Anomaly detection in metabarcoding amplicon reads 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;

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

Abstract

Metabarcoding has emerged as a critical tool in ecology and other scientific disciplines, facilitating species identification in diverse samples for biodiversity monitoring, community and microbiome analysis, dietary studies, and understanding species interactions. However, challenges arise from errors and artifacts introduced during laboratory processes such as PCR and sequencing. Manual inspection is impractical due to the vast amount of sequences, necessitating rapid algorithms to clean the data. Thorough bioinformatic data cleanup can reduce such mistakes by removal of low-quality sequences or such classified as non-fitting through alignments. However, in practice some anomalous sequences evade detection, while also normal sequences may be mistakenly removed.

Deep neural networks (DNNs) offer a promising solution by recognizing complex DNA sequence patterns. In this study I present a new software MetAnoDe (Metabarcoding Anomaly Detection), featuring development of novel deep-learning LSTM and CNN models for independent application and use as an ensemble model. MetAnoDe employs an alignment-free approach that complements existing tools, enhancing data cleanup efficiency. Here, the three models were trained for bacterial 16S-V4 and plant ITS2 markers which can be readily reused in other studies. Cross-validation and real-world data testing demonstrate high accuracy. Optimal integration into pipelines can also streamline overall runtime, synergizing effectively with current alignment-based methods. It is further adaptable for other markers due to the software's automated model training capability.

In conclusion, MetAnoDe enhances metabarcoding by efficiently identifying anomalous sequences. An integration of DNNs with traditional approaches enhances biodiversity estimates by reducing non-target sequence inclusion, ensuring more accurate and comprehensive results.

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 mitigate these issues, they cannot entirely eliminate them. Current efforts to clean up raw data 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). All of these methods are either sequence quality-, abundance- or alignment-based, or a combination thereof, and might find anomalous artifacts using sequencer device statistics or inherent metrics or reference data. Although these approaches significantly improve final data quality, residual reads in the final processed data are not entirely of true target, or even biological, origin given the aforementioned challenges, or valid data might be wrongly removed given e.g. rare occurrences, misclustering or quality of references (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 dereplicated reads (i.e., non-redundant sequences), 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 entirely or pragmatically inspect only the most abundant reads. This underscores the need for new algorithms that can rapidly identify anomalous sequences, ideally using approaches distinct from existing pipeline steps to detect sequences that current methods may miss or wrongly filter.

Deep neural networks (DNNs) are advanced machine learning models composed of multiple trainable 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 clean-up pipelines by automatically identifying and classifying anomalous reads.

Here, I propose the new software MetAnoDe (Metabarcoding Anomaly Detection) using novel character-level ensemble DNNs consisting of two major models (Minar & Naher, 2018; Mohammed & Kora, 2023). The first model is a recurrent neural network (RNN), specifically a long-short term memory (LSTM) model, which focuses on the sequential patterns of DNA sequences (Shiri et al., 2023; Yu et al., 2019). The second model is a convolutional neural network (CNN), which examines the spatial characteristics of a given DNA sequence (Alzubaidi et al., 2021). These models were individually and collectively as an ensemble benchmarked for biological validity, computational runtime, and potential integration points within a typical metabarcoding data processing pipeline. This software was applied separately to two commonly used markers in metabarcoding: bacterial 16S-V4 and plant ITS2. Pre-trained models for these markers are available for reuse, though the tool is applicable to other markers as well, provided new training is conducted — a process that can be automated but may be computationally intensive.

This tool is intended to complement existing alignment-, abundance- and quality-based approaches by applying machine learning predictions to detect and classify (or remove) anomalous sequences (Hakimzadeh et al., 2024). It can be integrated at various stages of a bioinformatic workflow. Ultimately, these classifications can assist researchers in making faster, more reliable decisions regarding sequence trustworthiness in large datasets, thereby enhancing the quality of metabarcoding data.

Material and Methods

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 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 ITS2, ensuring priming consistency and amplicon comparability. Sequences that did not fully span the amplicon region or were full-length duplicates on the sequence level were removed.

