced contrastive loss (WCE+BCL).

Table 2: Statistics for MetaCOXI, an integrated dataset of 5,608,848 metazoan COI sequences derived from ENA and BOLD. A trend of higher label uniqueness versus lower annotation coverage is observed at finer taxonomic resolutions.

TAXONOMIC LEVEL	UNIQUE LABELS	NON-NULL RECORDS
Domain	1	5,608,848
Phylum	31	5,608,839
Class	100	5,595,085
Order	615	5,543,122
Family	5,118	5,159,058
Genus	52,214	3,838,618
Species	743,671	4,339,019

Table 3: Sequence length summary statistics (bp) for eKOI and MetaCOXI.

STATISTIC	eKOI	MetaCOXI
Count	15,947	5,608,846
Mean	867.96	623.00
Std	343.42	133.66
Min	216	100
25%	655	579
50%	675	628
75%	1,078	658
Max	2,238	3,020

Unless stated otherwise, the constants reported below correspond to the default values used in our training code.

B.1 STANDARD CROSS-ENTROPY (CE)

Given logits $z \in \mathbb{R}^C$, predicted probabilities $p = \mathrm{softmax}(z)$, and ground-truth class $y \in \{1, \ldots, C\}$, the CE loss is

$$\mathcal{L}_{\mathrm{CE}}(x, y) = -\log p_y. \tag{1}$$

B.2 WEIGHTED CROSS-ENTROPY (WCE)

To counter class imbalance, we up-weight rare classes with inverse-frequency weights computed by streaming over the training split (to avoid loading full label vectors into RAM). Let

$$n_c = \sum_{i \in \mathcal{D}_{\mathrm{train}}} \mathbb{I}[y_i = c] \tag{2}$$

be the count of class c, and let $\varepsilon = 10^{-6}$ be a small stabilizer (default in code). We define

$$\tilde{w}_c = \frac{1}{n_c + \varepsilon}, \qquad \varepsilon = 10^{-6}, \tag{3}$$

and normalize weights to keep the mean weight close to 1:

$$w_c = \frac{C\, \tilde{w}_c}{\sum_{j=1}^{C} \tilde{w}_j}. \tag{4}$$

The resulting WCE objective is

$$\mathcal{L}_{\mathrm{WCE}}(x, y) = -w_y \log p_y. \tag{5}$$

B.3 HYBRID OBJECTIVE: WCE WITH BALANCED CONTRASTIVE LOSS (WCE+BCL)

To encourage a more discriminative feature geometry, we augment WCE with a balanced contrastive loss computed on the penultimate-layer features of the MLP. Let $h_i \in \mathbb{R}^d$ be the feature vector for sample i and $\tilde{h}_i = h_i/\|h_i\|_2$ its ℓ_2-normalized version. Define pairwise distances $d_{ij} = \|\tilde{h}_i - \tilde{h}_j\|_2$. For a minibatch, let $\mathcal{P} = \{(i,j) : y_i = y_j,\, i \neq j\}$ and $\mathcal{N} = \{(i,j) : y_i \neq y_j\}$ denote positive and negative pairs. The balanced contrastive loss is

$$\mathcal{L}_{\text{BCL}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} d_{ij}^2 + \frac{1}{|\mathcal{N}|} \sum_{(i,j)\in\mathcal{N}} \Big[\max\big(0, m - d_{ij}\big)\Big]^2, \qquad m = 1.0, \tag{6}$$

where the margin is fixed to $m = 1.0$ by default.

The hybrid objective is a weighted sum of WCE and BCL:

$$\mathcal{L}_{\text{WCE+BCL}} = \mathcal{L}_{\text{WCE}} + \lambda\, \mathcal{L}_{\text{BCL}}, \qquad \lambda = 0.1, \tag{7}$$

with default $\lambda = 0.1$.

Computational cap for BCL (default). Because $\mathcal{L}_{\text{BCL}}$ scales quadratically with batch size, we compute it on at most $S = 2048$ samples per minibatch by random subsampling when needed. In addition, for WCE+BCL runs we cap the effective minibatch size used for training to 8192 to avoid excessive pairwise computation and GPU memory pressure.

Practical details. For scalability (especially on MetaCOXI), class weights w_c are computed by streaming label counts from disk (avoiding loading all labels into memory). For WCE+BCL, the contrastive term is computed on a capped subset of samples from each minibatch (randomly subsampled when needed) to control the quadratic pairwise cost. Unless otherwise stated, we train a separate MLP for each target rank and embedding type. We report Macro-F1 (along with accuracy, balanced accuracy, and weighted-F1) to emphasize robustness under long-tailed label distributions.

