Received: 31 August 2024 | Accepted: 21 February 2025

DOI: 10.1111/2041-210X.70033

Methods in Ecology and Evolution

RESEARCH ARTICLE

Conservation, Ecology and Artificial Intelligence: Advances and Symbiotic Solutions

ORDNA: Deep-learning-based ordination for raw environmental DNA samples

Théophile Sanchez[1,2] | Steven Stalder[3] | Letizia Lamperti[2,4] | Sébastien Brosse[5] | Aline Frossard[1] | Flurin Leugger[1,2] | Romane Rozanski[1,2] | Shuo Zong[1,2] | Stéphanie Manel[4,6] | Laura Medici[1] | Fabienne Kuhn[1] | Xingguo Han[1] | Adrien Mestrot[7] | Camille Albouy[1,2] | Michele Volpi[3] | Loïc Pellissier[1,2]

[1]Ecosystems and Landscape Evolution, Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Birmensdorf, Switzerland; [2]Ecosystems and Landscape Evolution, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland; [3]Swiss Data Science Center, ETH Zürich, Zürich, Switzerland; [4]CEFE, University of Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France; [5]Laboratoire Évolution et Diversité Biologique (UMR5174 EDB), CNRS, IRD, Université de Toulouse 3 Paul Sabatier, Toulouse, France; [6]Institut Universitaire de France, Paris, France and [7]Institute of Geography and Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

Correspondence
Théophile Sanchez
Email: email@theophilesanchez.fr

Funding information
Swiss Data Science Center (SDSC), Grant/Award Number: C21-05 DNAi; ANR-SNF, Grant/Award Number: 205556 ShifteDNA

Handling Editor: Ruth Oliver

Abstract

1. Environmental DNA (eDNA) metabarcoding has revolutionized biodiversity monitoring, offering non-invasive tools to assess ecosystem health. The complexity of eDNA metabarcoding data poses major challenges for conventional ordination methods in understanding assemblage similarities and assessing biodiversity patterns.
2. Here, we introduce ORDNA (ORDination via Deep Neural Algorithm), a new deep learning method tailored for eDNA sample ordination. Leveraging artificial neural networks, ORDNA processes raw sequences from eDNA samples directly, bypassing potentially biased and cumbersome expert-based bioinformatic steps. The method is trained with a contrastive self-supervised learning approach, the triplet loss, to derive a two-dimensional representation of eDNA samples based on their read composition.
3. We apply ORDNA to four distinct eDNA datasets, demonstrating its robustness and superiority over traditional ordination techniques in capturing and visualizing ecological patterns.
4. Our results underline the potential of deep learning in advancing eDNA analysis, with ORDNA serving as a promising tool for more accurate and efficient biodiversity assessments.

KEYWORDS
deep learning, eDNA, sample ordination, self-supervised learning

Théophile Sanchez and Steven Stalder contributed equally to this work.

1 | INTRODUCTION

Biodiversity refers to the variety of life on Earth, encompassing the multitude of assemblages and species, and the genetic diversity within them. The aim of biodiversity science is to measure the richness and complexity of biological organisms in a given place and to understand the abiotic and biotic drivers that shape them (Purvis & Hector, 2000). Measuring biodiversity is crucial for understanding the health of ecosystems, their resilience to environmental changes and their capacity to provide services to human societies (Mace et al., 2012). Summarizing biodiversity into numbers is challenging due to the complex ecological relationships between species, whose co-occurrence patterns result from interactions between them and with their surrounding abiotic environment (Marshall et al., 2020). Despite these difficulties, quantifying biodiversity is essential for informed conservation and management efforts (Strange et al., 2024), as it provides a baseline for assessing environmental changes (Dornelas et al., 2013). In particular, essential biodiversity variables (e.g. allelic diversity, taxonomic diversity or habitat structure) are intended to summarize ecosystem complexity into numbers reflecting diversity or composition (Pereira et al., 2013). These tools are highly valuable to guide and evaluate management efforts aimed at sustaining the diverse functions and benefits of biodiversity (Jetz et al., 2019). However, species population indicators that address the distribution or composition of assemblages remain scarce. Therefore, the numerical characterization of biodiversity should be supported by methodological advancements in data collection, which require more efficient analytical approaches. These new approaches will in turn enhance the availability of such data for ecological management and policy decisions.