To create a balanced dataset for each marker for training and validation of models, the following datasets were generated and labelled with corresponding classes:

1. *Class 0 sequences (positive target amplicons):*
   All remaining sequences from the database were labelled as true biological target sequences. Further, to account for variability not covered in the database, additional sequences with random low-substitution (1 %) or insertion-deletion (indel, 1 %) variations were generated from the original database sequences. These were combined with the original sequences to form Class 0 sequences.
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 here developed software allows to also model further sequence classes if provided. This allows classifying known off-target 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. Class 4 or 5 sequences were each upscaled to match the amount of Class 0 sequences. This upscaling involved generating additional sequences with low substitution and indel variations as described for Class 0 sequences.

The complete workflow is visually outlined in Figure 1. After generating the separate datasets, sequences from all classes were concatenated, shuffled twice, and split into 75% training data and 25% cross-validation data. This process ensured a robust and comprehensive training dataset capable of improving model performance and accuracy in detecting anomalous sequences.

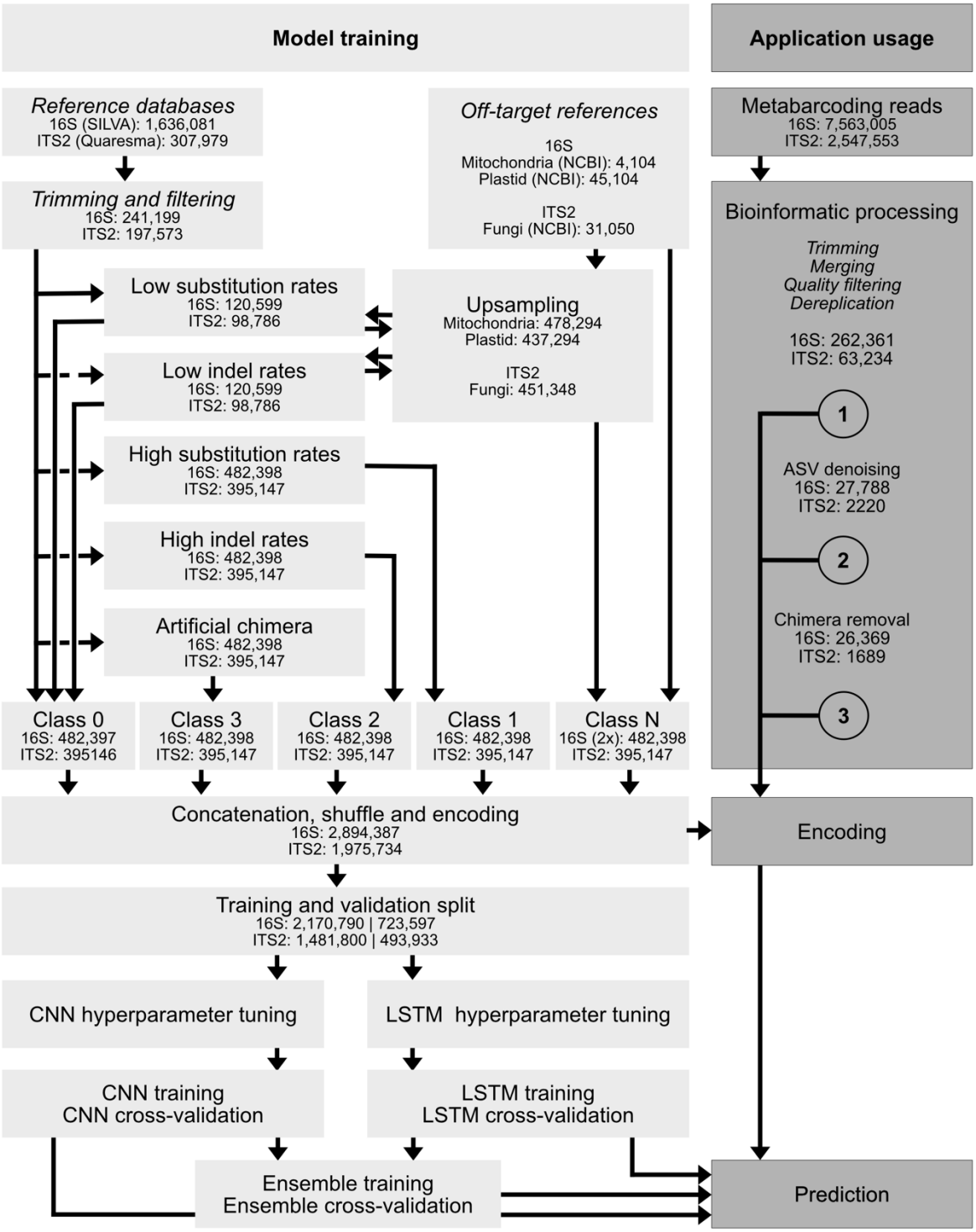

Figure 1: Workflow of MetAnoDe with the two possible pathways: left (light grey) side represents the training of new models, as here applied to ITS2 and 16S-V4. If pre-trained models are available, the right pathway (dark grey) is executed making predictions for query sequences. Pre-processing steps that need to be conducted prior to MetAnoDe application are highlighted in italic. Numbers represent volume of sequences used as input for the individual steps.

Model training

