Anomaly detection in metabarcoding sequences using an LSTM-CNN deep neural network ensemble (MetAnoDe)

Alexander Keller[1]

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

Data Accessibility and Benefit-Sharing

Data Accessibility Statement

All code, documentation, trained models, validation datasets and metadata, and example workflows used in this study are available from the MetAnoDe GitHub repository: https://github.com/chiras/MetAnoDe. A permanent snapshot of the repository at the time of submission, including code, documentation, trained models, processed datasets, and associated metadata, has been archived in Zenodo: https://doi.org/10.5281/zenodo.20313724. The GitHub repository will remain available for software development and future releases.

Raw metabarcoding sequence data analysed in this study originate from previously published studies listed in Table 1. These raw data are available through the repositories and accession numbers reported in the respective original publications. The combined datasets used for analyses presented here are included in both the GitHub repository and the archived Zenodo release.

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

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 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 (DNNs) 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), a software workflow combining LSTM and CNN models within an ensemble framework. MetAnoDe employs an alignment-free approach that complements existing tools, enhancing metabarcoding data cleanup efficiency. Cross-validation and independent real-world dataset testing demonstrated high classification accuracy across both bacterial 16S-V4 and plant ITS2 markers. The ensemble model achieved validation accuracies of up to 97%, while also identifying substantial proportions of anomalous sequences not detected by current alignment-based workflows. The software additionally supports automated generation of new models for other metabarcoding markers.

In conclusion, MetAnoDe enhances metabarcoding data cleanup by efficiently identifying anomalous sequences. Combining deep-learning and traditional bioinformatic approaches improves identification of residual non-target reads, thereby increasing robustness and reliability of downstream biodiversity analyses.

Keywords: machine learning, microbiome, metabarcoding, 16S, ITS2, outlier detection, convolutional neural network, Long Short-Term Memory, recurrent neural networks

Introduction

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 sequencing a specific region of DNA, typically a barcode gene (such as portions of the 16S rRNA gene for bacteria or the ITS2 rRNA gene 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 (cross-hybridization of DNA), off-target and random PCR products, primer hybrids, and other less well understood issues (Zinger et al., 2019). 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 (DNNs) 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. Leveraging DNN features, the aim of this study is to enhance metabarcoding data cleanup pipelines by automatically identifying and classifying anomalous reads.

Here, I introduce the new software MetAnoDe (Metabarcoding Anomaly Detection), a character-level ensemble deep neural network framework consisting of 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 sequence patterns (Shiri et al., 2023; Yu et al., 2019). The second is a convolutional neural network (CNN), which captures spatial sequence characteristics (Alzubaidi et al., 2021). Both architectures were newly designed and implemented specifically for DNA sequence data and anomalous metabarcoding sequence detection.

The models were benchmarked individually and jointly as an ensemble with respect to biological validity, computational runtime, and potential integration points within typical metabarcoding processing pipelines. The workflow was applied separately to two widely used metabarcoding markers, bacterial 16S-V4 and plant ITS2. Pre-trained models for both markers are provided for reuse, although the software is also applicable to other markers through automated retraining using user-provided reference datasets, a process that may become computationally intensive depending on dataset size and complexity.

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 can be integrated at multiple stages of a bioinformatic workflow and provides a complementary perspective to existing filtering strategies. 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

Software availability and installation

MetAnoDe is implemented in Python 3.11 and is publicly available through the GitHub repository: https://github.com/chiras/MetAnoDe. The software supports deployment

through Conda environments as well as Docker Compose-based containerization to facilitate reproducible and platform-independent deployment and execution. Detailed installation instructions, dependency specifications, and deployment configurations are provided within the repository. MetAnoDe was developed and tested using TensorFlow 2.17 and Keras 3.3.3 on Ubuntu 24.04 and MacOSX 12.3 systems with both CPU and GPU execution.

For model training and validation as reported here, Python 3.11 was used together with the following major dependencies: TensorFlow (2.17), Keras (3.3.3), Pandas (2.2.1), NumPy (1.23.5), Scikit-learn (1.4.2), BioPython (1.84), Matplotlib (3.8.4), and Keras-Tuner (1.4.7).

Software usage

A full list of command-line parameters and optional settings is provided in the GitHub repository documentation. In brief, MetAnoDe is executed from the command line and requires a query FASTA file and a model name when using pre-trained models:

```
python metanode.py -query <query.fasta> -p <model_name>
```

A minimal working example of the command-line workflow used in this study is provided below in the section “Independent real metabarcoding data prediction and manual validation”.

For training new marker-specific models, correctly trimmed and deduplicated reference sequences are additionally supplied with `-db`. Known off-target amplicon classes can optionally be added with `-ot`:

```
python metanode.py -query <query.fasta> \
  -p <model_name> \
  -db <ref.fasta> \
  -ot <ot1.fasta>,<ot2.fasta>,<ot3.fasta>
```

By default, MetAnoDe retains all query sequences and writes model classifications to the output table. Optional filtering can be enabled to remove sequences classified as anomalous.

