IT-UNIVERSITETET I KØBENHAVN

KISPECI1SE - Master Thesis in Data Science

DNABERT-H: Phylogenetic Contrastive Learning for Metagenomics

Prepared by: Anders Havbro Hjulmand (ahju@itu.dk), Andreas Frederik Flensted Olsen (frao@itu.dk), Eisuke Okuda (eiok@itu.dk)
Supervised by: Veronika Cheplygina (vech@itu.dk)
Date: June 2, 2025

DNABERT-H: Phylogenetic Contrastive Learning for Metagenomics

Anders Havbro Hjulmand, Eisuke Okuda, Andreas Flensted Olsen
{ahju,eiok,frao}@itu.dk

Supervisor: Veronika Cheplygina

***Abstract*—Metagenomics binning is a key task within metagenomics that identifies species present in a sample by grouping together similar DNA sequences. Genomic language models (gLMs) focus on single-genome applications, and their potential in metagenomics remains largely unexplored. A recent gLM, DNABERT-S, learns to separate species through contrastive learning, but does not utilize the hierarchical relationships between species. We introduce DNABERT-H, a novel gLM trained with a hierarchical multi-label contrastive loss that incorporates all eight taxonomic ranks to capture the evolutionary relationships between species. DNABERT-H achieves comparable performance to DNABERT-S in metagenomics binning on CAMI2 datasets but both gLMs are outperformed by state-of-the-art binners. We propose a Multiple Instance Learning (MIL) framework for phenotype classification using the genomes recovered from binning. Using this framework, DNABERT-H consistently outperforms DNABERT-S, while VAMB achieves the best performance. Our results highlight that explicitly learning the hierarchical structures between species improves downstream performance, yet current gLMs still lag behind state-of-the-art binning tools. Model and code is available at Github.**

***Index Terms*—Genomic language models, metagenomics binning, phenotype classification, multiple instance learning, hierarchical contrastive learning**

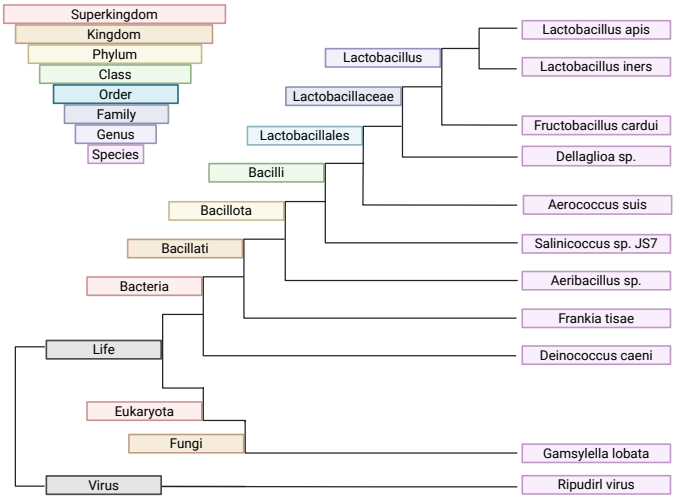

Fig. 1: Phylogenetic tree depicting all taxonomic ranks from species to superkingdom. DNABERT-H aims to capture the hierarchical structures of the phylogenetic tree by training on genomes from bacterial, fungal and viral species.

I. Introduction

Metagenomics is the study of the genetic material and interactions between microbial communities, offering insights into their diversity and functional roles in the environment [1]. Shotgun metagenome sequencing produces millions of fragmented DNA sequences (reads) from an environmental sample. These reads collectively capture the full genetic content of every organism present. The set of reads are typically assembled into longer contiguous sequences, known as contigs [2]. Metagenomics binning groups together similar contigs to obtain metagenome assembled genomes (MAGs) [1], [3], [4]. In this context it is necessary to obtain DNA representations that group contigs from the same species.

Genomic Language Models (gLMs) - inspired by natural language processing (NLP) - have recently been developed [5]–[9]. The majority of gLMs are pre-trained on large genomic datasets using Masked Language Modeling (MLM). However, our previous benchmark paper revealed that MLM-trained gLMs underperform simple baselines in metagenomics binning, indicating that a generic MLM objective may not align with metagenomics tasks [10].

DNABERT-S addresses this limitation by using contrastive learning to group sequences from the same species, and outperformed other gLMs in metagenomics tasks [11], [12]. However, DNABERT-S relies on SimCLR [11] which ignores higher-order relationships between species. As a result, negative sequence pairs sharing common ancestors are undesirably pushed apart in the embedding space [13].

The phylogenetic tree arranges organisms into hierarchical taxonomic ranks based on shared characteristics and evolutionary relationships [14] (see Figure 1). Species belonging to the same lineage, e.g. within the same genus, share more genetic material and common ancestors than those from more distant lineages [15].

In this paper, we introduce DNABERT-H, a specialized gLM that leverages the phylogenetic tree to learn hierarchical relationships between species. DNABERT-H incorporates labels from all taxonomic ranks during training by using Hierarchical Multi-label Contrastive Learning [16]. In this approach, sequences sharing lower-level ranks, e.g. genus, are drawn closer in the embedding space compared to sequences from distant lineages. DNABERT-H could improve generalization on downstream metagenomics tasks by explicitly preserving the phylogenetic structures.

We benchmark DNABERT-H in metagenomics binning against DNABERT-S and three state-of-the-art binners [17]–[19], using the widely adopted evaluation tool CheckM2 [20]. Metagenomics binning is conducted on eight datasets

including CAMI2, a leading benchmark dataset [21].

We propose a Multiple Instance Learning (MIL) framework for phenotype classification using the MAGs recovered from human gut microbiome samples. Using this framework, we evaluate DNABERT-H against DNABERT-S and VAMB [17] across two datasets.

Our contribution can be summarized as follows:

- We present DNABERT-H, a gLM trained with Hierarchical Multi-label Contrastive Learning to preserve phylogenetic relationships between species.
- We propose a Multiple Instance Learning (MIL) framework for phenotype classification that is compatible with any binning approach.
- We provide the first evaluation of gLMs against state-of-the-art metagenomic binners using the widely adopted evaluation tool CheckM2 [20].
- We incorporate a novel metagenomics dataset from patients undergoing treatment with the weight loss drug WEGOVY [22].

II. Background and related work

A. Genomic language models

Recent work in gLMs has gained attention in computational biology due to their wide range of applications within genomics such as transcription factor binding site prediction, gene regulatory element prediction [5], [7], [9], [23], [24] and genome wide association studies [25]. While most gLMs focus on single-genome applications, their potential in metagenomics remains largely unexplored.

Most gLMs utilize transformer-based architectures, drawing inspiration from advances in NLP and breakthroughs in protein modeling, such as AlphaFold [26] and ESMFold [27]. However, genomic sequences are considerably longer than typical protein sequences, often exceeding the context window limitations of standard transformers. This restricts the model's ability to capture long-range dependencies effectively [5], [28]. Still, the methodology holds the potential to advance the understanding of genomes and DNA sequences [28].

The pre-training data used for gLMs often consists of DNA sequences from a single species, such as the human reference genome which contains 3.2 billion pairs of DNA nucleotides, known as base-pairs (bp) [5], [23], [29]. Other gLMs include genomes from multiple species in their pre-training data, typically leading to better performance in various downstream genomic tasks [6], [30]–[32].

DNABERT-2 includes a multispecies dataset in their MLM pretraining. It adopts a BERT based architecture, where DNA sequences are tokenized using Sentence Piece and Byte-Pair Encoding (BPE). The model uses Attention with Linear Biases (ALiBi) [33] to handle longer input sequences during inference and uses FlashAttention [34] to improve computational efficiency.

DNABERT-S [12] uses the architecture of DNABERT-2, and includes bacterial and viral genomes to further pre-train the weights using a contrastive learning framework based on SimCLR [11]. This maximizes the distance between contigs from different species while minimizing the distance between contigs from the same species. DNABERT-S gains an increased performance in downstream tasks by regularizing the model with Manifold Instance Mix Up (MI-Mix) [35] where hidden representations of anchor instances are mixed at a random layer. DNABERT-S benchmarks against other existing gLMs in metagenomics binning and evaluates performance using ground truth labels instead of an evaluation tool such as CheckM2. State-of-the-art metagenomic binning tools [17]–[19], [36], [37] were not included in their benchmark.

B. Metagenomics binning tools

Metagenomics binning tools groups together similar contigs to recover the original genomes in a sample. Contig similarity is typically defined using two different approaches or a hybrid thereof [3]. The first is the composition based approach, where each contig is represented by sequence motifs, such as GC-content or tetranucleotide frequencies (TNFs), while the second approach relies on contig co-abundances across samples. Recent studies VAMB [17], Comebin [19], and METABAT2 [37], demonstrated that using both compositional encoding and co-abundances results in the best performance. In contrast, gLMs rely on learned representations of the tokenized sequences, thereby using a composition based approach.

VAMB (Variational Autoencoder for Metagenomics Binning) [17] trains an unsupervised bi-modal variational autoencoder to map the tetranucleotide frequencies and co-abundances into a latent encoding. The learned latent distributions from the VAE are used to draw representations of the contigs and cluster these directly to obtain bins using a K-medoid based algorithm.

TaxVAMB [18] is an extension of VAMB that includes hierarchical taxonomic annotations of the contigs as a third input to the VAE. The annotations are obtained using the tool Taxometer [38] that uses a hierarchical loss to learn the taxonomic labels of contigs. TaxVAMB employs a semi-supervised loss using the contig annotations as labels, such that the model jointly learns hidden states given the taxonomy, co-abundances and composition based inputs. TaxVAMB improves binning results from VAMB by incorporating taxonomic information, and especially found better performance for datasets with few samples i.e. less informative co-abundance vectors.

Two other deep learning approaches, Comebin [19] and Semibin2 [36], use a contrastive loss to learn the mapping from compositional and co-abundance feature vectors to metagenomic bins.

Comebin uses two FFNNs, one to merge the heterogeneous features (TNFs and co-abundances) and the second to learn the embeddings of contigs using unsupervised contrastive learning. Comebin generates multiple views for each contig by randomly sampling contiguous subsequences from the original contig. The contrastive loss then minimizes the distance between the anchor contig and its augmented views, while maximizing the distance to other contigs. Metagenomic bins are obtained using a Leiden community detection algorithm [39] on a K-nearest neighbor graph constructed from the learned contig embeddings. The authors find that Comebin outperforms other binners.

Semibin2 trains a Siamese network using feature vectors of co-abundances and TNFs by jointly optimizing two different objective functions. The first is the unsupervised reconstruction of the original inputs using two autoencoders, while the second is a supervised contrastive loss. Semibin2 relies on taxonomic annotation tools as part of their preprocessing pipeline to construct negative pairs of contigs if they originate from different species. Similar to Comebin, the positive pairs are constructed from splitting contigs into contiguous subsequences.

DNABERT-S, Comebin, and Semibin2, demonstrated that contrastive loss is useful for metagenomics binning. TaxVamb improved binning results of VAMB by including hierarchical taxonomic information in the input features. Here, we introduce a novel gLM, DNABERT-H, that utilizes both contrastive learning and hierarchical taxonomic information to learn biologically meaningful representations of DNA sequences.

C. Phenotype classification using human gut microbiome data

The human gut microbiota is known to contain approximately 10^{13} microorganisms and more than 100 times the genes of the human host [40], [41]. The extensive efforts in discovering the role of the human gut microbiome have established its vital involvement in physiology, nutrition, immune function, and human metabolism [42], [43]. Additionally, compositional changes in the gut microbiota has been linked to diseases such as CVDs, cancer, respiratory diseases, diabetes, IBD, brain disorders, chronic kidney diseases, and liver diseases [40], [44].

Machine learning models have been applied to explore the associations between the human gut microbiome and diseases, by mapping taxonomic profiles of the microbiome samples to a phenotype [45]–[47]. The taxonomic profiles are often obtained by aligning samples to known reference genomes [46]. A deep learning approach, DeepMicro [46], uses a VAE to learn hidden representations of these taxonomic profiles and subsequently classify phenotypes using different models. DeepMicro evaluates their method on six different disease datasets such as type 2 diabetes [43], [48], liver cirrhosis [49], and colectoral cancer [50]. Another approach, MetAML [51], achieved similar performance to DeepMicro on the six datasets without using dimensionality reduction. By performing permutation tests and comparing the classifier performance on true labels compared to randomly shuffled labels, the authors found considerable associations between microbiome samples and phenotypes.

Both DeepMicro and MetAML rely on reference databases by using tools such as MetaPhlAn2 [52] to obtain the taxonomic profiles, hence preventing novel microbial sequences to be considered in the classification. Metagenomics binning bypasses this limitation by profiling the metagenomic samples using composition and co-abundance based features, and serves as a promising avenue to discover associations between the human gut microbiome and disease.

D. Multiple Instance Learning for phenotype classification

A natural approach to model the associations between phenotype labels and metagenomic samples is using Multiple Instance Learning (MIL) [53], where a bag represents a sample and the instances contained in a bag is the taxonomic profile. In this scenario the bag labels can be considered weak labels as most organisms in the gut microbiota are beneficial to human health, and identifying key instances contributing to the bag labels is crucial for interpretability [40]. A recent study, IDMIL [54], uses an alignment free MIL framework for phenotype classification. Here, instances are constructed from grouping learned representations of sample reads into clusters. The cluster centroids are subsequently used as instance representations, and assigned a learnable weight to identify key instances. The authors find that two pathogens associated with liver cirrhosis coincided with the most influential instances. However, since clustering is performed independently for each sample, there is no consistent correspondence of instances across bags. As a result, the learned weights may not reflect the same biological entities between samples, raising concerns about the reliability of their interpretations.

We propose a MIL framework for phenotype classification with taxonomic profiles extracted from metagenomics binning. Our framework guarantees that biological entities are aligned across bags, which is crucial for interpretability of instances. Our framework does not rely on external reference databases and can be used with any binning tool.

III. Methods

The methods section consists of four parts: *A. Model*, which presents the pre-training of DNABERT-H; *B. Preprocessing metagenomics data*; *C. Metagenomics binning*, describing the binning approach for gLMs; and *D. Multiple Instance Learning for phenotype classification*. See Figure 2 for a visual overview of part *A* and Figure 3 for parts *B–D*.

A. Model

Conventional contrastive learning methods, such as SimCLR [11] and SupCon [55], do not capture the hierarchical relationships that often exist among labels, by relying on a flat label structure (SupCon) or augmented views of the data (SimCLR). This limitation may be problematic in metagenomic tasks where the taxonomy of microbial species follows a hierarchical structure.

To address this, we replace the contrastive learning framework from DNABERT-S [12] with Hierarchical Multi-label Contrastive Learning (HiMulConE) [16], aiming to learn representations that reflect the hierarchical structure of species. HiMulConE extends the standard supervised contrastive learning framework to the hierarchical multi-label setting, by incorporating the taxonomic labels of genomes at every level in the phylogenetic tree. DNABERT-H uses DNABERT-2 [6] as the backbone model architecture.

1) Pre-training pipeline: The pre-training pipeline, summarized in Figure 2, comprises three stages: *(a)* preprocessing each genome into non-overlapping $10{,}000$ bp sequence pairs annotated with multi-level taxonomic labels, *(b.1)* hierarchy-aware batch sampling to assemble anchors and their positive counterparts at every taxonomic rank, and *(b.2)* hierarchical multi-label contrastive learning utilizing the HiMulConE loss.

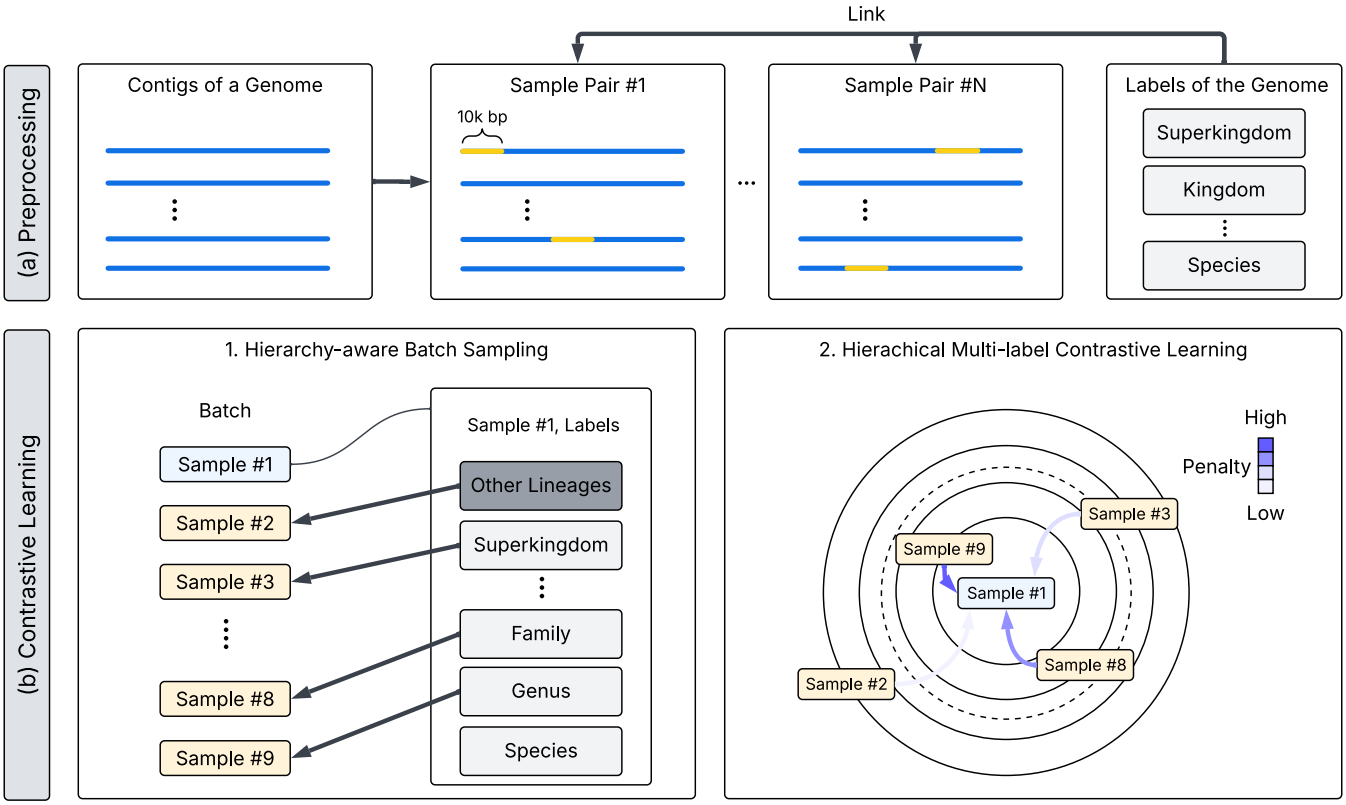

Fig. 2: Overview of the pretraining pipeline. **(a)** Each genome's contigs are split into non-overlapping chunks of 10,000 base-pairs, from which sequence pairs are randomly constructed. These pairs are annotated with hierarchical taxonomic labels. **(b.1)** Each batch centers on an anchor sample (blue), with additional samples selected to share labels at each taxonomic level, so that diverse hierarchical relationships can be represented within the batch. **(b.2)** HiMulConE encourages the model to assign larger loss values to pairs with greater taxonomic distances compared to those with smaller distances (HiConE). Additionaly, it applies higher penalty weights to pairs that are taxonomically closer, and lower weights to those that are more distant (HiMulCon).

(a) Preprocessing: Each reference genome consists of one or more contigs. For all genomes, contigs are partitioned into non-overlapping 10,000 bp length segments. A sample is created by pairing two randomly selected segments from the same genome, while ensuring that each segment appears exactly once across all samples for bacterial and fungal genomes. In contrast, unique segments from viral genomes can be paired multiple times to effectively up-sample the number of viral samples. Each sample is annotated with labels from all taxonomic ranks.

(b.1) Hierarchy-aware batch sampling and defining positive pairs: Each sample contains an instance i and its corresponding paired segment i^+. Each batch is constructed around a randomly chosen anchor sample that has not yet been used in the current epoch. Starting from this anchor, we trace its taxonomic lineage from superkingdom down to species. At each rank, we sample an additional, unused sample from a different genome that shares the same label at the immediate lower taxonomic rank. At the rank superkingdom, a sample from a different lineage is selected. Consequently, a batch contains an anchor sample and a unique sample from each taxonomic rank, resulting in a total of 9 samples.

These 9 samples are used to define positive pairs across all taxonomic ranks for an anchor sample in the batch. The paired segment contained in the anchor sample is used to define the positive pair at the lowest level (species level). For each taxonomic rank above species, other samples sharing the given rank are treated as positive pairs at that layer, such that the anchor sample is assigned one positive pair at each layer in the phylogenetic tree. In addition, samples sharing taxonomic labels at any rank in the tree are also considered positive pairs to each other.

