MT-MAG: Accurate and interpretable machine learning based taxonomic assignment of metagenome-assembled genomes, with a partial classification option

Wanxin Li[1,*], Lila Kari[1], Yaoliang Yu[1], Laura Hug[2]

1 School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
2 Department of Biology, University of Waterloo, Waterloo, Ontario

* wanxin.li@uwaterloo.ca

Abstract

We propose MT-MAG, a novel machine learning-based taxonomic assignment tool for hierarchically-structured local classification of metagenome-assembled genomes (MAGs). MT-MAG is capable of classifying large and diverse real metagenomic datasets, having analyzed for this study a total of 240 Gbp of data in the training set, and 7 Gbp of data in the test set. MT-MAG is, to the best of our knowledge, the first machine learning method for taxonomic assignment of metagenomic data that offers a "partial classification" option. MT-MAG outputs complete or a partial classification paths, and interpretable numerical classification confidences of its classifications, at all taxonomic ranks. MT-MAG is able to completely classify 48% more sequences than DeepMicrobes to the Species level (the only comparable taxonomic rank for DeepMicrobes), and it outperforms DeepMicrobes by an average of 33% in weighted accuracy, and by 89% in constrained accuracy.

Introduction

Recent advances in high-throughput metagenomic sequencing approaches demand increasing power from computational analysis tools [1]. Taxonomic assignment, one of the types of computational analysis tasks in metagenomics, refers to the process of classification of genome fragments to specific taxonomic ranks [2], and facilitates comparisons across studies and the extraction of additional information on the microorganisms present in the sequenced sample [3]. Metagenome-assembled genomes (MAGs) are obtained by grouping assembled contigs with similar characteristics (e.g., sequence composition, read coverage, time series abundances) into population genomes, or bins. The advances in computational analysis tools allow MAGs to be recovered from diverse environments [4]. As MAGs allow connections between predicted metabolisms or cellular functions and taxonomy [5], the design of efficient and accurate taxonomic assignment tools for MAGs has become an imperative need.

Though existing taxonomic assignment tools (e.g. CheckM [6], BERTax [7], GTDB-Tk [8]) have achieved good classification accuracies on benchmarked tasks, they suffer from various limitations.

Alignment-based tools (e.g., GTDB-Tk [8]) require DNA sequences to be aligned to reference sequences to obtain sequence similarities [9]. Alignment-based tools assume that homologous sequences are composed of a series of linearly arranged and more or less conserved sequence stretches, assumptions that may not always hold due to high

mutation rates, frequent genetic recombination events, etc. [10]. Another drawback of alignment-based tools is that they suffer from an often prohibitive consumption of runtime and computational memory.

Genetic marker-based tools such as IDTAXA [3] and GTDB-Tk [8] rely on taxonomic markers (e.g. 16S ribosomal RNA genes, internal transcribed spacers) to identify microorganisms. Genetic marker-based tools suffer from the fact that partial genomes frequently lack major markers. The absence of major markers can be caused either by the genome not being sequenced to a sufficient depth to assemble well (resulting in markers of interest possibly missing from the assembly [11]), or by fragments carrying the markers not binning with the rest of the genome, which is a frequent problem with 16S ribosomal RNA genes [12].

Finally, tools that are based on k-mer frequencies (e.g. DeepMicrobes [13], CLARK [14]) do not rely on alignment or genetic markers, and instead use k-mer frequencies as the input feature. However, existing k-mer based tools suffer from a lack of interpretability of their predicted taxonomic assignments, and the fact that they are only capable of taxonomic assignment at specific taxonomic levels (e.g., Genus, Species).

To address these limitations, in this paper we propose MT-MAG, a **m**achine learning-based **t**axonomic assignment tool for **m**etagenome-**a**ssembled **g**enomes. MT-MAG is an alignment-free and genetic marker-free tool, and is able to partially classify MAGs to the lowest taxonomic rank using a hierarchically-structured local classification, until it does not have enough confidence to classify more deeply. In addition, for a query genome, MT-MAG outputs not only a classification path, but also a confidence score of its classification at each taxonomic rank.

The main contributions of this paper are:

- **Partial Classification Option:** A novel feature of MT-MAG is the partial classification option which enables classification of genomes to the lowest taxonomic rank that MT-MAG has confidence in. The partial classification option is achieved by modifications to the Machine Learning with Digital Signal Processing (MLDSP) [15], an introduction of an (optimized) stopping thresholds picking ((O)STP) algorithm, and a combination of hierarchically-structured local classification with the decision from the (O)STP algorithm.

- **Superior Performance:** The accuracy statistics and partial classification rates enable an examination of the performance of MT-MAG at all taxonomic ranks. At the Species level (the only comparable taxonomic rank for DeepMicrobes), on two real and simulated datasets, MT-MAG outperforms DeepMicrobes [13] by an average of 33% in terms of weighted accuracy and 89% in terms of constrained accuracy. In addition, MT-MAG is able to completely classify an average of 48% more sequences than DeepMicrobes, at the Species level.

- **Interpretability:** MT-MAG outputs interpretable numerical classification confidences. Based on the classification confidences, it probabilistically interprets the final classification path, and uses reliability diagrams to assess the quality of the training sets and determine the reliability of the MT-MAG classification confidences.

- **Large Datasets:** MT-MAG is capable of classifying large and diverse datasets. In the two benchmark tasks performed in this paper, the total genome sizes of training sets analyzed range from 15 Gbp to 240 Gbp, and of the test sets range from 2 Gbp to 7 Gbp.

Materials and Methods

Materials: Datasets and task description

Two different datasets were analyzed in the computational experiments of this study, called **Task 1** and **Task 2**. The dataset analyzed in Task 1 was selected for direct performance comparison purposes, as being the dataset analyzed by DeepMicrobes [13]. More specifically, in Task 1 the training set comprises representative genomes from species in human gut microbiomes, and the test set comprises high-quality MAGs reconstructed from human gut microbiomes from a European Nucleotide Archive study [16]. Since the test set analyzed in Task 1 comprised simulated reads, a second dataset for selected for **Task 2**, whereby both the training set and the rest set consist of real DNA fragments. More specifically, in Task 2 the training set comprises representative and/or non-representative microbial genomes from GTDB r202, and the test set comprises 913 "draft" bacterial and archaeal genomes assembled from rumen metagenomic sequence data derived from 43 Scottish cattle [17].

The rationale behind the selection of DeepMicrobes for a benchmark comparison with MT-MAG is as follows. Like MT-MAG, DeepMicrobes is a machine learning-based alignment-free and genetic marker-free metagenomic taxonomic assignment tool that uses k-mer frequencies as input feature to predict taxonomic assigments of short reads at the Genus and Species level. DeepMicrobes has demonstrated better performance at the Species level classification, and better comparative accuracy in Species abundance estimation over other state-of-the-art tools, see [14, 18–22]. In addition, like MT-MAG, DeepMicrobes estimates classification confidences: The reads with classification confidences below the constant rejection threshold are considered to be *unclassified reads*, while the rest are considered to be *classified reads*. Within the set of classified reads, the reads whose classified Species taxa are the same as their ground-truth Species taxa are considered to be *correctly classified reads*. Lastly, to the best of our knowledge, DeepMicrobes is the only taxonomic assignment tool that enables probabilistic rejection using machine learning classifiers, similar to MT-MAG's design goals.

All genomes (MAGs) were provided in FASTA format: A FASTA file for a MAG typically contains multiple DNA sequences, which are the scaffolds generated from the original metagenome assembly that were subsequently binned into that MAG. As MT-MAG and DeepMicrobes have different requirements on their inputs, in that MT-MAG requires the training genomes to be $> 10{,}000$ bp (see "m-MLDSP as a subprocess"), while DeepMicrobes operates with short reads, the datasets were prepared separately for MT-MAG and DeepMicrobes.