Model data preparation

Data for training and validation of the models was sourced from curated public databases to ensure the inclusion of valid biological sequences. Specifically, the plant ITS2 dataset derived from curated public dataset (Quaresma et al., 2024), while the 16S-V4 dataset was obtained from the SILVA database (Quast et al., 2013). Both datasets encompass a broad range of biological sources and bioregions, thereby enhancing the generality of the models. Reference sequences were trimmed to match the amplicon region of (Kozich et al., 2013) for 16S-V4 and (Sickel et al., 2015) for plant ITS2, ensuring priming compatibility and amplicon comparability. Sequences that did not fully span the amplicon region or were full-length duplicates at the nulceotide level were removed.

To create balanced datasets for training and validation, the following sequence classes were generated for each marker. 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 (<1%) or insertion-deletion events (indels; <1%) were generated from the original reference sequences. These modified sequences were combined with the original sequences to form the final Class 0 dataset.

2. *Class 1 and 2 sequences (simulated errors):*
   High-substitution rate sequences (10%) and high-indel rate sequences (10%), simulating PCR or sequencing errors, were generated each in equal amounts to the Class 0 sequences. These constituted Class 1 and Class 2 sequences, respectively.
3. *Class 3 sequences (chimeras):*
   Chimeric sequences were created by merging parts of two randomly chosen sequences from the original database, forming Class 3 sequences also in equal amounts to the Class 0 sequences. Merged parts constituted at least 1/4 up to 3/4 each of the original sequences from the beginning and end of a sequence, respectively.
4. *Class N sequences (known off-target products):*
   The software additionally supports modelling of user-defined sequence classes, allowing inclusion of known off-target amplification products. For bacterial 16S-V4, mitochondrial and plastid 16S sequences were included as Class 4 and Class 5 sequences, respectively. For plant ITS2, fungal ITS2 sequences from the UNITE database (Nilsson et al., 2019) were included as Class 4 sequences. These datasets 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.

The complete workflow is outlined in Figure 1. After generation of all sequence classes, datasets were concatenated, shuffled twice, and split into 75% training data and 25% cross-validation data. Class balancing was maintained throughout the split, and no identical sequences were shared between training and cross-validation datasets to prevent data leakage between partitions. This process ensured a robust and comprehensive training dataset for accurate detection and classification of anomalous metabarcoding sequences.