For model training and validation, Python 3.9 was used along with several essential modules: TensorFlow (2.16.1), Keras (3.3.3), Pandas (2.2.1), NumPy (1.23.5), Scikit-learn (1.4.2), BioPython (1.78), Matplotlib (3.8.4), and Keras-Tuner (1.4.7). The following procedure was applied independently for each marker:

Data preparation: Sequences in the training and validation sets were tokenized and encoded at the character level using tokens defined from the full, i.e. train and validation, dataset. The resulting sequences were end-padded with zeros and reshaped to match the requirements of the CNN and LSTM input layers.

Model architectures: The architecture of the CNN, LSTM, and ensemble models is illustrated in Figure 2. In all models, dropout layers were included to prevent overfitting and enhance the generalizability of the models. Sparse categorical cross-entropy was utilized as the loss function, and L2 activation regularization was applied. For the ensemble model, the output layers of the LSTM and CNN models were concatenated. Fully connected layers were then added to form an ensemble model.

Hyperparameter tuning: Hyperparameter tuning for the LSTM and CNN models was conducted using a hyperband search as implemented in *keras-tuner* (Li et al., 2018). Optimal model parameters were chosen based on validation accuracy.

Technical cross-validation: Models were evaluated in their performing by calculating accuracy, precision and recall of predictions in the validation set.

Training of the 16S and ITS2 models was executed on an Ubuntu 24.04 system equipped with a Ryzen 7 (16-core) CPU, 64 GB RAM, and a GEFORCE 4070TiS GPU, utilizing the GPU for processing in TensorFlow. Training and validation code was also tested for compatibility on Ubuntu 24.04 and MacOSX 12.3 without GPU support using a smaller test set.

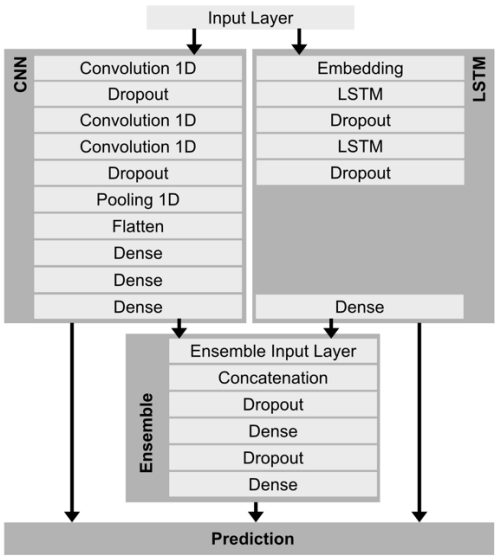

Figure 2: Layers of the individual models as defined in Tensorflow and Keras. All three models are used for predictions.