(b.2) Hierarchical Multi-label Contrastive Learning: HiMulConE combines two hierarchy-aware contrastive losses: the Hierarchical Multi-label Contrastive loss (HiMulCon) and the Hierarchical Constraint Enforcing loss (HiConE) [16], to effectively capture the hierarchical relationships in the phylogenetic tree.

Let $L = \{1, 2, \ldots, 8\}$ denote the eight taxonomic ranks, ordered from the lowest rank (species) at $l = 1$ to the highest rank (superkingdom) at $l = 8$. For every instance i, we define $P_l(i)$ as the set of its positive pairs at a specific taxonomic level $l \in L$ and denote a positive pair from this set as $p_l^i \in P_l(i)$. The pairwise supervised contrastive loss is then calculated as follows:

$$L^{\text{pair}}(i, p_l^i) = \log \frac{\exp(f_i \cdot f_{p_l^i}/\tau)}{\sum_{a \in A \setminus i} \exp(f_i \cdot f_a/\tau)} \tag{1}$$

Here, $A = \{1, 2, \ldots, 2B\}$ where B refers to the batch size, f denotes the embedding of an instance, and τ is a temperature parameter. The HiMulCon loss is then calculated by aggregating these pairwise losses across all hierarchical levels, each weighted by a penalty parameter λ_l:

$$L^{\text{HMC}} = \sum_{l \in L} \frac{1}{|L|} \sum_{i \in I} \frac{-\lambda_l}{|P_l(i)|} \sum_{p_l^i \in P_l(i)} L^{\text{pair}}(i, p_l^i)$$
$$\text{where,} \quad \lambda_l = exp(\frac{1}{l}). \tag{2}$$

However, using only HiMulCon loss may fail to maintain hierarchical consistency, by yielding smaller absolute loss values to positive pairs at higher levels compared to loss values at lower levels. To mitigate this, the HiConE loss is applied to enforce that the absolute loss value at each hierarchical level is at least as large as that of lower levels. The HiConE is defined as:

$$\sum_{l \in L} \frac{1}{|L|} \sum_{i \in I} \frac{-1}{|P_l(i)|} \sum_{p_l^i \in P_l(i)} \min\left(L^{\text{pair}}(i, p_l^i), L^{\text{pair}}_{\text{min}}(l-1)\right)$$
$$\text{where,} \quad L^{\text{pair}}_{\text{min}}(l) = \min_{(i, p_l^i)} L^{\text{pair}}(i, p_l^i). \tag{3}$$

Levels are processed sequentially from l = 1 to 8, ensuring that $L^{\text{pair}}_{\text{min}}(l-1)$ has already been computed at each step. Finally, the two losses are combined into a single hierarchy-aware loss (HiMulConE):

$$\sum_{l \in L} \frac{1}{|L|} \sum_{i \in I} \frac{-\lambda_l}{|P_l(i)|} \sum_{p_l^i \in P_l(i)} \min\left(L^{\text{pair}}(i, p_l^i), L^{\text{pair}}_{\text{min}}(l-1)\right) \tag{4}$$

B. Preprocessing metagenomic data

Shotgun sequencing produces millions of fragmented reads from metagenomic samples that are processed through several steps to obtain features used in metagenomics binning. These steps typically include cleaning, assembly, and alignment to reference sequences. Assembly merges groups of reads with significant overlap into the smallest possible superstring to construct contigs. In alignment, sample reads are mapped to the assembled contigs to estimate the abundance of each contig in the original samples. The resulting contigs and their abundances are often used as input features for metagenomics binning [1], [17], [56].

1) Three distinct preprocessing modes: In metagenomics binning, the assembly and alignment steps can be approached with three distinct frameworks: co-assembly binning, single-sample binning, and multi-sample binning. In co-assembly binning, all individual sample reads are pooled and jointly assembled into contigs. Sample reads are subsequently aligned to the assembled contigs, and binning is conducted collectively on the samples. In single-sample binning, each sample's reads are assembled, aligned, and binned independently of the others. In multi-sample binning each sample is individually assembled into contigs. All contigs are then concatenated into a contig catalogue, that is used as reference in the alignment step. Binning is performed collectively on all the samples. The three distinct frameworks influence the contig representations used in metagenomics binning. In this paper, we follow the multi-sample binning approach, which has demonstrated the best performance in recent works, although at an increased computational cost [17], [19], [36], [57], [58].

2) Preprocessing pipeline: Using the multi-sample binning approach, all samples $s_i \in \mathcal{S}$ from a dataset $\mathcal{D}$ were passed through the following preprocessing pipeline.

First, the reads were trimmed from adapter sequences and then mapped to the human genome, removing host DNA using Kneaddata (v.0.12.2) with the flag *–run-trim-repetitive* [59]. Before and after trimming, the phred quality scores of the reads were checked using FASTQC (v.0.12.1) [60]. The trimmed reads were assembled into contigs using metaSpades (v.4.1.0).

The multi-sample binning approach defines the contig catalogue as the concatenation of individually assembled contigs $\mathcal{C}_{s_i}$ such that $\mathcal{C} = \{\mathcal{C}_{s_1} || \ , \ldots, || \ \mathcal{C}_{s_{|\mathcal{S}|}}\}$. We used VAMBtools (v.4.1.3) to construct the contig catalogue $\mathcal{C}$ whilst removing contigs with sequence length < 2.000 base-pairs [17].

The contig catalogue was used as the reference for all samples during the alignment step. We mapped each read to the reference to obtain the read counts for each contig using Strobealign (v.0.16.0) [61] and subsequently sorted the read counts using Samtools (v.1.21) [62].

Following the preprocessing pipeline in VAMB [17], read counts were normalized by sequencing depth (total number of reads) and contig length to obtain RPKM normalized (reads per kilobase contig per million mapped reads) contig abundances using the python implementation of Coverm (pycoverm) (v.0.6.2) [17], [63]. The normalized contig abundances from each sample were merged into an abundance table of dimensions $|\mathcal{C}| \times |S|$ using VAMBtools (v.4.1.3) [17]. The abundance table was additionally normalized within each sample (column-wise) by the total number of mapped reads. Then, across samples (row-wise) by the total number of mapped sample reads to each contig. The sum of a contig's values across samples was taken prior to the column-wise normalization to compute contig sample-relative abundance.

For a dataset $\mathcal{D}$, the pipeline results in a contig catalogue $\mathcal{C}$ and contig co-abundances across samples.

C. Metagenomics binning

Metagenomics binning uses composition and co-abundance based features to cluster similar contigs, and is essentially a clustering problem where the number of clusters is unknown.

VAMB [17], TaxVAMB [18], MetaBAT [3], and DNABERT-S [12] use an iterative K-medoid based clustering algorithm, while Comebin [19] uses a Leiden community detection algrithm [39] to cluster contigs into bins. In this paper, we adopted the K-medoid clustering algorithm from DNABERT-S. The K-medoid algorithm was implemented on GPU to effectively handle the large contig catalogues produced by the multi-sample binning approach.

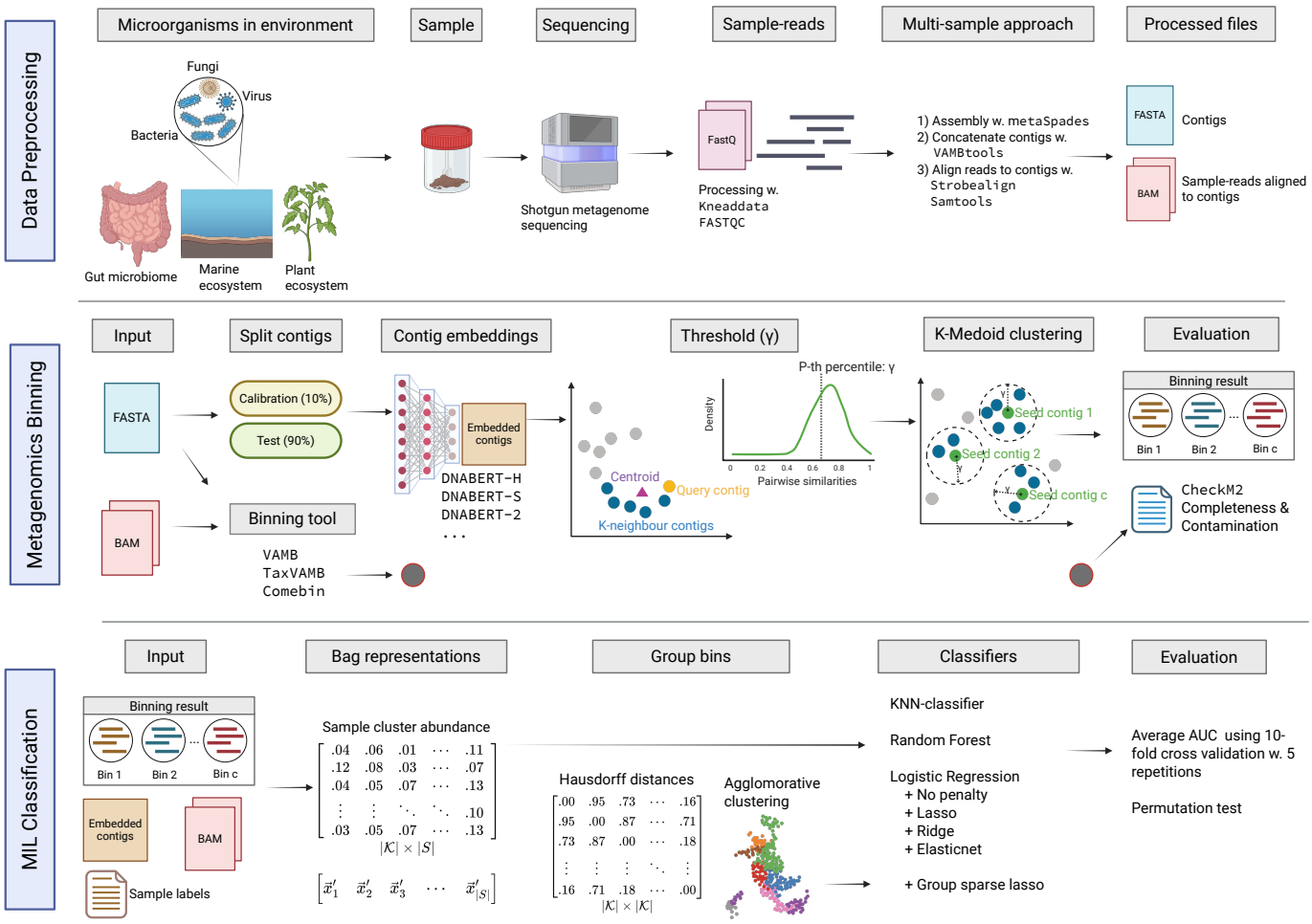

Fig. 3: Overview of methodology. **Top**: Reads from microbiome samples are processed using the multi-sample binning approach to obtain a contig catalogue along with co-abundances. **Middle**: Metagenomics binning is an unsupervised clustering problem. Contigs are embedded using a genomic language model; a similarity threshold is defined based on the percentile of distances within each K-nearest-neighbor centroid. Clustering is then performed using a K-medoid algorithm, and the resulting bins are evaluated using CheckM2. **Bottom**: Phenotype classification is framed as a Multiple Instance Learning (MIL) problem. Each bag contains $|\mathcal{K}|$ instances, and these are represented by the relative abundance of the k'th bin obtained in the metagenomics binning step. Classification is performed using KNN, Logistic Regression, Random Forest, and Sparse Group Lasso, with groups defined using agglomerative clustering on Hausdorff distances.

1) K-medoid clustering algorithm: The K-medoid clustering algorithm is unsupervised and does not require a predefined number of clusters making it well suited for metagenomics binning.

Given an embedded contig catalogue E, the K-medoid algorithm operates on the $|E| \times |E|$ similarity matrix $\mathcal{S}$, where each entry $\mathcal{S}_{i,j} = \cos\big(z_i, z_j\big)$ corresponds to the cosine similarity between a pair of embedded contigs i and j.

For Z steps, the algorithm picks the contig with the highest similarity to all other contigs as the seed centroid, and considers contigs within a similarity threshold γ to be the initial neighboring set of contigs $\mathcal{I}^t$ where $t = 0$.

At each iteration $t = 1, \ldots, T$, an average embedding of the previous neighbor set $\mathcal{I}^{t-1}$ is used to update the seed centroid, and the new neighboring set is defined as $\mathcal{I}^t$. The final set of neighboring contigs $\mathcal{I}^T$ is then assigned to the same cluster, and removed from the embedding space before continuing to the next step. After Z steps, clusters containing less contigs than the minimum bin size, m, are removed, and contigs from these clusters are not assigned a label. See Appendix A for the algorithmic implementation.

2) Threshold calibration: The threshold parameter $\gamma \in [0, 1]$ defines the decision boundary for neighboring contigs, and drastically influences the K-medoid clustering results.

Different models and datasets produce distinct embedding spaces, making it crucial to calibrate γ for each model on an independent calibration dataset $\mathcal{C}_{cal}$ prior to binning.

The following procedure was used to calibrate γ (see appendix B for the algorithmic implementation). For each embedded contig, we identified its top-K neighbors $\mathcal{H}$ using cosine similarity, and computed the centroid of $\mathcal{H}$. We then calculated the similarities between each neighbor in $\mathcal{H}$ and the centroid, and partitioned the similarities into $1,000$ equidistant bins. The P-th percentile of the binned similarities across all

contigs was used as the threshold parameter γ in the K-medoid algorithm. Using grid search, the hyperparameters K and P were optimized on the calibration dataset $\mathcal{C}_{cal}$ and the values resulting in the best metagenomics binning performance were used to recalculate the threshold parameter on the test set.

Other K-medoid based metagenomic binners use different methods to determine the threshold parameter. DNABERT-S [5] uses ground-truth labels to calculate distances between contigs and their species centroid on a calibration dataset. VAMB [17] calculates distributions of pairwise Pearson distances between a set of sampled seed contigs and all other contigs using the latent encoding. Assuming a bimodal distribution, the threshold is defined as the minimum between the initial local maximum and the global maximum. Without relying on ground-truth labels or direct distributional assumptions, our method offers an automated framework of finding an optimal threshold parameter by using an evaluation metric such as CheckM2.

D. Multiple Instance Learning for phenotype classification

Recent studies found substantial associations between the taxonomic profiles of gut microbiome samples and disease [46], [47]. We consider the recovered MAGs from metagenomics binning as the taxonomic profiles of the original samples, and propose a Multiple Instance Learning (MIL) framework to conduct phenotype classification. This evaluates the ability of a metagenomics binning model to differentiate between microbiome samples from healthy and diseased patients. Our proposed MIL framework can be used with any metagenomics binning approach and guarantees that biological entities are aligned across samples to ensure interpretability of the recovered genomes.

1) Multiple Instance Learning framework: Our MIL framework relies on a bag based method using the collective assumption, where bags with similar labels are assumed to contain similar instances, and that all instances contribute to the bag labels [64].

Let each patient sample $s_i \in \mathcal{S}$ be a bag B_i containing $|\mathcal{K}|$ instances where $\mathcal{K}$ is the set of bins obtained from the metagenomics binning step. A bag is defined as:

$$B_i = (x_{i1}, \ldots, x_{i|\mathcal{K}|}) \in \mathbb{R}^{|\mathcal{K}|} \tag{5}$$

where each instance x_{ik} is a scalar value representing the amount of bin k for sample i. A dataset $\mathcal{D}$ with $|\mathcal{S}|$ microbiome samples and corresponding labels $\mathcal{Y}$ is defined as:

$$\mathcal{D} = \{(B_i, y_i) | i = 1, \ldots, |\mathcal{S}|\} \tag{6}$$

where positive bags $y_i = 1$ indicates a diseased patient and $y_i = 0$ indicates a healthy control.

The instances within each bag vary across samples due to the differences in relative abundances of each bin k. To obtain this representation, we group contigs by their bin labels and sum their abundances. Given a contig catalogue $\mathcal{C}$, let $\vec{x}_i \in \mathbb{R}^{|\mathcal{C}|}$ denote the contig abundance vector for sample s_i. Each contig abundance vector is mapped into a lower-dimensional bin abundance vector $\vec{x}'_i \in \mathbb{R}^{|\mathcal{K}|}$, where $|\mathcal{K}| \ll |\mathcal{C}|$, using the following transformation:

$$x_{ik} = \sum_{a=1}^{n_k} \text{abundance}_{ia} \tag{7}$$

where n_k is the number of contigs a in bin k. We then normalize this vector to form a probability distribution over the bins:

$$\vec{x}'_i = (x_{i1}, \ldots, x_{i|\mathcal{K}|}) \cdot \frac{1}{\sum_{k=1}^{|\mathcal{K}|} x_{ik}} \tag{8}$$

Each entry in the normalized vector $\vec{x}'_i$ represents the relative abundance of a bin in sample i and this vector defines the instances in the bag B_i.

Under the collective assumption in our MIL framework, we are particularly interested in whether certain instances correlate with bag labels. Such a correlation would suggest that the relative abundances of certain species are linked to disease or health status. To map instances to species, each bin was taxonomically annotated with all taxonomic ranks using GTDB-tk (v.2.4.1) [65].

2) Classifiers: To map the bags into predicted labels, we define a classifier $f : \vec{x}'_i \rightarrow \hat{y}_i$ that takes the instances $\vec{x}'_i$ as input and outputs a predicted bag label $\hat{y}_i \in \{0, 1\}$.

As a baseline, we employed a k-nearest neighbors (KNN) classifier, assigning each test point to the majority class among its k nearest neighbors.

We further evaluated five logistic regression variants incorporating different regularization strategies: L1 (lasso), L2 (ridge), elastic net (with L1-ratio=0.5), and a non-regularized model.

Additionally, we implemented a sparse group lasso logistic regression, which extends the standard lasso by applying regularization at the individual predictor level and across predefined groups of predictors. This implies that the lasso feature selection can influence both entire groups of predictors and individual predictors [66]. Sparse group lasso has been utilized in genetic studies where the number of features often exceed the number of samples, and where predictors (e.g. genes) form groups according to characteristics such as pathways and functions [66]–[69].

The genomes recovered in metagenomics binning naturally form groups at higher taxonomic levels based on the phylogenetic tree. To determine these groups, we represented each bin by its embedded contigs and computed the 95th percentile Hausdorff distance between all bin pairs, resulting in a $|\mathcal{K}| \times |\mathcal{K}|$ distance matrix. Subsequently, an agglomerative clustering with Ward's method was applied on the Hausdorff distance matrix, with the number of clusters set to $\sqrt{|\mathcal{K}|}$. The resulting clusters defined the groups for the sparse group lasso model. This approach evaluates how well the metagenomic binners are able to capture higher-level structures between genomes.

Finally, we included a Random Forest classifier, as previous metagenomics phenotype studies have demonstrated strong performance using this classifier [46], [47].

Group	# Sequence-Pairs	# Species	# Genus	# Family	# Order	# Class	# Phylum	# Kingdom	# Superkingdom
Bacteria	1,500,001	20,234	3,836	747	282	115	48	4	1
Fungi	500,000	4,767	1,305	457	168	60	8	1	1
Virus	165,599	482	157	39	16	8	7	4	1

TABLE I: Summary statistics of the 2M pre-training dataset. All DNA sequences are $10,000$ bp long. The dataset was assembled in February 2025; as of May 2025, NCBI updated the rank name "Superkingdom" to "Domain" for bacteria and eukaryota, as well as "Acellular root" for virus, see NCBI updates.

To examine the instances that strongly contributed to the bag labels, we identified the features with the largest coefficients (Logistic Regression) or feature importance (Random Forest) averaged over all cross-validation splits. To ensure biological interpretability, we used the taxonomic annotations obtained with GTDB-tk [65] to map the most influential instances to their species.

IV. Experiments

A. Model

We trained three variants of DNABERT-H on two seperate pre-training datasets and used a single validation dataset to select the best model.

1) Pretraining data: Taxonomic annotations and reference genomes were obtained from the NCBI database [70]. Using the reference genomes, we created two pretraining datasets of varying sizes, one using 400k and the other using 2M sequence pairs from bacterial and fungal genomes. Both datasets also included sequence pairs from the same 165k viral genomes. See Appendix C for an overview of the 400k dataset.

Table I provides an overview of the constructed 2M training dataset, consisting of 1,500,001 pairs from 20,234 bacterial genomes, 500,000 pairs from 4,767 fungal genomes, and 165,599 pairs from 482 viral genomes, resulting in a total of 2,165,600 sequence pairs. Additionally, a summary of the tokenized training sequences for the 2M dataset is shown in Appendix E. Most sequences were tokenized into approximately 1,900 to 2,200 tokens.