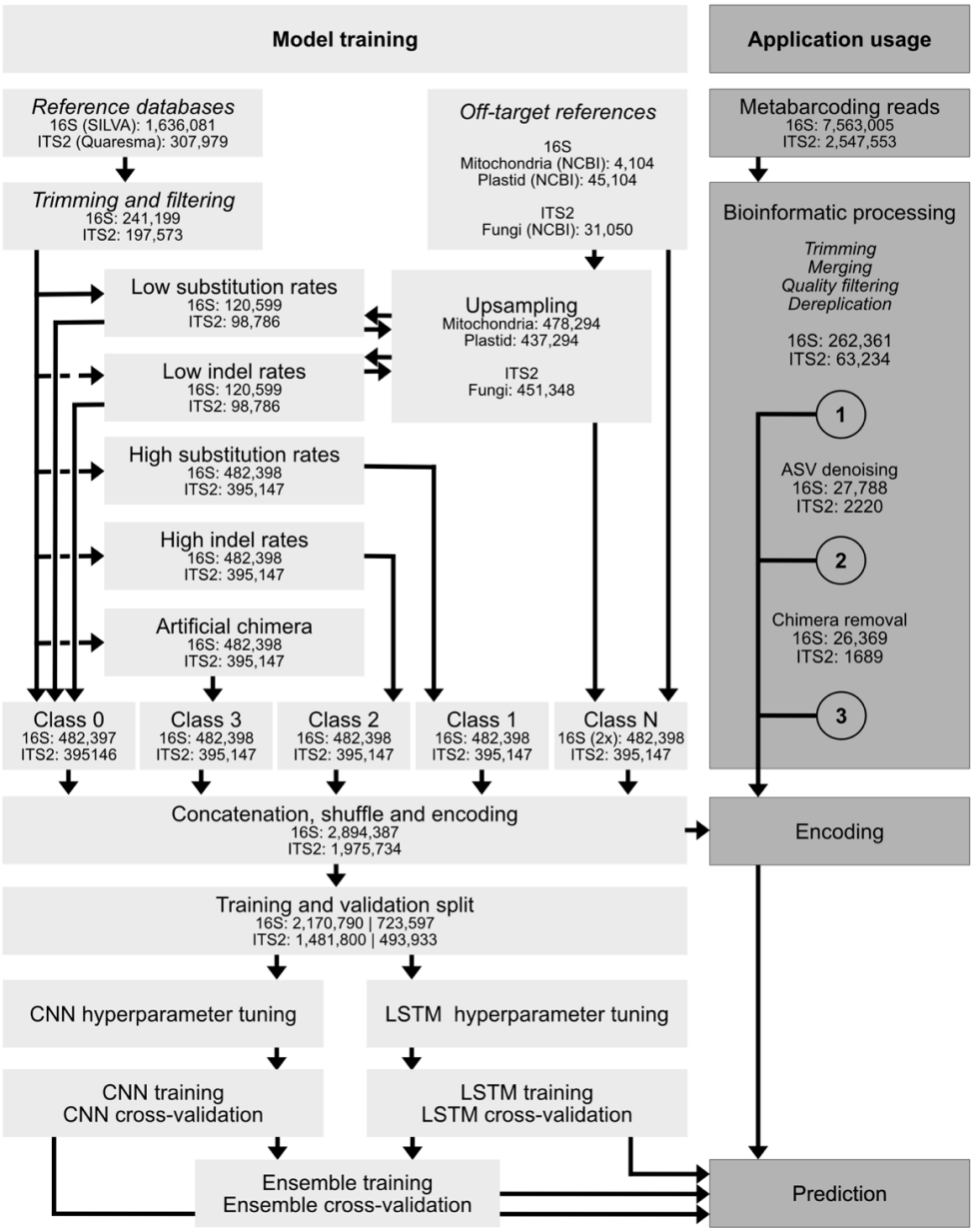

Figure 1: *Workflow of MetAnoDe illustrating the two principal processing pathways. The left pathway (light grey) represents automated training and validation of new models, here demonstrated for plant ITS2 and 16S-V4 metabarcoding markers. The right pathway (dark grey) represents application of pre-trained models for prediction on query datasets. Pre-processing steps required prior to MetAnoDe application are indicated in italics. Numbers denote sequence counts entering the respective workflow steps. Circled numbers indicate potential insertion points of MetAnoDe predictions within a standard metabarcoding bioinformatic pipeline.*

Model training

The following procedure was applied independently for each marker:

Data preparation: Sequences in the training and validation datasets were tokenized and encoded at the character level using a shared vocabulary generated from the complete

dataset. Encoded sequences were subsequently end-padded with zeros and reshaped to match the input requirements of the CNN and LSTM architectures.

Model architectures: The architectures of the CNN, LSTM, and ensemble models are illustrated in Figure 2. Dropout layers were incorporated throughout the models to reduce overfitting and improve generalization. Sparse categorical cross-entropy was used as the loss function, and L2 weight regularization was applied. For the ensemble model, latent output layers from the CNN and LSTM branches were concatenated and connected to additional fully connected layers for final classification.

Hyperparameter tuning: Hyperparameter optimization of the CNN and LSTM models was conducted using Hyperband search implemented in *keras-tuner* (Li et al., 2018). Optimal parameter combinations were selected based on validation accuracy.

Technical cross-validation: Model performance was evaluated on the validation datasets using overall accuracy, precision, recall, and F1 scores for individual classes.

Training of the 16S and ITS2 models was performed on an Ubuntu 24.04 system equipped with a Ryzen 7 16-core CPU, 64 GB RAM, and an NVIDIA GeForce RTX 4070 Ti SUPER GPU using TensorFlow GPU acceleration. Training and validation workflows were additionally tested for compatibility on Ubuntu 24.04 and MacOSX 12.3 systems without GPU support using reduced test datasets.

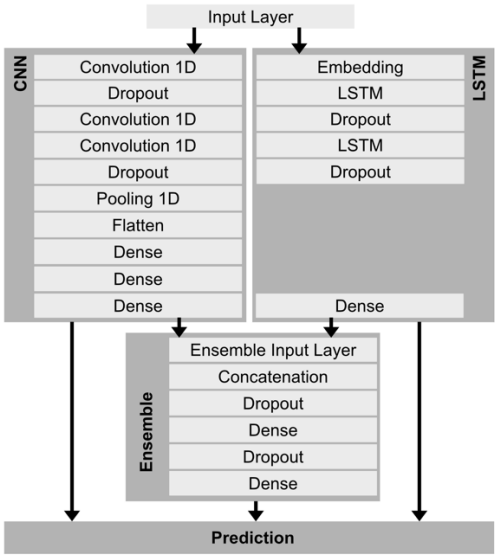

Figure 2: *Architecture of the CNN, LSTM, and ensemble models implemented in TensorFlow and Keras. The CNN branch captures spatial sequence characteristics through convolutional layers, whereas the LSTM branch models sequential dependencies within DNA sequences. Outputs from both architectures are concatenated within the ensemble model to generate final sequence classifications. All three models (CNN, LSTM, and Ensemble) were evaluated independently for prediction performance.*

Independent real metabarcoding data prediction and manual validation

For the plant ITS2 and bacterial 16S-V4 markers, several studies listed in Table 1 were selected based on compatibility of their amplicon and sequencing strategy with the developed models, public availability of raw data, and broad representation of biological contexts, geographical bioregions, and sample material types. This selection was intended

to evaluate the generalizability of the method across diverse metabarcoding applications. Bacterial 16S-V4 sequencing libraries were prepared according to Kozich et al. (2013), whereas plant ITS2 sequencing libraries followed Sickel et al. (2015). Other library preparation strategies may also be compatible with the models, provided that identical primer regions are targeted and adapter as well as primer sequences are removed prior to prediction. All datasets were sequenced on Illumina MiSeq, NovaSeq, or HiSeq platforms. Additional details regarding sequencing and library preparation are provided in the respective original publications. From each study, 20 random samples were selected for prediction analyses. The compiled datasets and metadata required to reproduce the analyses are archived in Zenodo (Keller, 2026). Sequence data were processed according to the workflow as used in Leonhardt et al. (2022; https://github.com/chiras/metabarcoding_pipeline), primarily utilizing VSEARCH (Rognes et al., 2016).