Task 1: Simulated reads for the test set, and sparse training set

Task 1 used simulated reads for the test set, and the training set makes up only 2.4 % of the GTDB at Species level, hence it will be referred to as *Task 1 (simulated/sparse)* in the remainder of the paper. The dataset for this task was specifically chosen so as to allow a direct comparison between the quantitative statistics (see "Performance metrics") of MT-MAG and those of DeepMicrobes.

The dataset for Task 1 was adapted from DeepMicrobes's first benchmark test [13]. The training set comprised 2,505 representative genomes of human gut microbial species identified previously by a large-scale assembling study of the species in human gut microbiomes, see [16] as well as the species labels assigned to these genomes from `ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/umgs_analyses`. The genomes in this dataset belong to 1,952 unclassified MAGs, and 553 microbial gut Species-level genome representatives from the human-specific reference (HR) database. These 2,505 genome database was referred to in [13] as *"HGR."* Starting from HGR, 150 basepair

(bp) paired-end reads were generated using the ART Illumina read simulator [23], and one read was simulated from a sequence. The simulated reads were used as the training set by DeepMicrobes. The total genome size of the training set of DeepMicrobes is 2.15 Mbp. The test set comprised 250 bp paired-end reads simulated using ART Illumina from 3,269 high-quality MAGs reconstructed from human gut microbiomes from a European Nucleotide Archive study titled "A new **g**enomic **b**lueprint of the **h**uman **g**ut **m**icrobiota" (GBHGM) [16]. The ground-truth species labels for the test set were assigned using the scripts available at `https://github.com/Finn-Lab/MGS-gut`. The total genome size of the test set of DeepMicrobes is 653.80 Mbp. For the assignments of ground-truth labels using the National Center for Biotechnology Information (NCBI) taxonomy above the Species level, for both the training and the test sets, CheckM [6] was used.

We noted that due to the lack of consensus on uncultivated taxa naming conventions [24], NCBI [25] labels are outdated, whereas in GTDB a consistent naming scheme was achieved by naming uncultivated taxa to be 'Genus name' sp1, 'Genus name' sp2 and so on [26]. Also, GTDB provides a complete taxonomic hierarchy, with no inconsistencies in naming based on standardized phylogenetic distances to define taxonomic ranks [27]. Thus, rather than using NCBI labels as the ground truth, we ran GTDB-Tk [8], to obtain new ground-truth labels for the training and test sets, based on GTDB release r202.

The training set for MT-MAG was prepared as follows. To address the challenge that the training genomes of MT-MAG are required to be $> 10{,}000$ bp, we used the original genomes in HGR as the training set. In addition, all sequences with length below 5,000 bp were discarded. The remaining 14,358 sequences comprised the training set of MT-MAG. The total genome size of the training set of MT-MAG is 15.47 Gbp. The ground-truth GTDB species labels are the same as for DeepMicrobes. Please refer to "Dataset details" in the Supplemental Materials for genome size, contig count and percent GC distributions of this dataset.

Regarding the test set for MT-MAG, similar to the training set, we used the original genomes for the 3,269 MAGs in GBHGM. The total genome size of the test set of MT-MAG is 7.42 Gbp. The ground-truth GTDB species labels are the same as for DeepMicrobes. Please refer to "Dataset details" in the Supplemental Materials for genome size, contig count and percent GC distributions of this dataset.

Task 2: Real DNA fragments for the test set, and dense training set

Task 2 used real DNA fragments as the test set, and the training set makes up 7.7% of GTDB taxonomy, hence it will be referred to as *Task 2 (real/dense)* in the remainder of the paper.

The training set for MT-MAG was taken from GTDB release r202, released on 27th April 2021. Note that the genome sizes for GTDB are significantly larger than those for HGR. Please refer to "Dataset details" in the Supplemental Materials for genome size, contig count and percent GC distributions of this dataset. Most GTDB MAGs contain multiple DNA sequences per genome. We concatenated all sequences listed for a genome, and selected 4 non-overlapping fragments of length 100,000 bp, using four random starts. We repeated the sampling process 20 times, resulting in 80 fragments per genome, to ensure we had a sufficient number of sets of fragments to perform cross-validation. The ground-truth labels were obtained directly from the GTDB taxonomy. The total genome size of the training set of MT-MAG is 239.53 Gbp.

To prepare the training set for DeepMicrobes, the training set of MT-MAG (consisting of GTDB full genomes) was used to simulate the training set for DeepMicrobes (consisting of reads), as follows. For every contig of a genome in the training set for MT-MAG, one 150 bp paired-end read was simulated and added to the

DeepMicrobes training set. The total genome size of the training set of DeepMicrobes is 24.54 Mbp, and approximately 3 million reads were simulated.

The test set for MT-MAG used 913 microbial genomes from metagenomic sequencing of the cow rumen, which were derived from 43 Scottish cattle [17]. We used all sequences as the training fragments, and the ground truth labels were obtained by running GTDB-Tk [8] using GTDB r202 data. The total genome size of the test set of MT-MAG is 2.18 Gbp. Please refer to "Dataset details" in the Supplemental Materials for genome size, contig count and percent GC distributions of this dataset.

The test set for DeepMicrobes (reads) was prepared from the test set for MT-MAG (genomes) using the same method as that used for the training set. The total genome size of the test set of DeepMicrobes is 47.43 Mbp, and approximately 300 thousand reads were simulated.

Table 1. Summary of total genome size (total number of basepairs analyzed), number of samples and number of contigs for training and test sets in Task 1 (simulated/sparse), and Task 2 (real/dense) for MT-MAG and DeepMicrobes.

Task ID	Dataset type	Tool	Total genome size	Number of samples	Number of contigs
Task 1 (simulated/sparse)	Training	MT-MAG	15.47 Gbp	2298	469537
		DeepMicrobes	2.15 Mbp	2298	14358
	Test	MT-MAG	7.42 Gbp	3269	245564
		DeepMicrobes	653.80 Mbp	3269	2615200
Task 2 (real/dense)	Training	MT-MAG	239.53 Gbp	635248	2540992
		DeepMicrobes	24.54 Mbp	218030	218030
	Test	MT-MAG	2.18 Gbp	913	158102
		DeepMicrobes	47.43 Mbp	1826	316204

Table 1 provides a summary of the total genome sizes (total number of basepairs analyzed), number of FASTA files, and number of contigs for training and test sets in Task 1 (simulated/sparse) and Task 2 (real/dense), for MT-MAG and DeepMicrobes. Note that, for both tasks, the total genome size of the training/test set of MT-MAG is 23,594 to 3,358,482 times larger than that for DeepMicrobes.

Methods: MT-MAG algorithm

This section describes the m-MLDSP subprocess that is at the core of MT-MAG, the hierarchically-structured local classification approach used by MT-MAG, and the two main phases of MT-MAG (training and classifying).

The m-MLDSP subprocess

MLDSP is an alignment-free software tool for ultrafast, accurate, and scalable genome classification at all taxonomic ranks, [15]. MLDSP combines supervised machine learning techniques with digital signal processing. MLDSP consists of two main steps: (a) *Pretraining*, whereby several different classifiers' performance is evaluated by 10-fold cross validation, and (b) *Classify*, whereby MLDSP first trains the classifiers using the entire training set *(Classify-Training)*, and then classifies new DNA sequences from the test set *(Classify-Classification)* (see "MT-MAG algorithm details" in Supplementary Materials).