We also constructed a validation dataset consisting of 50k sequence pairs (See Appendix D for an overview).

2) Model variants: We trained three variants of DNABERT-H: DNABERT-H-400k, DNABERT-H-2M, and DNABERT-H-2M-R. Both DNABERT-H-400k and DNABERT-H-2M were trained following the hierarchy-aware batch sampling strategy from [16]. However, we observed a potential limitation using this strategy. Specifically, as the sampler continuously seeks previously unused samples throughout an epoch, batches constructed in the later stages of training often deviate from the intended hierarchical structure. When subsequent batches use anchor samples from non-bacterial genomes, most of the samples included in a batch actually came from bacterial genomes (see Figure 4). This deviation likely resulted in the unstable validation loss observed during training (see Appendix F). To address this, we further developed DNABERT-H-2M-R, which was trained using a revised sampling strategy. In this strategy, we allowed non-viral genomes to be contained in multiple samples in contrast to the initial sampling strategy while ensuring that the proportion of the anchor genomes was uniformly higher throughout all batches (see Appendix G).

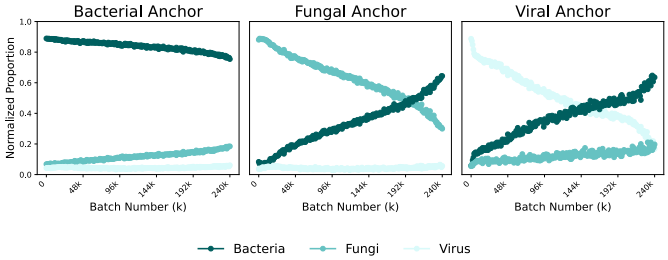

Fig. 4: Variation in the composition of pre-training batches over time when selecting anchors from different groups (Bacteria, Fungi, Virus). Each subplot shows the normalized proportion of samples from each group within batches as the pre-training progresses. The normalized proportion was calculated for every 10,000 batches.

3) Model training: All three variants of DNABERT-H were initialized using the pre-trained DNABERT-2 weights [6], and trained for three epochs using the loss in Eq. (4), with the temperature parameter τ of 0.07. We set a maximum token length to 2,000, and used the AdamW optimizer [71]. A LinearLR scheduler was used over the first 10% of the total training steps and switched to the CosineAnnealingLR scheduler [72] for the remainder of the training steps. We saved checkpoints at every 10,000 steps and selected the checkpoint with the lowest validation loss (see Appendix F).

We used a maximum learning rate of 8×10^{-6} for DNABERT-H-400k and 2×10^{-6} for DNABERT-H-2M and DNABERT-H-2M-R. We set a global batch size of 16 for DNABERT-H-400k and 18 for DNABERT-H-2M and DNABERT-H-2M-R.

Training time of DNABERT-H-400k was 40 hours on 2 NVIDIA A100 80GB GPUs whereas DNABERT-H-2M and DNABERT-2M-R took approximately 72 hours on 2 NVIDIA H100 94GB GPUs.

B. Metagenomics binning

We compared the binning performance of all DNABERT-H variants against three metagenomic binning tools, two gLMs, and two baseline models, on eight different datasets, including CAMI2, a leading metagenomics binning benchmark.

C. K-medoid clustering algorithm

We set the minimum bin size $m = 2$ and let the K-medoid algorithm pick $Z = 2000$ seed centroids. Following the setup

in DNABERT-S [12] we set the number of neighborhood seed updates $T = 3$.

Using a conservative bin size $m = 2$ results in many small bins that unlikely represents complete genomes. Following the setup in VAMB [17] we removed clusters with $\leq 250,000$ base-pairs for the metagenomics binning datasets, and $\leq 200,000$ for the phenotype classification datasets, to get the final bins.

1) Metagenomics binning datasets: We utilized seven single-end short-read synthetic datasets from the Critical Assessment of Metagenome Interpretation Challenge (CAMI2) [21] which are designed to mimic realistic microbiome samples. CAMI2 offers a standardized and comprehensive benchmark for metagenomics binning and is widely used to evaluate binning tools. The challenge provides gold-standard assemblies (GSA) along with reads.

The selected datasets span a diverse range of environments, including five human microbiomes: Airways ($n = 10$), Gastrointestinal ($n = 10$), Oral ($n = 10$), Urogenital ($n = 9$), and Skin ($n = 10$), as well as Marine ($n = 10$) and Plant ($n = 21$) microbiomes, resulting in a total of 80 samples. Each dataset was processed separately.

Following the multi-sample approach, the GSAs from each sample were concatenated into a contig catalogue $\mathcal{C}$, and used as reference when mapping reads during the alignment step to obtain contig co-abundances.

In addition to CAMI2, we included the MetaHIT "error-free" contig dataset from MetaBAT [3]. These contigs are considered "error-free" as they were obtained directly from 290 known reference genomes rather than being constructed from reads by an assembly algorithm. The Metahit dataset does not include co-abundances and thus VAMB, TaxVAMB, and Comebin were not benchmarked using this dataset. See detailed information on the dataset in Appendix H.

Each processed dataset contains between 81K to 439K contigs with a combined length between 1.1G to 3.6G base-pairs (see a detailed overview in Appendix I). Across all eight datasets, the contig lengths vary from 2K to 8.3M base-pairs with a heavy right-tailed distribution. Contigs with lengths ≥ 60K base-pairs were excluded for the gLMs due to memory constraints during embedding calculations, removing between 0% and 3.1% of contigs within each dataset (see Appendix J).

For a contig catalogue obtained from each dataset $\mathcal{C} \in \{\text{Airways}, \ldots, \text{Metahit}\}$, we randomly sampled 10% of its contigs to construct the calibration set $\mathcal{C}_{cal}$, and used the remaining 90% as a test set for metagenomics binning.

2) Benchmarked models: We benchmarked DNABERT-H variants against three metagenomic binning tools: VAMB, TaxVAMB, and Comebin, two gLMs: DNABERT-S, DNABERT-2, and two baseline models: TNF, DNA2Vec.

Tetranucleotide frequencies, **TNF**, is a composition based approach, that represents sequences by computing the relative frequency of each 4-mer, resulting in a 256-dimensional embedding.

DNA2Vec uses a Word2Vec approach to learn static embeddings of sequences and is pre-trained on the human genome using varying k-mer lengths [73]. We computed the dot product of the TNF vectors and the 4-mer DNA2Vec embeddings to obtain a 256-dimensional representation of sequences.

We also evaluated two supplementary models, TNF-Kernel and DNABERT-2-Random. **TNF-Kernel** reduces the 256 dimensional TNF vector to 103 dimensions by merging reverse complements (e.g., AGAT and ATCT). We computed the dot product of the TNF vectors and the `compliment kernel` from VAMBtools (v.4.1.3) to obtain TNF-Kernel. **DNABERT-2-Random** uses the DNABERT-2 architecture with randomly initialized weights. This design follows [74] who demonstrated that gLMs with randomly initialized weights can match or outperform pretrained ones in genomic benchmark tasks. We adopted the same weight initialization scheme as in the paper [74]. Both supplementary models performed poorly on the metagenomics binning task and were omitted from the main results (see Appendix K for these results).

The metagenomics binning tools take contigs and co-abundances as input and directly output metagenomic bins. In contrast, DNABERT-H, DNABERT-S, DNABERT-2, TNF, and DNA2Vec only use composition based features, and require a separate clustering step to form bins. For these, we used the K-medoid algorithm (see Appendix A).

Comebin incorporates single-copy marker genes (SCG) in a re-clustering step during binning. This is problematic because the evaluation tool CheckM2 [20] also relies on SCGs, as highlighted in TaxVAMB [18]. We obtained the bins prior to the SGC re-clustering step.

Our previous benchmark paper [10] revealed that among six gLMs [6]–[9], [12], [23], only DNABERT-S outperformed the baselines in metagenomics binning (see Appendix L for these results). Based on these findings, we only included DNABERT-S and DNABERT-2. DNABERT-2 was included as it serves as the backbone model for DNABERT-H.

We also note that other metagenomics binning tools are available [36], [75]–[77]. Due to time constraints, these were not included.

Detailed information on the commands used for the binning tools and checkpoints of the gLMs can be found in Appendix M.

3) Evaluation: We used CheckM2 (v.1.1.0) [20] to assess the quality of the identified bins. CheckM2 relies on universal single-copy marker genes (USCGs) - genes found exactly once in nearly every genome - to estimate the completeness and contamination of bins, which determine the quality of the recovered genome [78], [79].

CheckM2 searches each bin for their expected USCGs using a genome database. Completeness is estimated as the fraction of expected USCGs recovered in the bin, providing an estimate of how fully the target genome has been reconstructed. Contamination is estimated as the fraction of USCGs that exists more than once (extra copies), indicating whether bins contain DNA from other genomes. CheckM2 reports high completeness and low contamination if a bin has all its USCGs exactly once. Completeness is analogous to recall whereas contamination is analogous to precision.

Following previous works [19], [57], we measured performance as the number of bins achieving completeness $\geq 50\%, 60\%, 70\%, 80\%, 90\%$ with contamination $\leq 5\%$. We

denote MAGs with completeness $\geq 90\%$ as near-complete (NC) and MAGs with completeness $\geq 50\%$ as moderate quality (MQ). Near-complete MAGs can be treated as a genome in further downstream analysis [17], [18], [36].

D. Threshold calibration

For each model and calibration dataset $\mathcal{C}_{cal}$, we conducted a hyper-parameter search for $K \in \{100, 200, \ldots, 900, 1000\}$ and $P \in \{25, 50, 75\}$.

For each combination of (K, P), we computed a performance score $S_{K,P}$ as the weighted average of the number of recovered bins across contamination thresholds $c \in \{5, 10, 15, 20\}$ using the weights $w_c \in \{\frac{1}{1}, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$:

$$S_{K,P} = \sum_c w_c \cdot B_{K,P}(c) \tag{9}$$

where $B_{K,P}(c)$ is the number of bins obtained with completeness $\geq 50\%$ and contamination $\leq c\%$. This resulted in 30 different performance scores $S_{K,P}$. In the event of a tie between performance scores, we selected the K that minimized $|K - \sqrt{|\mathcal{C}_{cal}|}|$, as $K = \sqrt{N}$ is a commonly used heuristic for choosing the neighborhood size in KNN algorithms [80]. If a tie still existed given the selected K, we chose the minimum value of P. The selected combination of (K, P) values was subsequently used to recalculate the threshold parameter γ on the test set.

Appendix N presents the selected (K, P) parameters, threshold, and similarity distributions for each model and test dataset. When using the same model, the distribution shapes and corresponding thresholds are relatively consistent across datasets. For instance, DNABERT-H-2M yields thresholds ranging from 0.72 to 0.86. In contrast, applying different models to the same dataset leads to high variation in the similarity distributions and thresholds. For example, all the baseline models yields low-variance distributions whereas DNABERT-H and -S shows higher variance. These observations reflect the distinct structural properties of the embedding spaces across models, and highlight the importance of calibrating γ to each model and dataset.

The performance scores $S_{K,P}$ for every combination of (K, P) across models and calibration sets are presented in Appendix O. We observe no consistent pattern of the selected K and P values that yielded the highest performance scores among combinations of models and datasets.

E. Multiple Instance Learning for phenotype classification

We used the metagenomic bins from DNABERT-H, DNABERT-S and VAMB as taxonomic profiles in our MIL framework to evaluate their ability to distinguish between microbiomes from healthy and diseased patients using two different datasets: T2D-EUW and WEGOVY. In addition, we compared the results of our MIL framework on the T2D-EUW dataset to existing reported results from MetAML [47] and Deepmicro [46]. We used the DNABERT-H variant that had the best performance on the metagenomics binning task.

1) Datasets: The metagenomic study on type 2 diabetes, (**T2D-EUW**) [48], included $n = 145$ patients from a European women cohort partitioned into three categories: healthy controls, pre-diabetic and diagnosed with T2D. From the study, we included samples and their phenotype labels from all T2D diagnosed patients (53), and healthy controls (43), resulting in 96 samples. The samples were downloaded from the NCBI database [81].

The second metagenomic study, **WEGOVY**, is an ongoing study in Copenhagen on obesity and the treatment drug WEGOVY [22]. WEGOVY has recently gained a lot of attention due to its effects in obesity treatment. Understanding the influence of WEGOVY on the gut microbiome, particularly side-effects such as delayed gastric emptying, offers a promising direction for future research [82], [83]. The study aims to expose different patients to various dosage schemes while collecting gut microbiome samples before, during, and after treatment, to investigate the associations between changes in the gut microbiota and effective weight loss.

The study includes three different cohorts subjected to various stool sample testing schemes. Cohort 1 and 2 entail a baseline test before drug treatment, three tests during treatment, and one post treatment. Cohort 3 involves three baseline tests, bi-weekly tests during treatment and post treatment tests. Cohorts 1 and 2 only included 43 and eight baseline tests respectively, whereas the third cohort included samples from all 24 patients at all test times up until week 12.

We included two samples per patient in cohort 3: The third of the baseline tests, and a test at week 10 during treatment, resulting in $n = 48$ samples. Samples were assigned binary labels based on collection time: 0 for the baseline sample and 1 for the week 10 sample. This approach allows us to examine whether WEGOVY treatment leads to detectable changes in the gut microbiome.

Both datasets comprise short paired-end reads in FASTQ files from all participating patients. Both datasets were processed using the multi-sample approach (see Appendix I for datasets summaries).

Following the same approach as in the metagenomics binning experiment, 10% of the contigs from a contig catalogue $\mathcal{C} \in \{\text{T2D}-\text{EUW}, \text{WEGOVY}\}$, were used to determine the threshold parameter on a calibration set $\mathcal{C}_{cal}$. The remaining 90% of contigs were considered the test set and used for metagenomics binning to obtain the taxonomic profiles used in the phenotype classification.

2) Evaluation: We assessed classifier performance using the area under the ROC curve (AUC) which is commonly used in metagenomics phenotype classification studies [46], [47]. Performance was summarized using the mean and standard deviation over a stratified 10-fold cross-validation repeated 5 times. We ensured that each fold preserved the class proportions and that all classifiers were evaluated on the same splits.

Within each training fold, we conducted a nested 5-fold cross-validation to tune the hyperparameters using grid search. Once the best hyperparameter combination was identified, the classifier was retrained on the full training fold and evaluated on the held-out test fold. See Appendix P for the hyperparameter grids used for the classifiers.

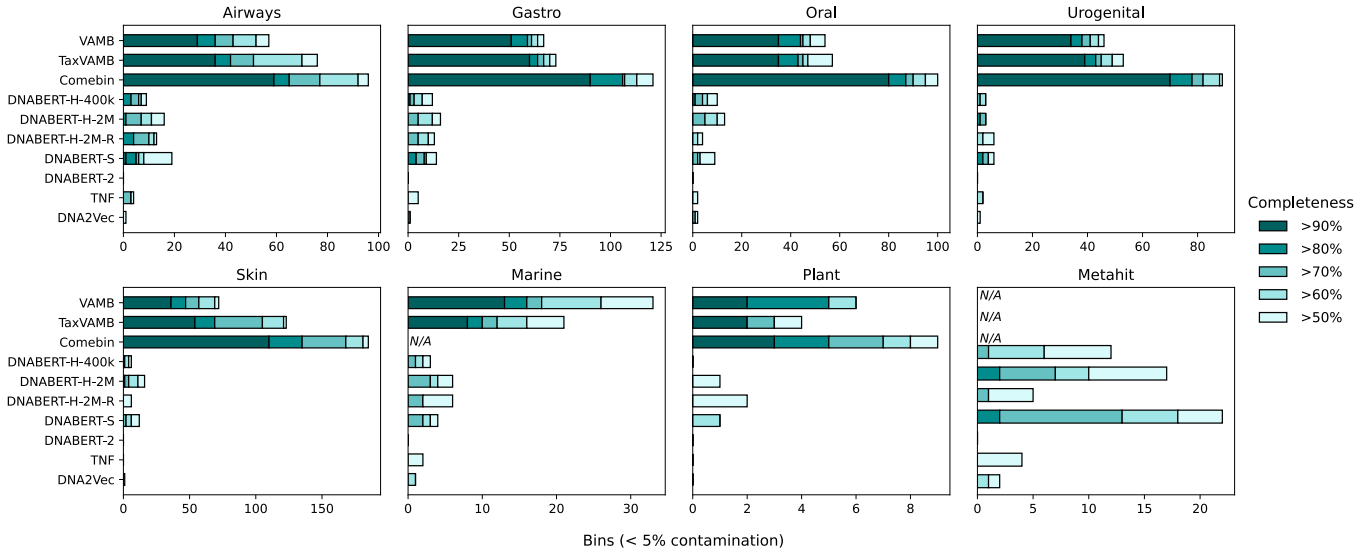

Fig. 5: The number of recovered bins with contamination $\leq 5\%$ at varying completeness thresholds across all metagenomics binning datasets. The benchmarked metagenomics binning tools VAMB, TaxVAMB and Comebin substantially outperformed genomic language models, including DNABERT-H. Comebin could not be run on the Marine dataset due to computational resource constraints, and neither VAMB, TaxVAMB, nor Comebin could be evaluated on the Metahit dataset.

To evaluate the classifiers' performance against a random learner, we conducted a permutation test on the labels to estimate the probability of our cross-validated score under the null-hypothesis, which assumes no relationship between features and target. The null distribution was generated by permuting the target labels and calculating the average 10-fold cross-validation score with 1,000 repetitions. A p-value below 0.05 indicates that the observed AUC is unlikely to be obtained given the null hypothesis [84]. The hyperparameter grid search was omitted in the permutation test. Instead, we used the most frequent hyperparameters identified in the optimized classifiers.

V. Results

A. Metagenomics binning

1) Overall binning results: We show the number of bins recovered by each model in Figure 5. The metagenomic binners VAMB, TaxVAMB, and Comebin consistently outperformed the gLMs and baseline models by a large margin, recovering up to 10 times more near-complete bins ($\geq 90\%$ completeness). Comebin had the best performance in six out of eight datasets, recovering more near-complete bins than both VAMB and TaxVAMB. TaxVAMB surpassed VAMB in the five human datasets - Airways, Gastro, Oral, Urogenital, Skin - whereas VAMB outperformed TaxVAMB in Marine and Plant datasets.

Across all eight datasets, the gLMs trained with a contrastive learning objective, DNABERT-H-400k, DNABERT-H-2M, DNABERT-H-2M-R, and DNABERT-S recovered more or equal number of bins at all completeness thresholds compared to the two baselines, TNF and DNA2Vec. In contrast, DNABERT-2 failed to identify any moderate quality bins, highlighting that a pure MLM pretraining objective is insufficient for differentiating species in the embedding space.

DNABERT-H-2M performed better than DNABERT-S in four out of eight datasets, recovering a total of 88 moderate-quality bins compared to DNABERT-S's 87 across all datasets, highlighting a comparable performance between the two models.

DNABERT-H-2M, trained on the full pretraining dataset, consistently outperformed its smaller variant DNABERT-H-400k. Additionally, DNABERT-H-2M consistently outperformed its other variant DNABERT-H-2M-R, indicating that our strategy of maintaining fixed proportions of fungal and viral sequences during training was ineffective.

2) Bin counts and performance: Across datasets, the best performing model, Comebin, yielded the highest average number of bins (510). Oppositely, DNABERT-2 (506) yielded a high bin count, yet failed to recover any genomes across datasets. In addition, the difference between the average number of bins in DNABERT-2 and DNABERT-2-Random (72) suggests that the MLM pretraining imposes some structure on the embedding space however, this structure fails to distinguish between species. See detailed results on bin counts in Appendix Q.

3) Computational resources: We report the runtimes for all model across datasets in Appendix R. The baseline methods were the fastest, with DNA2Vec and TNF taking 8 and 9 minutes on average for a dataset, followed by VAMB at 34 minutes. The gLMs — DNABERT-H, DNABERT-S, and DNABERT-2 — took substantially longer at 172 minutes on average for a dataset, while TaxVAMB and Comebin were the slowest at 215 and 275 minutes.

These runtimes highlight that gLMs demand ≈ 5 times more computational resources than VAMB, despite yielding lower performance. The strongest performing binners, Comebin and TaxVAMB, recover the highest number of bins but also require the longest runtimes. In a resource-constrained setting, VAMB

offers a good trade-off between efficiency and performance.

B. Multiple Instance Learning for phenotype classification

1) General results: The AUC performance of all classifiers is presented in Table II. On the T2D-EUW dataset, the two existing metagenomic approaches MetAML (76.2 AUC) and DeepMicro (89.9 AUC) outperformed all classifiers within our MIL framework.