Table 1*: Data origins that were used for independent real metabarcoding data prediction and manual validation.*

Marker	Sample type	Geographic origin	Reference
Bacterial 16S-V4	soil	Ecuador	(Metz et al., 2026)
Bacterial 16S-V4	flower & phylosphere	Germany	(Gaube et al., 2021)
Bacterial 16S-V4	frog skin & faeces	Caribbean	(F. Leonhardt et al., 2023)
Bacterial 16S-V4	bee nests	Germany	(S. D. Leonhardt et al., 2022)
Bacterial 16S-V4	wasp guts	Italy	(Ronchetti et al., 2022)
Bacterial 16S-V4	peat swamp soil	Malaysia	(Too et al., 2018)
Bacterial 16S-V4	soil	Tanzania	(Vogel et al., 2023)
Bacterial 16S-V4	bee guts, nests & pollen	Germany	(Voulgari-Kokota et al., 2019)
Bacterial 16S-V4	bee guts	Germany	(Weinhold et al., 2024)
Plant ITS2	honey, pollen	Brazil	(Martins et al., 2023)
Plant ITS2	bee nests	Germany	(Peters et al., 2022)
Plant ITS2	pollen	USA	(Vaudo et al., 2020)
Plant ITS2	gut contents	Tanzania	(Mayr et al., 2021)
Plant ITS2	pollen	Ecuador	(Villagómez et al., 2024)
Plant ITS2	bee nests	Australia	(Wilson et al., 2021)
Plant ITS2	gut contents	Germany	(König et al., 2022)

Anomaly predictions were conducted at multiple stages of the metabarcoding workflow (Figure 1) to identify suitable integration points for MetAnoDe within existing bioinformatic pipelines. This analysis aimed to assess overlap and complementarity between MetAnoDe predictions and alignment-based filtering approaches, while additionally evaluating whether early-stage identification and removal of anomalous sequences could improve downstream processing efficiency and reduce overall computational runtime. Runtime effects were assessed by comparing complete pipeline execution when MetAnoDe was applied at different workflow stages. Computationally intensive tasks such as denoising can substantially benefit from reduced sequence numbers, potentially making early-stage anomaly detection advantageous for runtime reduction (Coissac et al., 2012). Conversely, applying deep-learning predictions to larger intermediate datasets increases prediction runtime of the models themselves.

Finally, the sequences underwent manual inspection using BLAST (Altschul et al., 1990) against the NCBI GenBank database (excluding environmental samples) (Benson et al., 2018). Predictions by the Ensemble, LSTM, and CNN models were generated for all dereplicated sequences after VSEARCH merging, quality filtering, and dereplication (Figure 1; insertion point 1). Sequences flagged as anomalous by MetAnoDe were recorded for subsequent comparison. In parallel, the same datasets were processed through the complete VSEARCH workflow without MetAnoDe, including denoising, *de novo* chimera filtering, and taxonomic classification against reference databases. For 16S-V4, classifications were performed using either a bacterial, mitochondrial, and plastid reference database (BMP-DB) (Quast et al., 2013) or a curated bacterial-only database version thereof (B-DB), whereas ITS2 classifications were based on a plant reference database only (Quaresma et al., 2024).

Values reported here were obtained using the following calls on the command line, with the Test dataset and models available as stated here in the software repository:

- 16S: `python metanode.py -query data/TestSet/16S.all.merge.derep.fa -p 16S_2026_04_03 -t 8`
- ITS2: `python metanode.py -query data/TestSet/ITS2.all.merge.derep.pr.fa -p ITS2_2026_04_01 -t 8`

For final biological validation, the 100 most abundant sequences per dataset were manually inspected and compared against classifications generated by MetAnoDe and the VSEARCH workflows. Predictions were categorized as valid, non-critical, or critical. Valid classifications were those confirmed through manual inspection. Critical classifications represented transitions between true target sequences (Class 0) and anomalous sequence classes (Class 1-N), i.e. anomalous sequences incorrectly retained as valid targets or valid target sequences incorrectly flagged for removal. Non-critical classifications represented transitions among anomalous classes (Class 1-N), where sequences were correctly identified as anomalous but assigned to an incorrect anomaly category.