C PHYLUM-LEVEL CONFUSION MATRICES FOR k-MER BASELINES AND GLM EMBEDDINGS

To complement aggregate scores with a class-wise diagnostic, we report row-normalized confusion-matrix panels for both datasets (eKOI and MetaCOXI) under two representation families: (i) explicit k-mer frequency baselines, where we ablate the effective receptive field by increasing the maximum k from 1 to 6, and (ii) pretrained genomic foundation models (gLMs), where COI embeddings are extracted once and classifiers trained with the same weighted objective (WCE+BCL). Taken together, these panels make visible two recurring effects discussed in the results: increasing effective motif length concentrates probability mass toward the diagonal (improved separability at coarse ranks), and residual off-diagonal structure is dominated by long-tail taxa and domain-shift sensitivity, most clearly exposed when comparing eukaryote- vs. prokaryote-pretrained representations under otherwise similar tokenization.

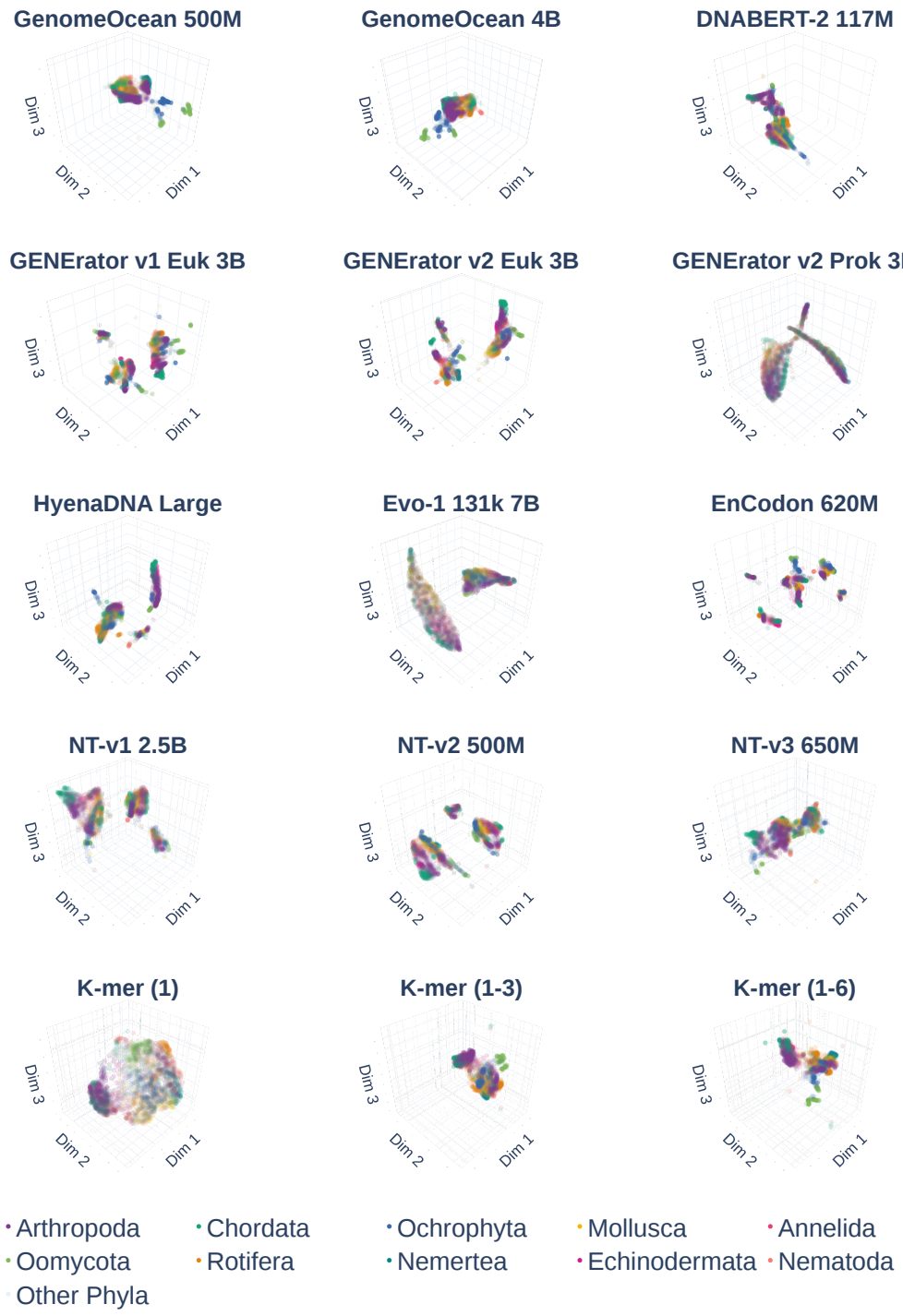

Figure 4: **3D UMAP of eKOI embeddings across representation methods.** Each panel shows a 3D UMAP projection computed from COI/COX1 sequence embeddings generated by a different representation approach (genomic language models and lightweight baselines). Points correspond to eKOI sequences and are colored by phylum, highlighting how well each embedding space separates broad eukaryotic lineages and revealing method-dependent clustering structure, overlap, and outliers.