Among the proposed MIL classifiers, we found the best performance using Random Forests, achieving an AUC of up to 73.1 with $p \leq 0.001$ using VAMB. This performance is slightly lower than MetAML.

The results generally showed high standard deviations across cross-validation splits, reaching up to 20.1 SD in the T2D-EUW dataset. We attribute this variability to the small sample size ($n = 96$), which limits the data available in each fold and leads to unstable performance.

2) VAMB outperformed DNABERT-S and DNABERT-H-2M: Among the taxonomic profiles extracted from the three binning approaches, using the recovered bins from VAMB to construct the instance representations resulted in the highest overall classification performance with a mean AUC of 73.1 on the T2D-EUW dataset. This corroborates with VAMB recovering a substantially higher number of bins at moderate quality on the T2D-EUW dataset (385) compared to DNABERT-S (9) and DNABERT-H-2M (3) (see detailed binning results on phenotype datasets in Appendix S). These findings indicate that the effectiveness of the MIL framework is influenced by the amount of biological signal captured in the metagenomics binning step. When high-quality bins are constructed, the MIL classifiers have better signal available to discern between phenotypes. This finding is also reflected in the superior performance of MetAML and Deepmicro that utilize less noisy taxonomic profiles obtained from reference databases.

However, the results also show that even poor binning results such as DNABERT-H-2M on T2D-EUW dataset, can achieve classification performance of up to 67.5 AUC. Notably, all three binning approaches produced a comparable number of bins, ranging from 1,569 to 2,613 across the two datasets (see Appendix Q). This suggests that informative signal persists even in low-quality bins.

3) Results on WEGOVY dataset: Despite VAMB recovering 231 bins of moderate quality from the WEGOVY dataset, none of its classifiers achieved a significant performance. This suggests that no discernible signal was present to separate microbiome samples after 10 weeks of treatment and prior to drug treatment. We believe the poor results are partly due to the limited sample size in each class (24 samples) and the nature of the outcome variable itself. In Appendix T, we present t-SNE visualizations of the bags obtained from the WEGOVY dataset encoded using two alternative outcome variables: *BMI Loss Percent* after 10 weeks and *Cumulative Dosage* of the drug over the treatment period. While some variation is visible between pre- and post-treatment bags, most of the observed differences appear to be driven by individual variability. In addition, the bags show no clear separation based on neither alternative outcome variable. The BMI Loss Percent and Cumulative Dosage variables were not used as outcome variables as these would have halved the number of samples in the training dataset. We exclude the WEGOVY dataset from further analyses given the limited performance observed.

TABLE II: Classification performance (mean ± SD AUC) on two phenotype datasets using our proposed MIL framework. Significance versus the permutation null distribution is denoted as $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***).

	Model	T2D-EUW	WEGOVY
	MetAML[a][47]	76.2 (11.1)[c]	–
	DeepMicro[b][46]	89.9 (8.0)[c]	
VAMB	**KNN classifier**	66.9 (16.5)**	41.7 (26.5)
	Logistic Regression		
	+ No penalty	67.2 (17.2)*	30.7 (24.1)
	+ Lasso	63.8 (18.7)*	47.0 (10.9)
	+ Ridge	67.6 (16.0)*	23.2 (21.7)
	+ Elastic-net	67.0 (17.9)*	46.7 (11.1)
	+ Group Sparse Lasso[d]	–	–
	Random Forest	**73.1 (18.5)*****	34.2 (27.0)
DNABERT-S	**KNN classifier**	55.6 (14.9)	48.6 (27.5)
	Logistic Regression		
	+ No penalty	61.9 (19.0)	40.2 (28.0)
	+ Lasso	62.9 (19.1)*	52.7 (25.0)
	+ Ridge	**64.5 (17.2)***	56.0 (31.1)
	+ Elastic-net	59.4 (19.1)	50.5 (27.9)
	+ Group Sparse Lasso	59.4 (17.8)	47.4 (31.2)
	Random Forest	64.0 (17.1)*	45.2 (24.9)
DNABERT-H-2M	**KNN classifier**	60.4 (16.5)	40.4 (26.4)
	Logistic Regression		
	+ No penalty	63.6 (18.6)*	39.5 (27.6)
	+ Lasso	**67.5 (19.4)****	44.7 (16.7)
	+ Ridge	65.1 (19.5)*	46.7 (26.0)
	+ Elastic-net	63.2 (20.1)*	43.3 (17.6)
	+ Group Sparse Lasso	64.7 (18.3)*	49.0 (30.2)
	Random Forest	67.0 (19.4)*	49.0 (30.8)

[a] MetAML uses a Random Forest.
[b] DeepMicro uses a shallow autoencoder combined with Random Forest.
[c] Reported results are copied from original papers.
[d] Group sparse lasso was not performed for VAMB because the contig embedding data could not be accessed.

4) DNABERT-H-2-M outperformed DNABERT-S: The results show that our model DNABERT-H-2M outperformed DNABERT-S across all classifiers on the T2D-EUW dataset. DNABERT-H-2M achieved the best AUC score of 67.5 using lasso regression while DNABERT-S achieved its best AUC score of 64.5 using ridge regression.

In the group sparse lasso regression, DNABERT-H-2M outperformed DNABERT-S by +5.3 AUC. In addition, the group sparse lasso regression ranked fourth among DNABERT-H-2M classifiers, while only ranking sixth among the DNABERT-S classifiers. These results indicate that DNABERT-H-2M obtained more meaningful groups in the agglomerative clustering algorithm compared to DNABERT-S. This suggests that DNABERT-H-2M better captures hierarchical structures between species compared to DNABERT-S.

Figure 6 provides an exploratory comparison of the hierarchical structures captured in DNABERT-S and DNABERT-H-2M. It shows the Hausdorff distance matrix sorted by the agglomerative clustering linkage, annotated with the tax-

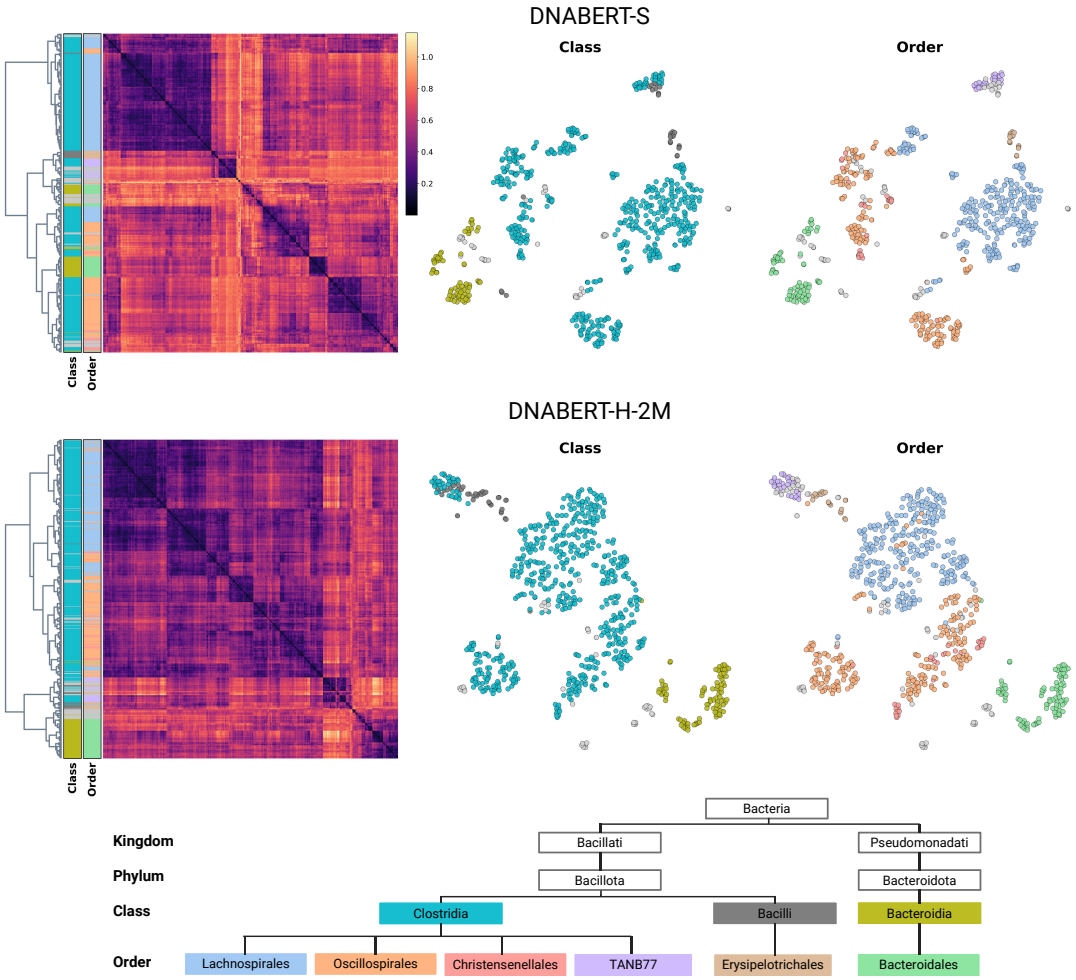

Fig. 6: Hierarchical taxonomy awareness of DNABERT-S and DNABERT-H-2M on the T2D-EUW dataset. **Left:** Hausdorff distance matrix with the dendrogram of the agglomerative clustering and corresponding labels at ranks class and order. **Right:** Hausdorff distances visualized using t-SNE with class and order labels. The labels are shown in the bottom as a phylogenetic tree. Only bins that could be assigned a label from GTDB-tk are included; DNABERT-S ($n = 648$), DNABERT-H-2M ($n = 751$). Bins from the 6 largest orders and corresponding 3 classes are colored, while the remaining bins are light grey.

onomies class and order, alongside a t-SNE visualization of the same distances. The taxonomic ranks are generally well separated in both models, but DNABERT-H-2M more clearly preserves higher-order phylogenetic relationships. For example, the classes *Clostridia* and *Bacteroidia* form coherent clusters in DNABERT-H-2M, but are dispersed into multiple clusters in DNABERT-S. A similar pattern is observed for the orders *Lachnospirales*, *Oscillospirales*, and *Bacteroidales*. Additionally, the distance matrix of DNABERT-H-2M show stronger hierarchical partitioning of the species.

While these findings are exploratory, they suggest that DNABERT-H-2M more faithfully preserves phylogenetic structures between species compared to DNABERT-S. Nonetheless, they also show that DNABERT-S captures some higher-order structures without explicitly incorporating taxonomic information into its training objective.

5) Influential instances: Figure 7 shows the ten most influential instances recovered with at least $\geq 50\%$ completeness from the best classifier for each metagenomics binner. Without the completeness threshold, the identified instances still rank among the top 25 most important features.

DNABERT-S and DNABERT-H-2M had substantial overlaps between their most influential species, yet none of the highly ranked features satisfied the contamination threshold $\leq 10\%$.

Four of the most influential instances identified by the Random Forest classifier using VAMB had contamination levels $\leq 10\%$. In contrast, the logistic regression variants of VAMB did not result in any high influential features meeting the same contamination threshold (see Appendix U). This aligns with the overall classification results which found that the combination of VAMB and Random Forest resulted in the best AUC of 73.1. This suggests that the Random Forest model is able to find salient relationships between MAGs and phenotypes.

VI. Discussion

In this paper, we introduced DNABERT-H, a specialized gLM that preserves the phylogenetic structures between

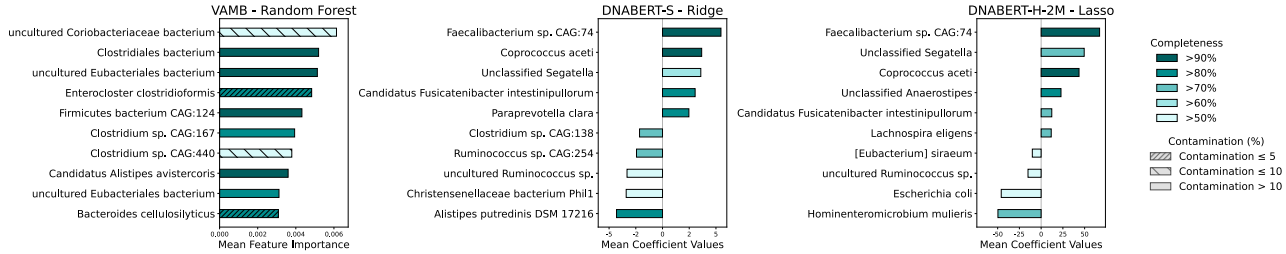

Fig. 7: Ten most influential instances recovered with at least $\geq 50\%$ completeness from the best classifier for each metagenomics binner on the T2D-EUW dataset. The instances are annotated with labels obtained using GTDB-tk and subsequently converted to their equivalent NCBI labels. Colors indicate bin completeness, while the hatches indicate the contamination thresholds.

species. Our approach incorporates labels from all eight taxonomic ranks during training by adopting the Hiearchical Multi-label Contrastive Learning framework [16].

A. Performance gaps in metagenomics binning

We benchmarked DNABERT-H in a zero-shot setting against several state-of-the-art binning tools across eight metagenomic datasets. To our knowledge, this is the first evaluation of gLMs against state-of-the-art metagenomic binners using the widely adopted evaluation tool CheckM2 [20]. DNABERT-S only compared their model to other gLMs and evaluated performance using ground-truth species labels.

Our results showed that while the binning tools VAMB, TaxVAMB, and Comebin consistently recovered substantially more near-complete genomes than gLMs, DNABERT-H achieved a performance on par with DNABERT-S while outperforming MLM-trained gLMs and baselines.

Other binning tools that rely exclusively on composition based features have been shown to yield sub-optimal performance in metagenomics binning compared to hybrid approaches [17], [19], [36], [56]. Similarly, DNABERT-H and DNABERT-S learn DNA sequence representations through pretraining on tokenized input, without incorporating co-abundance information. This difference in input modalities between state-of-the-art binning tools and DNABERT-H and DNABERT-S partly explains the performance gaps seen in Figure 5 and Appendix S.

In addition, binning using gLMs was only conducted on a subset of the data, as we reserved 10% of contigs for threshold calibration, while discarding contigs $\geq 60{,}000$ base-pairs, reducing the number of contigs by approximately 10 to 13%. This might have led to a modest decrease in binning performance, making it more difficult for gLMS to achieve high-completeness bins.

VAMB, TaxVAMB, and ComeBin follow the transduction learning framework, where parameters are optimized directly on the full dataset that is subsequently used for evaluation [85]. This implies that the binning tools guide their internal representations to each individual dataset, whereas DNABERT-H and DNABERT-S rely on representations learned during pre-training and do not optimize parameters to the evaluation datasets. We view this distinction between binning tools and gLMs to be the largest contributing factor to their performance gaps. To compete with existing state-of-the-art tools, designing finetuning tasks for gLMs that include contig co-abundances serve as a promising and unexplored direction for advancing the metagenomics binning performance of gLMs.

B. DNABERT-H efficiently learns hierarchical structures

DNABERT-H achieved metagenomic binning performance comparable to DNABERT-S and consistently performed better in the phenotype classification task. Both models effectively distinguished higher-order taxonomic ranks, but DNABERT-H more faithfully preserved the phylogenetic structures between species (see Figure 6). This difference is attributed to the Hierarchical Multi-label Contrastive Learning framework [16] in DNABERT-H, which explicitly learns phylogenetic structures. In contrast, because closely related species share more genetic material and thus are expected to produce similar embeddings, DNABERT-S implicitly learns phylogenetic structures by using a species level SimCLR based framework [14].

While both models rely on contrastive learning, DNABERT-S incorporates MI-Mix to regularize the model, which increases performance by about 2% in their downstream tasks [12]. DNABERT-H achieves similar performance without any explicit regularization, suggesting that contrastive learning is the primary contributor to metagenomics binning performance. Future work could explore regularization strategies for DNABERT-H to achieve similar performance improvements in downstream tasks.

Hierarchical Multi-label Contrastive learning serves as a promising framework that, by exploiting known taxonomic structures, not only learns to distinguish species but also preserve the hierarchical relationships between them.

C. Pretraining data

We found that DNABERT-H-2M outperformed its smaller variant DNABERT-H-400k, and that DNABERT-H-2M-R trained on fewer unique sequences underperformed compared to DNABERT-H-2M. These results suggest that pre-training on more DNA sequences results in richer representations that improves performance on downstream tasks.

Our best performing model DNABERT-H-2M had a higher exposure to bacterial genomes throughout its training compared to DNABERT-H-2M-R (see Figure 4). The increased bacterial exposure coincides with the taxonomic diversity in human microbiomes and may explain the improved performance of DNABERT-2M [42], [86], [87].

Future work could explore optimal strategies for balancing the ratio of bacterial, fungal and viral genomes seen during pretraining to improve downstream performance.

D. MLM transfer learning

Both DNABERT-S and DNABERT-H apply transfer learning using the pre-trained weights of DNABERT-2. However, there is no consensus on the effect of MLM pre-training in gLMs when applied to zero-shot or few shot settings within genomics [28], [88]. Two approaches that build on the gLM Nucleotide Transformer [30], AgroNT [89] and SegmentNT [90], achieve superior performance in downstream genomic tasks, while Vishniakov et al. [74] argues that the pre-trained gLMs [6], [7], [9], [23], [30], [91] perform equally well to their randomly initialized counterparts in downstream tasks.

We found that neither DNABERT-2 nor its randomly initialized counterpart recovered any bins in zero shot metagenomics binning. However DNABERT-2 produced 506 clusters in the K-medoid algorithm compared to just 72 clusters of DNABERT-2-Random, suggesting that MLM pretraining captures some structural features of DNA sequences.

E. Phenotype classification and Multiple Instance Learning

Metagenomics binning serve as key step in metagenome wide association studies (MWAS) by reconstructing the original genomes from microbiome samples, to uncover associations between diseases and the functional and taxonomic profiles of complex communities [92]. We evaluated the ability of our model DNABERT-H along with DNABERT-S and VAMB to find taxonomic profiles associated with type 2 diabetes using our proposed Multiple Instance Learning framework.

VAMB achieved the best classification performance on the T2D-EUW dataset, indicating that richer taxonomic profiles lead to better classification results Notably, DNABERT-H consistently outperformed DNABERT-S, indicating that the hierarchical loss in DNABERT-H results in more informative taxonomic profiles compared to the flat relationships learned by DNABERT-S.

The best AUC score of our proposed MIL framework was unable to match the performance of the existing results from MetAML [47] and DeepMicro [46] on the T2D-EUW dataset. In contrast to MetAML and DeepMicro, our proposed framework enables the discovery of associations between novel microbial species and diseases. Additionally, our MIL framework guarantees that instances consistently correspond to the same biological entities across bags, which is crucial for model interpretability.

DNABERT-H and DNABERT-S identified one novel species within the *Segatella* lineage among their top three most influential instances associated with type 2 diabetes (see Figure 7). This aligns with the finding that the closely related species *Segatella copri* (formerly *Prevotella copri* [93]) is more abundant in type 2 diabetics and aggravates insulin resistance [94]. The *Segatella copri* was also identified in the logistic regression variants of VAMB (see Appendix U).

In addition to finding novel species, our MIL framework identifies known markers for type 2 diabetes.

VAMB identified the MAG *Enterocloster clostridioformis* as one of the most contributing species to type 2 diabetes, which has also been linked to the disease in other studies [95]. High abundances of the genus *Alistipes*, has previously been associated with type 2 diabetes and was identified in the species *Candidatus Alistipes avistercoris* by VAMB [96], [97].

Both DNABERT-H and DNABERT-S identified *Faecalibacterium sp. CAG:74* as the most positively correlated species with type 2 diabetes, and the genus has been demonstrated to increase the risk of the disease [96].

Finally, DNABERT-S associated the species *Christensenellaceae bacterium Phil1* with healthy gut microbiome samples. This coincides with the previous finding, that high abundances of its genus *Christensenellaceae* has been linked to healthy patients [98], [99].

F. Implications and future work

Most gLMs focus on single-genome applications relying on MLM pretraining, but their applications in metagenomics remains largely unexplored. Our findings highlight that metagenomics binning and profiling of microbial communities requires specialized gLMs that have the ability to distinguish species and phylogenetic structures. However, current state-of-the-art metagenomics binners outperform DNABERT-H and DNABERT-S by up to an order of magnitude, highlighting that the practical utility of existing gLMs in metagenomics remains limited.

Future work could replace the composition based features with learnable gLM embeddings as inputs to existing metagenomics binning tools. This approach could offer a better understanding of the information contained in the learned representations of gLMs as opposed to TNFs. GLMS introduce computational overhead, but might offer richer feature representations that could enhance binning performance.

We evaluated MLM pretrained gLMs in metagenomics binning however, neither DNABERT-S nor DNABERT-H have been evaluated in single-genomic tasks such as predicting gene regulatory elements or transcription factor binding sites. Future work could benchmark DNABERT-H in these tasks to better understand the trade-offs between pre-training objectives.