AI Technology Disclosure

ChatGPT (GPT-5.5 Thinking) was used for language editing, grammar correction, and stylistic refinement during manuscript preparation. ChatGPT was also used for limited readability improvements and bug fixing of individual software code blocks during software preparation. The scientific concepts, study design, analyses, interpretation, manuscript content, and software implementation were developed by the author. No manuscript content or software code 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

Technical cross-validation

Overall, all models achieved high validation accuracy, exceeding 90% for both markers (Table 2). For the 16S dataset, overall accuracies ranged from 0.91–0.94, with the ensemble model performing best overall (0.94), followed by the LSTM (0.93) and CNN (0.91). For the plant ITS2 dataset, accuracies were even higher, ranging from 0.94–0.97, again with the ensemble model showing the best performance (0.97).

Fungal ITS2 sequences were predicted perfectly across all models, achieving precision, recall, and F1 scores of 1.00. Similarly, high indel and chimera classes generally showed very high classification performance in both markers, frequently with F1 scores above 0.95. Particularly strong improvements were observed for the ensemble models, which consistently outperformed or matched the individual CNN and LSTM architectures across most classes.

The most challenging classes were mitochondrial and plastid off-target sequences in the 16S models. Mitochondrial sequences showed comparatively low precision values (0.76–0.78), indicating confusion with other anomalous classes, while plastid sequences exhibited lower recall values (0.71–0.73), suggesting that a proportion of plastid reads remained difficult to distinguish reliably from true target sequences. Nevertheless, even these classes retained comparatively high overall F1 scores of approximately 0.80–0.87.

Importantly, class 0 ("positive") sequences, representing biologically valid target reads, achieved particularly high precision and recall values in the ensemble models for both markers, indicating a low rate of biologically critical misclassification.

Table 2: *Cross-validation results for the Ensemble, LSTM, and CNN models across all classes and overall validation accuracy. Within each cell, values are ordered as Ensemble | LSTM | CNN. Class "0 positive" (in bold) represents true target sequences and is therefore most closely associated with biologically critical errors, whereas misclassification among anomalous classes mainly reflects non-critical errors. Support indicates the sample size used for cross-validation.*

Model	**Class**	**Precision**	**Recall**	**F1-Score**
16S	0 positive	0.97 | 0.94 | 0.90	0.98 | 0.99 | 0.96	0.97 | 0.96 | 0.93
	1 high substitution	0.99 | 0.99 | 0.95	0.98 | 0.97 | 0.90	0.99 | 0.98 | 0.92
	2 high indel	0.99 | 0.98 | 0.99	0.99 | 0.99 | 0.97	0.99 | 0.99 | 0.98
	3 chimera	0.99 | 0.99 | 1.00	0.98 | 0.96 | 0.92	0.99 | 0.97 | 0.96
	4 mitochondria	0.78 | 0.77 | 0.76	0.99 | 1.00 | 0.97	0.87 | 0.87 | 0.85
	5 plastids	0.99 | 1.00 | 0.90	0.71 | 0.71 | 0.73	0.83 | 0.83 | 0.80
	Overall validation accuracy	0.94 | 0.93 | 0.91		
Model	**Class**	**Precision**	**Recall**	**F1-Score**
ITS2	0 positive	0.91 | 0.91 | 0.86	0.96 | 0.95 | 0.89	0.94 | 0.93 | 0.88
	1 high substitution	0.97 | 0.96 | 0.90	0.98 | 0.97 | 0.94	0.97 | 0.96 | 0.92
	2 high indel	0.98 | 0.97 | 0.99	0.98 | 0.96 | 0.95	0.98 | 0.96 | 0.97
	3 chimera	0.98 | 0.97 | 0.94	0.94 | 0.93 | 0.91	0.96 | 0.95 | 0.93
	4 fungi	1.00 | 1.00 | 1.00	1.00 | 1.00 | 1.00	1.00 | 1.00 | 1.00
	Overall validation accuracy	0.97 | 0.96 | 0.94		

Independent real data prediction and manual validation

Predictions for the independent 16S dataset (262,361 query sequences) showed that the deep-learning approaches performed substantially better than the standard VSEARCH workflow in avoiding critical classification errors (Figure 3). The ensemble model classified 69.70% of sequences as true biological sequences, while additionally identifying 17.73% as sequences with elevated substitution rates indicative of sequencing errors, 10.27% as mitochondrial sequences, 1.06% as sequences with high indel rates, 1.05% as chimeras, and 0.19% as chloroplast sequences. Total runtime for the independent prediction was 10 min 49 s. Manual validation showed that the LSTM and CNN models produced no critical misclassifications in the independent 16S dataset, whereas the ensemble model incorrectly retained a small number of fungal sequences as true biological sequences, corresponding to

a critical error rate of 2.4%. In contrast, the standard VSEARCH workflow produced 8% critical misclassifications when using an extended reference database containing mitochondrial and plastid sequences. Non-critical misclassification rates were 21.6% for the ensemble, 23.3% for the LSTM, and 24.0% for the CNN, compared to 18.4% for VSEARCH. These values represent proportions of unique sequence variants and therefore do not account for sequence abundances.

It is important to note that the reported VSEARCH performance was obtained using an extended reference database that already contained mitochondrial and plastid sequences, which still required downstream filtering. When these references were absent from the database, an additional 23.2% of sequences were critically misclassified, increasing the total critical error rate to 31.2% (Figure 3).