MT-MAG uses a modified version of MLDSP, called *m-MLDSP* (modified MLDSP) as a subprocess. The m-MLDSP subprocess augments MLDSP in several significant ways. First, it modifies MLDSP to handle the frequent special case where the parent taxon has only one child taxon, and it outputs classification confidences for its

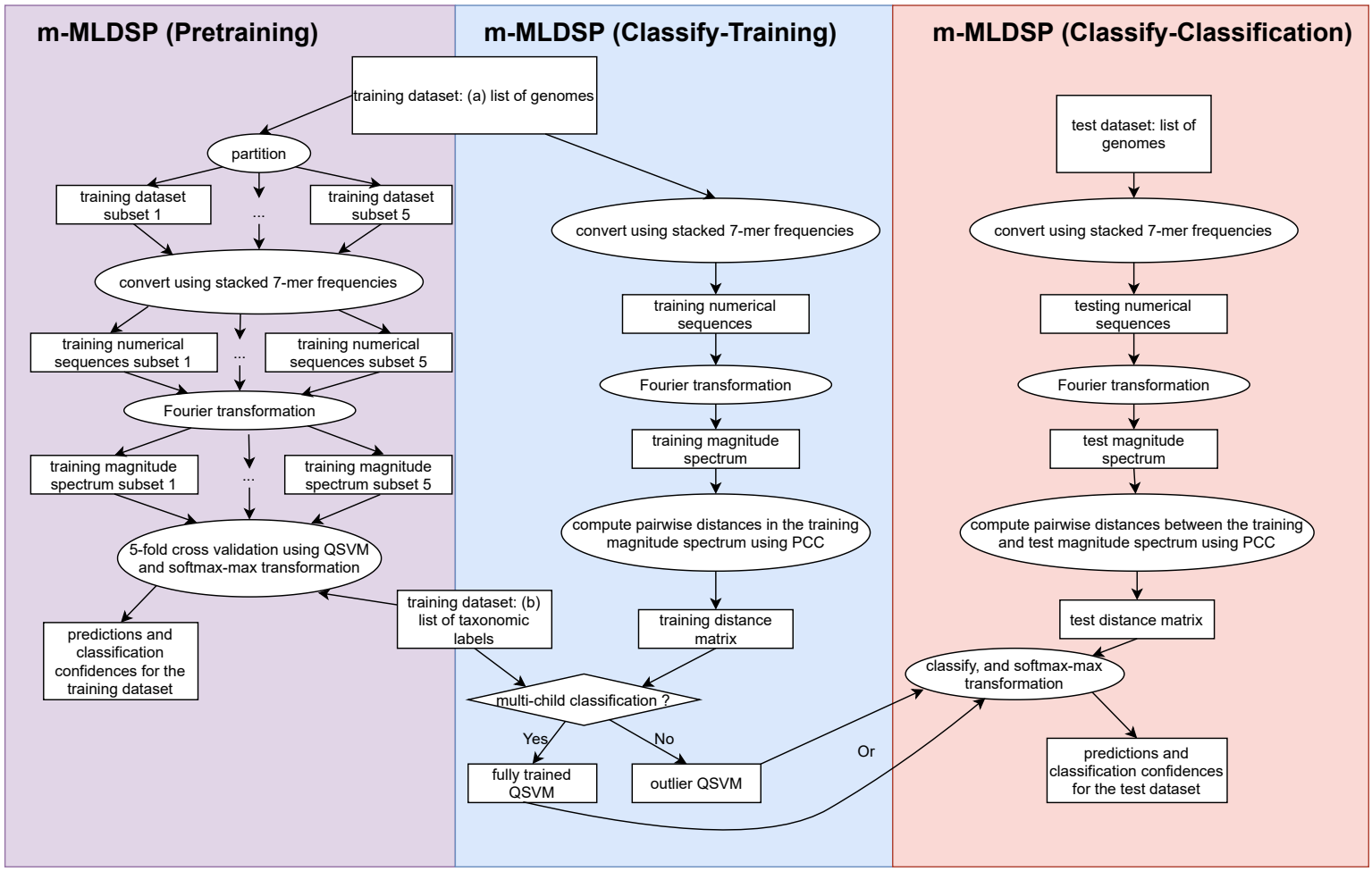


Fig 1. *Overview of m-MLDSP*, including the main steps that comprise m-MLDSP (Pretraining) (lavender box), m-MLDSP (Classify-Training) (blue box), m-MLDSP (Classify-Classification) (pink box). Ellipses represent computation steps. Rectangles represent inputs to, and outputs from, computation steps. The diamond represents a condition checking. Note that the training dataset consists of genomes together with their taxonomic labels.

classifications. Second, it adds an optimized stopping threshold picking algorithm called (O)STP, that is at the core of the partial classification option feature of MT-MAG. Third, it combines the hierarchically-structured local classification with the output of the (O)STP algorithm to output "uncertain classification," if the classification confidence is below the stopping threshold.

More specifically, the modifications and augmentations to MLDSP include: selecting the parameter $k = 7$ for the purpose of k-mer representations, adding a QSVM classifier to handle the aforementioned special case, choosing the QSVM classifier from the list of MLDSP classifiers, adding classification confidences to the MLDSP, and using the modified MLDSP (m-MLDSP) as a subprocess in MT-MAG. Figure 1 provides an overview of m-MLDSP, including the main steps to accomplish m-MLDSP (Classify-Training), m-MLDSP (Classify-Classification) and m-MLDSP (Pretraining).

The MLDSP implementation of the algorithms assumes that the input DNA sequences belong to multiple child taxa (*multi-child classification*). If this is the case, in the m-MLDSP (Classify-Training) step, a QSVM classifier called *fully trained QSVM* is trained, using the entire training set. In the m-MLDSP (Classify-Classification) step,

m-MLDSP computes classifications (taxonomic assignments) for the DNA sequences in the test set by using the fully trained QSVM, and the classification confidences of these classifications by using a softmax-max transformation. In contrast with its precursor, m-MLDSP then applies five-fold cross-validation to obtain classifications, and uses a softmax-max transformation to compute classification confidences, for the entire training set. These results are later used for determining the stopping thresholds for each pair of parent and child taxon.

The case where the training/test sequences belong to a single child taxon (*single-child classification*) has not been considered, and cannot be handled by MLDSP. In this case, in the m-MLDSP (Classify-Training) step, a QSVM classifier called *outlier QSVM* is trained, using the entire training set, with an optional user-specified outlier fraction (default: 10%). In the m-MLDSP (Classify-Classification) step, m-MLDSP computes classifications for the DNA sequences in the test set by classifying using the outlier QSVM, and computes the classification confidences of these classifications by utilizing a softmax-max transformation. The m-MLDSP (Pretraining) step is not applicable here, since there is no need for picking stopping thresholds in the case of single-child classifications.

A hierarchically-structured local classification approach

Taxonomic assignment is a problem of hierarchical classification, whereby instances are grouped according to a hierarchy. A hierarchy can be formalized as an direct acyclic graph where every node can be reached by a unique path from the root node (see Figure 2). In machine learning, there are generally three types of approaches to hierarchical classification [28]. The simplest approach is *flat classification* where all parent nodes are ignored, and a single classifier is trained to classify each instance directly into a leaf node. The second approach is the so-called *big bang* classification where a single classifier is trained for all nodes in the hierarchy. The third approach is the *hierarchically-structured local classification*, whereby one multi-class classifier is trained for each parent-to-child relationship. This third approach is an iterative classifying process where instances classified to a child node are then further classified with the next-level classifier, where the child node is now the parent node for the next-level classifier.

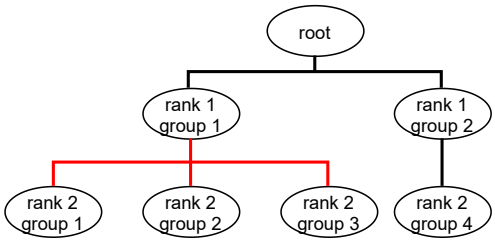


Fig 2. A sample hierarchy (taxonomy) with three parent-to-children relationships. The node below, and connected by a path, to a given node, is called its child node; the given node is called the child node's parent. A parent node with all its children nodes forms a parent-to-child relationship. A node without a child node is called a leaf node. The level of a node is the length of path from that node to the root node. The part highlighted in red is a multi-child classification, while the part with parent "rank 1 group 2" is a single-child classification.