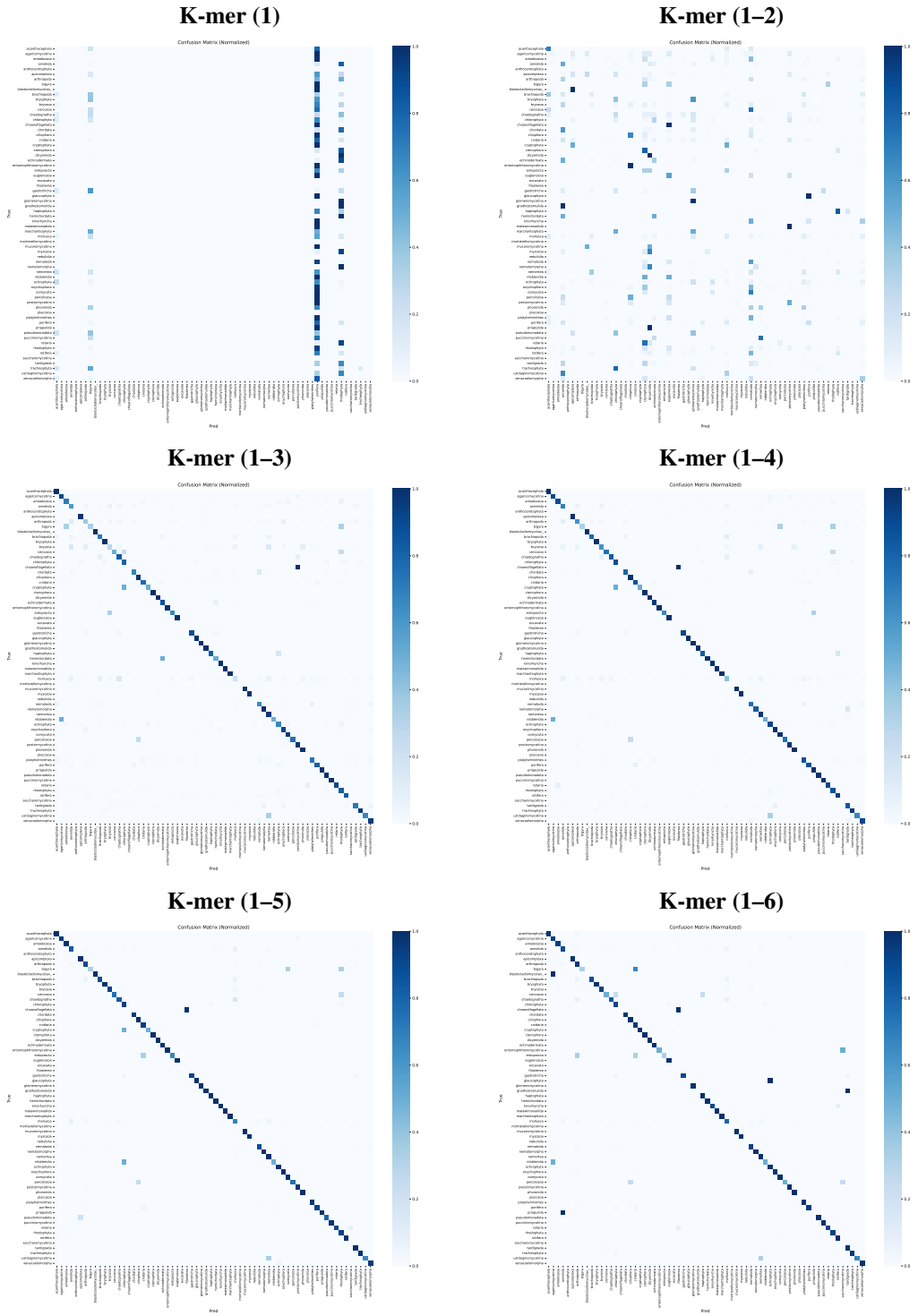

Figure 5: **Phylum-level confusion matrices for k-mer baselines on eKOI.** Row-normalized confusion matrices for MLP classifiers trained on length-normalized k-mer frequency features, with panels ordered by increasing k (top-left to bottom-right; $k = 1, \ldots, 6$). At $k = 1$, predictions exhibit a strong "collapse" toward a few dominant phyla, yielding broad off-diagonal mass and poor separation for many rare groups. Increasing k sharpens the diagonal substantially (by $k \geq 3$ most major phyla become near-uniquely identified), indicating that longer motifs capture the discriminative signal missing in short k-mers.