VII. Conclusion

In this paper, we introduced a novel gLM, DNABERT-H, that leverages the phylogenetic tree to learn the hierarchical relationships between species. DNABERT-H was trained on more than 2 million sequences from bacterial, fungal and viral reference genomes, using the adopted Hierarchical Multi-label Contrastive Learning framework. This enables DNABERT-H to distinguish between species while preserving the evolutionary relationships between them.

We compared the performance of DNABERT-H to existing state-of-the-art binning tools, DNABERT-S, the MLM pretrained DNABERT-2, and two baseline models in metagenomics binning on datasets from CAMI2.

We showed that DNABERT-H achieved similar performance to DNABERT-S in metagenomics binning without using regularization techniques. Additionally, we demonstrated that ex-

isting state-of-the-art binning tools substantially outperformed DNABERT-H, DNABERT-S and DNABERT-2.

We proposed a Multiple Instance Learning (MIL) framework for phenotype classification that utilizes the taxonomic profiles from any metagenomics binning tool to explore the associations between human gut microbiome samples and disease. Our proposed MIL framework ensures interpretability by identifying the most influential instances contributing to the bag labels. Although our framework performed worse than existing approaches that rely on reference genome databases, it enables the discovery of associations between novel microorganisms and phenotypes.

We found that richer taxonomic profiles led to better classification results, by showing that VAMB resulted in the best overall performance. We also showed that our proposed model DNABERT-H yielded richer taxonomic profiles than DNABERT-S by consistently achieving superior performance in the phenotype classification task.

To conclude, this study highlights that explicitly learning the hierarchical structures between species increases performance in downstream tasks, but that gLMs still underperform compared to state-of-the-art binners. Future directions for advancing gLMs in metagenomics include fine-tuning on metagenomics tasks while incorporating co-abundance information.

REFERENCES

[1] R. Sleator, C. Shortall, and C. Hill, "Metagenomics," *Letters in Applied Microbiology*, vol. 47, no. 5, pp. 361–366, Nov. 2008, ISSN: 0266-8254. DOI: 10.1111/j.1472-765X.2008.02444.x. [Online]. Available: https://doi.org/10.1111/j.1472-765X.2008.02444.x (visited on 05/15/2025).

[2] J. C. Wooley, A. Godzik, and I. Friedberg, "A Primer on Metagenomics," en, *PLOS Computational Biology*, vol. 6, no. 2, e1000667, Feb. 2010, Publisher: Public Library of Science, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000667. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000667 (visited on 05/15/2025).

[3] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," en, *PeerJ*, vol. 3, no. 8, e1165, 2015, ISSN: 2167-8359. DOI: 10.7717/peerj.1165. [Online]. Available: https://escholarship.org/uc/item/58v471ws (visited on 11/29/2024).

[4] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, *et al.*, "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea," *Nature Biotechnology*, vol. 35, no. 8, pp. 725–731, 2017, ISSN: 1087-0156. DOI: 10.1038/nbt.3893. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6436528/ (visited on 05/19/2025).

[5] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, Aug. 2021, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab083. [Online]. Available: https://doi.org/10.1093/bioinformatics/btab083 (visited on 12/01/2024).

[6] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, *DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome*, arXiv:2306.15006, Mar. 2024. DOI: 10.48550/arXiv.2306.15006. [Online]. Available: http://arxiv.org/abs/2306.15006 (visited on 12/01/2024).

[7] H. Dalla-Torre, L. Gonzalez, J. M. Revilla, *et al.*, *The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics*, en, Pages: 2023.01.11.523679 Section: New Results, Jan. 2023. DOI: 10.1101/2023.01.11.523679. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.01.11.523679v1 (visited on 12/01/2024).

[8] M. Sanabria, J. Hirsch, P. M. Joubert, and A. R. Poetsch, "DNA language model GROVER learns sequence context in the human genome," en, *Nature Machine Intelligence*, vol. 6, no. 8, pp. 911–923, Aug. 2024, Publisher: Nature Publishing Group, ISSN: 2522-5839. DOI: 10.1038/s42256-024-00872-0. [Online]. Available: https://www.nature.com/articles/s42256-024-00872-0 (visited on 12/01/2024).

[9] V. Fishman, Y. Kuratov, A. Shmelev, *et al.*, *GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences*, en, Pages: 2023.06.12.544594 Section: New Results, Aug. 2024. DOI: 10.1101/2023.06.12.544594. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.06.12.544594v3 (visited on 12/01/2024).

[10] *Benchmarking-DNA-Foundation-Models-on-Binning-Human-Gut-Microbial-Strains*, Jan. 2025. [Online]. Available: https://github.com/eisuke119/Research-Project/blob/main/Benchmarking-DNA-Foundation-Models-on-BInning-Human-Gut-Microbial-Strains.pdf (visited on 05/04/2025).

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, *A Simple Framework for Contrastive Learning of Visual Representations*, arXiv:2002.05709, Jul. 2020. DOI: 10.48550/arXiv.2002.05709. [Online]. Available: http://arxiv.org/abs/2002.05709 (visited on 12/05/2024).

[12] Z. Zhou, W. Wu, H. Ho, *et al.*, *DNABERT-S: Pioneering Species Differentiation with Species-Aware DNA Embeddings*, arXiv:2402.08777 [q-bio], Oct. 2024. DOI: 10.48550/arXiv.2402.08777. [Online]. Available: http://arxiv.org/abs/2402.08777 (visited on 01/27/2025).

[13] J. Li, P. Zhou, C. Xiong, and S. C. H. Hoi, *Prototypical Contrastive Learning of Unsupervised Representations*, arXiv:2005.04966 [cs], Mar. 2021. DOI: 10.48550/arXiv.2005.04966. [Online]. Available: http://arxiv.org/abs/2005.04966 (visited on 06/01/2025).

[14] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya.," en, *Proceedings of the National Academy of Sciences*, vol. 87, no. 12, pp. 4576–4579, Jun. 1990, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.87.12.4576. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.87.12.4576 (visited on 05/13/2025).

[15] H. Philippe, H. Brinkmann, D. V. Lavrov, *et al.*, "Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough," en, *PLOS Biology*, vol. 9, no. 3, e1000602, Mar. 2011, Publisher: Public Library of Science, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000602. [Online]. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000602 (visited on 05/13/2025).

[16] S. Zhang, R. Xu, C. Xiong, and C. Ramaiah, *Use All The Labels: A Hierarchical Multi-Label Contrastive Learning Framework*, arXiv:2204.13207 [cs], Apr. 2022. DOI: 10.48550/arXiv.2204.13207. [Online]. Available: http://arxiv.org/abs/2204.13207 (visited on 05/06/2025).

[17] J. N. Nissen, J. Johansen, R. L. Allesøe, *et al.*, "Improved metagenome binning and assembly using deep variational autoencoders," en, *Nature Biotechnology*, vol. 39, no. 5, pp. 555–560, May 2021, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/s41587-020-00777-4. [Online]. Available: https://www.nature.com/articles/s41587-020-00777-4 (visited on 11/29/2024).

[18] S. Kutuzova, P. Piera, K. N. Nielsen, *et al.*, *Binning meets taxonomy: TaxVAMB improves metagenome binning using bi-modal variational autoencoder*, en, Pages: 2024.10.25.620172 Section: New Results, Oct. 2024. DOI: 10.1101/2024.10.25.620172. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.10.25.620172v1 (visited on 04/27/2025).

[19] Z. Wang, R. You, H. Han, W. Liu, F. Sun, and S. Zhu, "Effective binning of metagenomic contigs using contrastive multi-view representation learning," en, *Nature Communications*, vol. 15, no. 1, p. 585, Jan. 2024, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-023-44290-z. [Online]. Available: https://www.nature.com/articles/s41467-023-44290-z (visited on 04/29/2025).

[20] A. Chklovski, D. H. Parks, B. J. Woodcroft, and G. W. Tyson, "CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning," en, *Nature Methods*, vol. 20, no. 8, pp. 1203–1212, Aug. 2023, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-023-01940-w. [Online]. Available: https://www.nature.com/articles/s41592-023-01940-w (visited on 04/29/2025).

[21] F. Meyer, A. Fritz, Z.-L. Deng, *et al.*, "Critical Assessment of Metagenome Interpretation: The second round of challenges," en, *Nature Methods*, vol. 19, no. 4, pp. 429–440, Apr. 2022, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-022-01431-4. [Online]. Available: https://www.nature.com/articles/s41592-022-01431-4 (visited on 04/27/2025).

[22] Guldbergsen, "The Predict Study," Quality Assurance, Embla A/S, Jun. 2025.

[23] E. Nguyen, M. Poli, M. Faizi, *et al.*, *HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution*, arXiv:2306.15794, Nov. 2023. DOI: 10.48550/arXiv.2306.15794. [Online]. Available: http://arxiv.org/abs/2306.15794 (visited on 12/01/2024).

[24] F. I. Marin, F. Teufel, M. Horlacher, *et al.*, "BEND: Benchmarking DNA Language Models on Biologically Meaningful Tasks," en, Oct. 2023. [Online]. Available: https://openreview.net/forum?id=uKB4cFNQFg (visited on 11/25/2024).

[25] G. Benegas, C. Albors, A. J. Aw, C. Ye, and Y. S. Song, "GPN-MSA: An alignment-based DNA language model for genome-wide variant effect prediction," *bioRxiv*, p. 2023.10.10.561776, Apr. 2024, ISSN: 2692-8205. DOI: 10.1101/2023.10.10.561776. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10592768/ (visited on 05/15/2025).

[26] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with AlphaFold," en, *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. [Online]. Available: https://www.nature.com/articles/s41586-021-03819-2 (visited on 05/15/2025).

[27] Z. Lin, H. Akin, R. Rao, *et al.*, *Language models of protein sequences at the scale of evolution enable accurate structure prediction*, en, Pages: 2022.07.20.500902 Section: New Results, Jul. 2022. DOI: 10.1101/2022.07.20.500902. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1 (visited on 05/15/2025).

[28] M. E. Consens, B. Li, A. R. Poetsch, and S. Gilbert, "Genomic language models could transform medicine but not yet," en, *npj Digital Medicine*, vol. 8, no. 1, pp. 1–4, Apr. 2025, Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-025-01603-4. [Online]. Available: https://www.nature.com/articles/s41746-025-01603-4 (visited on 05/15/2025).

[29] *Homo sapiens genome assembly GRCh38*, en. [Online]. Available: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/ (visited on 11/25/2024).

[30] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, *et al.*, "Nucleotide Transformer: Building and evaluating robust foundation models for human genomics," en, *Nature Methods*, vol. 22, no. 2, pp. 287–297, Feb. 2025, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-024-02523-z. [Online]. Available: https://www.nature.com/articles/s41592-024-02523-z (visited on 05/15/2025).

[31] E. Nguyen, M. Poli, M. G. Durrant, *et al.*, *Sequence modeling and design from molecular to genome scale*

with Evo, en, Pages: 2024.02.27.582234 Section: New Results, Feb. 2024. DOI: 10.1101/2024.02.27.582234. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.02.27.582234v1 (visited on 05/15/2025).

[32] D. Zhang, W. Zhang, Y. Zhao, *et al.*, *DNAGPT: A Generalized Pre-trained Tool for Versatile DNA Sequence Analysis Tasks*, arXiv:2307.05628 [q-bio], Aug. 2023. DOI: 10.48550/arXiv.2307.05628. [Online]. Available: http://arxiv.org/abs/2307.05628 (visited on 05/27/2025).

[33] O. Press, N. A. Smith, and M. Lewis, *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation*, arXiv:2108.12409 [cs], Apr. 2022. DOI: 10.48550/arXiv.2108.12409. [Online]. Available: http://arxiv.org/abs/2108.12409 (visited on 12/11/2024).

[34] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*, arXiv:2205.14135, Jun. 2022. DOI: 10.48550/arXiv.2205.14135. [Online]. Available: http://arxiv.org/abs/2205.14135 (visited on 12/05/2024).

[35] V. Verma, A. Lamb, C. Beckham, *et al.*, *Manifold Mixup: Better Representations by Interpolating Hidden States*, arXiv:1806.05236, May 2019. DOI: 10.48550/arXiv.1806.05236. [Online]. Available: http://arxiv.org/abs/1806.05236 (visited on 12/05/2024).

[36] S. Pan, X.-M. Zhao, and L. P. Coelho, "SemiBin2: Self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing," *Bioinformatics*, vol. 39, no. Supplement_1, pp. i21–i29, Jun. 2023, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad209. [Online]. Available: https://doi.org/10.1093/bioinformatics/btad209 (visited on 04/29/2025).

[37] D. D. Kang, F. Li, E. Kirton, *et al.*, "MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, e7359, Jul. 2019, ISSN: 2167-8359. DOI: 10.7717/peerj.7359. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6662567/ (visited on 05/19/2025).

[38] S. Kutuzova, M. Nielsen, P. Piera, J. N. Nissen, and S. Rasmussen, *Taxometer: Improving taxonomic classification of metagenomics contigs*, en, Pages: 2023.11.23.568413 Section: New Results, Nov. 2023. DOI: 10.1101/2023.11.23.568413. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.11.23.568413v1 (visited on 05/27/2025).

[39] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: Guaranteeing well-connected communities," en, *Scientific Reports*, vol. 9, no. 1, p. 5233, Mar. 2019, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z. [Online]. Available: https://www.nature.com/articles/s41598-019-41695-z (visited on 04/29/2025).

[40] K. Hou, Z.-X. Wu, X.-Y. Chen, *et al.*, "Microbiota in health and diseases," en, *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, pp. 1–28, Apr. 2022, Publisher: Nature Publishing Group, ISSN: 2059-3635. DOI: 10.1038/s41392-022-00974-4. [Online]. Available: https://www.nature.com/articles/s41392-022-00974-4 (visited on 05/21/2025).

[41] M. J. Bull and N. T. Plummer, "Part 1: The Human Gut Microbiome in Health and Disease," *Integrative Medicine: A Clinician's Journal*, vol. 13, no. 6, pp. 17–22, Dec. 2014, ISSN: 1546-993X. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4566439/ (visited on 05/21/2025).

[42] C. A. Lozupone, J. I. Stombaugh, J. I. Gordon, J. K. Jansson, and R. Knight, "Diversity, stability and resilience of the human gut microbiota," *Nature*, vol. 489, no. 7415, pp. 220–230, Sep. 2012, ISSN: 0028-0836. DOI: 10.1038/nature11550. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3577372/ (visited on 05/30/2025).

[43] J. Qin, Y. Li, Z. Cai, *et al.*, "A metagenome-wide association study of gut microbiota in type 2 diabetes," en, *Nature*, vol. 490, no. 7418, pp. 55–60, Oct. 2012, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature11450. [Online]. Available: https://www.nature.com/articles/nature11450 (visited on 05/31/2025).

[44] I. Cho and M. J. Blaser, "The human microbiome: At the interface of health and disease," en, *Nature Reviews Genetics*, vol. 13, no. 4, pp. 260–270, Apr. 2012, Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: 10.1038/nrg3182. [Online]. Available: https://www.nature.com/articles/nrg3182 (visited on 05/15/2025).

[45] G. Roy, E. Prifti, E. Belda, and J.-D. Zucker, "Deep learning methods in metagenomics: A review," *Microbial Genomics*, vol. 10, no. 4, p. 001 231, 2024, Publisher: Microbiology Society, ISSN: 2057-5858. DOI: 10.1099/mgen.0.001231. [Online]. Available: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001231 (visited on 11/29/2024).

[46] M. Oh and L. Zhang, "DeepMicro: Deep representation learning for disease prediction based on microbiome data," en, *Scientific Reports*, vol. 10, no. 1, p. 6026, Apr. 2020, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-020-63159-5. [Online]. Available: https://www.nature.com/articles/s41598-020-63159-5 (visited on 05/10/2025).

[47] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights," en, *PLOS Computational Biology*, vol. 12, no. 7, e1004977, Jul. 2016, Publisher: Public Library of Science, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004977. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977 (visited on 05/10/2025).

[48] F. H. Karlsson, V. Tremaroli, I. Nookaew, *et al.*, "Gut metagenome in European women with normal, impaired and diabetic glucose control," eng, *Nature*,

vol. 498, no. 7452, pp. 99–103, Jun. 2013, ISSN: 1476-4687. DOI: 10.1038/nature12198.

[49] J. S. Bajaj, D. M. Heuman, P. B. Hylemon, *et al.*, "Altered profile of human gut microbiome is associated with cirrhosis and its complications," eng, *Journal of Hepatology*, vol. 60, no. 5, pp. 940–947, May 2014, ISSN: 1600-0641. DOI: 10.1016/j.jhep.2013.12.019.

[50] G. Zeller, J. Tap, A. Y. Voigt, *et al.*, "Potential of fecal microbiota for early-stage detection of colorectal cancer," *Molecular Systems Biology*, vol. 10, no. 11, p. 766, Nov. 2014, ISSN: 1744-4292. DOI: 10.15252/msb.20145645. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299606/ (visited on 05/31/2025).

[51] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights," en, *PLOS Computational Biology*, vol. 12, no. 7, e1004977, Jul. 2016, Publisher: Public Library of Science, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004977. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004977 (visited on 05/10/2025).

[52] D. T. Truong, E. A. Franzosa, T. L. Tickle, *et al.*, "MetaPhlAn2 for enhanced metagenomic taxonomic profiling," en, *Nature Methods*, vol. 12, no. 10, pp. 902–903, Oct. 2015, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/nmeth.3589. [Online]. Available: https://www.nature.com/articles/nmeth.3589 (visited on 05/21/2025).

[53] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, May 2018, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2017.10.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320317304065 (visited on 05/28/2025).

[54] M. A. Rahman and H. Rangwala, "IDMIL: An alignment-free Interpretable Deep Multiple Instance Learning (MIL) for predicting disease from whole-metagenomic data," *Bioinformatics*, vol. 36, no. Supplement_1, pp. i39–i47, Jul. 2020, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa477. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa477 (visited on 05/10/2025).

[55] P. Khosla, P. Teterwak, C. Wang, *et al.*, *Supervised Contrastive Learning*, arXiv:2004.11362 [cs], Mar. 2021. DOI: 10.48550/arXiv.2004.11362. [Online]. Available: http://arxiv.org/abs/2004.11362 (visited on 05/06/2025).

[56] C. M. K. Sieber, A. J. Probst, A. Sharrar, *et al.*, "Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy," en, *Nature Microbiology*, vol. 3, no. 7, pp. 836–843, Jul. 2018, Publisher: Nature Publishing Group, ISSN: 2058-5276. DOI: 10.1038/s41564-018-0171-1. [Online]. Available: https://www.nature.com/articles/s41564-018-0171-1 (visited on 05/24/2025).

[57] H. Han, Z. Wang, and S. Zhu, "Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes," en, *Nature Communications*, vol. 16, no. 1, p. 2865, Mar. 2025, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-025-57957-6. [Online]. Available: https://www.nature.com/articles/s41467-025-57957-6 (visited on 04/27/2025).

[58] J. Mattock and M. Watson, "A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination," en, *Nature Methods*, vol. 20, no. 8, pp. 1170–1173, Aug. 2023, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-023-01934-8. [Online]. Available: https://www.nature.com/articles/s41592-023-01934-8 (visited on 04/27/2025).

[59] L. J. McIver, G. Abu-Ali, E. A. Franzosa, *et al.*, "bioBakery: A meta'omic analysis environment," eng, *Bioinformatics (Oxford, England)*, vol. 34, no. 7, pp. 1235–1237, Apr. 2018, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx754.

[60] S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger, and S. Wingett, *FastQC*, Jan. 2012.

[61] K. Sahlin, "Strobealign: Flexible seed size enables ultra-fast and accurate read alignment," *Genome Biology*, vol. 23, no. 1, p. 260, Dec. 2022, ISSN: 1474-760X. DOI: 10.1186/s13059-022-02831-7. [Online]. Available: https://doi.org/10.1186/s13059-022-02831-7 (visited on 04/27/2025).

[62] H. Li, B. Handsaker, A. Wysoker, *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/ (visited on 04/27/2025).

[63] S. T. N. Aroney, R. J. P. Newell, J. N. Nissen, A. P. Camargo, G. W. Tyson, and B. J. Woodcroft, "CoverM: Read alignment statistics for metagenomics," *Bioinformatics*, vol. 41, no. 4, btaf147, Apr. 2025, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btaf147. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaf147 (visited on 05/07/2025).

[64] V. Cheplygina, D. M. J. Tax, and M. Loog, "Dissimilarity-based Ensembles for Multiple Instance Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1379–1391, Jun. 2016, arXiv:1402.1349 [stat], ISSN: 2162-237X, 2162-2388. DOI: 10.1109/TNNLS.2015.2424254. [Online]. Available: http://arxiv.org/abs/1402.1349 (visited on 05/25/2025).

