MetAnoDe: **Met**abarcoding **Ano**maly **De**tection using an LSTM-CNN deep neural network ensemble

Alexander Keller[1,2,§]

[1]Cellular and Organismic Networks, Faculty of Biology, Ludwig-Maximilians-Universität Munich, 82152 Planegg-Martinsried, Germany; keller@bio.lmu.de

[2]School of Biology and Environmental Science, Queensland University of Technology, Brisbane, QLD, Australia

[§]Corresponding author: Alexander Keller, keller@bio.lmu.de

Data Accessibility and Benefit-Sharing

Data Accessibility Statement

All code, documentation, pre-trained models, and example workflows used in this study are available from the *MetAnoDe* GitHub repository (https://github.com/chiras/*MetAnoDe*). The release used in this study (commit ca2fb57), including code, documentation and pre-trained models, is archived in Zenodo (https://doi.org/10.5281/zenodo.20313724). The mock-community benchmark, including scripts, configuration files, simulated data, intermediate results and a description of all parameters, is archived separately in Zenodo (https://doi.org/10.5281/zenodo.23918277). The GitHub repository will remain available for software development and future releases.

Reference sequences were obtained from public databases *SILVA* (Quast et al., 2013), *UNITE* (Abarenkov et al., 2025), BOLD (Ratnasingham & Hebert, 2007), *GenBank* (Benson et al., 2018) and Quaresma et al., (2024). Versions and download dates are documented in the archives. Previously published metabarcoding datasets were used only to calibrate the mock simulation. (Elliott et al., 2021; Gaube et al., 2021; König et al., 2022; F. Leonhardt et al., 2023; S. D. Leonhardt et al., 2022; Martins et al., 2023; Mayr et al., 2021; Metz et al., 2026; Peters et al., 2022; Ronchetti et al., 2022; Staab et al., 2025; Too et al., 2018; Vaudo et al., 2020; Villagómez et al., 2024; Vogel et al., 2023; Voulgari-Kokota et al., 2019; Weinhold et al., 2024; Wilson et al., 2021). Their data are available from the repositories reported in the original publications.

Benefit-Sharing Statement

No new biological material was collected for this study.

Funding: This study was supported by the Excellence Fonds of Ludwig-Maximilians-Universität München (grant number VIII.2/bk 865104-8).

Conflict of Interest: The author declares no conflict of interest.

Ethics Approval: No new biological material, animals, human participants, or human subject data were collected for this study. The study used previously published and publicly available sequence datasets only; therefore, no additional ethics approval was required.

Patient Consent: Not applicable.

Permission to Reproduce Material from Other Sources: No third-party copyrighted figures, tables, or other material requiring permission were reproduced.

Clinical Trial Registration: Not applicable.

Abstract

Amplicon metabarcoding has emerged as a critical tool in ecology and other scientific disciplines, facilitating species identification in mixed samples for biodiversity monitoring, community and microbiome analysis, dietary studies, and understanding species interactions. However, challenges arise from errors and artifacts introduced during sampling and laboratory processes such as polymerase chain reaction (PCR) and sequencing. Manual inspection is impractical due to the vast number of sequences, necessitating rapid algorithms for data cleanup. Thorough bioinformatic processing can reduce such errors through removal of low-quality or non-target sequences using quality-, abundance-, and alignment-based approaches. However, in practice, some anomalous sequences evade detection, while valid sequences may also be incorrectly removed.

Deep neural networks offer a promising complementary solution to alignment-based methods by recognizing complex DNA sequence patterns. This study introduces *MetAnoDe* (**Met**abarcoding **Ano**maly **De**tection), an alignment-free tool that classifies sequences of a defined amplicon as valid, anomalous (accumulated substitutions or indels, chimeras) or off-target, using a convolutional neural network (CNN) and a recurrent long short-term memory (LSTM) network combined into one model. Pre-trained models are provided for 13 markers across bacteria, fungi, plants and animals, and models for further markers can be trained with the same workflow. To evaluate MetAnoDe independently of its training data, I simulated mock communities for all markers, including pseudogene-like divergent copies, chimeras and off-target sequences, and processed them with a standard workflow, so that the origin of every resulting sequence was known. Added after *de novo* chimera removal and taxonomic filtering, *MetAnoDe* reduced the proportion of retained anomalous sequences from 22.5-58.1% to 10.5-25.1%, while additionally removing 0.2 to 3.6% of valid sequences. It detected most divergent copies and many off-target sequences missed by taxonomic filtering, whereas chimeras were detected mainly already by *de novo* chimera removal. MetAnoDe is therefore suited as an additional quality-control layer after standard processing.

Keywords: deep learning, anomaly detection, amplicon metabarcoding, pseudogenes, quality control, outlier detection, convolutional neural network, long short-term memory, recurrent neural networks, mock communities

Introduction

Amplicon-based metabarcoding has emerged as a vital tool in ecology and other scientific disciplines for assessing biodiversity (Cristescu, 2014; Hajibabaei, 2012; Taberlet et al., 2012). It is a molecular biology technique used to identify and quantify species within a complex sample based on DNA sequences (Deiner et al., 2017). It involves amplifying and sequencing a specific region of DNA, typically a standardised marker region (such as portions of the 16S ribosomal DNA for bacteria or the internal transcribed spacer 2 (ITS2) of the nuclear ribosomal DNA for plants), from a mixed sample containing DNA from multiple taxa (Keller et al., 2015; Kozich et al., 2013). Metabarcoding is particularly useful in monitoring ancient and current biodiversity in environmental samples (like soil or water) (Deiner et al., 2017; Pedersen et al., 2015), assessing microbiomes (Kozich et al., 2013; Schmidt et al., 2013; Shanahan et al., 2021), studying diet composition through food or fecal analysis (Pompanon et al., 2012), ecological interactions (Bell et al., 2023) and various other ecological and biological applications. In general, metabarcoding follows a systematic workflow that begins with field sampling, followed by laboratory procedures, including amplicon library preparation and sequencing, resulting in raw amplicon sequencing reads (Zinger et al., 2019). Subsequently, bioinformatic analyses are performed to bring such data into a meaningful format that allows scientific interpretation (Coissac et al., 2012; Zinger et al., 2019).

The strategy employed in bioinformatics processing is crucial for obtaining reliable results (Coissac et al., 2012; Hakimzadeh et al., 2024; Zinger et al., 2019). Despite substantial progress in developing pipelines to transform raw sequencing data into meaningful biological information, significant challenges remain that can introduce biases and affect estimates of diversity, community composition, and species interactions (Coissac et al., 2012; Zinger et al., 2019). Sources of erroneous amplicon reads include sequencing and polymerase chain reaction (PCR) errors, PCR chimeras (formed when incompletely extended products prime on templates of other taxa in subsequent PCR cycles), off-target and random PCR products, primer hybrids, and other less well understood issues (Zinger et al., 2019). Their consequences are particularly pronounced in metabarcoding, because each amplified molecule is sequenced individually. Erroneous products, including chimeras between templates of different taxa, thus appear as separate sequence variants instead of being absorbed into a consensus sequence, as in barcoding of single specimens. While laboratory procedures, such as selecting low-error sequencing technologies or using proofreading polymerases, can partly mitigate issues, they cannot entirely eliminate them. Current efforts to clean up raw data thus involve quality filtering or trimming, amplicon sequence variant (ASV) denoising or operational taxonomic unit (OTU) clustering, as well as the removal of singletons and chimeras (Hakimzadeh et al., 2024). These methods are primarily based on sequence quality statistics, abundance patterns, sequence alignments, or combinations thereof. They identify anomalous reads using sequencing metrics, inherent sequence properties, or comparisons against reference databases. Although these approaches substantially improve data quality, residual anomalous sequences may still persist in final datasets, while valid sequences can also be incorrectly removed due to rare occurrence, misclustering, or limitations of reference databases (Zinger et al., 2019).

In an ideal workflow, each read would undergo manual inspection to determine its trustworthiness. However, this process is extremely time-consuming and requires considerable expertise (Coissac et al., 2012). Manual inspection of a single sequence can

take minutes to hours, depending on researchers' decisions, and is not always reproducible or consistent within and across studies. Considering that metabarcoding studies often involve millions of reads, complete manual verification becomes impractical (Bálint et al., 2016). While the aforementioned tools can reduce this number to few thousands up to hundreds of thousands of representative sequences to consider (Bálint et al., 2016), the volume remains daunting. Consequently, many studies either omit manual verification, pragmatically inspect only the most abundant reads, or apply abundance-based filtering under the assumption that erroneous sequences are mostly rare. Although such filtering can remove many artifacts, it may also discard valid low-abundance taxa and retain abundant artifacts derived from PCR, sequencing, or off-target amplification. This underscores the need for new algorithms that can rapidly identify anomalous sequences using approaches distinct from existing pipeline steps, thereby detecting sequences that current methods may miss or wrongly filter.

Deep neural networks are advanced machine learning models composed of multiple layers of interconnected nodes, designed to recognize patterns and make decisions from large datasets (Christin et al., 2019). The biological target amplicon in metabarcoding is usually well-defined by the primer sequences used during PCR, as well as by its overall layout due to evolutionary constraints (Valentini et al., 2009), even when sequence patterns are very complex, divergent between markers and taxonomic groups, and therefore may not be obvious. Such complex patterns are predestined to be modelled using DNNs. DNNs have already been applied to the taxonomic classification of DNA (meta-)barcode sequences (Yang et al., 2022; Ye et al., 2026), but to my knowledge, not yet for data cleanup procedures. The aim of this study is to enhance metabarcoding data cleanup pipelines by automatically identifying sequences consistent with valid target amplicons and separating them from anomalous sequences, whatever their origin, rather than explaining every individual artifact.