In contrast with DeepMicrobes which uses flat classification, MT-MAG uses hierarchically-structured local classification for the reasons detailed below. First, in the case of flat classification, an erroneous classification of a DNA sequence directly at the

Species level is more likely, due to the very large number of classes at the Species level. This, in turn, results in a higher likelihood of placing the sequence into an erroneous higher-level taxonomic rank, e.g., Phylum. Such a serious misplacement is less likely to happen with hierarchically-structured local classification, which passes through multiple classifications, from higher to lower taxonomic ranks, thus providing multiple check-points for the identification of an incorrect classification. For example, an incorrect Phylum classification could be prevented if any of the classifications prior to and including this level are deemed "uncertain."

In addition, in the case of flat classification, if the classification confidence of a sequence into a Species taxon does not meet the required confidence level, this sequences is simply deemed "unclassified," with no further information being provided. In contrast, the hierarchically-structured local classification provides the option of partial classification and can output partial classification paths for such sequences, even if their Species level classification is uncertain.

Finally, flat classification requires significantly more computational time and memory resources, because it involves a single big classification task wherein all the training sequences are loaded into memory simultaneously. In contrast, a hierarchically-structured local classification approach involves multiple smaller classification tasks and, for each classification task, one only needs to load into memory the sequences pertaining to the specific parent taxon being classified at this step in the hierarchy. In particular, for classifications at higher taxonomic ranks, one can use, e.g., only representative genomes as opposed to all the genomes available for that parent taxon.

The MT-MAG training phase and classifying phase

MT-MAG comprises two phases, training and classifying, as described below.

The MT-MAG training phase (of the training set – genomes with known ground truth labels) comprises multiple training processes: For each parent taxon, after preprocessing its training set (see "Materials: dataset and task description"), two situations can occur, depending on the number of child taxa:

- *Multi-child classification.* For every parent-child pair, a stopping threshold is determined, based on the confidences calculated by m-MLDSP (Pretraining) with the training set as input. MT-MAG selects the stopping threshold from a list of candidate stopping thresholds, and searches for the stopping threshold T which results in the fewest number of genomes with classification confidences lower than T, while at the same time resulting in the classification accuracies of the other genomes being higher than a user-specified accuracy parameter.

  More specifically, a stopping threshold is the result of subtracting a "variability" parameter from the maximum of *(a)* the minimum of the candidate thresholds (numbers between 0 and 1) that result in a "constrained accuracy" being greater than a user-specified parameter (default: 90%), and *(b)* the average of classification confidences for the genomes with correct m-MLDSP (Pretraining) classifications. Note that, in contrast to DeepMicrobes, which uses the same threshold for all child-classes, the MT-MAG stopping thresholds are class-specific.

  Subsequently, a QSVM classifier (the fully trained QSVM) is trained with the entire training set of this parent taxon, as part of the m-MLDSP (Classify-Training) step.

- *Single-child classification.* A QSVM classifier (the outlier QSVM) is trained in the m-MLDSP (Classify-Training) step, by setting a fraction (default 10%) of the training set as a second child-class (called outlier taxon).

The MT-MAG classifying phase (of the test set – genomes with known ground truth labels –, and unknown genomes) proceeds as follows. When, in the process of hierarchically-structured local classification, MT-MAG has classified a test/unknown genome into a parent taxon, and attempts to classify it further into one of its child taxa, two possibilities can occur:

- *Multi-child classification.* If the parent taxon has multiple child taxa, then the fully trained QSVM is used to classify the test/unknown genome into one of the child taxa, and this result is also used to compute a classification confidence as part of the m-MLDSP (Classify-Classification) step. If this classification confidence is below the stopping threshold for this parent-child pair, then this classification is considered uncertain, and no further attempts are made to classify this test/unknown genome from the child taxon into its own child taxa.

- *Single-child classification.* If the parent taxon has a single child taxon, then the outlier QSVM is used to classify the test/unknown genome into either the child taxon or the outlier taxon as part of the m-MLDSP (Classify-Classification) step, and the result is used to compute a classification confidence. If the output is the outlier taxon, then this classification is considered uncertain and no further classifications are attempted.

Given a test/unknown genome, the output of MT-MAG is either (a) a complete classification path down to the Species level, if all the intermediate classification confidences exceed the stopping thresholds, or (b) a partial classification path, up to the lowest taxonomic rank with high enough classification confidence, otherwise. In either case, the output of MT-MAG also includes the classification confidence for each taxon along the classification path.

Figure 3 illustrates the MT-MAG training phase and the MT-MAG classifying phase for classifying two genomes, genome a and genome b, belonging to a given parent taxon, into one of its two child taxa (multi-child classification). Please refer to "MT-MAG algorithm details" in Supplementary Materials for an illustration of the single-child classification scenario, namely the training phase and the classifying phase for classifying two genomes from a parent taxon into its single-child taxon (single-child classification).

Results

In this section, we first describe the performance metrics used for measuring the performance of MT-MAG, and for comparing it with DeepMicrobes in terms of constrained accuracy, absolute accuracy, and weighted accuracy (higher is better), and partial classification rate (lower is better). Second, we present a detailed analysis of the novel features of MT-MAG: *(i)* the capability to classify a DNA sequence at all taxonomic ranks, *(ii)* the capability to output an interpretable classification confidence for the classification at each taxonomic rank along the classification path, and *(iii)* the capability to output a "partial classification" option when the classification confidence of a classification does not meet a given threshold. Third, since DeepMicrobes is able to classify only at the Species level, we summarize the results of a comparative analysis of the performance of MT-MAG with that of DeepMicrobes at the Species level.

Performance metrics

In this section, we define the terminology and the performance metrics used to discuss and assess the performance of MT-MAG's classification of test and unknown genomes.

A classification of a genome x from taxonomic rank tr_1 to taxonomic rank tr_2, is called *a classification at* tr_2. If m-MLDSP classifies genome x at taxonomic rank tr_2 as