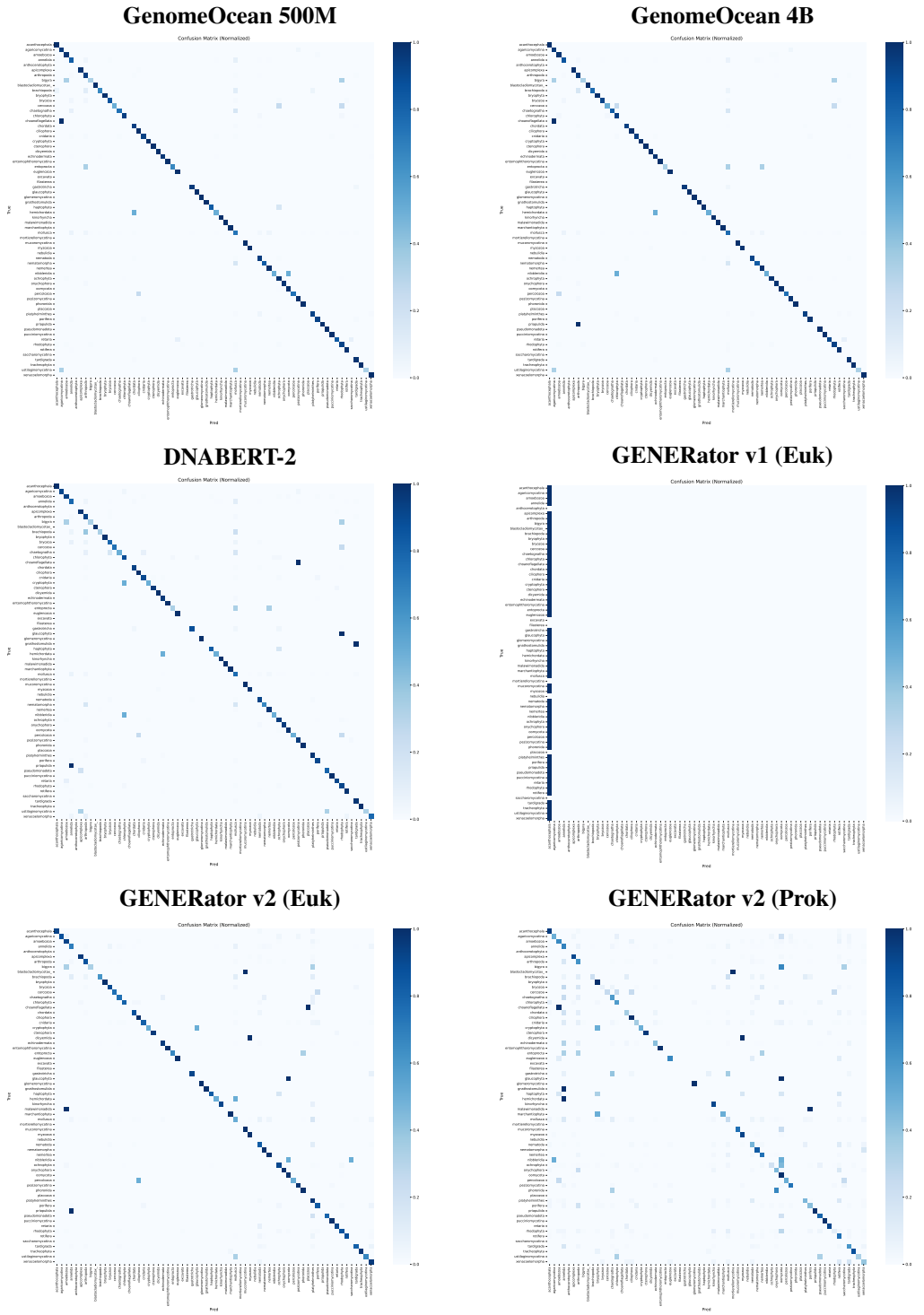

Figure 6: **Phylum-level confusion matrices for genomic foundation models on eKOI (part I).** Normalized confusion matrices for phylum prediction on eKOI using classifiers trained on fixed-length COI embeddings extracted from six pretrained gLMs (panel titles). Panels are arranged top-left to bottom-right as: GenomeOcean 500M, GenomeOcean 4B, DNABERT-2, GENERator v1 (Euk), GENERator v2 (Euk), and GENERator v2 (Prok). Overall diagonal dominance indicates that multiple architectures recover strong phylum-level signal, while structured off-diagonal pockets highlight remaining confusions concentrated in rare lineages; notably, the eukaryote-trained GENERator variants reduce some of these errors relative to the prokaryote-trained counterpart, consistent with domain-shift effects between pretraining data and protist-rich mitochondrial barcodes.