[65] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "GTDB-Tk v2: Memory friendly classification with the genome taxonomy database," *Bioinformatics*, vol. 38, no. 23, pp. 5315–5316, Dec. 2022, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btac672. [Online].

Available: https://doi.org/10.1093/bioinformatics/btac672 (visited on 05/25/2025).

[66] J. Friedman, T. Hastie, and R. Tibshirani, *A note on the group lasso and a sparse group lasso*, arXiv:1001.0736 [math], Jan. 2010. DOI: 10.48550/arXiv.1001.0736. [Online]. Available: http://arxiv.org/abs/1001.0736 (visited on 05/10/2025).

[67] J. Li, K. Liang, and X. Song, "Logistic regression with adaptive sparse group lasso penalty and its application in acute leukemia diagnosis," *Computers in Biology and Medicine*, vol. 141, p. 105 154, Feb. 2022, ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2021.105154. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521009483 (visited on 05/10/2025).

[68] S. Ma, X. Song, and J. Huang, "Supervised group Lasso with applications to microarray data analysis," *BMC Bioinformatics*, vol. 8, no. 1, p. 60, Feb. 2007, ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-60. [Online]. Available: https://doi.org/10.1186/1471-2105-8-60 (visited on 05/10/2025).

[69] B. Liquet, P. L. de Micheaux, B. P. Hejblum, and R. Thiébaut, "Group and sparse group partial least square approaches applied in genomics context," *Bioinformatics*, vol. 32, no. 1, pp. 35–42, Jan. 2016, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv535. [Online]. Available: https://doi.org/10.1093/bioinformatics/btv535 (visited on 05/10/2025).

[70] N. A. O'Leary, M. W. Wright, J. R. Brister, *et al.*, "Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation," eng, *Nucleic Acids Research*, vol. 44, no. D1, pp. D733–745, Jan. 2016, ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

[71] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, arXiv:1711.05101 [cs], Jan. 2019. DOI: 10.48550/arXiv.1711.05101. [Online]. Available: http://arxiv.org/abs/1711.05101 (visited on 05/16/2025).

[72] I. Loshchilov and F. Hutter, *SGDR: Stochastic Gradient Descent with Warm Restarts*, arXiv:1608.03983 [cs], May 2017. DOI: 10.48550/arXiv.1608.03983. [Online]. Available: http://arxiv.org/abs/1608.03983 (visited on 05/28/2025).

[73] P. Ng, *Dna2vec: Consistent vector representations of variable-length k-mers*, arXiv:1701.06279, Jan. 2017. DOI: 10.48550/arXiv.1701.06279. [Online]. Available: http://arxiv.org/abs/1701.06279 (visited on 11/30/2024).

[74] K. Vishniakov, K. Viswanathan, A. Medvedev, *et al.*, *Genomic Foundationless Models: Pretraining Does Not Promise Performance*, en, Pages: 2024.12.18.628606 Section: New Results, Dec. 2024. DOI: 10.1101/2024.12.18.628606. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.12.18.628606v1 (visited on 05/17/2025).

[75] P. Zhang, Z. Jiang, Y. Wang, and Y. Li, *CLMB: Deep contrastive learning for robust metagenomic binning*, arXiv:2111.09656 [cs], Nov. 2021. DOI: 10.48550/arXiv.2111.09656. [Online]. Available: http://arxiv.org/abs/2111.09656 (visited on 05/04/2025).

[76] C.-C. Liu, S.-S. Dong, J.-B. Chen, *et al.*, "MetaDecoder: A novel method for clustering metagenomic contigs," *Microbiome*, vol. 10, no. 1, p. 46, Mar. 2022, ISSN: 2049-2618. DOI: 10.1186/s40168-022-01237-8. [Online]. Available: https://doi.org/10.1186/s40168-022-01237-8 (visited on 05/04/2025).

[77] Z. Wang, P. Huang, R. You, F. Sun, and S. Zhu, "MetaBinner: A high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities," *Genome Biology*, vol. 24, no. 1, p. 1, Jan. 2023, ISSN: 1474-760X. DOI: 10.1186/s13059-022-02832-6. [Online]. Available: https://doi.org/10.1186/s13059-022-02832-6 (visited on 05/04/2025).

[78] S. Wang, M. Ventolero, H. Hu, and X. Li, "A revisit to universal single-copy genes in bacterial genomes," en, *Scientific Reports*, vol. 12, no. 1, p. 14 550, Aug. 2022, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-022-18762-z. [Online]. Available: https://www.nature.com/articles/s41598-022-18762-z (visited on 04/29/2025).

[79] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Research*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, ISSN: 1088-9051. DOI: 10.1101/gr.186072.114. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4484387/ (visited on 05/21/2025).

[80] O. D. Richard, E. H. Peter, and David G. Stork, "Pattern Classification," en, *Journal of Classification*, vol. 18, no. 2, pp. 273–275, 2001, ISSN: 0176-4268, 1432-1343. [Online]. Available: https://www.elibrary.ru/item.asp?id=816540 (visited on 05/04/2025).

[81] N. A. O'Leary, E. Cox, J. B. Holmes, *et al.*, "Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets," en, *Scientific Data*, vol. 11, no. 1, p. 732, Jul. 2024, Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/s41597-024-03571-y. [Online]. Available: https://www.nature.com/articles/s41597-024-03571-y (visited on 05/26/2025).

[82] G. Singh, M. Krauthamer, and M. Bjalme-Evans, "Wegovy (Semaglutide): A New Weight Loss Drug for Chronic Weight Management," EN, *Journal of Investigative Medicine*, vol. 70, no. 1, pp. 5–13, Jan. 2022, Publisher: SAGE Publications, ISSN: 1081-5589. DOI: 10.1136/jim-2021-001952. [Online]. Available: https://doi.org/10.1136/jim-2021-001952 (visited on 05/15/2025).

[83] A. Chaudhry, B. Gabriel, J. Noor, S. Jawad, and S. R. Challa, "Tendency of Semaglutide to Induce Gastroparesis: A Case Report," *Cureus*, vol. 16, no. 1, e52564, ISSN: 2168-8184. DOI: 10.7759/cureus.52564. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10874596/ (visited on 05/26/2025).

[84] M. Ojala and G. C. Garriga, "Permutation Tests for Studying Classifier Performance," in *2009 Ninth IEEE International Conference on Data Mining*, ISSN: 2374-8486, Dec. 2009, pp. 908–913. DOI: 10.1109/ICDM.2009.108. [Online]. Available: https://ieeexplore.ieee.org/document/5360332 (visited on 05/20/2025).

[85] P. Schneider and F. Xhafa, "Chapter 8 - Machine learning: ML for eHealth systems," in *Anomaly Detection and Complex Event Processing over IoT Data Streams*, P. Schneider and F. Xhafa, Eds., Academic Press, Jan. 2022, pp. 149–191, ISBN: 978-0-12-823818-9. DOI: 10.1016/B978-0-12-823818-9.00019-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128238189000195 (visited on 05/29/2025).

[86] A. Almeida, A. L. Mitchell, M. Boland, *et al.*, "A new genomic blueprint of the human gut microbiota," en, *Nature*, vol. 568, no. 7753, pp. 499–504, Apr. 2019, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-019-0965-1. [Online]. Available: https://www.nature.com/articles/s41586-019-0965-1 (visited on 05/30/2025).

[87] S. A. Sankar, J.-C. Lagier, P. Pontarotti, D. Raoult, and P.-E. Fournier, "The human gut microbiome, a taxonomic conundrum," *Systematic and Applied Microbiology*, Taxonomy in the age of genomics, vol. 38, no. 4, pp. 276–286, Jun. 2015, ISSN: 0723-2020. DOI: 10.1016/j.syapm.2015.03.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0723202015000454 (visited on 05/30/2025).

[88] M. E. Consens, C. Dufault, M. Wainberg, *et al.*, *To Transformers and Beyond: Large Language Models for the Genome*, arXiv:2311.07621 [q-bio], Nov. 2023. DOI: 10.48550/arXiv.2311.07621. [Online]. Available: http://arxiv.org/abs/2311.07621 (visited on 05/15/2025).

[89] J. Mendoza-Revilla, E. Trop, L. Gonzalez, *et al.*, *A Foundational Large Language Model for Edible Plant Genomes*, en, Pages: 2023.10.24.563624 Section: New Results, Oct. 2023. DOI: 10.1101/2023.10.24.563624. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.10.24.563624v1 (visited on 05/31/2025).

[90] B. P. d. Almeida, H. Dalla-Torre, G. Richard, *et al.*, *SegmentNT: Annotating the genome at single-nucleotide resolution with DNA foundation models*, en, Pages: 2024.03.14.584712 Section: New Results, Mar. 2024. DOI: 10.1101/2024.03.14.584712. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.03.14.584712v1 (visited on 05/29/2025).

[91] Y. Schiff, C.-H. Kao, A. Gokaslan, T. Dao, A. Gu, and V. Kuleshov, "Caduceus: Bi-directional equivariant long-range DNA sequence modeling," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24, vol. 235, Vienna, Austria: JMLR.org, Jul. 2024, pp. 43 632–43 648. (visited on 05/29/2025).

[92] J. Wang and H. Jia, "Metagenome-wide association studies: Fine-mining the microbiome," en, *Nature Reviews Microbiology*, vol. 14, no. 8, pp. 508–522, Aug. 2016, Publisher: Nature Publishing Group, ISSN: 1740-1534. DOI: 10.1038/nrmicro.2016.83. [Online]. Available: https://www.nature.com/articles/nrmicro.2016.83 (visited on 05/30/2025).

[93] A. Blanco-Míguez, E. J. C. Gálvez, E. Pasolli, *et al.*, "Extension of the *Segatella copri* complex to 13 species with distinct large extrachromosomal elements and associations with host conditions," *Cell Host & Microbe*, vol. 31, no. 11, 1804–1819.e9, Nov. 2023, ISSN: 1931-3128. DOI: 10.1016/j.chom.2023.09.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1931312823003797 (visited on 06/02/2025).

[94] H. K. Pedersen, V. Gudmundsdottir, H. B. Nielsen, *et al.*, "Human gut microbes impact host serum metabolome and insulin sensitivity," en, *Nature*, vol. 535, no. 7612, pp. 376–381, Jul. 2016, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature18646. [Online]. Available: https://www.nature.com/articles/nature18646 (visited on 06/01/2025).

[95] Humphrey, Suzanne, *Investigating the role of Enterocloster clostridioformis bacteriophages in gut microbiome dysbiosis associated with metabolic disease states.* en. [Online]. Available: https://pureportal.strath.ac.uk/en/projects/investigating-the-role-of-enterocloster-clostridioformis-bacterio (visited on 06/01/2025).

[96] L. Fu, A. Baranova, H. Cao, and F. Zhang, "Gut microbiome links obesity to type 2 diabetes: Insights from Mendelian randomization," *BMC Microbiology*, vol. 25, no. 1, p. 253, Apr. 2025, ISSN: 1471-2180. DOI: 10.1186/s12866-025-03968-8. [Online]. Available: https://doi.org/10.1186/s12866-025-03968-8 (visited on 06/01/2025).

[97] G. Gradisteanu Pircalabioru, M.-C. Chifiriuc, A. Picu, L. M. Petcu, M. Trandafir, and O. Savu, "Snapshot into the Type-2-Diabetes-Associated Microbiome of a Romanian Cohort," *International Journal of Molecular Sciences*, vol. 23, no. 23, p. 15 023, Nov. 2022, ISSN: 1422-0067. DOI: 10.3390/ijms232315023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9741184/ (visited on 06/01/2025).

[98] O. Ignatyeva, D. Tolyneva, A. Kovalyov, *et al.*, "Christensenella minuta, a new candidate next-generation probiotic: Current evidence and future trajectories," *Frontiers in Microbiology*, vol. 14, p. 1 241 259, Jan. 2024. DOI: 10.3389/fmicb.2023.1241259.

[99] J. L. Waters and R. E. Ley, "The human gut bacteria Christensenellaceae are widespread, heritable, and associated with health," *BMC Biology*, vol. 17, no. 1, p. 83, Oct. 2019, ISSN: 1741-7007. DOI: 10.1186/s12915-019-0699-4. [Online]. Available: https://doi.org/10.1186/s12915-019-0699-4 (visited on 06/01/2025).

[100] S. D. Ehrlich, "MetaHIT: The European Union Project on Metagenomics of the Human Intestinal Tract," en, in *Metagenomics of the Human Body*, K. E. Nelson, Ed., New York, NY: Springer, 2011, pp. 307–316, ISBN: 978-1-4419-7089-3. DOI: 10.1007/978-1-4419-7089-3_15. [Online]. Available: https://doi.org/10.1007/978-1-4419-7089-3_15 (visited on 11/29/2024).

[101] *ERP000108 : Study : SRA Archive : NCBI*. [Online]. Available: https://www.ncbi.nlm.nih.gov/Traces/index.html?view=study&acc=ERP000108 (visited on 11/29/2024).

[102] E. W. Sayers, E. E. Bolton, J. R. Brister, *et al.*, "Database resources of the national center for biotechnology information," eng, *Nucleic Acids Research*, vol. 50, no. D1, pp. D20–D26, Jan. 2022, ISSN: 1362-4962. DOI: 10.1093/nar/gkab1112.

[103] J. Kim and M. Steinegger, "Metabuli: Sensitive and specific metagenomic classification via joint analysis of amino acid and DNA," en, *Nature Methods*, vol. 21, no. 6, pp. 971–973, Jun. 2024, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-024-02273-y. [Online]. Available: https://www.nature.com/articles/s41592-024-02273-y (visited on 04/29/2025).

APPENDIX A
K-MEDOID ALGORITHM FOR METAGENOMICS BINNING

Algorithm 1 describes the K-medoid clustering algorithm as implemented in [3], [12]. The algorithm outline is adopted from [12]. The threshold γ is obtained from algorithm 2.

Algorithm 1 Modified K-Medoid Clustering

Require: Threshold γ, minimum bin size m, embeddings $E \in \mathbb{R}^{N \times d}$, number of steps Z, number of iterations T

Initialize: Predictions $p \in \mathbb{R}^N$, $p_i = -1$ for $i = 1, \ldots, N$, similarity matrix $S = EE^\top$ with $S_{ij} = 0$ if $S_{ij} < \gamma$, density vector $d \in \mathbb{R}^N$ with $d_i = \sum_{j=1}^{N} S_{ij}$
for step $z = 1$ to Z **do**
  Select seed index $s = \arg\max_{s'} d_{s'}$ and corresponding seed E_s
  for iteration $t = 1$ to T **do**
    Find neighborhood indices $\mathcal{I}$ of E_s where $s(E_i, E_s) > \gamma$ and $p_i = -1$ for each $i \in \mathcal{I}$
    Update seed: $E_s \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} E_i$
  end for
  Set $p_i \leftarrow z$, $d_i \leftarrow 0$ for each $i \in \mathcal{I}$
  Update density: $d_x \leftarrow d_x - \sum_{i \in \mathcal{I}} S_{xi}$ for each $x \in \{1, 2, \ldots, N\}$
end for
for step $z = 1$ to Z **do**
  Find indices $\mathcal{I}$ where $p_i = z$ for each $i \in \mathcal{I}$
  if $|\mathcal{I}| < m$ **then**
    Set $p_i \leftarrow -1$ for each $i \in \mathcal{I}$
  end if
end for
Return: Predictions p

APPENDIX B
SIMILARITY THRESHOLD ALGORITHM

Algorithm 2 outlines the procedure for estimating the similarity threshold γ based on top-K neighborhood centroids.

Algorithm 2 Estimating Similarity Threshold via Percentile of Neighbor-Centroid Similarities

Require: Embeddings $E \in \mathbb{R}^{N \times d}$, number of neighbors K, percentile $P \in [1, 100]$

Initialize: Similarity matrix $S = EE^\top$, empty list of similarities $s = [\,]$
for $i = 1$ to N **do**
  Find top-K neighborhood indices $\mathcal{H}_i$ from S_i and corresponding embeddings $E_{\mathcal{H}_i}$
  Compute centroid $\mu_i = \frac{1}{K} \sum_{j=1}^{K} E_{\mathcal{H}_{i,j}}$
  Compute pairwise similarities to centroid: $s_{i,j} = \langle E_{\mathcal{H}_{i,j}}, \mu_i \rangle$ for $j = 1, \ldots, K$ and append $s_{i,j}$ to s
end for
Sort and normalize similarities: $s \leftarrow \mathrm{sort}(s)$; $s \leftarrow s / \sum s$
Compute cumulative distribution: $c \leftarrow \mathrm{cumsum}(s)$ and find index l such that $c_l \geq \frac{P}{100}$
Determine threshold: $\gamma \leftarrow s[l]$
Return: threshold γ

APPENDIX C
SUMMARY OF THE 400K PRE-TRAINING DATASET

Group	# Pairs	# Species	# Genus	# Family	# Order	# Class	# Phylum	# Kingdom	# Superkingdom
Bacteria	300,000	20,234	3,836	747	282	115	48	4	1
Fungi	100,000	4,766	1,305	457	168	60	8	1	1
Virus	165,599	482	157	39	16	8	7	4	1

TABLE III: Summary statistics for the 400k pre-training dataset. All DNA sequences are 10,000 bp long. The dataset was assembled in February 2025; as of May 2025, NCBI updated the rank name "Superkingdom" to "Domain" for bacteria and "Acellular root" for virus, see NCBI updates.

APPENDIX D
SUMMARY OF THE VALIDATION DATASET

Group	# Pairs	# Species	# Genus	# Family	# Order	# Class	# Phylum	# Kingdom	# Superkingdom
Bacteria	29,997	18,266	3,621	712	271	111	47	4	1
Fungi	10,002	4,601	1,272	449	165	59	8	1	1
Virus	11,724	48	38	16	11	7	6	4	1

TABLE IV: Summary statistics for the validation dataset. All DNA sequences are 10,000 bp long. The dataset was assembled in February 2025; as of May 2025, NCBI updated the rank name “Superkingdom” to ”Domain” for bacteria and ”Acellular root” for virus, see NCBI updates.

APPENDIX E
TOKEN LENGTH STATISTICS FOR TRAINING SEQUENCES

Group	Mean	25th pct.	Median	75th pct.
Bacteria	2,152	2,065	2,180	2,239
Fungi	2,082	2,058	2,089	2,113
Virus	1,961	1,916	1,935	1,976

TABLE V: Token length statistics for the training sequences. All sequences are tokenized using Byte-Pair Encoding (BPE).

APPENDIX F
VALIDATION LOSS ACROSS TRAINING STEPS FOR DNABERT-H VARIANTS

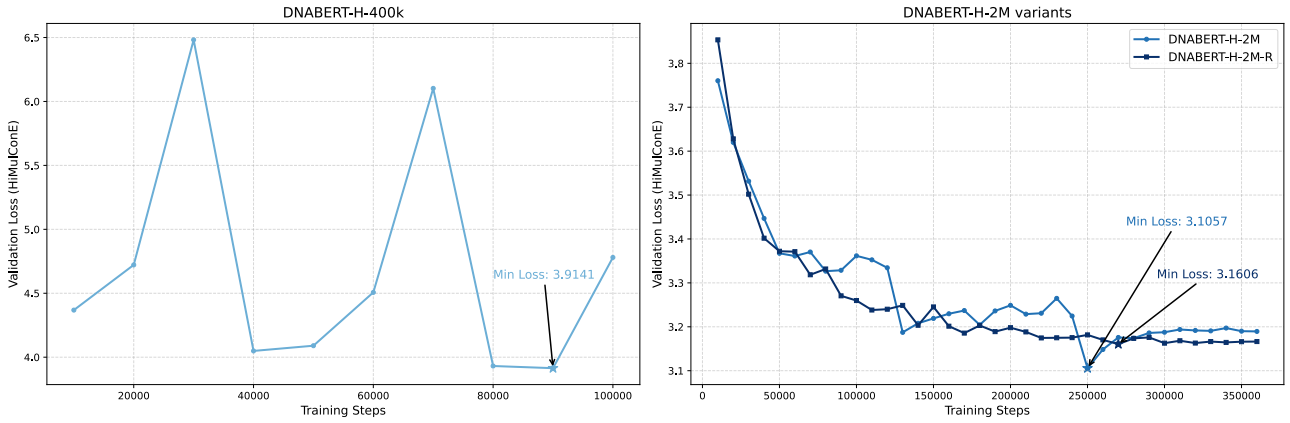

Fig. 8: Validation loss (HiMulConE) over training steps, comparing DNABERT-H variants (DNABERT-H-400k, DNABERT-H-2M and DNABERT-H-2M-R). The validation loss is calculated at every 10,000 steps with a batch size of 90. Arrows indicate the minimum loss achieved by each variant.