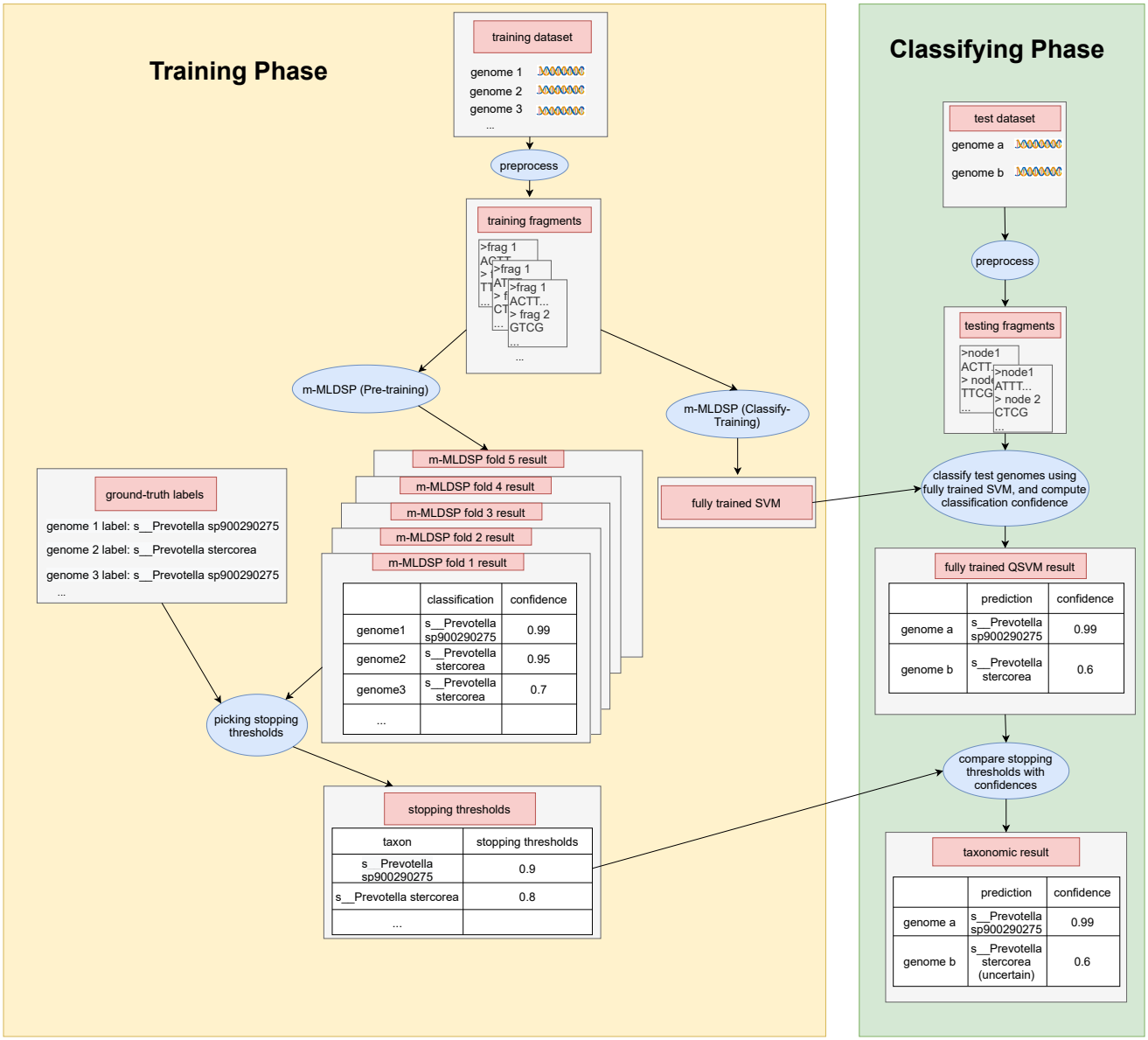


Fig 3. *MT-MAG pipeline for classifying two genomes, genome a and genome b, from the parent taxon Genus Prevotella into its two child taxa, Species Prevotella sp900290275, and Species Prevotella stercorea (multi-child classification).* Blue ellipses represent computation steps. Gray rectangles represent inputs to, and outputs from, computation steps. In the MT-MAG training phase (yellow box), the training set is preprocessed and given as the input to m-MLDSP (Pretraining). The classifications and classification confidences outputted in m-MLDSP (Pretraining) for the training set from all folds are used for determining the stopping thresholds for every child taxon of this parent taxon. Furthermore, in m-MLDSP (Classify-Training), a fully trained QSVM is trained by using the entire training data. In the MT-MAG classifying phase (green box), the test set is preprocessed and given as the input to m-MLDSP (Classify-Classification), together with the fully trained SVM from the training phase. m-MLDSP (Classify-Classification) outputs a classification and a classification confidence for each genome in the test set. Then, the classification confidence from m-MLDSP (Classify-Classification) to classify a genome from the parent taxon into a child taxon is compared with the stopping threshold of the parent taxon and child taxon. If the classification confidence is lower than its stopping threshold, then the output is "uncertain classification" and further classification into children of this child taxon will not be attempted.

belonging to taxon t, but the confidence of this classification does not pass the stopping threshold for taxon t, then we say that the classification of x *stops at* tr_2. We say that a classification of genome x as belonging to taxon t at taxonomic rank tr_2, is a *correct classification at* tr_2, if t matches the ground-truth taxon of x at tr_2. For example, in Figure 4, the classification of genome x from root to "rank 1 group 1" is a classification of x at *rank 1*, and the classification of x from "rank 1 group 1" to "rank 2 group 2" is a classification of genome x at *rank 2*. The classification of genome x stops at rank 2. Additionally, if the ground truth is that x belongs to "rank 1 group 1," we say that the classification of x from root to rank 1 is a correct classification at rank 1.

Given a taxonomic rank tr, we call the classification of a genome x a *partial classification at* tr if the classification stops at tr, or at a rank higher than tr; otherwise, the classification is called a *complete classification at* tr. We say a classification is a *correct classification up to* tr if it is a complete classification at tr, and a correct classification at all taxonomic ranks higher than, and including, tr.

For example, in Figure 4, the classification of genome x is a complete classification at rank 1, but a partial classification at rank 2. In addition, if the ground-truth taxon of genome x at rank 1 is "rank 1 group 1", then the classification of genome x is a correct classification at rank 1, as well as a correct classification up to rank 1.

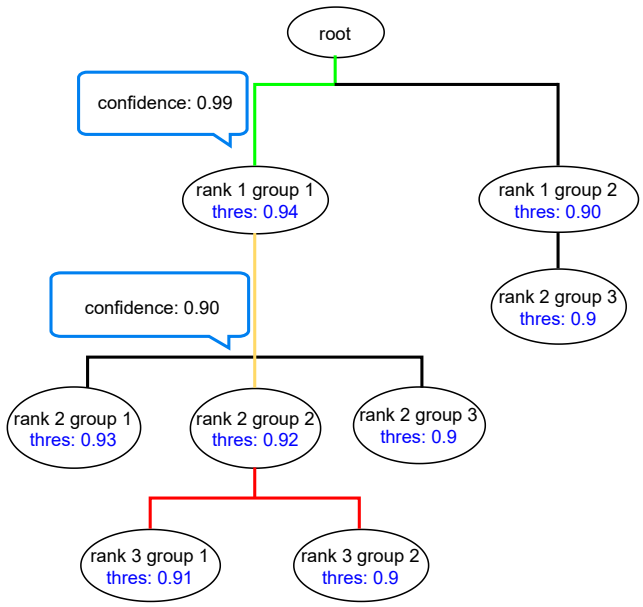


Fig 4. An example of a classification path for an unknown genome x. The pre-calculated stopping thresholds are noted in blue within the taxon nodes. First, x must belong to root, and MT-MAG classifies x into rank 1 group 1; the confidence 0.99 exceeds the stopping threshold for rank 1 group 1 (0.94), so MT-MAG continues classification. In the next iteration, MT-MAG classifies x into rank 2 group 2, however, since the confidence 0.90 is below the pre-calculated threshold for rank 2 group 2 (0.90), MT-MAG does not have enough confidence to continue classification. The path in green represents complete classifications. The path in yellow represents partial classification(s). The part in red represents classifications that were not attempted.

With this terminology, for a given taxonomic rank tr we define the following performance metrics (the subscript g indicates that these metrics refer to genomes):

- $CA_g(tr)$ (constrained accuracy): the proportion of the test genomes with correct classifications up to tr, to the test genomes with complete classifications at tr.

- $AA_g(tr)$ (absolute accuracy): the proportion of the test genomes with correct classifications up to tr, to all test genomes.

- $WA_g(tr)$ (weighted accuracy): the weighted proportion of the test genomes with correct classifications up to tr (with weight 1) together with the test genomes with partial classifications at tr (with weight 0.5), to all test genomes. The weight parameter 0.5 indicates the assumption that a partial classification lies "halfway" between a correct classification and an incorrect classification. This design decision will hereafter be called the *neutrality assumption*.

- $PCR_g(tr)$ (partial classification rate): the proportion of the test genomes with partial classifications at tr, to all test genomes.