For this purpose, I introduce the new software *MetAnoDe* (Metabarcoding Anomaly Detection), a deep neural network framework that reads sequences nucleotide by nucleotide and combines two complementary architectures (Minar & Naher, 2018; Mohammed & Kora, 2023). The first is a recurrent neural network (RNN), specifically a long short-term memory (LSTM) model, designed to capture sequential patterns along the sequence (Shiri et al., 2023; Yu et al., 2019). The second is a convolutional neural network (CNN), which captures spatial sequence characteristics as sequence motifs and their positions (Alzubaidi et al., 2021). Both architectures were newly designed and implemented specifically for amplicon sequences and combined into one classifier.

Pre-trained models were developed for 13 widely used markers: four regions of bacterial 16S (V3-V4, V4, V4-V5, V5-V7), fungal ITS1 and ITS2, plant ITS2 and the chloroplast gene for the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (rbcL), three variants of the 5’ region of the mitochondrial cytochrome c oxidase subunit I gene (COI-5P) for invertebrates, metazoans and vertebrates, and the mitochondrial 12S and 16S rRNA genes of vertebrates. The software is also applicable to further markers through automated training with user-provided reference sequences. To evaluate the models independently of their training data, I simulated mock communities for all markers and processed them with a standard metabarcoding workflow, so that the origin of every resulting sequence was known and the contribution of each workflow step, and of *MetAnoDe*, could be traced for each type of anomaly.

This tool is intended to complement existing alignment-, abundance-, and quality-based approaches by using machine learning predictions to detect, classify, and optionally remove anomalous sequences (Hakimzadeh et al., 2024). It does not replace established steps such as denoising, *de novo* chimera removal or taxonomic classification, but adds an assessment of whether each sequence is consistent with the marker region, independent of reference databases at the time of prediction. Ultimately, the resulting classifications can support faster and more reliable assessment of sequence trustworthiness in large datasets, thereby improving the robustness and quality of downstream metabarcoding analyses.

Material and Methods

Conceptual framework

MetAnoDe classifies each unique sequence of a defined amplicon region (e.g. dereplicated reads, ASVs or OTUs) into one of several classes. One class represents valid target sequences. The other classes represent anomalies, each standing for a group of technical and biological sources:

1. S*equences with accumulated substitutions* arise from polymerase errors during PCR and base-calling errors during sequencing, which on Illumina platforms are dominated by substitutions (Schirmer et al., 2015). Biologically, similar patterns result from divergent non-functional copies of the marker, such as nuclear mitochondrial pseudogenes (NUMTs) or non-functional ribosomal DNA copies (Porter & Hajibabaei, 2021). Different mechanisms of biological and technical substitutions also accumulate.
2. *Sequences with accumulated insertions and deletions* arise technically from polymerase slippage, particularly in homopolymer and repeat regions, and from sequencing errors, and biologically from pseudogenes, in which indels are no longer constrained by function (Keller et al., 2009).
3. *Chimeras* form during PCR when incompletely extended products prime on templates of other taxa in subsequent cycles.
4. *Off-target sequences* result from primer binding to non-target templates present in the same sample, such as organellar DNA or DNA of co-occurring non-target organisms (Zinger et al., 2019). In addition, primers, particularly highly degenerate ones, can bind to similar sites in unrelated genomic regions of any organism in the sample, producing non-specific amplicons that do not derive from the marker gene. Known off-target groups are modelled as separate classes, whereas non-specific amplicons are expected to deviate strongly from the learned structure of the target region and thus to be assigned to one of the anomaly classes.

A separate model is trained for each marker. Each model consists of a long short-term memory network (LSTM) and a convolutional neural network (CNN), combined into one ensemble classifier, hereafter the *MetAnoDe* model. *MetAnoDe* does not assign taxonomy within the target group; this remains the task of taxonomic classifiers applied to sequences retained as valid.

Markers and reference data

Pre-trained models were developed for 13 markers covering bacteria, fungi, plants and animals (Table 1). For bacterial 16S, four commonly used regions were included (V3-V4, V4,

V4-V5 and V5-V7), each defined by its primer pair. Their target and off-target sequences were taken from the SILVA small-subunit rRNA reference database, non-redundant at 99% identity (SSU Ref NR99), release 144 (Quast et al., 2013), and assigned by the *SILVA* taxonomy to the target group (Bacteria and Archaea) and to the off-target groups (mitochondria, plastids and, for V4-V5, whose forward primer also binds eukaryotic 18S, eukaryotes). For the other markers, target sequences were obtained from curated databases where these were available for the marker (Abarenkov et al., 2025; Quaresma et al., 2024; Ratnasingham & Hebert, 2007) and otherwise from *GenBank* (Benson et al., 2018). Off-target sequences were obtained from the same sources. All sequences were reduced to the amplicon defined by the primer pair: a sequence was retained only if it contained both primer sites with at most two mismatches each (primer sequences in Table S1), ensuring that every training sequence covered the complete amplicon with the same boundaries as real data after primer removal. Sequences with ambiguous nucleotides, sequences outside the expected length range and sequences not belonging to the respective group were removed, and identical sequences were kept only once. For fungal ITS, amplicons from *GenBank* were complemented with *UNITE* sequences (Abarenkov et al., 2025).

Intensively sequenced taxa are strongly overrepresented in public databases. To retain taxonomic breadth while limiting this bias, sequences were selected in a taxonomically balanced way. For sequences from *GenBank* and *SILVA*, at most five distinct sequence variants were retained per species, and at most 45 per genus (for *SILVA*, per deepest named taxon) for records not identified to species. For the larger COI dataset from the Barcode of Life Data System (BOLD), one sequence was retained per species, and up to three per genus or family for records identified only to these levels. ITS2 and *UNITE* were deduplicated after download as the original datasets already impose similar restrictions: The plant ITS2 data already comprise up to nine sequences per species (Quaresma et al., 2024), and *UNITE* sequences are represented by one sequence per species.

Table 1. *Markers, primers and training data of the pre-trained MetAnoDe models. For each marker, the primer pair defines the amplicon; all training sequences were trimmed to this amplicon and span it completely. The target class contains the valid sequences of the marker. Off-target classes contain known non-target groups amplified by the same primers, each forming a separate class. Numbers are unique sequences in the training sets after taxonomically balanced selection and before generation of the anomaly classes. All classes were brought to equal size for training. Primer sequences are given in Table S1.*

Marker	**Primers (forward / reverse)**	**Amplicon length (median, bp)**	**Target class: source, sequences**	**Off-target classes (sequences)**
Bacterial 16S-V3-V4	341F / 785R (Klindworth et al., 2013)	422	*SILVA* 144; 133,858	mitochondria (56), plastids (788)
Bacterial 16S-V4	515F / 806R (Caporaso et al., 2011)	253	*SILVA* 144; 113,184	mitochondria (128), plastids (346)
Bacterial 16S-V4-V5	515F-Y / 926R (Parada et al., 2016)	373	*SILVA* 144; 126,677	mitochondria (68), plastids (611), eukaryotes (10,477)
Bacterial 16S-V5-V7	799F / 1193R (Bodenhausen et al., 2013; Chelius & Triplett, 2001)	377	*SILVA* 144; 120,491	mitochondria (51), plastids (57)

Fungal ITS1	ITS1F / ITS2 (Gardes & Bruns, 1993; White et al., 1990)	260	*GenBank* + *UNITE*; 99,426	plants (2,632)
Fungal ITS2	fITS7 / ITS4 (Ihrmark et al., 2012; White et al., 1990)	278	*GenBank* + *UNITE*; 156,235	plants (21,748)
Plant ITS2	ITS-S2F / ITS4R (Sickel et al., 2015)	333	Quaresma et al. (2024); 197,573	fungi (31,050)
Plant rbcL	F52 / R193 (Little, 2014)	140	*GenBank*; 8,597	bryophytes (1,015), green algae (123)
Invertebrate COI-5P	mlCOIintF / jgHCO2198 (Leray et al., 2013)	313	BOLD; 252,710	vertebrates (20,182), plants (2,295), protists (425), fungi (369)
Metazoan COI-5P	mlCOIintF / jgHCO2198 (Leray et al., 2013)	313	BOLD; 272,892	plants (2,295), protists (425), fungi (369)
Vertebrate COI-5P	mlCOIintF / jgHCO2198 (Leray et al., 2013)	313	BOLD; 20,182	invertebrates (20,182), plants (2,295), protists (425), fungi (369)
Vertebrate 12S	V12S-U (Wang et al., 2023)	160	*GenBank*; 17,415	invertebrates (2,038), bacteria (5)
Vertebrate 16S	V16S-U (Wang et al., 2023)	208	*GenBank*; 48,450	invertebrates (3,407)

Construction of training classes

To create balanced datasets for training and validation, the following sequence classes were generated for each marker (Figure 1). The same workflow is also automatically applied when training models for additional user-provided metabarcoding markers (see below).

1. *Class 0 sequences (true target amplicons):*
   All remaining sequences from the database were labelled as true biological target sequences. To account for natural variation not fully represented within the reference databases, additional sequences containing random low-frequency substitutions (probability of 0.5% per position) or insertion-deletion events (indels; 0.5% per position) were generated from the original reference sequences. Low-frequency variants of reference sequences represent the variation of closely related taxa not covered by reference databases, intra-specific variation or minor sequencing errors that usually do no affect downstream analyses.
2. *Class 1 and 2 sequences (simulated errors):* Sequences with high substitution rates (Class 1) and high indel rates (Class 2) were generated in equal amounts to the Class 0 sequences. To represent a range of error intensities rather than a single level, the error probability was drawn separately for each simulated sequence from a uniform distribution. For substitutions, the probability per position ranged from 6.7% to 20%. Because a random replacement can reproduce the original nucleotide, this corresponds to about 5 to 15% realised sequence divergence as confirmed for the simulated sequences of the pre-trained models. For indels, the insertion and

deletion probabilities per position each ranged from 2.5% to 5%. These rates define the classes as accumultions of biological and technical artifacts; they are not intended to reproduce the error rates of particular polymerases or sequencing platforms.