Independent real data prediction and manual validation

For the ITS2 and 16S-V4 markers, several own studies listed in Table 1 were selected based on the amplicon and sequencing strategy used and their compatibility with the models, their public availability, and their diverse biological contexts, geographical bioregions, and input material types. This selection aimed to test the generalizability of the method. Bacterial 16S-V4 sequencing libraries were prepared according to (Kozich et al., 2013), while plant ITS2 sequencing libraries were generated following (Sickel et al., 2015). Data obtained from other library preparation strategies might be equally used, given usage of the same primers as well as adapter and primer trimming. All data was sequenced either on Illumina MiSeq or HiSeq devices. More details are provided in the respective publications of studies and library generation. From each of these studies, 20 random samples were chosen for predictions. Data from these subsets were processed according to the pipeline outlined by (S. D. Leonhardt et al., 2022)(https://github.com/chiras/metabarcoding_pipeline) utilizing mainly VSEARCH (Rognes et al., 2016).

Table 1: Data origins that were used for real data prediction.

Marker	Sample type	Geographic origin	Reference
Bacterial 16S-V4	soil	Ecuador	(Garcia et al., 2024)
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)

Predictions were conducted at various stages of the metabarcoding pipeline (Figure 1) to determine the optimal integration point for the new tool within existing workflows. This approach aimed to evaluate the overlap and exclusivity between MetAnoDe classifications and those generated by alignment-based filtering methods. Additionally, it assessed whether applying MetAnoDe early in the pipeline would enhance the quality of subsequent processing steps and reduce overall runtime, or if it would be sufficient to apply MetAnoDe only to the final, fully processed data.

Finally, sequences underwent manual inspection using BLAST (Altschul et al., 1990) against the NCBI GenBank database (excluding environmental samples) (Benson et al., 2018). Classifications were validated by examining their taxonomy and alignments, as well as verifying the integrity of the underlying GenBank records. This inspection was applied to the classifications from all three models after VSEARCH merging, quality filtering and dereplication (Figure 1: insertion point 1), as well as the entire VSEARCH data analysis pipeline without MetAnoDe. Classifications from all four approaches were evaluated as valid, non-critical, or critical. Valid classifications were those confirmed through manual inspection. Non-critical classifications were those where anomalies were correctly identified, i.e., sequences flagged for removal, but not with the correct anomaly class. Critical classifications referred to instances where anomalous sequences were incorrectly classified as valid targets or valid sequences were incorrectly flagged for removal.

The optimal insertion points for the models were evaluated by examining the runtime of the entire pipeline when the models were applied at different stages. Computationally intensive tasks, such as denoising, can significantly benefit from a reduced number of sequences, making early-stage identification of anomalous sequences advantageous for reducing runtime (Coissac et al., 2012). Conversely, an increase in the number of sequences processed by the deep-learning models can lengthen their prediction runtime.

Results

Technical cross-validation

Overall, all models achieved high validation accuracy, exceeding 90% for both markers. Fungal sequences in the ITS2 models were predicted perfectly. However, predictions for mitochondria and plastid sequences as off-targets in 16S were less precise (low as 0.77) or with lower recall values (low as 0.71), respectively. During technical validation, individual models performed slightly better than the ensemble.

Table 2: Cross-validation results of all three models overall and for the individual classes. Within each cell, values are ordered as Ensemble | LSTM | CNN. Support indicates the sample size used for cross-validation.