The three accuracies $CA_g(tr), AA_g(tr)$ and $WA_g(tr)$ are numbers between 0 and 1, with $CA_g(tr) \geq AA_g(tr)$, and where higher values indicate better performance. The partial classification rate $PCR_g(tr)$ is a number between 0 and 1, and a lower value indicates a higher proportion of classified genomes, indicating thus a better performance.

MT-MAG novel features

Classifications at all taxonomic ranks

In contrast with DeepMicrobes which only classifies reads at the Species level, a significant feature of MT-MAG is its capability to classify genomes at all taxonomic ranks. Table 2 provides a summary of MT-MAG's performance metrics, at all taxonomic ranks, for both Task 1 (simulated/sparse) and Task 2 (real/dense).

In Task 1 (simulated/sparse), MT-MAG achieves an excellent performance at taxonomic ranks from Phylum to Genus, with a lower performance the Species level. Specifically, the MT-MAG constrained accuracies $CA_g(tr)$ are above 98% at all taxonomic ranks, except at the Species level $CA_g(Species)$, where they drop to 80.54%. The increase in the number of incorrect classifications at the Species level explains, in part, the 15.01% drop in weighted accuracy $WA_g(tr)$ from the Genus to the Species level.

In Task 2 (real/dense), the MT-MAG has an excellent performance all around, with constrained accuracies $CA_g(tr)$ above 97% at all taxonomic ranks. However, the partial classification rates $PCR_g(tr)$ (lower is better) increase by 10.31% from the Domain level to Phylum level, indicating that there is a large number of test genomes not classified beyond at Phylum level. This provides a possible explanation for the 5.51% drop in weighted accuracy $WA_g(tr)$ from the Domain to the Phylum level.

Numerical classification confidences

In addition to the final classification path, MT-MAG also outputs numerical classification confidences along the classification path, indicating how confident MT-MAG is in the classification, at each taxonomic rank. For example, the final classification path for genome x, illustrated in Figure 4 is intepreted as MT-MAG being 99% confident in classifying x from "root" to "rank 1 group 1", and 90% confident in classifying x from "rank 1 group 1" to "rank 2 group 2". However, since the confidence of the latter classification does not exceed the pre-calculated stopping threshold of 92%, this classification is deemed "uncertain" and, in addition, no further classifications are attempted.

Table 2. Summary of MT-MAG performance metrics at all taxonomic ranks, for Task 1 (simulated/sparse) and Task 2 (real/dense): constrained accuracy $CA_g(tr)$, absolute accuracy $AA_g(tr)$, weighted accuracy $WA_g(tr)$ (higher is better); partial classification rate $PCR_g(tr)$ (lower is better).

Task ID	**Taxonomic Rank**	$CA_g(tr)(\%)$	$AA_g(tr)(\%)$	$WA_g(tr)(\%)$	$PCR_g(tr)(\%)$
Task 1 (simulated/sparse)	Phylum	100.00	94.01	97.00	5.99
	Class	100.00	94.01	97.00	5.99
	Order	98.50	92.18	95.86	7.36
	Family	100.00	81.77	90.89	18.23
	Genus	99.35	76.93	88.21	25.57
	Species	80.54	61.18	73.20	24.04
	Average	96.56	83.35	90.36	14.03
Task 2 (real/dense)	Domain	99.75	96.16	97.96	3.60
	Phylum	99.30	85.49	92.45	13.91
	Class	99.30	85.13	92.27	14.27
	Order	99.57	84.17	91.91	15.47
	Family	99.71	82.97	91.37	16.79
	Genus	97.44	73.02	85.55	25.06
	Species	98.63	68.82	83.93	30.22
	Average	99.10	82.25	90.78	17.04

Determining the reliability of the MT-MAG classification confidences

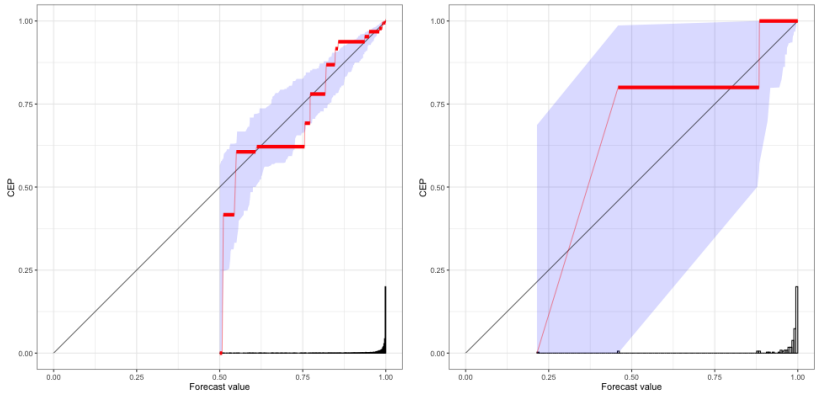
The selection of training sets is important for MT-MAG, since the composition of the training sets affects both *(i)* the parameters of the classifier that is trained on the training set and then used to classify unknown genomes, and *(ii)* the computation of stopping thresholds used to stop classifications once they become uncertain. A tool that can be used to determine whether the training sets are well selected and that the MT-MAG classification confidences are reliable estimates are the so-called *reliability diagrams*, [29].

Reliability diagrams plot the observed frequency of an event against the predicted probability of that event (see Figure 5 for an example). Here, an event is a match between the ground truth and the class assignment predicted by the classifier under consideration. In a reliability diagram, the step-wise lines (red in Figure 5) represent the reliability curves, that is, the observed frequency of the event against the predicted probability of that event. The diagonal indicates a perfect match between the observed frequency and predicted probability, and a curve that is closer to the diagonal indicates a better selected training set and a more reliable classifier. The shaded band (purple in Figure 5) is called the *confidence band*, and a narrower confidence band also indicates a better selected training set and a higher reliability of the classifier. The bar graph (bottom right in Figure 5) is the histogram of predicted probabilities. Usually, a reliability diagram of a classification is accompanied by a *reliability score*, which is a number greater or equal to zero (the smaller the number, the better the reliability, and 0 is the best reliability score).

In the case of MT-MAG, the predicted probability is the classification confidence, and a reliability diagram is an indicator of the suitability of the training set and of the reliability of the classification confidences computed by MT-MAG. Consider for example the scenario whereby MT-MAG classifies a training genome to be Domain Bacteria with classification confidence 0.9. If the classification confidence matches the observed frequency, then out of all the training genomes classified by MT-MAG as Domain Bacteria with classification confidence 0.9, we would expect roughly 90% to have their

ground-truth taxa to be Domain Bacteria. In the reliability diagram, this match will be indicated by a point on the reliability curve that falls on or near the diagonal. If the classification confidences do not match the observed frequencies, this will also be observed on the reliability diagram (e.g., by a wide confidence band), and could be an indicator that the training set, imbalanced at some parent-to-child classification (some of the child taxa have too few or too many training sequences compared to the others) [30]. Note that for MT-MAG, we used an extended version of the reliability diagrams [31] to generate the diagrams. See Supplemental Materials for details.

Figure 5 displays the reliability diagrams and reliability scores of two parent-to-child classifications for Task 2 (real/dense): (a) the parent taxon GTDB root into its two child taxa, Domain Bacteria and Domain Archaea, and (b) the parent taxon Family Campylobacteraceae into its five child taxa, Genus Campylobacter, Genus Campylobacter_A, Genus Campylobacter_B, Genus Campylobacter_D, and Genus Campylobacter_E. If we compare the two reliability diagrams we first observe that the reliability curve for the classification of Family Campylobacteraceae deviates more from the diagonal than that for the classification of the GTDB root. Second, we note that the confidence band for the classification of Family Campylobacteraceae is wider that than of the classification the GTDB root. Lastly, we observe that the 0.001 reliability score of the classification of Family Campylobacteraceae is smaller than the 0.005 reliability score of the classification of the GTDB root. These three indicators suggest that the training set of the GTDB root into its Domains is better selected, and its classification confidences are higher, than those of Family Campylobacteraceae into its five genera. Thus, for test/unknown genomes, we can expect more accurate classifications of the GTDB root to its child taxa than from Family Campylobacteraceae to its child taxa.