3. *Class 3 sequences (chimeras):* Chimeras were generated by joining the beginning of one randomly chosen reference sequence with the end of another, each contributing between one quarter and three quarters of its length.
4. *Class N sequences (known off-target products):*
   The software additionally supports modelling of user-defined sequence classes (Table 1), allowing inclusion of known off-target amplification products. These sets were balanced to the size of the Class 0 dataset by generating additional sequences with low-frequency substitutions and indels, as described above for Class 0 sequences.

Sequences of the target class outside 0.75 to 1.25 times the median amplicon length were removed; for markers with naturally variable length (ITS, 12S, vertebrate 16S), the range was 0.5 to 1.5 times the median. Chimeras outside 0.5 to 1.5 times the median length were replaced.

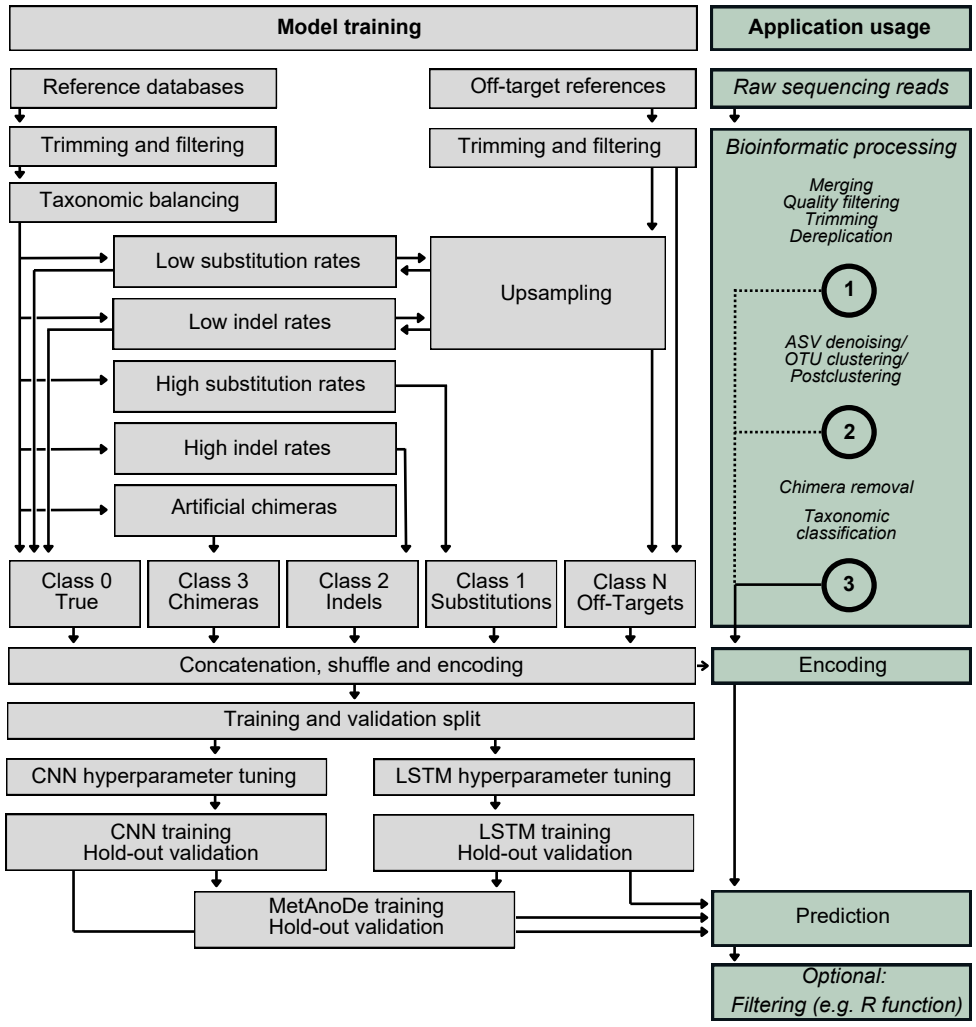


Figure 1: *Workflow of MetAnoDe. Left (light grey): training of a marker-specific model. Reference sequences of the target group and of known off-target groups are reduced to the amplicon defined by the primer pair and selected in a taxonomically balanced way. From these, the training classes are constructed (target, sequences with accumulated substitutions or indels, chimeras, off-target), and a*

CNN and an LSTM are trained and combined into the MetAnoDe model. Right (dark grey): application of a trained model within a standard amplicon metabarcoding pipeline. MetAnoDe classifies the unique sequences after denoising and chimera removal (recommended position, solid arrow). Application at earlier steps is possible (dashed arrows). Sequences classified as valid are passed to taxonomic classification. Predictions can be applied to community data with the provided R *function. Steps outside MetAnoDe are shown in italics.*

Model architectures and training

Sequences are encoded nucleotide by nucleotide (Figure 1). The CNN detects local sequence motifs and their positions, whereas the LSTM reads the sequence position by position and captures dependencies along it (Figure 2). Both start with a learned representation of the nucleotides (embedding), use dropout to reduce overfitting and are trained with a cross-entropy loss. The *MetAnoDe* model combines the class probabilities predicted by the trained CNN and LSTM, whose weights are kept fixed, using a small fully connected network.

The data of each marker were randomly split into 75% training and 25% validation data, with identical sequences never shared between the two. Models were trained with the Adam optimiser for up to 400 epochs. The learning rate was halved when the validation loss did not improve for three epochs (minimum 10^{-5}), and training stopped when the validation accuracy had not improved for ten epochs; the weights of the best epoch were retained. For the LSTM, gradients were clipped to a maximum norm of 0.5 to stabilise training. Hyperparameters were determined by a Hyperband search (Li et al., 2018).

Hold-out validation

Performance on the validation data for each marker was summarised as precision, recall and F1 score per class and overall accuracy, separately for the CNN, the LSTM and the *MetAnoDe* model (Table S2). Because the validation data were used to stop training and to select hyperparameters, and because close relatives of validation sequences occur in the training data, performance on these data is an upper bound. Independent performance was assessed with the mock communities described below.

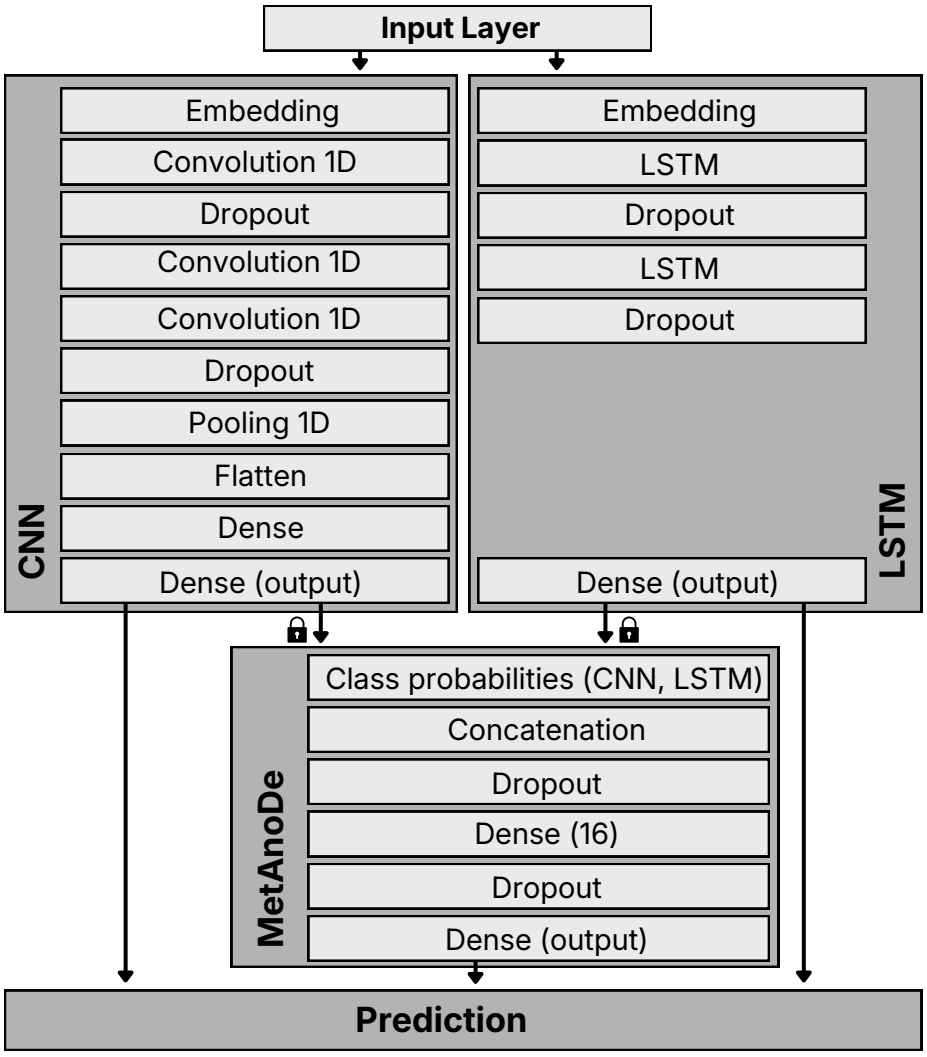


Figure 2: *Architecture of the CNN, the LSTM and the combined MetAnoDe model (implemented in TensorFlow and Keras). Both branches start with a learned representation of the nucleotides (embedding). The CNN detects local sequence motifs and their positions through convolutional layers, and the LSTM captures dependencies along the sequence. In the MetAnoDe model, the class probabilities predicted by the trained CNN and LSTM, whose weights are kept fixed (lock symbols), are concatenated and passed through a small fully connected network, which learns how to combine the two predictions for each class. Predictions of all three models are reported.*

Mock-community benchmark

The hold-out validation tests the models on sequences generated by the same procedure as the training data. To evaluate them on independently generated anomalies in a more realistic metabarcoding workflow, I simulated mock communities for all 13 markers. In contrast to real datasets, the origin of every resulting ASV is known, allowing each *MetAnoDe* prediction to be checked against its true origin and the detection of different anomaly classes to be tracked across workflow steps.