Environmental DNA (eDNA) has emerged as a revolutionary tool for monitoring and assessing biodiversity (Deiner et al., 2017). This technology enables the capture and sequencing of DNA fragments shed by organisms into their environment, providing a snapshot of the biological communities present in a particular habitat without the need for direct observation, thereby causing minimal disturbance (Polanco Fernández et al., 2021). Utilizing different primers allows targeted DNA sequencing of specific barcodes of a more or less specific group of organisms, enabling their identification within water, soil and air samples. This makes eDNA a fast, cost-effective, non-invasive and systematic method to assess the health status of an ecosystem (Cordier et al., 2019). eDNA datasets typically comprise millions of DNA reads per sample, and errors may be introduced during various steps of the sequencing process, such as amplification or base calling errors. These errors can, in turn, introduce biases or inaccuracies in the derived data, affecting the subsequent ecological analysis and potentially leading to misinterpretations of ecological signals (Burian et al., 2021). Moreover, ecosystem properties are embedded in complex non-linear relationships within and between eDNA samples. This unprecedented data complexity and noisiness represent a major challenge for conventional ordination methods developed in an era when ecological datasets were comparatively less complex (e.g. species occurrence matrices). eDNA studies usually circumvent these issues by performing several bioinformatic steps to transform the reads into data exploitable by ordination methods (Marques et al., 2020). Specifically, bioinformatic pipelines use denoising (without clustering) to produce amplicon sequence variants (ASVs) or combine denoising with clustering to produce molecular operational taxonomic units (MOTUs). From these, one can then create a table of taxa occurrences (up to the species level) by comparing the obtained sequences to genetic reference databases (Mathon et al., 2021). Finally, these tables can be converted to appropriate distance or (dis)similarity metrics before ordination (Lamperti et al., 2023). Each of these steps has the potential to introduce additional biases, inaccuracies and loss of information. For instance, preferentially filtering or clustering certain sequences might affect the representation of the true biological diversity and abundance. In the case of detecting species presences, the incompleteness of reference databases can lead to important information loss (Burian et al., 2021). This underscores the necessity of adapting or developing methods tailored for eDNA metabarcoding that are capable of efficiently extracting and processing relevant information directly from raw data while being robust to noise.

The study of biodiversity typically relies on various measurements of species abundances, presences/absences or traits, which are often summarized using ordinations to facilitate their ecological interpretation (Borcard et al., 2018). Ordination methods are particularly valuable for analysing beta diversity, as they simplify complex ecological datasets into interpretable dimensions, revealing patterns and gradients in species composition and their relationships with environmental variables (Dray et al., 2012). They enable the discernment of patterns and gradients in species composition and environmental variables, offering a robust framework to explore the multidimensional nature of ecological interactions and their spatial organization (Borcard et al., 2004). Ordination reduces raw ecological data into a few interpretable dimensions that represent the main gradients of variation, thereby facilitating the visualization and understanding of underlying ecological processes and informing strategies for ecosystem management and conservation (Arranz et al., 2022; Leprieur et al., 2016). With a rich history of development over many decades, ordination methods, such as principal coordinates analysis (PCoA) (Gower, 1966) and non-metric multidimensional scaling (NMDS) (Kruskal, 1964), have played a central role in ecological analyses. They are not directly applicable to raw eDNA data, however, due to their complex representation, which lies in a high-dimensional space. To apply traditional ordination methods, eDNA sequence reads first need to be transformed into MOTU or ASV tables (Cilleros et al., 2019) by following a bioinformatic workflow. Since this process involves many complex steps and user inputs, the results can suffer from biases and information loss (Lamperti et al., 2023). Therefore, developing alternative ordination approaches that directly use raw reads could improve the final dataset representation and better automatize the analysis of the growing number of eDNA datasets.