(a) Reliability diagrams for the classification of the GTDB root into Domains Bacteria and Archaea; Reliability score is 0.001.

(b) Reliability diagram for the classification the Family Campylobacteraceae into genera. Reliability score is 0.005.

Fig 5. Reliability diagrams for the classification of (a) the GTDB root to its child taxa Domain Archaea and Domain Bacteria, and (b) Family Campylobacteraceae to its five child taxa (i.e., Genus Campylobacter, Genus Campylobacter_A, Genus Campylobacter_B, Genus Campylobacter_D, Genus Campylobacter_E). The larger deviation of the reliability curve (red) from the diagonal, the narrower confidence band (purple), and smaller reliability score (0.001 vs. 0.005) all indicate a higher reliability of classification of the GTDB root (left) than that of the classification of Family Campylobacteraceae (right).

Species level comparison of MT-MAG with DeepMicrobes

The quantitative results at all taxonomic ranks benchmarked the performance of MT-MAG at all taxonomic ranks. In this following, we benchmark MT-MAG against the performance of DeepMicrobes at the Species level (the only level at which DeepMicrobes classifies). The performance metrics below are used to assess the quality of DeepMicrobes' classification (the subscript r indicates that these metrics refer to reads, and the exact definitions of the terms used can be found in "Materials: Datasets and task description"):

- CA_r (constrained accuracy): the proportion of correctly classified test reads, to classified test reads.
- AA_r (absolute accuracy): the proportion of correctly classified reads, to all test reads.
- WA_r (weighted accuracy): the weighted proportion of correctly classified test reads (with weight 1) together with the unclassified test reads (with weight 0.5), to all test reads. The weight parameter 0.5 indicates the neutrality assumption.
- UCR_r (unclassified rate): the proportion of unclassified test reads to all test reads.

The three accuracies CA_r, AA_r and WA_r are numbers between 0 and 1, with $CA_r \geq AA_r$, and where higher values indicate better performance. The unclassified rate UCR_r is a number between 0 and 1, where lower values indicate a higher proportion of classified reads, thus indicating a better performance.

The difference between the DeepMicrobes's and MT-MAG's performance metrics is the type of training input. DeepMicrobes uses simulated short reads from every genome as input, and makes classifications for each read. As a consequence, DeepMicrobes can have multiple Species classifications for a genome (one of each read form that genome). In contrast, MT-MAG takes uses genomes as input, and makes a single classification for each genome. For a direct comparison between DeepMicrobes and MT-MAG, the reads used as input for DeepMcrobes were simulated from the same training set and test set of the genomes that were used as input for MT-MAG. Obviously, due to their respective length characteristics, the total number of simulated reads used as input to DeepMicrobes is significantly larger than the total number of genomes used as input to MT-MAG.

Since DeepMicrobes only makes classifications at Species level, to compare its performance with that of MT-MAG, we set the parameter tr (taxonomic rank) to Species in MT-MAG, and proceeded to compare CA_r with CA_g(Species), AA_r with AA_g(Species), WA_r with WA_g(Species), and UCR_r with PCR_g(Species).

The comparison procedure is as follows. Firstly, we check if AA_r and AA_g(Species) are both high ($> 90\%$) guaranteeing good performance of both DeepMicrobes and MT-MAG on the entire test set. If this is the case, we compare AA_r with AA_g(Species), with the higher of the two being the better performing method. Otherwise, we compare CA_r with CA_g(Species) and UCR_r with PCR_g(Species), with the better performing method being the one with higher CA_r/CA_g(Species) but lower UCR_r/PCR_g(Species). Recall that a high CA_r/CA_g(Species) indicates that a high proportion of classified test reads/genomes are correctly classified by the method, and a low UCR_r/PCR_g(Species) indicates that there is a sufficient number of test reads/genomes that were classified. In addition, we compare WA_r with WA_g(Species) to determine which of DeepMicrobes and MT-MAG performs better under the neutrality assumption.

Table 3 summarizes the results of DeepMicrobes's and MT-MAG's constrained accuracies, absolute accuracies, weighted accuracies and rejection/partial classification rates at the Species level, for both Task 1 (simulated/sparse) and Task 2 (real/dense).

For Task 1 (simulated/sparse), since the absolute accuracies for DeepMicrobes and MT-MAG are below 90%, we base our conclusions on the comparison between the constrained accuracies. As seen in Table 3, MT-MAG demonstrates a significantly better performance than DeepMicrobes, having a constrained accuracy 80.54% higher than that of DeepMicrobes, and a partial classification rate 85.97% lower than the unclassified rate of DeepMicrobes. In addition, under the neutrality assumption of a rejection/partial classification, MT-MAG performs better than DeepMicrobes since its weighted accuracy is 23.20% greater than that of DeepMicrobes.

For Task 2 (real/dense), it is also the case that the absolute accuracies for DeepMicrobes and MT-MAG are below 90%, and thus we base our conclusions on the comparison between the constrained accuracies. Also in this scenario, Table 3 shows that MT-MAG demonstrates a significantly better performance than DeepMicrobes, having a constrained accuracy 98.27% higher than that of DeepMicrobes, and a partial classification rate 27.18% lower than the unclassified rate of DeepMicrobes. In addition, under the neutrality assumption of a rejection/partial classification, MT-MAG performs better than DeepMicrobes since its weighted accuracy is 51.27% greater than that of DeepMicrobes.

In summary, MT-MAG demonstrates a significantly better performance on both Task 1 (simulated/sparse) and Task 2 (real/dense), by all metrics.

Table 3. Summary of MT-MAG and DeepMicrobes accuracy statistics and partial classification/unclassified rates at the Species level (genome inputs in the case of MT-MAG, and reads inputs in the case of DeepMicrobes). Since for both tasks the absolute accuracies (AA_r and AA_g(Species)) are less than 90%, we interpret the results based on the comparisons on the constrained accuracies (CA_r/CA_g(Species)) and partial/rejection rates (UCR_r/PCR_g(Species)), with a better performance indicated by the method with a higher CA_r/CA_g but lower UCR_r/PCR_r. Additionally, when comparing weighted accuracies (WA_r/WA_g(Species)), a higher value indicates better performance under the neutrality assumptions. MT-MAG demonstrates significantly better performance than DeepMicrobes, by all metrics.

Task ID	**Metric**	**DeepMicrobes(%)**	**MT-MAG(%)**
Task 1 (simulated/sparse)	CA_r/CA_g(Species)	0.00	80.54
	AA_r/AA_g(Species)	0.00	61.18
	WA_r/WA_g(Species)	50.00	73.20
	UCR_r/PCR_g(Species)	100.00	14.03
Task 2 (real/dense)	CA_r/CA_g(Species)	0.36	98.63
	AA_r/AA_g(Species)	0.13	68.82
	WA_r/WA_g(Species)	32.66	83.93
	UCR_r/PCR_g(Species)	65.07	30.22

Discussion

We introduced MT-MAG, a novel machine learning-based taxonomic assignment tool for metagenome-assembled genomes. This is, to the best of our knowledge, the first machine learning method for taxonomic assignment of metagenomic data that has a partial classification option, and outputs a partial classification path in cases where it does not have enough confidence for a complete classification to the lowest taxonomic rank. In addition, MT-MAG outputs interpretable numerical classification confidences of its classifications, at each taxonomic rank. Reliability diagrams confirmed the quality of the training sets and the overall reliability of the MT-MAG classification confidences. Compared with DeepMicrobes (the only other machine learning tool for taxonomic

assignment of metagenomic data, with confidence scores), MT-MAG obtains a significantly better performance, at the only comparable taxonomic rank for DeepMicrobes (i.e. Species), by an average of 33% in weighted accuracy, and 89% in constrained accuracy. In addition, MT-MAG is able to (completely) classify 48% more sequences than DeepMicrobes.