Target and off-target templates were obtained from *GenBank* (accessed 10. October 2026; Benson et al., 2018) and for the bacterial 16S regions from *SILVA* 144 (Quast et al., 2013), and reduced to the amplicon as described for the reference data. To distinguish the recognition of known sequences from generalisation to new ones, about half of the target

templates in each community were identical to training sequences of the model (seen), while the remainder were not (unseen). For each marker, 25 communities were assembled from 50 to 200 target templates with log-normally distributed abundances, and off-target templates contributing 2 to 20% of the community. PCR was simulated with *Grinder* (Angly et al., 2012), with 10% chimeric molecules and a polymerase error rate of 0.1% per base, and sequencing was simulated with *InSilicoSeq* (Gourlé et al., 2019), using the Illumina MiSeq error model and paired-end reads of the length used for the respective marker. Both tools are independent of the simulation used to construct the training classes. To represent pseudogenes and similar divergent copies, which arise before PCR and are therefore amplified like genuine sequences, each community additionally contained modified copies with one to three short insertions or deletions (frameshifts for the protein-coding markers), each at a lower abundance than its parent. To reflect the composition of real sequencing libraries, non-specific amplification products, primer dimers and low-quality reads were added. Their proportions were calibrated so that read-merging rates and the proportion of reads passing quality filtering fell within the interquartile ranges of real datasets processed with the same pipeline for the respective makers (see data availability for details).

Simulated reads were processed with a standard VSEARCH-based pipeline (Leonhardt et al., 2022; Rognes et al., 2016; https://github.com/chiras/metabarcoding_pipeline): read merging, quality filtering, dereplication, denoising, *de novo* chimera removal and taxonomic classification against reference databases restricted to the target group of each marker (Abarenkov et al., 2025; Benson et al., 2018; Quaresma et al., 2024; Quast et al., 2013; Ratnasingham & Hebert, 2007). For the bacterial regions, classification against an earlier SILVA release than the one providing the templates ensured that sequences added in the later release were new to the database, as in real studies. In this reference workflow, ASVs classified as off-target or not classified to family level were removed (taxonomic filter). *MetAnoDe* was applied to the final ASVs after chimera removal. Each ASV was labelled by its known origin and its fate was traced along the workflow (Figure S1): ASVs identical to a template as target or off-target; ASVs derived from a chimeric molecule, according to the simulation record, as chimeras; the modified copies according to their construction; remaining ASVs according to the predominant type of difference from their closest template, and ASVs without a template of at least 50% identity as unrelated.

For the reference workflow and for the reference workflow combined with *MetAnoDe*, I determined the proportion of target ASVs removed (false removal) and the proportion of anomalous ASVs retained, for each anomaly class and as the mean across classes. All data, scripts, parameters and results of the benchmark are archived in Zenodo (https://doi.org/10.5281/zenodo.23918277).

Final model selection

For each marker, several models were trained with different random seeds and, for some markers, alternative training settings. The model released for each marker was selected by its performance in the mock-community benchmark, with hold-out validation and the performance of the CNN and LSTM as secondary criteria (candidates and their values in Table S3). Within a marker, candidates differed by at most one to two percentage points so that the selection has little influence on the reported values.

Downstream filtering

To apply the predictions in community analyses, an R function (*MetAnoDe_filter*) filters phyloseq objects (McMurdie & Holmes, 2013) based on the *MetAnoDe* classes. Filtering thresholds and criteria can be set manually during the function call. By default, sequences of off-target classes are removed regardless of their abundance, because off-target amplicons such as chloroplast sequences are often abundant and widespread. Sequences of the other anomaly classes can be removed directly or only when they are also rare across samples (below user-defined thresholds of prevalence and abundance), thereby reducing the risk of removing misclassified genuine taxa.

Implementation and availability

MetAnoDe is implemented in Python and available at https://github.com/chiras/*MetAnoDe*. It is provided as a Docker image with all package versions fixed (based on the NVIDIA TensorFlow container 25.01; TensorFlow 2.17, Keras 3.3); installation in a Conda environment is supported as an alternative. Prediction requires a query FASTA file and a model name; training of a new marker-specific model requires reference sequences of the target group and, optionally, of off-target groups. By default, all query sequences are retained and annotated with the predicted class and class probabilities from the *MetAnoDe* model and the individual CNN and LSTM. Removal of sequences classified as anomalous is optional. Parameters and usage examples are given in the repository documentation. The models and results reported here were generated with release v1.0 which is archived together with the pre-trained models in Zenodo (https://doi.org/10.5281/zenodo.23916955). The mock-community benchmark including scripts, data and full results is archived separately (https://doi.org/10.5281/zenodo.23918277). Models were trained on Ubuntu 24.04 workstations with graphics processing units (GPUs) by NVIDIA (RTX 4070 Ti Super, 4090 or 5090). Prediction was tested on Ubuntu and macOS with and without GPU support. Mock-community benchmarks were performed on macOS without GPU support.

AI Technology Disclosure

ChatGPT (GPT-5.5, OpenAI) and Claude (Opus 5.5, Anthropic) were used for language editing, grammar correction, and stylistic refinement during manuscript preparation and revision. They were also used for limited readability improvements, for review of code changes and bug fixing of individual software code blocks during software preparation. Claude Code (Opus 5.5, Anthropic) was used for implementing the Mock-community benchmark pipeline. The scientific concepts, study design, analyses, interpretation, manuscript content, and software implementation were developed by the author. No manuscript content or software code beyond the benchmark pipeline was drafted by AI and all AI-assisted edits were reviewed, verified, and approved by the author, who remains fully responsible for the manuscript and software. No AI tool was used to generate or change data, perform analyses, or determine scientific conclusions.

Results

Hold-out validation

Overall, all models achieved high accuracy on the validation data, between 0.96 and 0.99 (Table 2). The precision of the target class, i.e. the proportion of sequences classified as valid that were valid, ranged from 0.90 to 0.97, and its recall from 0.93 to 0.99. For every marker,

the target class was the class most difficult to separate, with the lowest values for markers of naturally variable length (fungal ITS1 and ITS2, vertebrate 12S), whose natural variation overlaps more with the simulated anomaly classes. The *MetAnoDe* model achieved a mean F1 score across classes of 0.960 to 0.995, the highest of the three models for all markers except bacterial 16S V5-V7, where it equalled the CNN (0.988), with small gains over the better of the CNN and LSTM (Table S2).

Table 2: *Performance of the* MetAnoDe *models. Hold-out validation: overall accuracy and precision and recall of the target class on the validation data (an upper bound, see Methods). Values for individual classes and for the CNN and LSTM are given in Table S2. Mock communities: proportion of target ASVs removed and of anomalous ASVs retained by the reference workflow (*de novo *chimera removal and taxonomic filter) and by the reference workflow combined with* MetAnoDe*, based on all ASVs after denoising. The proportion of retained anomalies is the mean across anomaly classes. Bacterial 16S templates were drawn from SILVA 144 and classified against SILVA 138.2. Templates of the other markers were drawn from GenBank.*

Marker	**Hold-out**		**Mock: false removal**		**Mock: anomalous ASVs retained**	
	accuracy	**target precision / recall**	**reference (%)**	**reference with *MetAnoDe* (%)**	**reference (%)**	**reference with *MetAnoDe* (%)**
Bacterial 16S V3-V4	0.99	0.96 / 0.98	17.5	19.1	27.1	16.4
Bacterial 16S V4	0.98	0.95 / 0.96	9.1	11.5	30.9	16.1
Bacterial 16S V4-V5	0.99	0.96 / 0.98	19.1	20.4	22.9	14.3
Bacterial 16S V5-V7	0.99	0.96 / 0.98	16.7	17.7	22.5	15.9
Fungal ITS1	0.96	0.90 / 0.93	3.6	7.2	29.8	12.2
Fungal ITS2	0.97	0.91 / 0.94	3.6	6.3	33.7	19.2
Plant ITS2	0.98	0.94 / 0.97	0.1	1.2	58.1	21.8
Plant rbcL	0.98	0.92 / 0.95	5.5	6.3	31.3	19.3
Invertebrate COI-5P	0.99	0.97 / 0.99	3.1	5.0	30.2	20.4
Metazoan COI-5P	0.99	0.97 / 0.99	2.9	4.7	35.9	25.1
Vertebrate COI-5P	0.99	0.96 / 0.99	1.8	2.0	23.5	13.6
Vertebrate 12S	0.96	0.90 / 0.93	4.8	6.3	31.7	10.5
Vertebrate 16S	0.96	0.92 / 0.94	0.9	1.5	40.6	13.3

Mock community benchmark

The mock communities of all markers were simulated with the same design, but their composition after denoising differed between markers, depending on template pools, amplicon length and the number of chimeras surviving denoising (Figure S1). This yielded 1,169 to 2,578 target ASVs and 435 to 1,131 anomalous ASVs per marker after denoising, about 35,600 ASVs in total over all markers, each with known origin. The reference workflow, i.e. chimera removal and taxonomic filter, removed 0.1 to 19.1% of the target ASVs and retained between 22.5 and 58.1% of the anomalous ASVs (class-balanced mean; Table 2). Adding *MetAnoDe* reduced the proportion of retained anomalies to 10.5 to 25.1%, while it increased the removal of target ASVs by 0.2 to 3.6 percentage points. The largest reductions were found for plant ITS2, vertebrate 16S and vertebrate 12S (by 36.3, 27.3 and

21.2%), whereas the highest additional flagging of target ASVs occurred for the two fungal ITS markers (2.7 and 3.6%).