Model	Class	Precision	Recall	F1-Score	Support
16S	0 positive	0.91 | 0.94 | 0.96	0.94 | 0.97 | 0.97	0.92 | 0.96 | 0.97	120,347 | 120,347 | 362,050
	1 high substitution	0.89 | 0.99 | 0.98	0.92 | 0.96 | 0.98	0.90 | 0.97 | 0.98	120,720 | 120,720 | 361,678
	2 high indel	0.94 | 0.97 | 0.98	0.94 | 0.98 | 0.99	0.94 | 0.98 | 0.99	120,502 | 120,502 | 361,896
	3 chimera	1.00 | 0.98 | 0.99	0.94 | 0.96 | 0.97	0.97 | 0.98 | 0.98	120,157 | 120,157 | 362,241
	4 mitochondria	0.80 | 0.77 | 0.78	0.93 | 1.00 | 0.98	0.86 | 0.87 | 0.87	120,478 | 120,478 | 361,920
	5 plastids	0.92 | 0.99 | 0.98	0.76 | 0.71 | 0.72	0.83 | 0.83 | 0.83	120,649 | 120,649 | 361,749
	Overall validation accuracy		0.90 | 0.93 | 0.94		722,853 | 722,853 | 2,171,534
Model	Class	Precision	Recall	F1-Score	Support
ITS2	0 positive	0.81 | 0.84 | 0.82	0.89 | 0.96 | 0.87	0.85 | 0.90 | 0.85	103,355 | 103,355 | 310,598
	1 high substitution	0.86 | 0.96 | 0.88	0.90 | 0.96 | 0.93	0.88 | 0.96 | 0.91	103,646 | 103,646 | 310,308
	2 high indel	0.98 | 0.98 | 0.98	0.92 | 0.96 | 0.95	0.95 | 0.97 | 0.96	104,014 | 104,014 | 309,940
	3 chimera	0.96 | 0.98 | 0.96	0.88 | 0.86 | 0.88	0.92 | 0.91 | 0.92	103,478 | 103,478 | 310,476
	4 fungi	1.00 | 1.00 | 1.00	1.00 | 1.00 | 1.00	1.00 | 1.00 | 1.00	103,538 | 103,538 | 310,416
	Overall validation accuracy		0.92 | 0.95 | 0.93		518,031 | 518,031 | 1,551,738

Independent real data prediction and manual validation

Predictions for the 16S independent dataset showed that none of the sequences were critically misclassified by the three deep-learning models, whereas the standard VSEARCH pipeline misclassified 8% of the sequences critically. Regarding non-critical classifications, the ensemble, LSTM, and CNN models had wrong assignment rates of 12.8%, 23.2%, and 24.0%, respectively, compared to 18.4% for VSEARCH. Keep in mind that these results represent proportions of unique read numbers, not reflecting their abundances.

It's important to note that the values reported for VSEARCH were obtained using an extended reference database that included mitochondrial and plastid sequences, which need to be removed in subsequent steps. Using a database that does not include such references resulted in an additional 23.2%, leading to a total of 31.2% critical errors.

For the ITS2 independent dataset, critical mistakes were observed across all methods: 6% for the Ensemble, LSTM, and CNN models, and 7% for VSEARCH. Non-critical errors rates were also nearly identical, with the Ensemble and LSTM models each at 11% and the CNN and VSEARCH methods both at 12%. In contrast to the cross-validation, during independent validation, the ensemble model outperformed the individual models for the 16S marker and classified with equal quality for ITS2. Overlap between errors in classifications of models and VSEARCH ranged between 35% and 60%. Overall overlap in all predicted classifications between the ensemble and VSEARCH was 78.4%.