APPENDIX G
COMPOSITION OF BATCHES USED IN DNABERT-H-2M-R

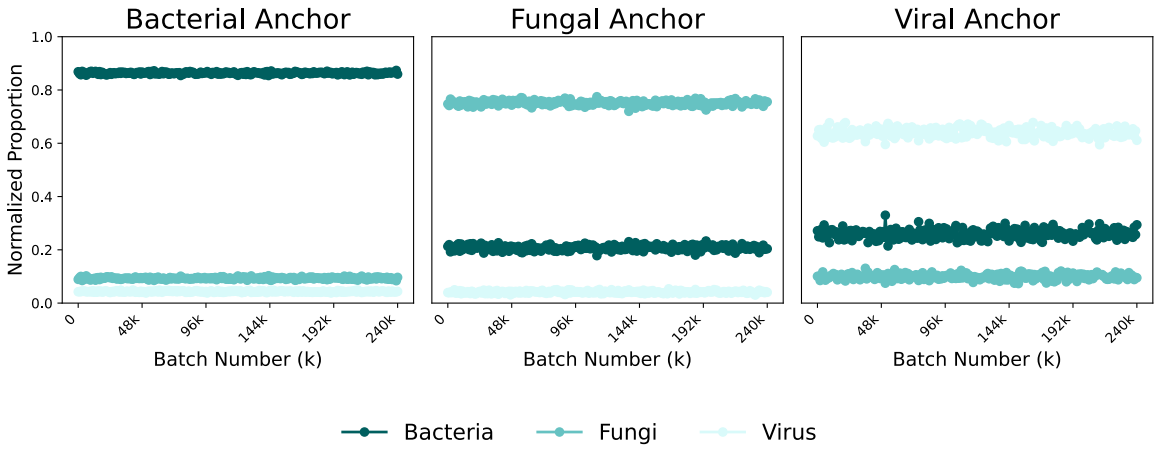

Fig. 9: Stable taxonomic composition of batches used for training DNABERT-H-2M-R. Each subplot shows the consistent proportions of Bacteria, Fungi, and Virus samples across batches for different anchor groups.

APPENDIX H
METAHIT DATASET DETAILS

The MetaHIT "error-free" contig dataset is obtained from MetaBAT [3]. The dataset was derived from reads obtained by the MetaHIT consortium [100], coming from 264 human gut microbiome samples [101]. MetaBAT mapped these reads to known reference genomes using the NCBi-database at strain level resolution [102], and selected 290 genomes with the highest coverage across samples. The genomic sequences varied in length from 1.9M to 10.5M base pairs. To create the dataset, the genomic sequences were shredded into contigs with 31 overlapping base pairs at both ends. There is between 41 and 1,617 contigs per reference strain genome.

APPENDIX I
OVERVIEW OF DATASETS

Task	Dataset	# Contigs	Total Length	Mean Length	Median Length	Max Length
Metagenomics binning	Airways	187.7K	1.7G	8.8K	3.3K	6.2M
	Gastro	81.6K	1.7G	21.1K	4.0K	6.5M
	Oral	201.6K	1.9G	9.6K	3.2K	5.5M
	Urogenital	57.8K	1.0G	17.4K	3.3K	7.3M
	Skin	173.9K	1.8G	10.4K	3.6K	5.6M
	Marine	439.0K	3.3G	7.4K	3.1K	5.2M
	Plant	300.2K	3.6G	12.1K	3.6K	8.3M
	Metahit	179.5K	1.1G	6.1K	4.0K	64.6K
Phenotype classification	T2D-EW	1.1M	6.9G	6.3K	3.4K	716.0K
	WEGOVY	1.4M	10.2G	7.1K	3.8K	425.3K

TABLE VI: Overview of the contigs in the eight datasets used for metagenomics binning, and the two datasets used for phenotype classification. Length refers to the number of base-pairs. The minimum length is 2K for all datasets. $G = 10^9$.

APPENDIX J
CONTIG LENGTH DISTRIBUTIONS OF DATASETS

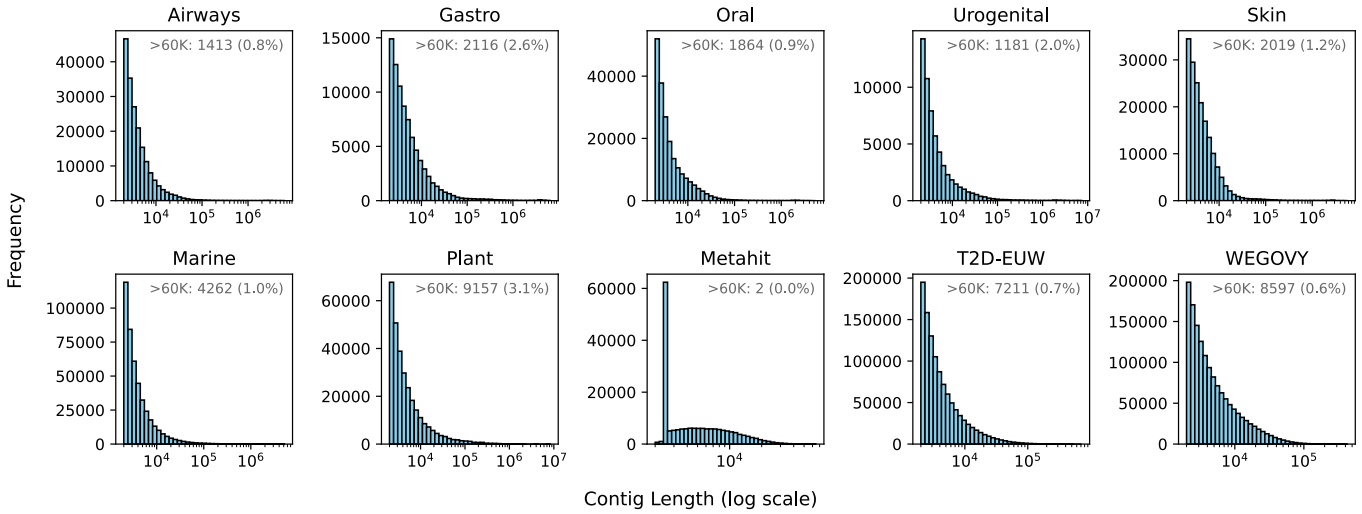

Fig. 10: Histograms of contig lengths in the seven CAMI2 datasets, the Metahit "error-free" dataset, and the two datasets used for phenotype classification. Contigs with lengths $\geq$ 60K base-pairs were excluded for the genomic language models, removing between 0% and 3.1% of contigs within each dataset.

APPENDIX K
SUPPLEMENTARY METAGENOMICS BINNING RESULTS

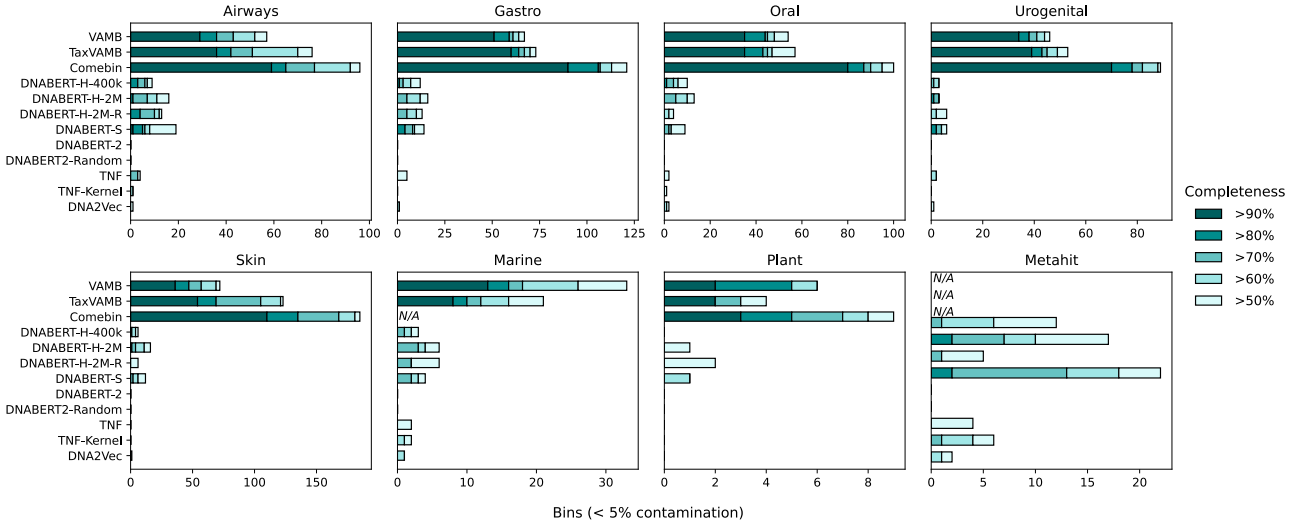

Fig. 11: The number of recovered bins with contamination $\leq 5\%$ and varying completeness thresholds including TNF-Kernel and DNABERT-2 Random. Comebin could not be run on the Marine dataset, and neither VAMB, TaxVAMB, nor Comebin could be evaluated on the Metahit dataset.

APPENDIX L
RESULTS FROM "BENCHMARKING DNA FOUNDATION MODELS ON BINNING HUMAN GUT MICROBIAL STRAINS" [10]

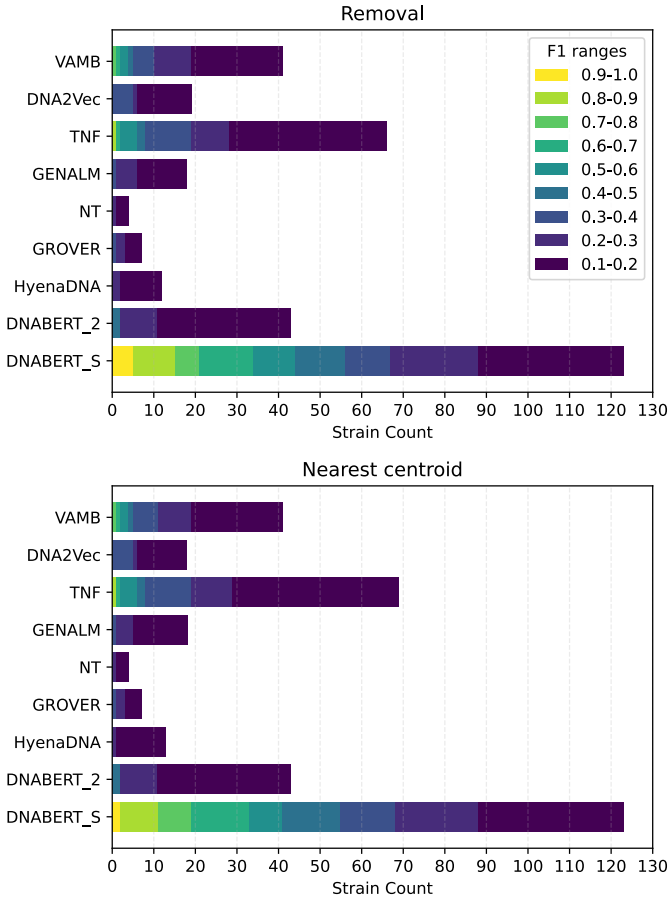

Fig. 12: Metagenomics binning results on the MetaHIT "error-free" contig dataset from MetaBAT [3]. DNABERT-S outperforms the baselines (DNA2Vec and TNF), but the remaining genomic language models perform worse than baselines. Note that F1-scores are computed based on the ground truth contig labels, in contrast to using CheckM2. The results are shown for two scenarios: (top) excluding unpredicted contigs, and (bottom) assigning them to the nearest cluster centroid.

TABLE VII: Overview of the previously benchmarked genomic language models in metagenomics binning [10]

Model	Architecture	Pre-Training strategy	Tokenizer	Window size	Sequence length	Pretraining data	Source	Benchmarked in research project
DNABERT	BERT	MLM	Overlapping k-mer	512	512	Human genome	[5]	✗
DNABERT-2	BERT	MLM	BPE	512	$10,000^{a}$	Multispecies	[6]	✓
DNABERT-S	BERT	C^2LR, MI-Mix	BPE	512	$10,000^{a}$	Multispecies	[12]	✓
NT	BERT	MLM	Non-overlapping 6-mer	1000	6,000	Multispecies	[7]	✗
NT V2	BERT	MLM	Non-overlapping 6-mer	2,048	12,000	Multispecies	[7]	✓
GROVER	BERT	MLM	BPE	512	$8,192^{b}$	Human genome	[8]	✓
GENA-LM	BERT	MLM	BPE	512	4,500	Multispecies	[9]	✓
Hyena-DNA	Hyena	NTP	Single nucleotide	1,000,000	1,000,000	Human genome	[23]	✓

[a] DNABERT-2 and DNABERT-S adopted Attention with Linear Biases ALiBi [34] which was showcased to handle sequences of length up to 10,000 at inference. This context length was also used in the evaluation of DNABERT-2 and DNABERT-S. In addition, the sequence length limit of DNABERT-2 and DNABERT-S is considered to be 3,000 in [7], and 1,000-4,000 in [9].

[b] No explicit length was reported in GROVER [8]. We derived the sequence length by assuming that the 512 BPE tokens had an average length of 16, following a previous benchmark study BEND [24].

APPENDIX M
COMMANDS FOR BENCHMARKED METAGENOMIC BINNERS

VAMB [17] (v.5.0.3) was run with command *vamb bin default* with default parameters. TaxVAMB [18] (v.5.0.3) was run with the command *vamb bin taxvamb* with default parameters. Taxanomic annotations for TaxVAMB was obtained from Metabuli [103] (v.1.1.0) using the command *metabuli classify* with *-seq-mode 1* for single-end reads. Metabuli was configured to use GTDB (v.214.1) as reference database, obtained with the command *metabuli databases GTDB*. Taxconverter was used to preprocess Metabuli classifications with command *taxconverter metabuli*. Comebin [19] (v.1.0.4) was run with command *run_comebin.sh* using the default parameters *-n 6*, defining the number of augmented views of a contig. DNABERT-S was obtained from checkpoint *https://huggingface.co/zhihan1996/DNABERT-S*. DNABERT-2 was obtained from checkpoint *https://huggingface.co/zhihan1996/DNABERT-2-117M*, using transformers (v.4.50.3).

APPENDIX N
THRESHOLD SIMILARITY DISTRIBUTIONS

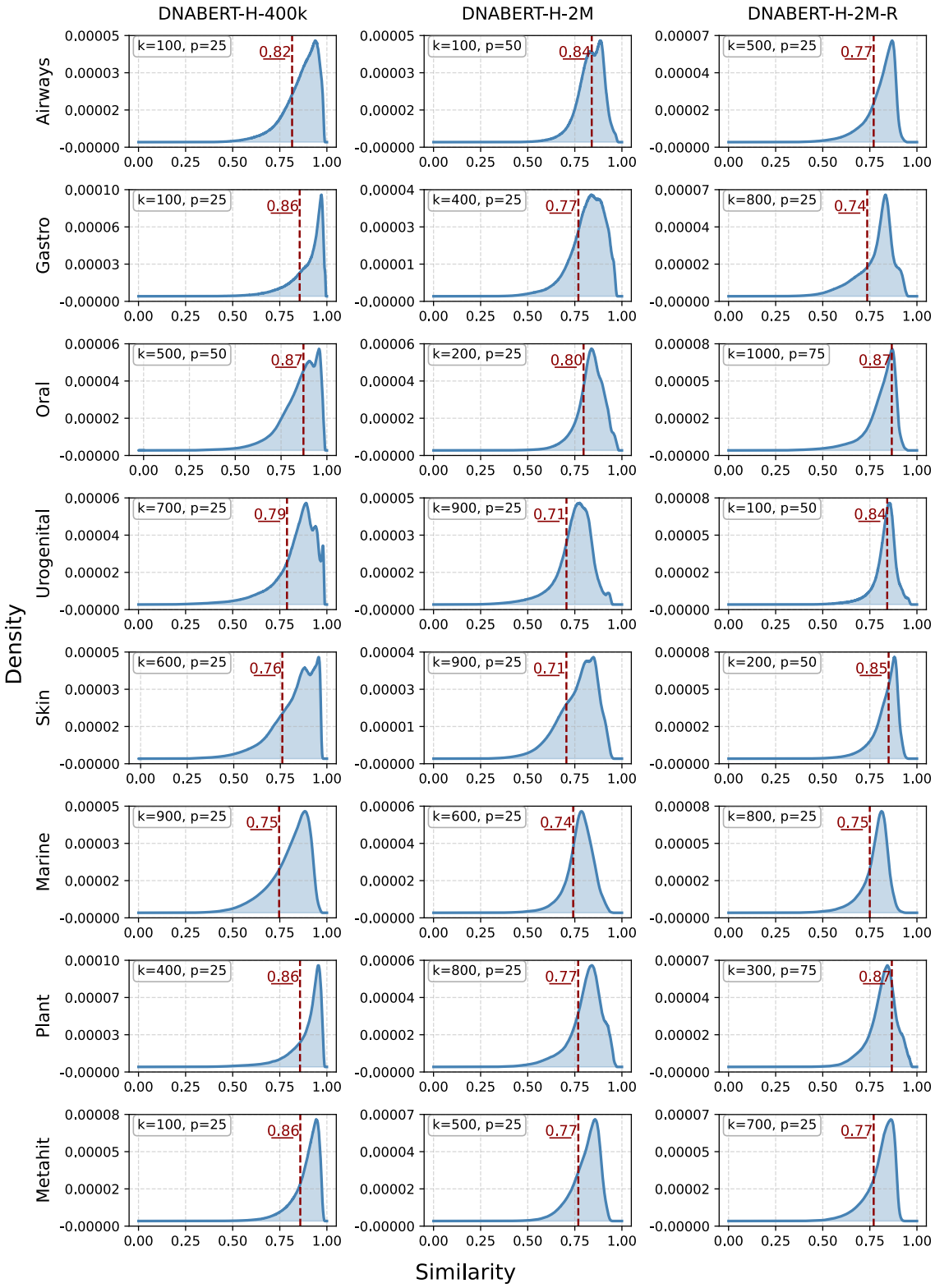

Fig. 13: Threshold similarity distributions for DNABERT-H-400k, DNABERT-H-2M, and DNABERT-H-2M-R