Among the final ASVs, *MetAnoDe* and the taxonomic filter flagged largely different subsets of anomalies (Figure 3). Depending on the marker, 34 to 70% of the anomalous final ASVs were flagged by both, 10 to 46% by *MetAnoDe* only and 0 to 29% by the taxonomic filter only, while 9 to 25% were flagged by neither. The false removal of the reference workflow depended on amplicon length. For the four bacterial 16S regions, which shared template source and reference database, the taxonomic filter removed 8.9% of the target ASVs for V4 (253 bp) but 16.7 to 19.1% for the longer regions (V3-V4, V4-V5, V5-V7; 373 to 422 bp), because fewer valid sequences reached the identity required for classification over a longer amplicon. *MetAnoDe*, in contrast, flagged 1.3 to 2.5% of the target ASVs in all four regions. On a second benchmark for bacterial 16S-V4 with templates of cultured type strains (NCBI Reference Sequence database, RefSeq), *MetAnoDe* flagged fewer target ASVs (0.2% instead of 2.4%), consistent with the higher false-removal rate for lineages less represented in the training data (Table S3).

Detection by anomaly class

The contribution of *MetAnoDe* differed strongly between anomaly classes. Divergent copies with accumulated substitutions were detected by the reference workflow at 13.6 to 79.5%, depending on the marker (Table S4), compared with 87.6 to 96.3% by the reference workflow combined with *MetAnoDe*. *MetAnoDe* alone detected 86.1 to 96.3% of them. Copies with insertions or deletions were rarely detected by the reference workflow (0 to 26.5%). When *MetAnoDe* was added, detection rose to 14.3 to 76.5%, with the lowest values for COI and the longer bacterial regions.

For the 19 off-target groups of the non-bacterial markers (Table S4), MetAnoDe alone detected 84.0 to 100% of the off-target ASVs, all of them for 11 groups and at least 95% for 16. Its detection was equal to or higher than that of the taxonomic filter for every group, and higher for 15. The difference was largest where the reference database did not cover the off-target group, as for fungal sequences in plant ITS2 (89.7% against 32.8%). The lowest detection occurred for bryophytes in plant rbcL, the off-target group most closely related to the target group (84.0% by MetAnoDe, 85.2% in combination with the taxonomic filter). For the bacterial 16S regions, the taxonomic filter flagged mitochondrial, plastid and eukaryotic sequences at 80.4 to 100%, because these were not classified to family against a database restricted to Bacteria and Archaea. MetAnoDe alone detected mitochondrial and plastid sequences at 79.8 to 98.9% for V3-V4, V4 and V4-V5, raising the combined detection to 99.0 to 100%, but only 26.0 and 40.7% for V5-V7, and eukaryotic sequences in V4-V5 at 17.9%. On the bacterial 16S V4 benchmark with cultured type strains and databases lacking mitochondria, MetAnoDe detected all mitochondrial sequences, against 44.4% for the taxonomic filter (see benchmark archive: https://doi.org/10.5281/zenodo.23918277). Non-specific products unrelated to the marker were removed almost completely in the combined workflow (97.8 to 100%).

Confusion between mitochondria and their closest bacterial relatives was not observed: of 9,804 bacterial ASVs in the four 16S regions, including 32 of Rickettsiales and 935 of other Alphaproteobacteria, none was classified as mitochondrial by MetAnoDe, and none of the Rickettsiales ASVs was flagged as anomalous, whereas the taxonomic filter flagged 18.8% of them (Figure S2).

For chimeras, in contrast, *de novo* chimera removal was decisive (Figure S1). Applied to all denoised sequences, *UCHIME3* (Edgar, 2016) detected 48 to 87% of the chimeric ASVs and *MetAnoDe* 15 to 39%, of which 50 to 93% were also detected by *UCHIME3*. *MetAnoDe* thus detected 5.9 to 37.8% of the chimeras that *UCHIME3* missed, while 8 to 37% were detected by neither method.

The classification of the final ASVs by origin and predicted class is shown in Figure S3. Most misclassifications concerned copies with insertions or deletions classified as valid, particularly for the COI markers and the longer bacterial 16S regions, consistent with their low detection (Table S4).

Seen and unseen templates

To distinguish the recognition of known sequences from generalisation, I compared ASVs derived from templates identical to training sequences (seen) with ASVs derived from templates absent from the training data (unseen; Figure S4). The reference workflow retained anomalies derived from unseen templates far more often than those derived from seen templates (28.7 to 75.9% versus 4.5 to 40.0%), because its taxonomic filter depends on related reference sequences. With *MetAnoDe*, the proportion of retained anomalies derived from unseen templates fell to 11.2 to 18.3%, whereas the reduction was smaller for anomalies derived from seen templates. *MetAnoDe* alone flagged 0.3 to 7.3% of the target ASVs derived from unseen templates, compared with 0.1 to 2.5% of those derived from seen templates, with the highest values for the two fungal ITS markers.

Runtime

Prediction with a trained model took 1 to 5 s for 955 to 7,280 unique sequences on a single workstation GPU, for all three models together (about 1,000 to 2,700 sequences per second). Training a new model took 0.6 to 22.7 h, increasing mainly with the number of reference sequences: from 0.6 h for plant rbcL (8,597 target sequences) to 22.7 h for metazoan COI (272,892).

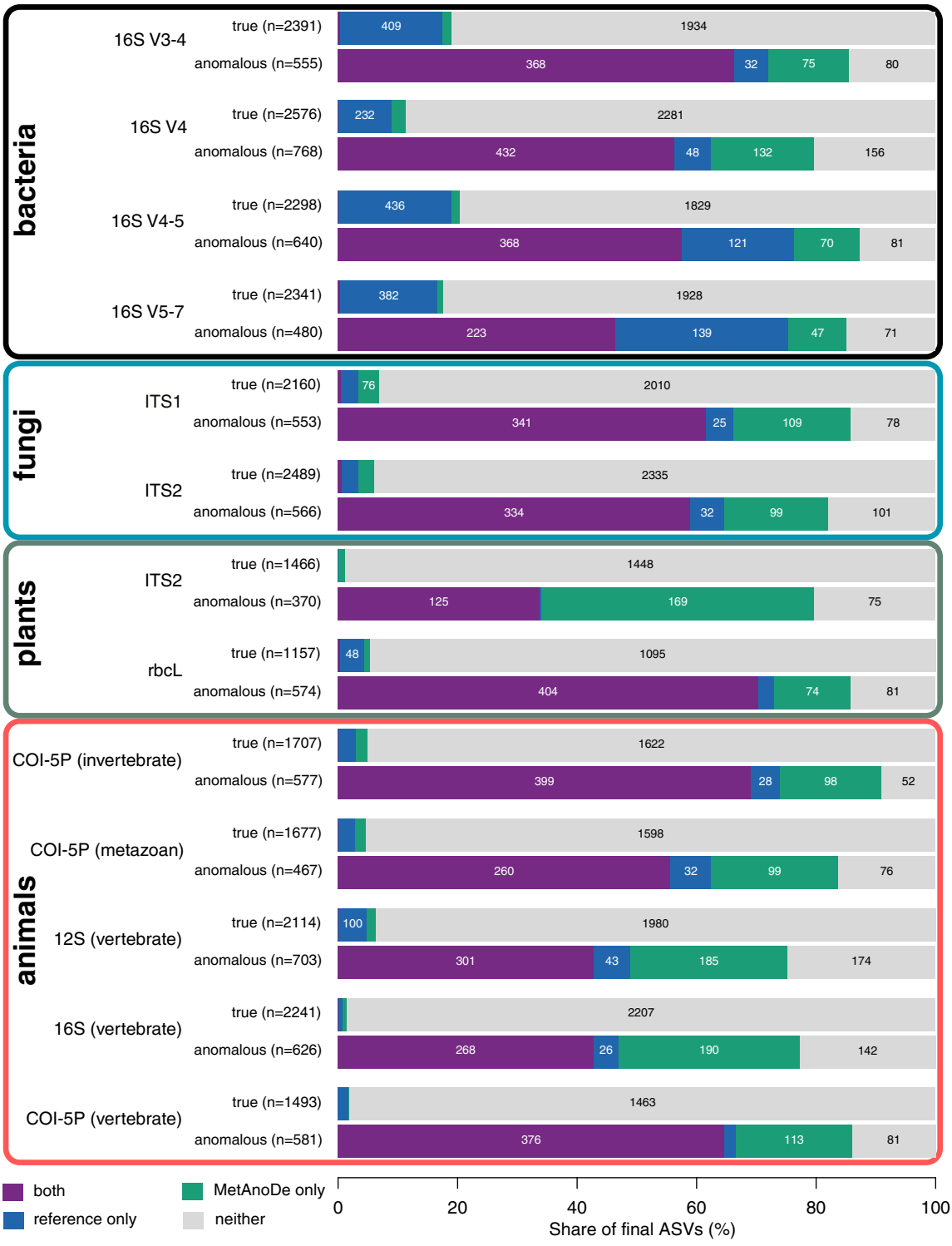


Figure 3: *Final ASVs of the mock communities flagged by the reference filter and by* MetAnoDe*, for each marker, grouped by organism group. For each marker, the upper bar shows the ASVs derived from target templates and the lower bar the ASVs derived from anomalies (n is the number of ASVs). Colours indicate whether an ASV was flagged by both the taxonomic filter and* MetAnoDe*, by the reference filter only, by* MetAnoDe *only, or by neither. For target ASVs, flagging represents a false removal, for anomalous ASVs, a correct detection. Shown are the final ASVs after denoising and* de novo *chimera removal; chimeras removed at that step are not included, the full tracing is available in Figure S1. Numbers in the bars are ASV counts (omitted for small segments).*

Discussion

MetAnoDe as a complementary anomaly detection layer

Here, I introduce *MetAnoDe*, a workflow and software for identifying and classifying anomalous sequences in amplicon metabarcoding data using deep-learning models. The software provides two principal usage pathways. First, pre-trained models can be applied directly to query datasets, allowing rapid prediction without further training. Such models are already provided for 13 markers covering bacteria, fungi, plants and animals (Table 1). Second, if no suitable model exists for a given target region, new models can be trained within the same automated software environment using user-provided reference sequences and optional off-target classes. This design allows *MetAnoDe* to be used both as an immediately applicable tool for established markers and as a flexible framework for extending anomaly detection to additional regions.