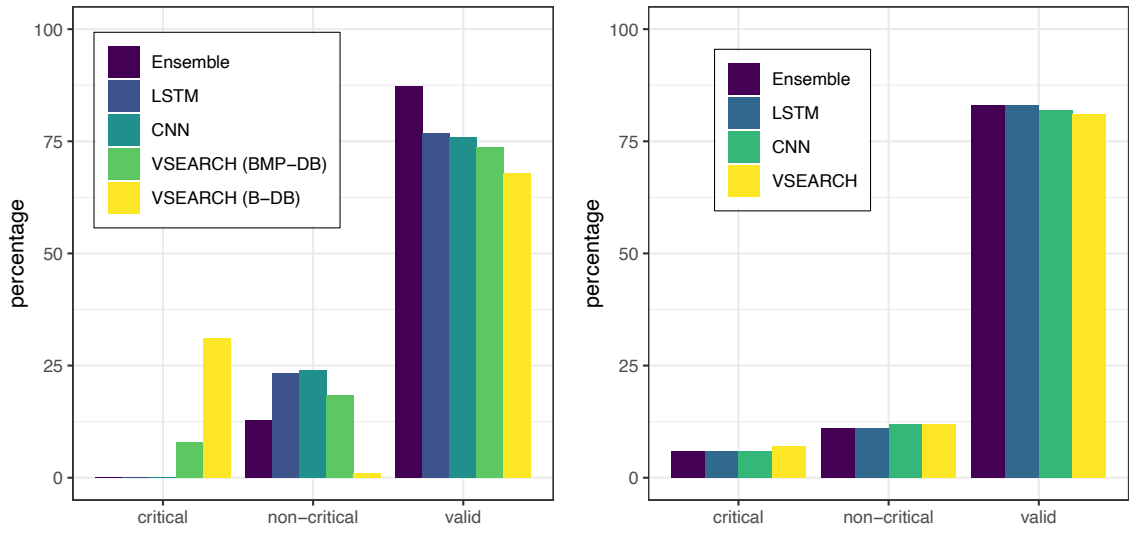

Figure 3: Manual validation results with real metabarcoding data for 16S-V4 (left), ITS2 (right). Critical errors are considered such that lead to either to missing to remove anomalous sequences or incorrect filtering of valid sequencing. Non-critical errors are considered such that detect an anomaly 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

Overall, there was a runtime improvement in both datasets (16S: 5.1%, ITS2: 3.4% when applying the models before denoising instead of after. There was no runtime difference in total runtime applying the models before or after chimeric filtering.

Discussion

In this manuscript, I introduce a novel workflow and software designed for identifying and classifying anomalous sequences in metabarcoding data using deep-learning models. The software offers two distinct pathways that can be executed either independently or concurrently with a single command. If only data predictions are required (i.e., utilizing pre-trained models provided or previously generated), no further training is necessary. This ensures fast and reliable processing, typically completing within minutes with high accuracy. In cases where no pre-existing model covers the target region, training can be conducted seamlessly within the same software environment. Detailed instructions on executing the software via the command line are provided in the corresponding dedicated sections below.

To the best of my knowledge, no other workflow or software has been proposed to date that addresses the detection of anomalous metabarcoding reads using deep-learning methodologies, e.g., see a recent review of available methods (Hakimzadeh et al., 2024). The approach is alignment-free and independent of sequencing quality statistics, providing a complementary perspective on data compared to existing tools.

Validation

The technical cross-validation reported very high accuracy, precision, and recall overall and for individual classes. Consequently, errors and artifacts simulated here, as well as off-target products, were detected and classified with high confidence. The validation using true target metabarcoding data, i.e., the independent validation, performed slightly less well but still maintained high accuracy. This slight decrease might be due to the fact that the data used for cross-validation was generated using random distributions for errors, indels, and chimeras, which may not entirely reflect real PCR and sequencing errors and artifacts (Schirmer et al., 2015). Additionally, non-functional copies of ribosomal DNA and pseudogenes were not included in training as a separate class and might fall out of patterns of random degradation (Porter & Hajibabaei, 2021).

For the 16S data, all mistakes made by the deep-learning models were non-critical, meaning anomalies were identified but not always correctly classified. In the ITS2 dataset, mistakes were comparable in frequency to those made by current alignment-based methods. It is noteworthy that mistakes (both critical and non-critical) often affected different sequences between VSEARCH (Rognes et al., 2016) and the models, highlighting the complementary nature of these approaches. This suggests that combining deep-learning and alignment-based methods can provide more comprehensive detection of anomalous sequences. Comparing the models, the ensemble performed significantly better on the 16S data than the other models, indicating it might be the best choice for sequence filtering.

Overall, this investigation confirms that a significant number of anomalous sequences, ranging from 25 to 30%, are not detected by classical methods. As a result, many non-target sequences are being included in biodiversity estimates and analyses that lack a thorough follow-up sequence analysis of residuals. This underscores the importance of identifying such anomalies, where the proposed deep-learning models can make a valuable contribution to identifying potential targets for removal.