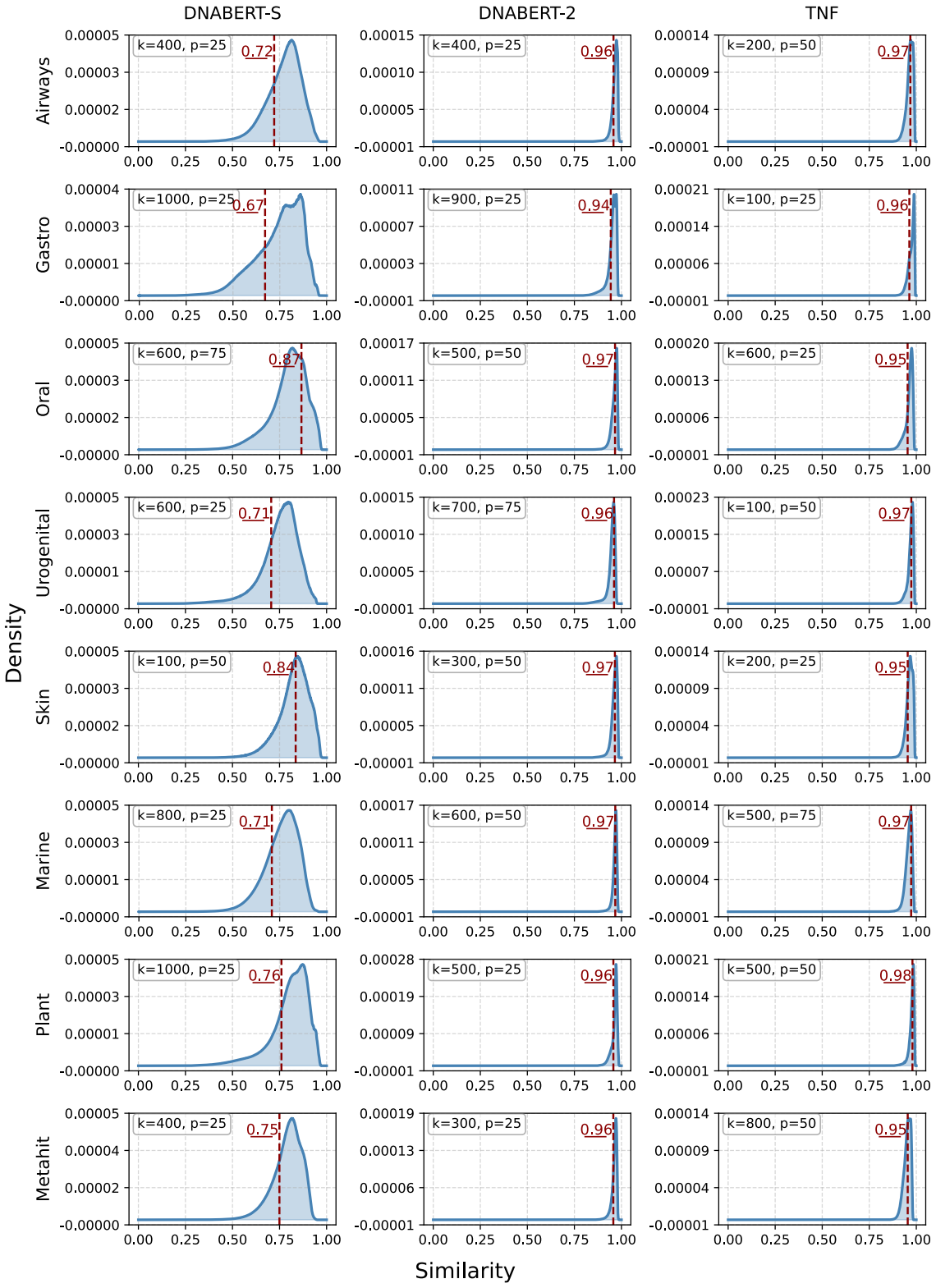

Fig. 14: Threshold similarity distributions for DNABERT-S, DNABERT-2, TNF

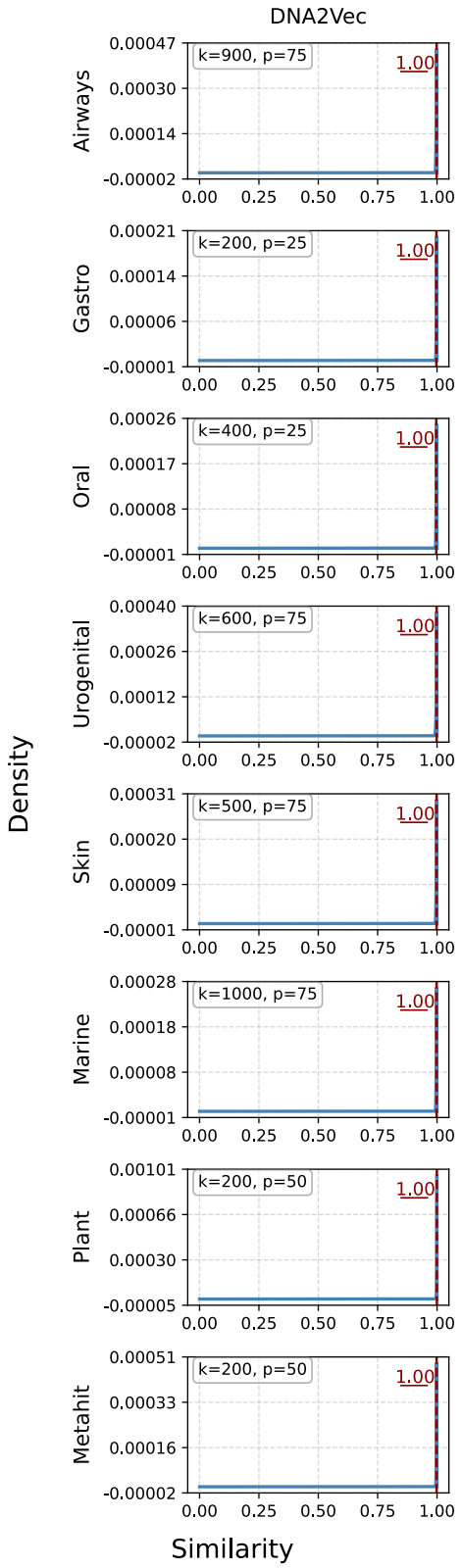

Fig. 15: Threshold similarity distributions for DNA2Vec.

APPENDIX O
PERFORMANCE SCORES $S_{K,P}$ FOR COMBINATIONS OF SELECTED (K, P) VALUES

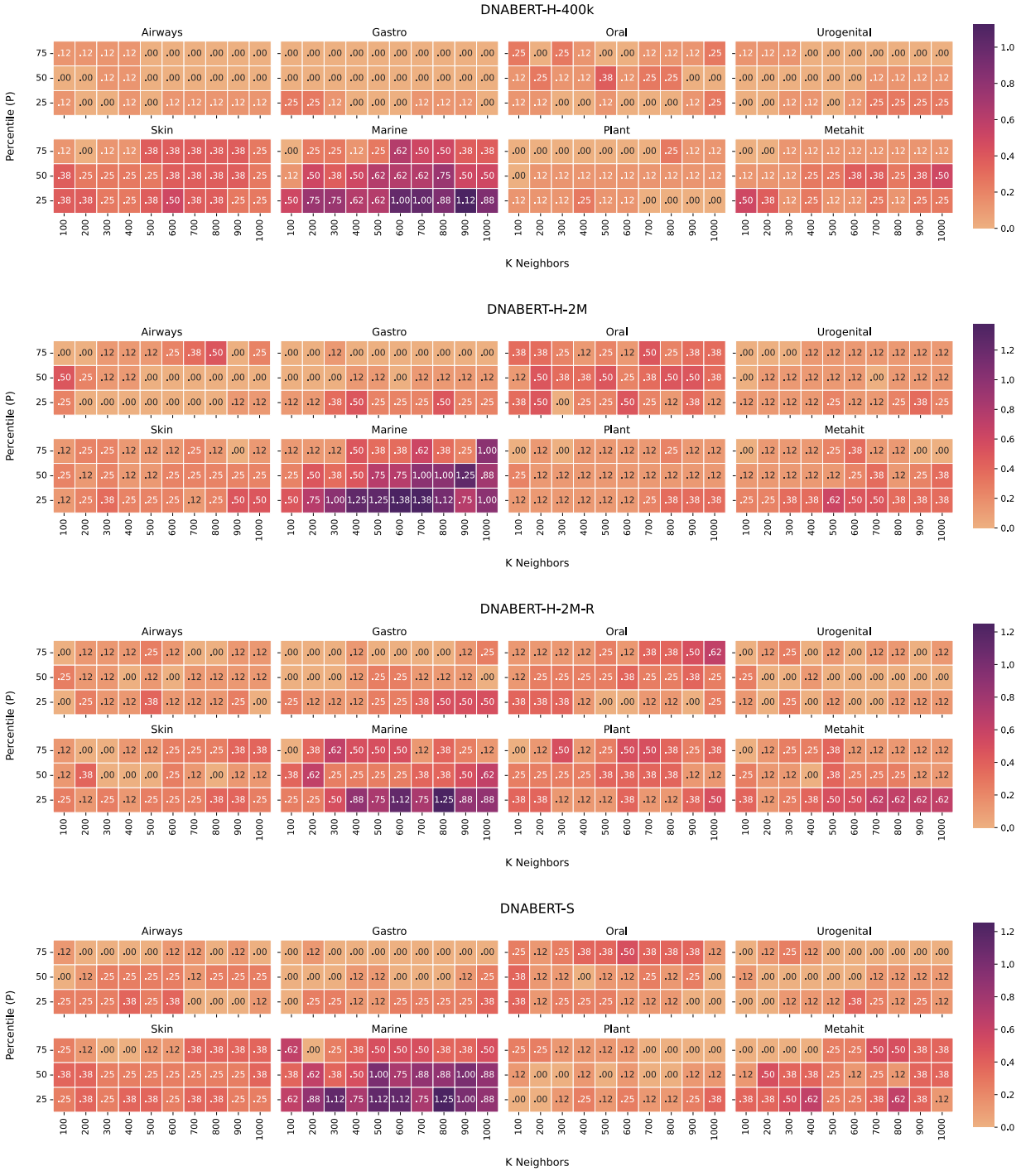

Fig. 16: Comparison of performance scores $S_{K,P}$ for combinations of selected (K, P) values for DNABERT-H-400k, DNABERT-H-2M, DNABERT-H-2M-R, and DNABERT-S.

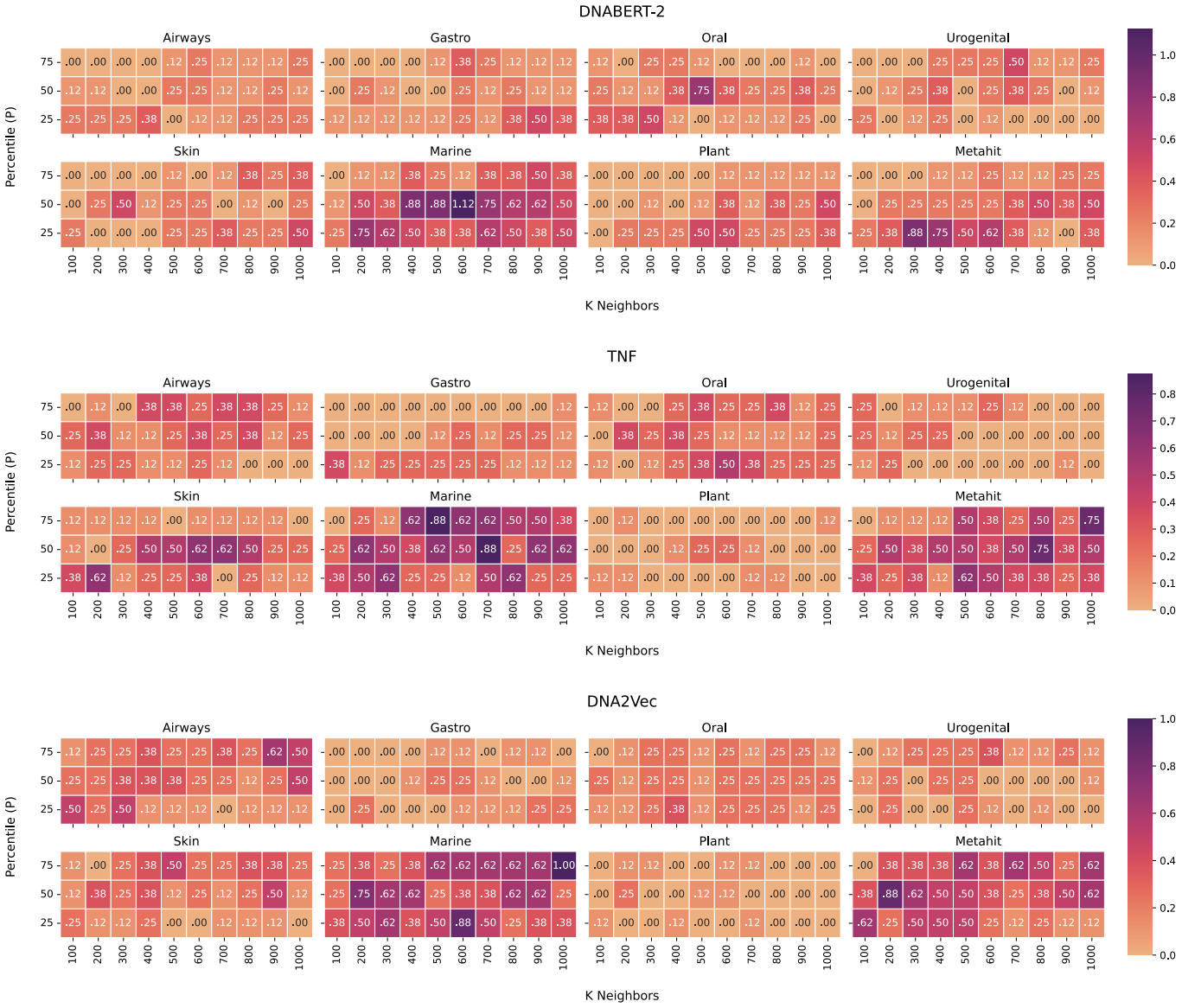

Fig. 17: Comparison of performance scores $S_{K,P}$ for combinations of selected (K, P) values for DNABERT-2, TNF, and DNA2Vec.

APPENDIX P
HYPERPARAMETER GRIDS FOR CLASSIFIERS

Model	Hyperparameters
KNN	$k \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
Logistic Regression	$\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$
Sparse Group Lasso	Group penalty: $\lambda_g \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$ L1 penalty: $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$
Random Forest	n estimators $\in \{100, 300, 500, 700, 900\}$ max features $\in \{\texttt{sqrt}, \texttt{log2}\}$ min samples leaf $\in \{1, 2, 3, 4, 5\}$ criterion $\in \{\texttt{gini}, \texttt{entropy}\}$

TABLE VIII: Hyperparameter grids used in model selection. The Random Forest grid follows a previous metagenome classification study [46].

APPENDIX Q
NUMBER OF BINS BEFORE AND AFTER POSTPROCESSING

Model	Airways		Gastro		Oral		Urogenital		Skin		Marine		Plant		Metahit	
VAMB	31579	251	1981	176	35771	300	5784	169	24178	279	32975	116	152998	412	152998	412
TaxVAMB	16658	291	1730	189	21716	331	2504	176	14240	348	22951	333	20180	423	20180	423
Comebin	575	572	448	447	848	845	330	328	652	651	583	576	2000	410	2000	251
DNABERT-H-400k	2000	269	2000	123	2000	306	1296	70	2000	216	2000	368	2000	269	2000	312
DNABERT-H-2M	2000	338	1491	140	2000	346	700	87	1681	216	2000	354	2000	296	2000	296
DNABERT-H-2M-R	2000	299	980	143	2000	453	2000	167	2000	423	2000	564	2000	329	2000	329
DNABERT-S	2000	269	818	97	2000	326	1001	90	2000	297	2000	343	2000	522	2000	275
DNABERT-2	2000	399	1020	182	2000	539	1921	200	2000	481	2000	728	2000	1043	2000	479
DNABERT2-Random	1999	55	1979	39	2000	65	1433	22	1998	72	2000	99	2000	132	1961	91
TNF	2000	188	2000	132	2000	206	2000	77	2000	215	2000	255	2000	410	2000	251
TNF-Kernel	2000	231	2000	98	2000	224	2000	69	2000	206	2000	285	2000	425	2000	301
DNA2Vec	2000	353	2000	178	2000	312	2000	127	2000	389	2000	442	2000	761	2000	478

TABLE IX: Number of bins before (left column) and after (right column) removing bins with $\leq 250,000$ base-pairs for each model across the eight metagenomics binning datasets.

Model	T2D-EUW		WEGOVY	
VAMB	193111	1802	140883	2613
DNABERT-S	2000	1569	2000	1935
DNABERT-H-2M	2000	1680	2000	1971

TABLE X: Number of bins before (left column) and after (right column) removing bins with $\leq 200,000$ base-pairs for VAMB, DNABERT-S, and DNABERT-H-2M on the two phenotype datasets.

APPENDIX R
METAGENOMICS BINNING RUNTIMES

Model	Airways	Gastro	Oral	Urogenital	Skin	Marine	Plant	Metahit
VAMB	33	19	33	9	30	61	55	–
TaxVAMB	134	65	143	45	128	525	468	–
Comebin	315	192	321	117	292	–	412	–
DNABERT-H / -S / -2	135	79	166	52	123	342	332	148
TNF	8	4	9	3	7	21	15	8
TNF-Kernel	7	4	7	2	7	17	13	7
DNA2Vec	6	3	7	2	6	17	13	7

TABLE XI: Runtime (in minutes) for each model on the eight metagenomics binning datasets, measured on an single NVIDIA L40 GPU. Comebin could not be run on the Marine dataset and neither VAMB, TaxVAMB, nor Comebin could be evaluated on the Metahit dataset.

APPENDIX S
METAGENOMICS BINNING RESULTS FOR CLASSIFICATION DATASETS

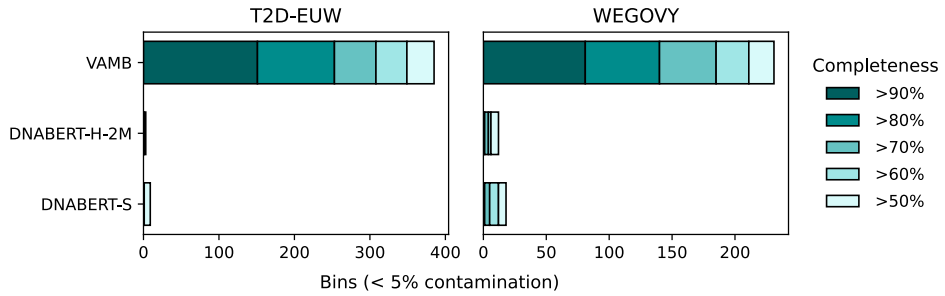

Fig. 18: The number of recovered bins with contamination $\leq 5\%$ at varying completeness thresholds across the two phenotype classification datasets.

APPENDIX T
T-SNE VISUALIZATION OF WEGOVY BAGS

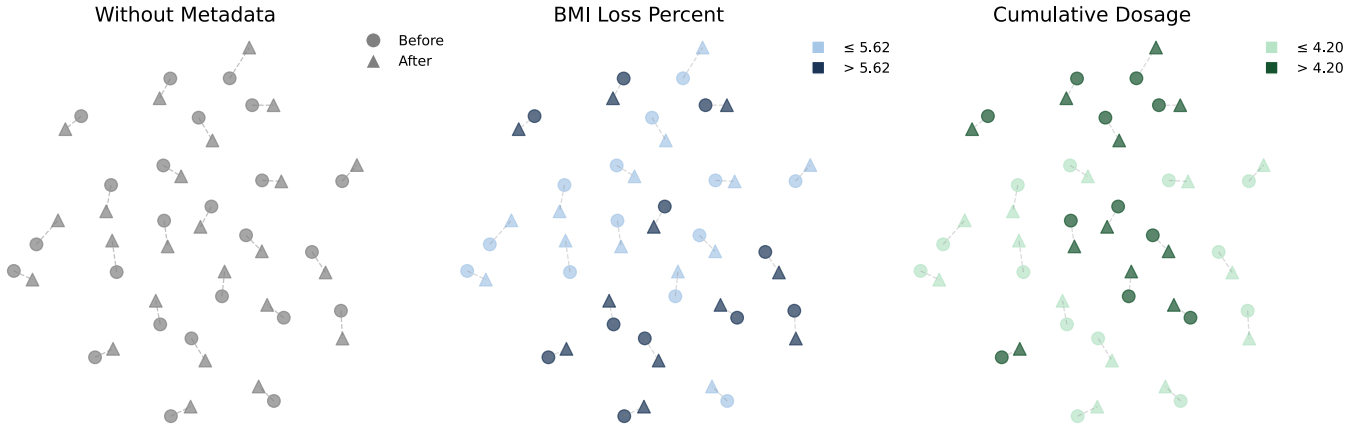

Fig. 19: WEGOVY dataset using t-SNE. Bags from the same patient are connected by an edge. The left subplot corresponds to the phenotype classification task in our main results, showing the small differences in bags before and after treatment. The middle and right plots show the partitioning of the samples by their median *BMI Loss Percent* and *Cumulative Dosage*. None of the visualizations show any trend between before and after bags. Neither do they show any clear separation using the other variables.

APPENDIX U
TEN MOST INFLUENTIAL SPECIES FOR THE T2D CLASSIFICATION TASK USING VAMB

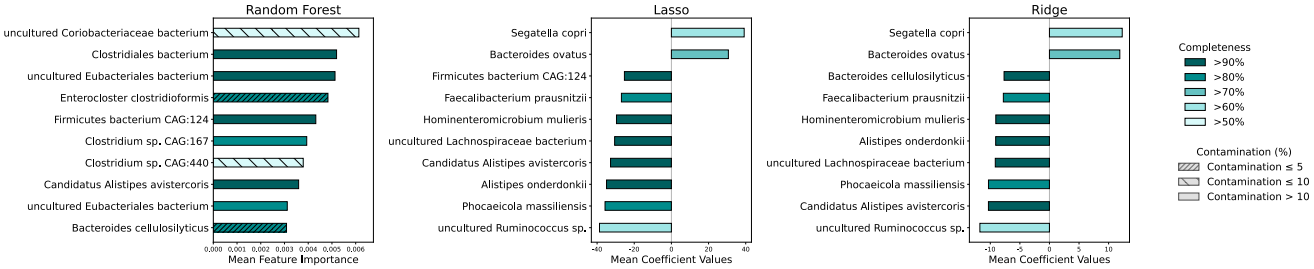

Fig. 20: Ten most influential instances recovered with at least $\geq 50\%$ completeness from the Random Forest, Lasso, and Ridge classifiers using VAMB. The instances are annotated with labels obtained using GTDB-tk and subsequently converted to their equivalent NCBI labels. Colors indicate bin completeness, while the hatches indicate the contamination thresholds.