Deep-learning models have previously been applied to the taxonomic classification of DNA (meta-)barcode sequences (Badirli et al., 2023; Yang et al., 2022; Ye et al., 2026). These approaches, however, assume that the classified sequences are valid representatives of the target region. To the best of my knowledge, no other workflow or software has been proposed to date that specifically addresses detection of anomalous metabarcoding sequences using deep-learning methods, also considering recent reviews of available metabarcoding software (Hakimzadeh et al., 2024). *MetAnoDe* is alignment-free and does not rely on sequencing quality statistics or on matching against a reference database during prediction. It therefore provides a complementary perspective to existing quality-, abundance-, and alignment-based filtering strategies rather than replacing these steps.

What MetAnoDe detects that reference-based filtering misses

The mock communities allowed tracing every ASV to its origin, and thereby showed which anomalies each workflow step detects. Across markers, *MetAnoDe* reduced the proportion of anomalous ASVs retained by a standard workflow from 23.5 - 58.1% to 10.5 - 25.1%, at the cost of 0.2 to 3.6 percentage points of additionally removed target ASVs. Among the final ASVs, *MetAnoDe* flagged 10 to 46% of the anomalies that the taxonomic filter retained, whereas a taxonomic filter added 0 to 29% to *MetAnoDe*.

The underlying reason lies in the different questions both approaches ask. A alignment-based filter asks whether a sequence resembles a reference sequence of the target group closely enough. *MetAnoDe* asks whether a sequence conforms to the structure of the marker region as represented in the training data. Anomalies that resemble a genuine taxon therefore pass the alignment-based filter but not *MetAnoDe*. This applies above all to divergent copies of genuine sequences, such as pseudogenes or nuclear mitochondrial copies, which arise before PCR, are amplified like functional sequences and survive denoising. The reference workflow detected 13.6 to 79.5% of such copies with accumulated substitutions, *MetAnoDe* 86.1 to 96.3%. The same principle applies to off-target groups missing from the reference database: mitochondrial sequences in a bacterial database and fungal sequences in a plant database were classified as unassigned or as plausible targets by the alignment-based filter, while *MetAnoDe* detected them in 100% and 89.7% of cases.

The four bacterial 16S regions illustrate the same principle. The false removal of the reference workflow doubled from V4 to the longer regions, because the identity required for classification is reached less often over a longer and more variable amplicon, whereas

the false removal added by *MetAnoDe* remained low in all regions. Reference-based filtering thus depends on how well the database represents the community at the resolution of the chosen region, while *MetAnoDe* depends on how well its training data represent the region as a whole.

More generally, errors in metabarcoding data arise mainly at two levels: in the sequences themselves, which *MetAnoDe* addresses, and in their taxonomic assignment. Assignment errors are typically reduced by restricting reference databases to taxa that are plausible for a given region and season (Keller et al., 2020; Quaresma et al., 2024). Quality control at both levels is complementary.

Chimeras and the recommended position in the workflow

For chimeras, the picture is reversed. *De novo* chimera removal detected 62 to 87% of the chimeric ASVs, *MetAnoDe* alone 15 to 39%, and most chimeras detected by *MetAnoDe* were also detected by UCHIME3. Chimeras of closely related parents differ little from genuine sequences of either parent, and abundance-based *de novo* detection, which exploits the presence of both parents in the same sample, is better suited to them than an assessment of the sequence alone. *MetAnoDe* added 2 to 14% of the chimeras missed by UCHIME3 (Edgar, 2016), while 8 to 32% remained undetected by both.

This defines the recommended position of *MetAnoDe*: after denoising and *de novo* chimera removal, as an additional layer on the retained sequences. Applied before chimera removal, it would largely duplicate UCHIME3 while missing most chimeras. Applied before denoising, hard removal of sequences classified as anomalous could alter denoising, because sequences with irregularities may still carry information that denoising algorithms use to reconstruct ASVs from patterns of sequence similarity and abundance. I therefore recommend to first perform the standard workflow of the respective pipeline and then apply *MetAnoDe* to the final sequences.

Within this framework, the predicted classes can be used flexibly, depending on study goals and desired stringency. Flagged sequences can be inspected or removed selectively. The provided *R* function applies the predictions to phyloseq objects: Off-target sequences are removed independent of their abundance by default, because off-target amplicons (e.g. such as chloroplast sequences) are often abundant and widespread. Sequences of the other anomaly classes can be removed either directly or only when they are also rare across samples. The latter option reduces the risk of removing misclassified genuine taxa, but retains artefacts that recur across samples, such as chimeras of abundant parent sequences.

Reuse, scope and marker transferability

Pre-trained models can be applied directly to query data, provided that adapter and primer sequences have been removed and the analysed amplicon matches the model. By default, the software retains all query sequences and adds the predicted classes to the output. In multi-marker studies, sequences of different markers cannot be linked to the same organism within a mixed sample, and each marker is therefore processed separately: after assigning sequences to markers, a step supported by common metabarcoding pipelines (e.g. Leonhardt et al., 2022)), *MetAnoDe* is applied per marker with the corresponding model.

Models for further markers can be trained with the same workflow. This requires unique reference sequences of the target group trimmed to the amplicon and ideally balanced for taxonomic breadth within the target group plus, optionally, known off-target groups. There

is no fixed limit on the number of reference sequences, sequence length or off-target classes, but memory and training time increase with data size (0.6 to 22.7 h on a single GPU for the models presented here), so that a GPU is recommended for training, whereas prediction is fast on systems with central processing units (CPUs) only as well as with GPUs.

The approach requires amplicons of a defined marker region. It is therefore not applicable to data without defined amplicon boundaries, such as shotgun metagenomic (Quince et al., 2017) or target enrichment data (e.g. Angiosperms353; Benkendorf et al., 2025) which consist of randomly fragmented reads with varying boundaries and are subject to different error sources. In principle, *MetAnoDe* could also be applied to high-throughput sequencing of single specimen barcoding data amplified with the same primers, but this has not been tested here, and some anomaly classes, such as chimeras between taxa, would not be expected in such data.

Limitations and future development

Several limitations should be considered when interpreting the results. First, both the training classes and the benchmark rely on simulation. The benchmark used tools and error processes independent of the training data, and its read populations were calibrated against real datasets. But it cannot capture all context-dependent PCR and sequencing errors. Performance was therefore summarised per class and as a class-balanced mean, and should be interpreted per class. In the training classes, substitutions are placed uniformly across positions. Empirically derived error profiles, such as the preference for transitions in polymerase misincorporation, could be incorporated as they become better characterised.

Second, the current models do not explicitly include all possible sources of anomalous sequences. Non-functional ribosomal DNA copies, pseudogenes, primer hybrids, and unusual degradation products may not always fit the trained anomaly classes (Bálint et al., 2016; Porter & Hajibabaei, 2021; Schirmer et al., 2015). Short frameshifts in COI were detected least well, by *MetAnoDe* alone in 12.5 to 20.8% of cases, although still far more often than by the reference workflow (1.8 to 8.3%), indicating that the models do not explicitly exploit the reading frame. A frame-aware representation is a concrete direction for protein-coding markers. Future versions could be extended with open-set or out-of-distribution recognition, as developed to recognise taxa absent from training data in DNA barcoding and image-based identification (Badirli et al., 2023; Fujisawa & Imai, 2026; Impiö & Raitoharju, 2024). Applied to MetAnoDe, such methods would allow flagging sequences that resemble none of the trained classes, and thus dataset-, platform- or run-specific artefacts that are difficult to anticipate.

Third, performance depends on the breadth of the training data and on the resolution of the marker. Genuine sequences of lineages absent from the training data are flagged more often, and as reference databases grow, retraining will extend the taxonomic coverage of the models. Independent of the training data, short and conserved amplicons provide fewer positions at which anomalies and related off-target groups differ from valid sequences. This was most evident for the plant rbcL mini-barcode (about 140 bp; Little, 2014), for which MetAnoDe reduced the proportion of retained anomalies less than for plant ITS2 (by 12.0 against 36.3 percentage points) and separated bryophytes, the off-target group most closely related to the target group, least reliably (84.0%). For bacterial 16S, the corresponding case are mitochondria and their closest bacterial relatives, the Rickettsiales, which include

endosymbionts such as *Wolbachia*. In the benchmark, no true bacterial ASV was classified as mitochondrial (Figure S2), including *Wolbachia* and other Alphaproteobacteria.

Finally, transferability requires a consistent marker definition. Pre-trained models should only be applied to data generated with primers that define the same amplicon after adapter and primer trimming. Datasets generated with different primers or substantially different amplicon boundaries require retraining.