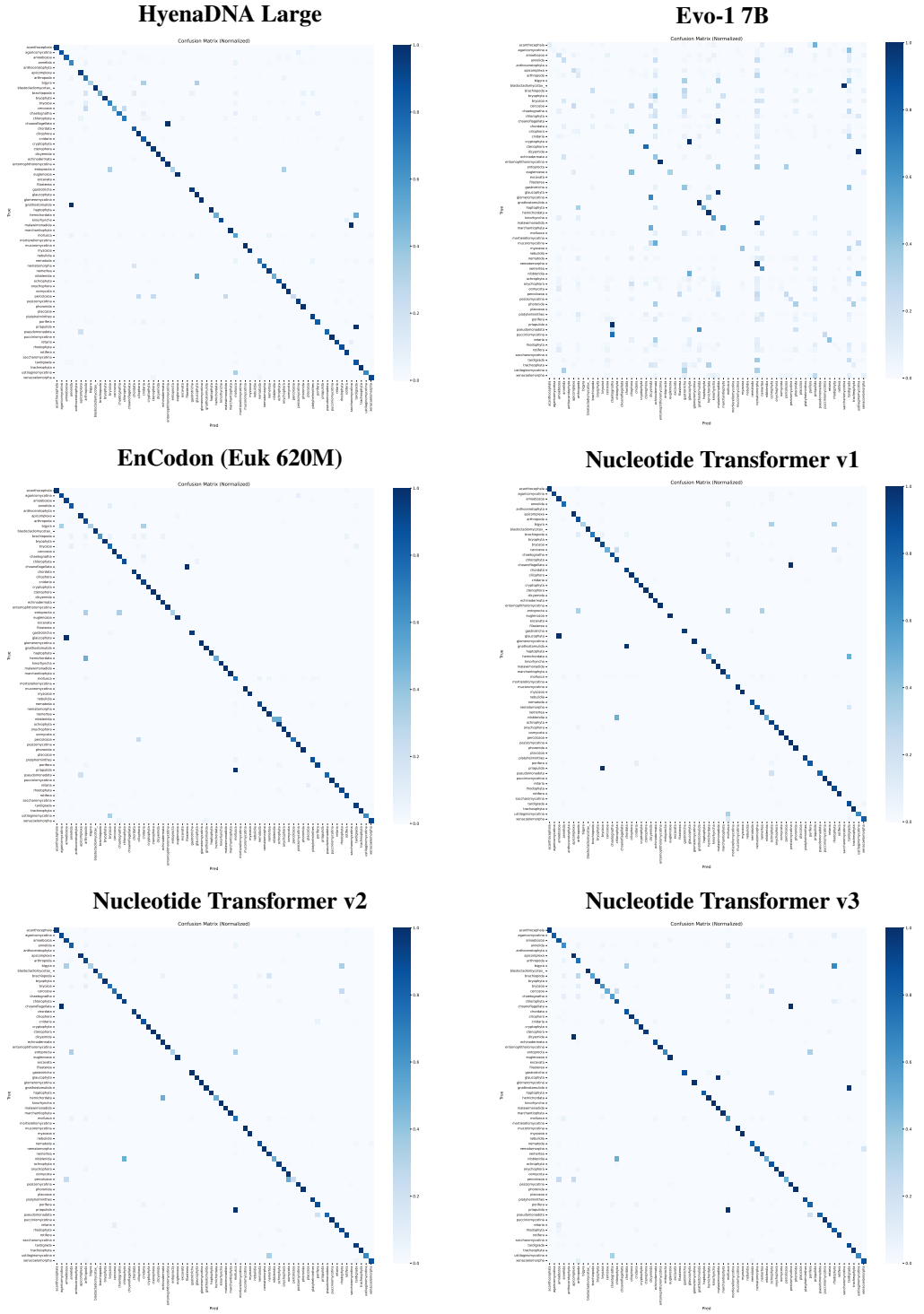

Figure 7: **Phylum-level confusion matrices for additional genomic foundation models on eKOI (part II).** Normalized confusion matrices for phylum prediction on eKOI using classifiers trained on fixed-length COI embeddings extracted from HyenaDNA Large, Evo-1 (7B), EnCodon (Euk 620M), and Nucleotide Transformer (v1–v3). Across panels, strong diagonal mass indicates that broad phylum-level signal is largely recoverable despite the dataset's diversity-driven difficulty, while residual off-diagonal structure is concentrated in rare or compositionally similar lineages under the long-tail regime. Differences between architectures and pretraining paradigms qualitatively mirror the rank-wise trends discussed in the main results, highlighting the role of representation inductive biases and corpus alignment when transferring to protist-rich mitochondrial barcodes.