Insertion points and runtime

Overall, I observed small runtime benefits when applying the models prior to denoising the data, but no significant difference when applied before or after chimera filtering. This is likely because denoising is a highly computationally intensive task, whereas chimera filtering is much less so, and thus a reduced data volume has a more substantial impact on runtime on the first. However, for early-stage application of the models (e.g., stages 1 and 2 in Figure 1) to achieve runtime benefits, it is necessary to directly remove anomalous sequences by the model, not just flag them (optional parameters). This approach is similar to current alignment methodologies that involve direct removal and might be suitable for fully automated workflows. Given that all methods, both alignment-based and deep-learning, have inherent flaws, it may be advisable to work with flagged sequences when working with new data, followed by their subsequent inspection and potential removal. This corresponds to a flagging only insertion at point 3 in Figure 1, where only the residual sequences that have passed all previous filters are considered for deep-learning predictions. Subsequently, flagged sequences can be manually inspected within this comparatively smaller volume of final sequences.

Predictions using pre-trained models

Predictions can be promptly generated using the pre-trained models available in the repository. This corresponds to the dark grey workflow in Figure 1. Adapter as well as primer sequences however need to be removed from data prior to analysis to match the model, as this varies between different amplicon library generation strategies. The software is called by the command line, requiring only the inclusion of the pre-trained models alongside the query data.

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

By default, the software retains all sequences in the query data but annotates them based on their classification from each of the three models in the output. However, an option for sequence removal is also available. Additional customizable options can be explored by running the script without additional arguments. The software generates two output files stored in the 'predictions' subfolder: a comma-separated file (CSV) presenting classification results in tabular format, and a second file containing flagged sequences (or a subset if removal is opted) in FASTA format.

All dependencies, as specified above, need to be installed for proper execution of the code. Installation guidelines for these dependencies are provided in the repository. The script supports both GPU and CPU data processing, with notable runtime improvements

achievable when utilizing GPUs. The reported predictions were conducted on Ubuntu 24.04 with GPU support, but have also been tested on Ubuntu 24.04 and MacOSX 12.3 without GPU support.

Predictions with other target regions and new training of models

The workflow is entirely automated and can be adapted for different target regions, however necessitating complete training of models from scratch in such cases. In case no pre-trained model is specified, every required component of the light grey workflow depicted in Figure 1 is executed, encompassing data pre-processing, hyperparameter optimization, and final model generation. Milestones are set during execution, allowing to skip parts in case they are already present when needed.

To initiate the process, correctly trimmed and deduplicated reference sequences must be provided using the parameter **-db <ref.fasta>**. Optionally, multiple known off-target amplicon regions can be incorporated using **-ot <ot1.fasta>,<ot2.fasta>,<ot3.fasta>[,...]**, ensuring each type is included separately in the model. A designated model name must be specified to consolidate all pertinent models and parameters.

An illustrative example of the software's call that involves both training models on new data and predicting query data in a unified execution:

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

Once a model is trained, it can be reused for new data by specifying the corresponding model name. The script supports both GPU and CPU processing; however, it is important to note that CPU processing significantly extends the duration of model training. Therefore, for efficient training, GPU utilization is strongly recommended here. There is no strict limit on the number of reference sequences and their lengths or off-target classes that can be incorporated. However, the memory required for encoding and training could potentially be a constraint depending on the available hardware resources.

Conclusions

In conclusion, the introduced workflow and software effectively identify and classify anomalous sequences in metabarcoding data using deep-learning models. With high validation accuracy, these models offer a complementary perspective to existing alignment-based methods. Ultimately, the deep-learning models, in synergy with traditional methods, can significantly enhance the detection of anomalous sequences, reducing the inclusion of non-target sequences in biodiversity estimates. This approach underscores the importance of integrating multiple methods for a comprehensive analysis, where deep-learning models add substantial value by identifying potential targets for removal.

Acknowledgments

I appreciate funding by the LMU Munich by the Excellence Fonds (number VIII.2/bk 865104-8) providing necessary means for this study.

Data and code availability

All code and data as used for this study is available on GitHub (https://github.com/chiras/MetAnoDe). Raw metabarcoding data is obtainable from the individual studies as listed in Table 1. Processed data at the stages investigated in this study are included in the repository.

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., 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

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