Conclusions

MetAnoDe provides an alignment-free deep-learning based anomaly detection layer for amplicon metabarcoding workflows. It comes with pre-trained models for 13 widely used markers across bacteria, fungi, plants and animals, and a workflow for training models for further markers. In mock communities with known origin of every sequence, *MetAnoDe* reduced the proportion of anomalous sequences retained by a standard workflow from 23.5–58.0% to 10.5–25.1%, while removing 0.2 to 3.6 percentage points more valid sequences. Its contribution was largest for anomalies that reference-based filtering cannot recognise: divergent copies of genuine sequences such as pseudogenes, off-target groups missing from the reference database, and anomalies derived from lineages absent from it. *De novo* chimera removal remains necessary, as chimeras of related parents were detected mainly by UCHIME3. *MetAnoDe* is therefore best applied after the standard workflow, as an additional layer that flags residual anomalous sequences for inspection or selective removal before taxonomic and ecological analyses.

Acknowledgments

This study was supported by the Excellence Fonds of Ludwig-Maximilians-Universität München (grant number VIII.2/bk 865104-8).

References

Abarenkov, K., Zirk, A., Piirmann, T., Pöhönen, R., Ivanov, F., Nilsson, R. H., & Kõljalg, U. (2025). *UNITE general FASTA release for Fungi 2* [Application/gzip]. UNITE Community. https://doi.org/10.15156/BIO/3301230

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., & Tyson, G. W. (2012). Grinder: A versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, *40*(12), e94. https://doi.org/10.1093/nar/gks251

Badirli, S., Picard, C. J., Mohler, G., Richert, F., Akata, Z., & Dundar, M. (2023). Classifying the unknown: Insect identification with deep hierarchical Bayesian learning. *Methods in Ecology and Evolution*, *14*(6), 1515–1530. https://doi.org/10.1111/2041-210X.14104

Bálint, M., Bahram, M., Eren, A. M., Faust, K., Fuhrman, J. A., Lindahl, B., O'Hara, R. B., Öpik, M., Sogin, M. L., Unterseher, M., & Tedersoo, L. (2016). Millions of reads, thousands of taxa: Microbial community structure and associations analyzed via marker genes. *FEMS Microbiology Reviews*, *40*(5), 686–700. https://doi.org/10.1093/femsre/fuw017

Bell, K. L., Turo, K. J., Lowe, A., Nota, K., Keller, A., Encinas-Viso, F., Parducci, L., Richardson, R. T., Leggett, R. M., Brosi, B. J., Burgess, K. S., Suyama, Y., & de Vere, N. (2023). Plants, pollinators and their

interactions under global ecological change: The role of pollen DNA metabarcoding. *Molecular Ecology*, *32*(23), 6345–6362. https://doi.org/10.1111/mec.16689

Benkendorf, R. C., Woodworth, E. J., CaraDonna, P. J., Ogilvie, J. E., Taddeo, S., & Fant, J. B. (2025). Improving plant DNA metabarcoding accuracy with ecological filters and Angiosperms353: Field and pollen microscopy validation. *Applications in Plant Sciences*, *13*(5), e70026. https://doi.org/10.1002/aps3.70026

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. (2018). GenBank. *Nucleic Acids Research*, *46*(D1), D41–D47. https://doi.org/10.1093/nar/gkx1094

Bodenhausen, N., Horton, M. W., & Bergelson, J. (2013). Bacterial Communities Associated with the Leaves and the Roots of Arabidopsis thaliana. *PLOS ONE*, *8*(2), e56329. https://doi.org/10.1371/journal.pone.0056329

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, *108*(supplement_1), 4516–4522. https://doi.org/10.1073/pnas.1000080107

Chelius, M. K., & Triplett, E. W. (2001). The Diversity of Archaea and Bacteria in Association with the Roots of Zea mays L. *Microbial Ecology*, *41*(3), 252–263. https://doi.org/10.1007/s002480000087

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*(10), 1632–1644. https://doi.org/10.1111/2041-210X.13256

Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, *21*(8), 1834–1847. https://doi.org/10.1111/j.1365-294X.2012.05550.x

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, *29*(10), 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Edgar, R. C. (2016). UCHIME2: Improved chimera prediction for amplicon sequencing. *bioRxiv*, 074252. https://doi.org/10.1101/074252

Elliott, B., Wilson, R., Shapcott, A., Keller, A., Newis, R., Cannizzaro, C., Burwell, C., Smith, T., Leonhardt, S. D., Kämper, W., & Wallace, H. M. (2021). Pollen diets and niche overlap of honey bees and native bees in protected areas. *Basic and Applied Ecology*, *50*, 169–180. https://doi.org/10.1016/j.baae.2020.12.002

Fujisawa, T., & Imai, T. (2026). Performance and Limitations of Out-Of-Distribution Detection for Insect DNA Barcoding. *Ecology and Evolution*, *16*(2), e73112. https://doi.org/10.1002/ece3.73112

Gardes, M., & Bruns, T. D. (1993). ITS primers with enhanced specificity for basidiomycetes—Application to the identification of mycorrhizae and rusts. *Molecular Ecology*, *2*(2), 113–118. https://doi.org/10.1111/j.1365-294X.1993.tb00005.x

Gaube, P., Junker, R. R., & Keller, A. (2021). Changes amid constancy: Flower and leaf microbiomes along land use gradients and between bioregions. *Basic and Applied Ecology*, *50*, 1–15. https://doi.org/10.1016/j.baae.2020.10.003

Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., & Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, *35*(3), 521–522. https://doi.org/10.1093/bioinformatics/bty630

Hajibabaei, M. (2012). The golden age of DNA metasystematics. *Trends in Genetics*, *28*(11), 535–537. https://doi.org/10.1016/j.tig.2012.08.001

Hakimzadeh, A., Abdala Asbun, A., Albanese, D., Bernard, M., Buchner, D., Callahan, B., Caporaso, J. G., Curd, E., Djemiel, C., Brandström Durling, M., Elbrecht, V., Gold, Z., Gweon, H. S., Hajibabaei, M., Hildebrand, F., Mikryukov, V., Normandeau, E., Özkurt, E., M. Palmer, J., … Anslan, S. (2024). A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. *Molecular Ecology Resources*, *24*(5), e13847. https://doi.org/10.1111/1755-0998.13847

Ihrmark, K., Bödeker, I. T. M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J., Brandström-Durling, M., Clemmensen, K. E., & Lindahl, B. D. (2012). New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology*, *82*(3), 666–677. https://doi.org/10.1111/j.1574-6941.2012.01437.x

Impiö, M., & Raitoharju, J. (2024). Improving Taxonomic Image-based Out-of-distribution Detection With DNA Barcodes. *2024 32nd European Signal Processing Conference (EUSIPCO)*, 1272–1276. https://doi.org/10.23919/EUSIPCO63174.2024.10715139

Keller, A., Danner, N., Grimmer, G., Ankenbrand, M., von der Ohe, K., von der Ohe, W., Rost, S., Härtel, S., & Steffan-Dewenter, I. (2015). Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biology (Stuttgart, Germany)*, *17*(2), 558–566. https://doi.org/10.1111/plb.12251

Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., & Ankenbrand, M. J. (2020). BCdatabaser: On-the-fly reference database creation for (meta-)barcoding. *Bioinformatics*, *36*(8), 2630–2631. https://doi.org/10.1093/bioinformatics/btz960

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T., & Wolf, M. (2009). 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene*, *430*(1), 50–57.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., & Glöckner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, *41*(1), e1. https://doi.org/10.1093/nar/gks808

König, S., Krauss, J., Keller, A., Bofinger, L., & Steffan-Dewenter, I. (2022). Phylogenetic relatedness of food plants reveals highest insect herbivore specialization at intermediate temperatures along a broad climatic gradient. *Global Change Biology*, *28*(13), 4027–4040. https://doi.org/10.1111/gcb.16199

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology*, *79*(17), 5112–5120. https://doi.org/10.1128/AEM.01043-13

Leonhardt, F., Keller, A., Arranz Aveces, C., & Ernst, R. (2023). From Alien Species to Alien Communities: Host- and Habitat-Associated Microbiomes in an Alien Amphibian. *Microbial Ecology*, *86*(4), 2373–2385. https://doi.org/10.1007/s00248-023-02227-5

Leonhardt, S. D., Peters, B., & Keller, A. (2022). Do amino and fatty acid profiles of pollen provisions correlate with bacterial microbiomes in the mason bee Osmia bicornis? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1853), 20210171. https://doi.org/10.1098/rstb.2021.0171

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. https://doi.org/10.1186/1742-9994-10-34

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, *18*(185), 1–52.