Limitations of MT-MAG include the fact that, being a supervised machine learning classification algorithm, its performance relies on the availability of ground-truth taxonomic labels for the DNA sequences in the training set. In addition, any incorrect or unstable ground-truth labels in the training set may cause erroneous future classifications. This limitation could be addressed, e.g., by extending our supervised machine learning approach to semi-supervised machine learning (where some, but not all, information about the training set is available), or even to unsupervised machine learning (where the training process does not require any ground-truth taxonomic labels, see, e.g., [32]).

Second, even though MT-MAG significantly outperforms DeepMicrobes in Task 1 (simulated test set, and sparse training set) and Task 2 (real test set, and dense training set), there is still room for improvement in accuracies and partial classification rates. An analysis of Task 1's (simulated/sparse) training set suggests two possible reasons contributing to incorrect classifications. One reason is the fact that the training set was the HGR database, which constitutes a very small subset of the GTDB taxonomy, in terms of both the number of representative genomes and of coverage of the GTDB taxonomy. This could be addressed by requiring a specific level of coverage for known taxa, to ensure that feature characteristics are reasonably well-represented. Another reason is the fact that, due to computational requirements of MLDSP, the training set had to exclude any contigs shorter than 5,000 bp, and this selection process resulted in the removal of 93% of the available basepairs. This could be addressed by finding ways to relax the selection criteria for the training set, to allow more sequences to participate in the training process without compromising the classification performance.

Third, the interpretability of classification could be further enhanced by exploring the last layer of the classifier. For example, the process of computing classification confidences could be used to derive information regarding which pairs of child taxa are difficult to distinguish from each other, which could potentially be biologically relevant.

Lastly, even though MT-MAG achieves superior performance on the datasets analyzed in this paper, it would be desirable to obtain mathematical proofs of the optimality of the classifier, such as the Bayes optimality proofs in [33].

Acknowledgments

We thank DeepMicrobes's author Qiaoxing Liang for the assistance on DeepMicrobes, Dr. Gurjit Randhawa for the guidance on MLDSP, and Fatemeh Alipour for replicating the benchmark results.

References

1. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Molecular Cell. 2015;58(4):586-97. doi:10.1016/j.molcel.2015.05.004.

2. Sharon I, Banfield JF. Genomes from metagenomics. Science. 2013;342(6162):1057-8. doi:10.1126/science.1247023.

3. Murali A, Bhargava A, Wright ES. IDTAXA: A novel approach for accurate taxonomic assignment of microbiome sequences. Microbiome. 2018;6(1):1-4. doi:10.1186/s40168-018-0521-5.

4. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology. 2017;2(11):1533-42. doi:10.1038/s41564-017-0012-7.

5. Frioux C, Singh D, Korcsmaros T, Hildebrand F. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. Computational and Structural Biotechnology Journal. 2020;18:1722-34. doi:10.1016/j.csbj.2020.06.028.

6. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research. 2015;25(7):1043-55. doi:10.1101/gr.186072.114.

7. Mock F, Kretschmer F, Kriese A, Böcker S, Marz M. BERTax: taxonomic assignment of DNA sequences with deep neural networks. BioRxiv. 2021. doi:10.1101/2021.07.09.451778.

8. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019;36(6):1925–1927. doi:10.1093/bioinformatics/btz848.

9. Eisenhofer R, Weyrich LS. Assessing alignment-based taxonomic classification of ancient microbial DNA. PeerJ. 2019;7:e6594. doi:10.7717/peerj.6594.

10. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology. 2017;18(1):1-7. doi:10.1186/s13059-017-1319-7.

11. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods. 2012;9(8):811-4. doi:10.1038/nmeth.2066.

12. Zhou F, Olman V, Xu Y. Barcodes for genomes and applications. BMC Bioinformatics. 2008;9(1):1-1. doi:10.1186/1471-2105-9-546.

13. Liang Q, Bible PW, Liu Y, Zou B, Wei L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics and Bioinformatics. 2020;2(1):lqaa009. doi:10.1093/nargab/lqaa009.

14. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16(1):1-3. doi:10.1186/s12864-015-1419-2.

15. Randhawa GS, Hill KA, Kari L. ML-DSP: machine learning with digital signal processing for ultrafast, accurate, and scalable genome classification at all taxonomic ranks. BMC Genomics. 2019;20(1):267. doi:10.1186/s12864-019-5571-y.

16. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. A new genomic blueprint of the human gut microbiota. Nature. 2019;568(7753):499-504. doi:10.1038/s41586-019-0965-1.

17. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, Liachko I, Snelling TJ, Dewhurst RJ, Walker AW, Roehe R. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nature Communications. 2018;9(1):1-1. doi:10.1038/s41467-018-03317-6.

18. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15(3):1-2. doi:10.1186/gb-2014-15-3-r46.

19. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Research. 2016;26(12):1721-9. doi:10.1101/gr.210641.116.

20. Ounit R, Lonardi S. Higher classification sensitivity of short metagenomic reads with CLARK-S. Bioinformatics. 2016;32(24):3823-5. doi:10.1093/bioinformatics/btw542.

21. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic assignment for metagenomics with Kaiju. Nature Communications. 2016;7(1):1-9. doi:10.1038/ncomms11257.

22. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Computational Biology. 2016;12(6):e1004957. doi:10.1371/journal.pcbi.1004957.

23. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593-4. doi:0.1093/bioinformatics/btr708.

24. Murray AE, Freudenstein J, Gribaldo S, Hatzenpichler R, Hugenholtz P, Kämpfer P, Konstantinidis KT, Lane CE, Papke RT, Parks DH, Rossello-Mora R. Roadmap for naming uncultivated Archaea and Bacteria. Nature Microbiology. 2020;5(8):987-94. doi:10.1038/s41564-020-0733-x.

25. Federhen S. The NCBI taxonomy database. Nucleic Acids Research. 2012;40(D1):D136-43. doi:10.1093/nar/gkr1178.

26. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology. 2018;36(10):996-1004. doi:10.1038/nbt.4229.

27. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete Domain-to-Species taxonomy for Bacteria and Archaea. Nature Biotechnology. 2020;38(9):1079-86. doi:10.1038/s41587-020-0501-8.

28. Babbar R, Partalas I, Gaussier E, Amini MR. On flat versus hierarchical classification in large-scale taxonomies. In: Annual Conference on Neural Information Processing Systems; 2013; p. 1824-1832.

29. Hartmann HC, Pagano TC, Sorooshian S, Bales R. Confidence builders: evaluating seasonal climate forecasts from user perspectives. Bulletin of the American Meteorological Society. 2002;83(5):683-98. doi:10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2.

30. Wallace BC, Dahabreh IJ. Class probability estimates are unreliable for imbalanced data (and how to fix them). In: IEEE International Conference on Data Mining; 2012; p. 695-704. doi:10.1109/ICDM.2012.115.

31. Dimitriadis T, Gneiting T, Jordan AI. Stable reliability diagrams for probabilistic classifiers. Proceedings of the National Academy of Sciences; 2021;118(8). doi:10.1073/pnas.2016191118.

32. Arias PM, Alipour F, Hill K, Kari L. (in press). DeLUCS: deep learning for unsupervised classification of DNA sequences. PLoS ONE. 2021. doi:10.1101/2021.05.13.444008.

33. Ramaswamy H, Tewari A, Agarwal S. Convex calibrated surrogates for hierarchical classification. In: Proceedings of Machine Learning Research; 2015; p. 1852-1860.