For the independent plant ITS2 dataset (63,234 query sequences), the ensemble model classified 87.97% of sequences as true biological sequences, while predicting 4.81% as fungal contamination, 3.40% as chimeras, 3.04% as sequences with elevated substitution rates indicative of sequencing errors, and 0.79% as sequences with high indel rates. Runtime for the full independent ITS2 prediction was 3 min 28 s. Manual validation showed that the ensemble model produced no critical misclassifications in the ITS2 dataset, whereas critical error rates reached 6% for both the LSTM and CNN models and 7% for VSEARCH (Figure 3). Non critical error rates were also lowest for the ensemble model (1.01%), compared to 11% for the LSTM and 12% for both the CNN and VSEARCH approaches.

In contrast to the cross-validation results, independent validation revealed distinct trade-offs among the deep learning approaches. For the 16S dataset, the LSTM and CNN models completely avoided critical misclassifications, whereas the ensemble model reduced non critical errors relative to the individual models while retaining only a small number of fungal sequences as false positives. For plant ITS2, the ensemble model showed the strongest overall performance, producing no critical misclassifications and substantially fewer non critical errors than the alternative approaches.

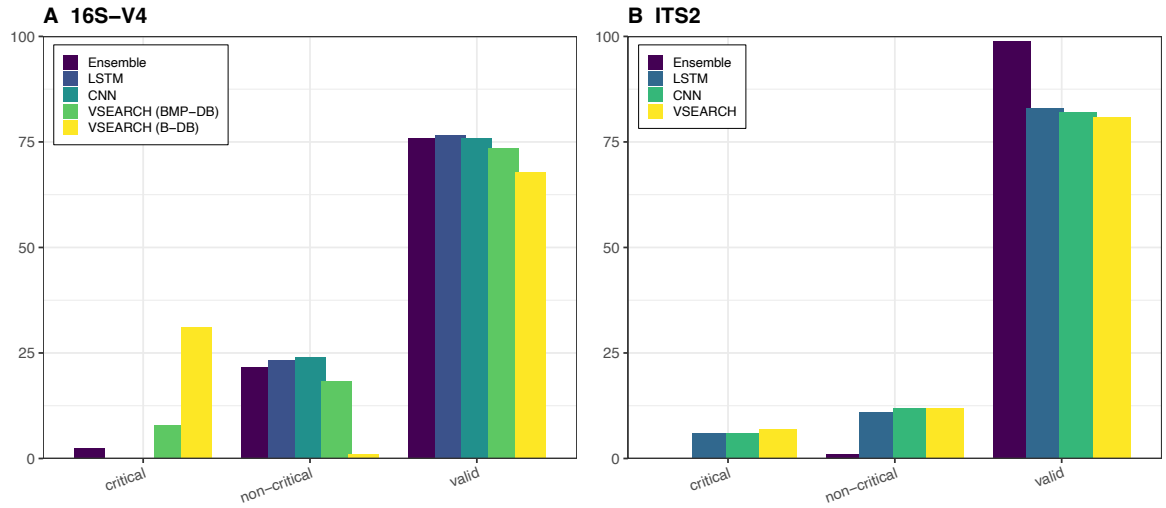

Figure 3: *Manual validation results with real metabarcoding data for 16S-V4 (left), plant ITS2 (right). Critical errors are considered such that lead to either to missing to remove anomalous sequences or incorrect filtering of valid sequencing. Non-critical errors are considered such that detect an anomaly but not the correct class of anomaly. In 16S-V4, BMP-DB indicates that a follow-up taxonomic classification against a reference database that includes bacterial, mitochondrial and plastid*

sequences has been applied additionally to Figure 1 in the standard metabarcoding pipeline to classify off-targets. The same applies to B-DB, where Mitochondria and Plastids are not included. For ITS2, only a plant database was used.

Insertion point

Runtime improvements can be achieved by applying MetAnoDe prior to denoising and subsequently hard removing sequences predicted as anomalous before ASV inference. Under this strategy, total runtime of the entire workflow decreased by 5.1% for the 16S dataset and 3.4% for the ITS2 dataset compared to application after denoising, because fewer sequences entered the computationally intensive denoising step. In contrast, no substantial runtime differences were observed when the models were applied before versus after chimera filtering.

No differences in sequence classifications were observed when MetAnoDe was inserted at different stages of the workflow, indicating stable prediction behaviour across pipeline positions. However, overlap between classification errors of MetAnoDe models and VSEARCH ranged only between 35% and 60%, indicating that both approaches frequently identified different anomalous sequences. Overall overlap between Ensemble and VSEARCH classifications was 78.4%. This demonstrates that deep-learning based predictions provide a complementary perspective to conventional alignment-, abundance-, and rule-based filtering methods.

Discussion

MetAnoDe as a complementary anomaly detection resource

Here, I introduce MetAnoDe, a workflow and software for identifying and classifying anomalous sequences in 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 bacterial 16S-V4 and plant ITS2 metabarcoding studies. 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 metabarcoding regions.