Little, D. P. (2014). A DNA mini-barcode for land plants. *Molecular Ecology Resources*, *14*(3), 437–446. https://doi.org/10.1111/1755-0998.12194

Martins, A. C., Proença, C. E. B., Vasconcelos, T. N. C., Aguiar, A. J. C., Farinasso, H. C., de Lima, A. T. F., Faria, J. E. Q., Norrana, K., Costa, M. B. R., Carvalho, M. M., Dias, R. L., Bustamante, M. M. C., Carvalho, F. A., & Keller, A. (2023). Contrasting patterns of foraging behavior in neotropical stingless bees using pollen and honey metabarcoding. *Scientific Reports*, *13*(1), 14474. https://doi.org/10.1038/s41598-023-41304-0

Mayr, A. V., Keller, A., Peters, M. K., Grimmer, G., Krischke, B., Geyer, M., Schmitt, T., & Steffan-Dewenter, I. (2021). Cryptic species and hidden ecological interactions of halictine bees along an elevational gradient. *Ecology and Evolution*, *11*(12), 7700–7712. https://doi.org/10.1002/ece3.7605

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Metz, T., Farwig, N., Dormann, C. F., Schaefer, H. M., Guevara-Andino, J. E., Brehm, G., Burneo, S., Chao, A., Chazdon, R. L., Colwell, R. K., Diniz, U. M., Donoso, D. A., Endara, M.-J., Erazo, S., Escobar, S., Falconí-López, A., Feldhaar, H., Villamarin, M. G., Grella, N., … Blüthgen, N. (2026). Biodiversity resilience in a tropical rainforest. *Nature*, *652*(8112), 1232–1239. https://doi.org/10.1038/s41586-026-10365-2

Minar, M. R., & Naher, J. (2018). *Recent Advances in Deep Learning: An Overview*. https://doi.org/10.13140/RG.2.2.24831.10403

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 757–774. https://doi.org/10.1016/j.jksuci.2023.01.014

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, *18*(5), 1403–1414. https://doi.org/10.1111/1462-2920.13023

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., Spens, J., Thomsen, P. F., Bohmann, K., Cappellini, E., Schnell, I. B., Wales, N. A., Carøe, C., Campos, P. F., Schmidt, A. M. Z., Gilbert, M. T. P., Hansen, A. J., Orlando, L., & Willerslev, E. (2015). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1660), 20130383. https://doi.org/10.1098/rstb.2013.0383

Peters, B., Keller, A., & Leonhardt, S. D. (2022). Diets maintained in a changing world: Does land-use intensification alter wild bee communities by selecting for flexible generalists? *Ecology and Evolution*, *12*(5), e8919. https://doi.org/10.1002/ece3.8919

Pompanon, F., Deagle, B. E., Symondson, W. O. C., Brown, D. S., Jarman, S. N., & Taberlet, P. (2012). Who is eating what: Diet assessment using next generation sequencing. *Molecular Ecology*, *21*(8), 1931–1950. https://doi.org/10.1111/j.1365-294X.2011.05403.x

Porter, T. M., & Hajibabaei, M. (2021). Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. *BMC Bioinformatics*, *22*(1), 256. https://doi.org/10.1186/s12859-021-04180-x

Quaresma, A., Ankenbrand, M. J., Garcia, C. A. Y., Rufino, J., Honrado, M., Amaral, J., Brodschneider, R., Brusbardis, V., Gratzer, K., Hatjina, F., Kilpinen, O., Pietropaoli, M., Roessink, I., van der Steen, J., Vejsnæs, F., Pinto, M. A., & Keller, A. (2024). Semi-automated sequence curation for reliable reference datasets in ITS2 vascular plant DNA (meta-)barcoding. *Scientific Data*, *11*(1), 129. https://doi.org/10.1038/s41597-024-02962-5

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844. https://doi.org/10.1038/nbt.3935

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Ronchetti, F., Polidori, C., Schmitt, T., Steffan-Dewenter, I., & Keller, A. (2022). Bacterial gut microbiomes of aculeate brood parasites overlap with their aculeate hosts', but have higher diversity and specialization. *FEMS Microbiology Ecology*, *98*(12), fiac137. https://doi.org/10.1093/femsec/fiac137

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, *43*(6), e37. https://doi.org/10.1093/nar/gku1341

Schmidt, P.-A., Bálint, M., Greshake, B., Bandow, C., Römbke, J., & Schmitt, I. (2013). Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry*, *65*, 128–132. https://doi.org/10.1016/j.soilbio.2013.05.014

Shanahan, E. R., McMaster, J. J., & Staudacher, H. M. (2021). Conducting research on diet–microbiome interactions: A review of current challenges, essential methodological principles, and recommendations for best practice in study design. *Journal of Human Nutrition and Dietetics*, *34*(4), 631–644. https://doi.org/10.1111/jhn.12868

Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). *A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU* (arXiv:2305.17473). arXiv. https://doi.org/10.48550/arXiv.2305.17473

Sickel, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Härtel, S., Lanzen, J., Steffan-Dewenter, I., & Keller, A. (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecology*, *15*(1), 20. https://doi.org/10.1186/s12898-015-0051-y

Staab, M., Keller, A., Achury, R., Hilpert, A., Hölzel, N., Prati, D., Weisser, W. W., & Blüthgen, N. (2025). Experimental reduction of land use increases invertebrate abundance in grasslands. *Basic and Applied Ecology*, *88*, 62–71. https://doi.org/10.1016/j.baae.2025.08.006

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, *21*(8), 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Too, C. C., Keller, A., Sickel, W., Lee, S. M., & Yule, C. M. (2018). Microbial Community Structure in a Malaysian Tropical Peat Swamp Forest: The Influence of Tree Species and Depth. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.02859

Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution*, *24*(2), 110–117. https://doi.org/10.1016/j.tree.2008.09.011

Vaudo, A. D., Biddinger, D. J., Sickel, W., Keller, A., & López-Uribe, M. M. (2020). Introduced bees (Osmia cornifrons) collect pollen from both coevolved and novel host-plant species within their family-level phylogenetic preferences. *Royal Society Open Science*, *7*(7), 200225. https://doi.org/10.1098/rsos.200225

Villagómez, G. N., Keller, A., Rasmussen, C., Lozano, P., Donoso, D. A., Blüthgen, N., & Leonhardt, S. D. (2024). Nutrients or resin? – The relationship between resin and food foraging in stingless bees. *Ecology and Evolution*, *14*(2), e10879. https://doi.org/10.1002/ece3.10879

Vogel, C., Poveda, K., Iverson, A., Boetzl, F. A., Mkandawire, T., Chunga, T. L., Küstner, G., Keller, A., Bezner Kerr, R., & Steffan-Dewenter, I. (2023). The effects of crop type, landscape composition and agroecological practices on biodiversity and ecosystem services in tropical smallholder farms. *Journal of Applied Ecology*, *60*(5), 859–874. https://doi.org/10.1111/1365-2664.14380

Voulgari-Kokota, A., Grimmer, G., Steffan-Dewenter, I., & Keller, A. (2019). Bacterial community structure and succession in nests of two megachilid bee genera. *FEMS Microbiology Ecology*, *95*(1), fiy218. https://doi.org/10.1093/femsec/fiy218

Wang, Z., Liu, X., Liang, D., Wang, Q., Zhang, L., & Zhang, P. (2023). VertU: Universal multilocus primer sets for eDNA metabarcoding of vertebrate diversity, evaluated by both artificial and natural cases. *Frontiers in Ecology and Evolution*, *11*. https://doi.org/10.3389/fevo.2023.1164206

Weinhold, A., Grüner, E., & Keller, A. (2024). Bumble bee microbiota shows temporal succession and increase of lactic acid bacteria when exposed to outdoor environments. *Frontiers in Cellular and Infection Microbiology*, *14*. https://doi.org/10.3389/fcimb.2024.1342781

White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). AMPLIFICATION AND DIRECT SEQUENCING OF FUNGAL RIBOSOMAL RNA GENES FOR PHYLOGENETICS. In *PCR Protocols* (pp. 315–322). Elsevier. https://doi.org/10.1016/B978-0-12-372180-8.50042-1

Wilson, R. S., Keller, A., Shapcott, A., Leonhardt, S. D., Sickel, W., Hardwick, J. L., Heard, T. A., Kaluza, B. F., & Wallace, H. M. (2021). Many small rather than few large sources identified in long-term bee pollen diets in agroecosystems. *Agriculture, Ecosystems & Environment*, *310*, 107296. https://doi.org/10.1016/j.agee.2020.107296

Yang, C.-H., Wu, K.-C., Chuang, L.-Y., & Chang, H.-W. (2022). DeepBarcoding: Deep Learning for Species Classification Using DNA Barcoding. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(4), 2158–2165. https://doi.org/10.1109/TCBB.2021.3056570

Ye, B., Xia, J., Wan, X., Wu, M., & Chiba, S. (2026). Multi-barcoding-based species identification using a hierarchical attention network with staged curriculum learning. *Methods in Ecology and Evolution*, *17*(4), 1143–1160. https://doi.org/10.1111/2041-210x.70264

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, *31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862. https://doi.org/10.1111/mec.15060

Data Accessibility and Benefit-Sharing

Data Accessibility Statement

All code, documentation, pre-trained models, and example workflows used in this study are available from the *MetAnoDe* GitHub repository (https://github.com/chiras/*MetAnoDe*). The release used in this study (commit ca2fb57), including code, documentation and pre-trained models, is archived in Zenodo (https://doi.org/10.5281/zenodo.20313724). The mock-community benchmark, including scripts, configuration files, simulated data, intermediate results and a description of all parameters, is archived separately in Zenodo (https://doi.org/10.5281/zenodo.23918277). The GitHub repository will remain available for software development and future releases.

Reference sequences were obtained from public databases *SILVA* (Quast et al., 2013), *UNITE* (Abarenkov et al., 2025), BOLD (Ratnasingham & Hebert, 2007), *GenBank* (Benson et al., 2018) and Quaresma et al., (2024). Versions and download dates are documented in the archives. Previously published metabarcoding datasets

were used only to calibrate the mock simulation. (Elliott et al., 2021; Gaube et al., 2021; König et al., 2022; F. Leonhardt et al., 2023; S. D. Leonhardt et al., 2022; Martins et al., 2023; Mayr et al., 2021; Metz et al., 2026; Peters et al., 2022; Ronchetti et al., 2022; Staab et al., 2025; Too et al., 2018; Vaudo et al., 2020; Villagómez et al., 2024; Vogel et al., 2023; Voulgari-Kokota et al., 2019; Weinhold et al., 2024; Wilson et al., 2021). Their data are available from the repositories reported in the original publications.

Benefit-Sharing Statement

No new biological material was collected for this study.

Author Contributions

A.K. conceived the study, developed the software, performed the analyses, interpreted the results, and wrote the manuscript.