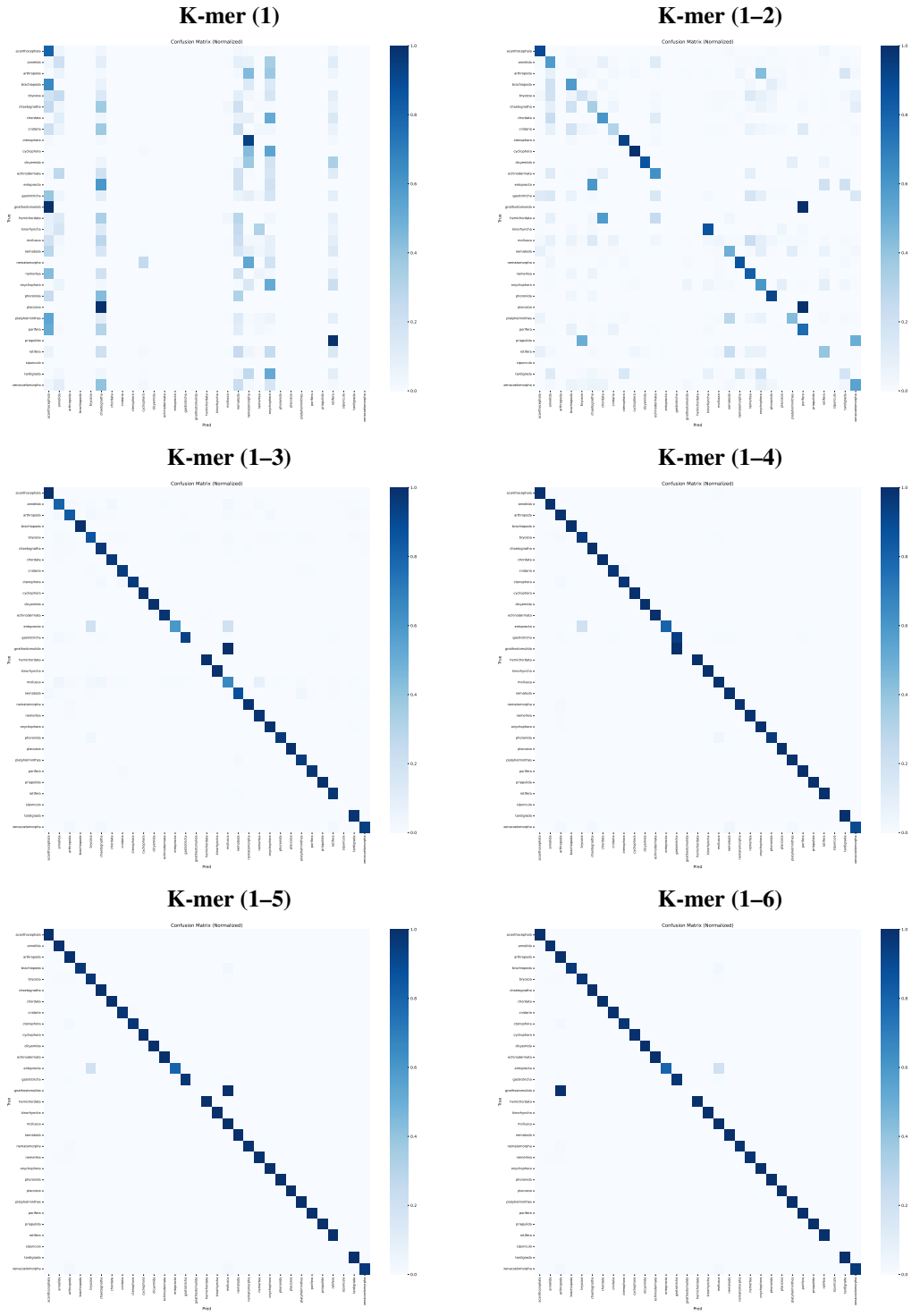

Figure 8: **Phylum-level confusion matrices for k-mer baselines on MetaCOXI.** Normalized confusion matrices for MLP classifiers trained on length-normalized k-mer frequency features, with panels ordered by increasing k (top-left to bottom-right; $k = 1, \ldots, 6$). At $k = 1$, predictions exhibit a strong "collapse" toward a few dominant phyla, yielding broad off-diagonal mass and poor separation for many rare groups. Increasing k sharpens the diagonal substantially (by $k \geq 3$ most major phyla become near-uniquely identified), indicating that longer motifs capture the discriminative signal missing in short k-mers. However, a small set of rare phyla remains systematically unstable even at large k—most notably *Gnathostomulida*, which is repeatedly absorbed by larger neighboring phyla depending on k, and *Entoprocta*, which retains residual confusion with compositionally similar groups.