To the best of my knowledge, no other workflow or software has been proposed to date that specifically addresses anomalous metabarcoding read detection using deep-learning methods, also considering recent reviews of available metabarcoding software (Hakimzadeh et al., 2024). MetAnoDe is alignment-free and independent of sequencing quality statistics. It therefore provides a complementary perspective to existing quality-, abundance-, and alignment-based filtering strategies rather than replacing them. This distinction is important because many currently used approaches rely on overlapping assumptions, whereas deep-learning models can detect complex inherent sequence patterns without explicit alignments or reference database matching during prediction.

Validation and biological interpretation

Technical cross-validation showed high classification performance for both markers. Overall validation accuracy exceeded 90% in all models and was highest for the ensemble models, reaching 94% for 16S-V4 and 97% for plant ITS2. This indicates that combining LSTM and CNN architectures improved overall classification performance relative to the individual model branches. Fungal ITS2 off-target sequences were predicted perfectly across all models. In contrast, mitochondrial and plastid off-target sequences in the 16S-V4 models remained more challenging. Mitochondrial sequences showed reduced precision, whereas plastid sequences showed reduced recall, suggesting partial overlap in learned sequence features particularly among mitochondrial, and plastid 16S reads.

Independent validation using real metabarcoding datasets showed that performance patterns differed slightly from technical cross-validation. For 16S-V4, exact class-level agreement was lower, mainly because of non-critical misclassification among anomalous classes. In contrast, the updated ITS2 ensemble validation showed very high agreement, with almost all manually inspected sequences correctly classified and no critical errors. These differences are expected because the simulated cross-validation data were generated using defined random distributions for substitutions, indels, and chimeras, whereas real PCR and sequencing artifacts may follow more complex and context-dependent processes (Schirmer et al., 2015). In addition, non-functional ribosomal DNA copies and pseudogenes were not included as separate training classes, although such sequences may occur in real datasets and deviate from the simulated anomaly classes (Porter & Hajibabaei, 2021).

Importantly, almost all classification errors made by MetAnoDe were non-critical. Errors involving transitions between true target sequences and anomalous classes are biologically critical because they either retain anomalous sequences as valid targets or remove valid biological sequences. In contrast, misclassifications among anomalous classes are non-critical in many downstream applications because the sequence is still correctly flagged as problematic. When considering only biologically critical errors, performance was high: critical accuracy ranged from 97.6-100% for the deep-learning models in 16S-V4 and from 94.0-100% in ITS2. By comparison, critical accuracy for VSEARCH (Rognes et al., 2016) was 92.0% in 16S-V4 when using the BMP-DB reference database, 68.8% when using the bacterial-only database, and 93.0% in ITS2. From a practical data-cleaning perspective, this distinction is central: detecting that a sequence is unlikely to represent a valid target amplicon is often more important than correctly resolving whether the anomaly originated from a chimera, an indel-rich sequence, a substitution-rich artifact, or an off-target amplification product.

Complementarity with VSEARCH and conventional workflows

The independent validation also showed that MetAnoDe and conventional VSEARCH-based workflows did not simply identify the same problematic sequences. Overlap between classification errors of the deep-learning models and VSEARCH ranged between 35% and 60%, while overall overlap between Ensemble and VSEARCH classifications was 78.4%. This incomplete overlap indicates that both approaches detect partly distinct subsets of anomalous sequences.

This complementarity is central to the proposed use of MetAnoDe. Alignment-based workflows are highly effective for many standard filtering tasks, including similarity-based taxonomic assignment and chimera detection. However, they may miss anomalous reads

that remain plausible under alignment-based criteria or that are difficult to classify due to incomplete reference databases. Conversely, deep-learning models detect deviations from the learned target-region structure and may therefore be sensitive to inconsistencies in primer trimming, amplicon boundaries, or biological sequences that fall outside the trained target definition. This risk is likely reduced for broadly sampled markers such as 16S-V4 trained on broad and taxonomically rich SILVA database data, but may become more relevant for less comprehensively represented markers or reference-poor taxonomic groups.

Combining both approaches therefore provides a more comprehensive quality assessment than either strategy alone. Overall, the results indicate that substantial numbers of anomalous sequences may remain undetected by classical workflows alone. In the datasets analysed here, MetAnoDe identified residual anomalous sequences in the range of approximately 25–30% of dereplicated reads, depending on marker and workflow context. If such sequences are not identified through additional inspection or filtering, they may enter downstream biodiversity estimates, community analyses, or interaction network reconstruction. MetAnoDe can therefore make a valuable contribution by identifying candidate sequences for removal or manual review.

Insertion points and recommended workflow

Runtime analyses showed that applying MetAnoDe *prior* to denoising and directly removing sequences predicted as anomalous can reduce total pipeline runtime. Under this strategy, total runtime decreased by 5.1% for the 16S-V4 dataset and 3.4% for the ITS2 dataset, because fewer sequences entered the computationally intensive denoising step. In contrast, no substantial runtime differences were observed when the models were applied before versus after chimera filtering. Sequence classifications were stable across insertion points, suggesting that model predictions were not strongly affected by the tested pipeline stage.