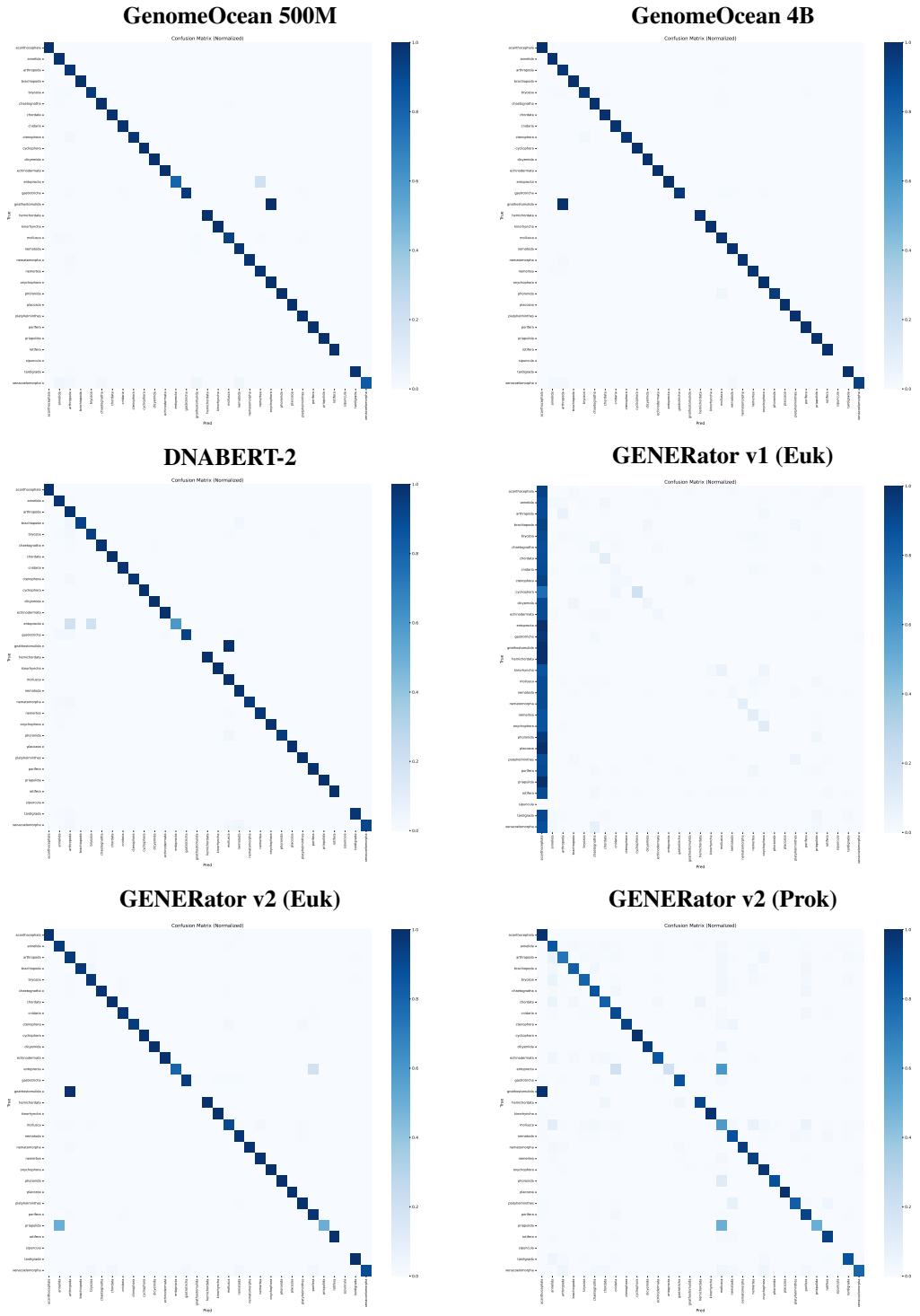

Figure 9: **Phylum-level confusion matrices for genomic foundation models on MetaCOXI (part I).** Normalized confusion matrices for phylum prediction on MetaCOXI using classifiers trained on fixed-length COI embeddings extracted from the six pretrained backbones shown in the panel titles. In this web-scale metazoan setting, most models exhibit a strongly diagonal structure, consistent with near-complete coarse-rank coverage and the fact that COI carries robust motif- and codon-adjacent signal that is recoverable when tokenization captures sufficient context (e.g., multi-scale BPE decoders such as GenomeOcean and DNABERT-2, and 6-mer GENERator variants). Residual off-diagonal pockets are concentrated in rare or closely related phyla, reflecting the long-tail regime. Consistent with corpus-alignment effects, eukaryote-pretrained GENERator models show cleaner separation than the prokaryote-pretrained counterpart under identical 6-mer tokenization, highlighting domain-shift sensitivity at coarse ranks.

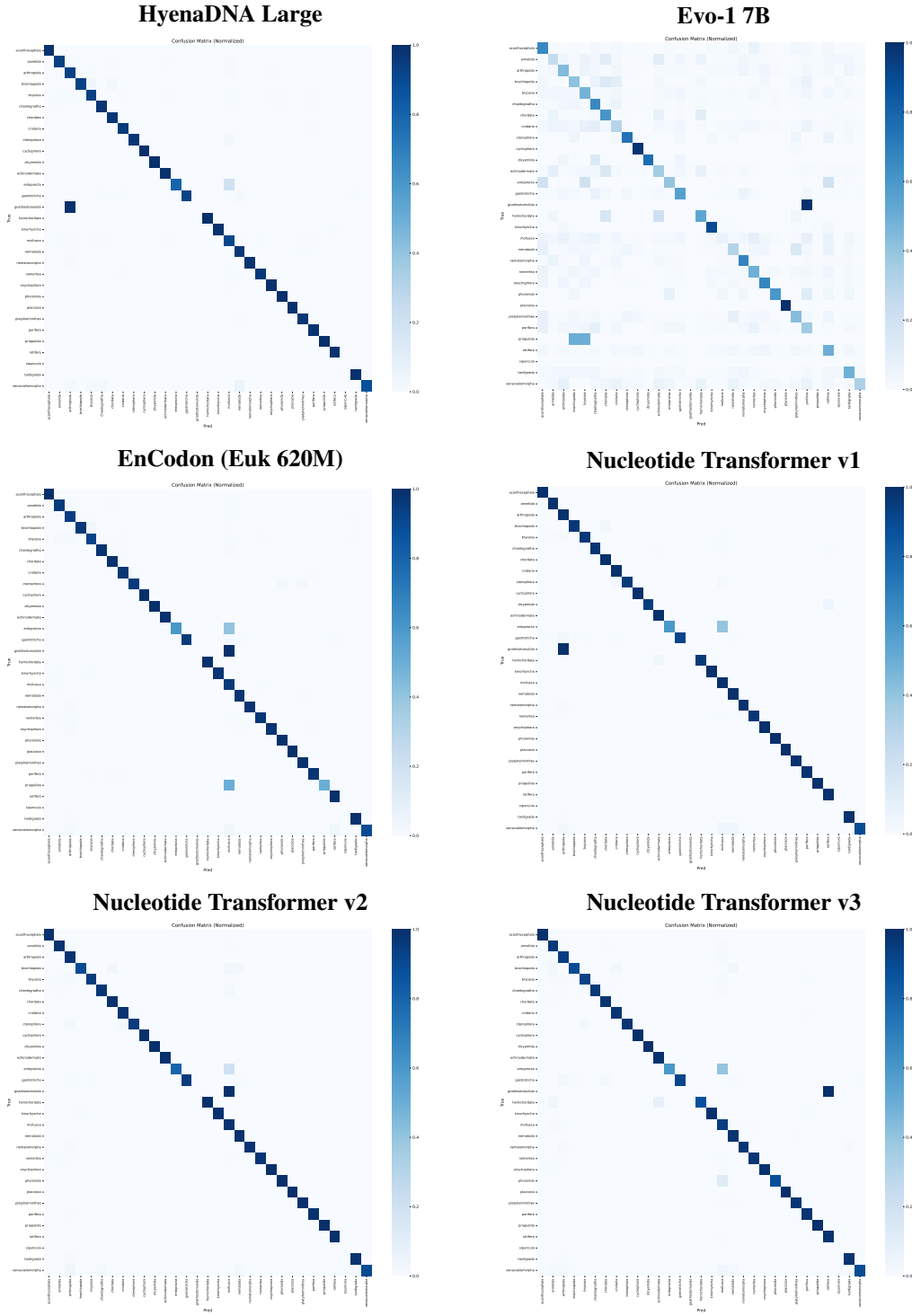

Figure 10: **Phylum-level confusion matrices for additional genomic foundation models on MetaCOXI (part II).** Normalized confusion matrices for phylum prediction on MetaCOXI using a downstream MLP trained on fixed-length COI embeddings extracted from HyenaDNA Large, Evo-1 (7B), EnCodon (Euk 620M), and Nucleotide Transformer (v1–v3) (panel titles). In this web-scale metazoan setting, the MLM encoder family (EnCodon and Nucleotide Transformer v1–v3; 6-mer/codon-level units) yields highly concentrated diagonal structure with only sparse, localized off-diagonal pockets, indicating robust coarse-rank separability. In contrast, HyenaDNA and especially Evo-1 display more diffuse error mass, suggesting stronger sensitivity to representation inductive biases and/or optimization under the ultra long-tail.