However, the optimal insertion point for anomaly detection should not be considered solely from a computational perspective. Although early hard removal can reduce runtime, sequences with elevated substitution rates or other irregularities may still contain information relevant for denoising algorithms that reconstruct ASVs from patterns of sequence similarity and abundance structure. Premature removal of such sequences could therefore alter denoising behaviour or affect inference of low-abundance biological variants. For this reason, I do not generally recommend mandatory hard filtering of sequences classified as anomalous prior to denoising. A more conservative and broadly applicable strategy is to first perform the standard metabarcoding workflow using the established denoising and filtering procedures of the respective pipeline, and then apply MetAnoDe as an additional anomaly detection and quality assessment layer to the retained sequence set. Under this framework, predicted anomaly classes can be interpreted as independent quality indicators and used flexibly depending on study goals, marker systems, and desired stringency levels.

This approach may be particularly useful for novel datasets, understudied marker systems, or reference-poor taxonomic groups, where both alignment-based and deep-learning based filtering approaches may produce erroneous classifications. Rather than enforcing direct automated removal, flagged sequences can be inspected manually or removed selectively depending on downstream analytical goals. To support such workflows, MetAnoDe provides

accompanying R functions compatible with *phyloseq* objects, allowing optional downstream filtering, subsetting, or annotation of sequences assigned to specific anomaly classes.

Software deployment, reuse, and marker transferability

MetAnoDe is designed for practical reuse in metabarcoding workflows. Pre-trained models for bacterial 16S-V4 and plant ITS2 are provided and can be applied directly to query data, provided that adapter and primer sequences have been removed and the analysed amplicon region matches the model. By default, the software retains all query sequences and adds model-based classifications to the output. Optional sequence removal is available for workflows where automated hard filtering is desired. The main outputs are a tabular prediction file and a FASTA file containing flagged or retained sequences.

The workflow can also be adapted to other metabarcoding markers through automated training of new models. This requires correctly trimmed and deduplicated target reference sequences and, optionally, additional known off-target sequence classes. Once trained, a marker-specific model can be reused for future prediction tasks. Although there is no fixed limit on the number of reference sequences, sequence length, or off-target classes, memory requirements for encoding and model training may become limiting depending on hardware resources. GPU use is therefore strongly recommended for training, whereas prediction can be performed on both CPU and GPU systems.

To facilitate reproducible use, the repository provides installation instructions, dependency specifications, Conda-based deployment options, and Docker Compose based containerization. This should make the software accessible both for local analyses and reproducible computational environments.

Limitations and future development

Several limitations should be considered when interpreting the results. First, the technical validation relied on simulated substitution, indel, and chimera classes. Although these simulations provide controlled benchmarks, they cannot fully capture the complexity of real PCR and sequencing errors, which are currently also not well understood.

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). For new marker systems, model performance will therefore depend on the quality and breadth of both target reference sequences and included off-target classes.

Third, transferability requires consistent marker definition. Pre-trained models should only be applied to data generated from compatible primer regions after adapter and primer trimming. Datasets generated with different primers or substantially different amplicon boundaries will require retraining.

Finally, the independent manual validation was restricted to the 100 most abundant dereplicated sequences per dataset. These sequences represented a substantial fraction of total read abundance, accounting for 44.0% of reads in the 16S-V4 dataset and 52.7% in the ITS2 dataset. The validation therefore focused on sequences with the greatest potential influence on abundance-weighted downstream analyses. However, it does not represent exhaustive manual curation of all dereplicated sequences.

Conclusions

MetAnoDe provides an alignment-free deep-learning based anomaly detection layer for metabarcoding workflows. The software identifies and classifies anomalous sequences with high validation accuracy and provides information that is complementary to conventional alignment-, abundance-, and quality-based filtering methods. Its strongest value is not as a replacement for established workflows, but as an additional quality assessment layer that can flag residual anomalous sequences for inspection or selective removal. In combination with traditional methods, MetAnoDe can improve the robustness, transparency, and reliability of downstream biodiversity analyses.

Acknowledgments

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

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

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

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

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

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

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

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

Keller, A. (2026). *MetAnoDe: Metabarcoding Anomaly Detection* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.20313724

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

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

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.

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

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

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, *47*(D1), D259–D264. https://doi.org/10.1093/nar/gky1022

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

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

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

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

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

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, trained models, processed validation datasets, and example workflows used in this study are available from the MetAnoDe GitHub repository: https://github.com/chiras/MetAnoDe. A permanent snapshot of the repository at the time of submission, including code, documentation, trained models, processed datasets, and associated metadata, has been archived in Zenodo: https://doi.org/10.5281/zenodo.20313724. The GitHub repository will remain available for software development and future releases.

Raw metabarcoding sequence data analysed in this study originate from previously published studies listed in Table 1. These raw data are available through the repositories and accession numbers reported in the respective original publications. The combined datasets used for analyses presented here are included in both the GitHub repository and the archived Zenodo release.

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.