Artificial neural networks (ANNs) excel at identifying complex patterns within large datasets, often outperforming traditional statistical methods in tasks such as classification, regression, clustering and data generation. ANNs rely on computational graphs structured as sequences of mathematical functions, called layers. An entire ANN architecture typically encompasses between thousands and billions of 'trainable' parameters that make up these layers, where each of them can be updated to minimize a loss function measuring the fit of the model's output to training data. To tailor an ANN to a specific task, one must define a loss function that quantifies the network's performance on a training dataset for the specific problem. Then, optimization based on backpropagation and gradient descent adjusts the trainable parameters to decrease this error. In essence, deep learning algorithms (with 'deep' referring to ANNs with several layers between input and output) are intended to automatically optimize complex non-linear functions parameterized by learnable weights and biases. Since the advances in parallel computing brought about by graphics processing units (GPUs), ANNs have proven to be a valuable tool in many domains involving high-dimensional data, such as computer vision, natural language processing and signal processing (Borowiec et al., 2022). Consequently, ANNs have also become a go-to tool for processing -omics data, approaching tasks such as protein binding site prediction, deducing protein structures and deciphering population genetics (Battey et al., 2020; Sanchez et al., 2021; Yan & Wang, 2023). These breakthroughs have also led ecologists to employ ANNs to predict species habitat distributions, analyse audio signals and automate the annotation of images from remote sensing or camera traps (Lang et al., 2023; LeBien et al., 2020; Vélez et al., 2023). These examples illustrate that, alongside their ability to extract information from complex data, ANNs are also flexible regarding the type of data that can be processed. That is why ecologists have also leveraged them to extract information from eDNA metabarcoding data (Flück et al., 2022; Lamperti et al., 2023). Finally, ANNs have facilitated considerable advances in deep metric learning, where one seeks to optimize an embedding space that follows a distance metric by pulling similar data points together and pushing dissimilar points further apart with respect to that metric (Kaya & Bilge, 2019). This naturally leads to an ordering of the data points along the embedding axes, which makes deep learning methods well suited for ordination, where the condensed representations of data should also result in an ordering along the most important gradients of variation.

Here, we leverage the ability of ANNs to process high-dimensional data with complex underlying structures to summarize the main ecological information in raw eDNA samples, bypassing possibly biased bioinformatic steps. Since eDNA datasets are unlabelled collections of raw sequences, the problem of ordination cannot be approached with standard supervised learning based on explicit targets. We therefore use a relatively new family of techniques grouped under the term self-supervised learning (SSL), which allows us to define suitable learning objectives based on automatically created targets from unlabelled data. *Contrastive* SSL achieves this by defining similar or dissimilar data points from prior knowledge of the unlabelled data, as opposed to relying on an explicit human-annotated definition. By focusing on the dissimilarities and similarities between samples, the SSL framework inherently aligns with the ecological concept of beta diversity, which emphasizes variation in species composition between communities. Here, we present the ORDNA (ORDination via Deep Neural Algorithm) method, a SSL algorithm for the ordination of eDNA samples. Based on a contrastive SSL approach called triplet loss (Ge et al., 2018), ORDNA provides a representation of eDNA samples in a two-dimensional space based on the (dis)similarity of their read compositions. In our setting, each data point is an eDNA sample, and we consider two eDNA samples to be similar when they have similar read compositions. Analogous to ordination based on MOTUs or species detection, we use read composition as a proxy for the species community at a given sample location and time. We benchmark the method using four distinct eDNA datasets: fish assemblages recovered from water eDNA from (1) French Guiana and (2) Brittany and eukaryote assemblages recovered from soil samples from (3) several Swiss forests and (4) a polluted industrial area located in the town of Visp in south-western Switzerland (Coutant et al., 2023; Frossard et al., 2018; Rozanski et al., 2022). In particular, we ask the following questions:

1. How does the structure of ORDNA correlate with traditional post-bioinformatic ordinations? We compare ORDNA to PCoA applied after bioinformatic processing to assess their similarity in reducing the dimensionality in the original dataset. We expect that ORDNA distances show a good correlation with PCoA distances, but that gradients are better represented by the non-linear embedding provided by ORDNA.
2. Does ORDNA extract more relevant ecological information from the eDNA metabarcoding data? We extract ecological information from the different datasets and test how well it correlates with ORDNA and with the two principal dimensions of PCoA. Because of its non-linearity, we expect ORDNA to result in higher Pearson correlations with ecological data representing the environments in which the samples were taken.
3. Can we use ORDNA to define an ordination space in which we can ordinate newly collected data to obtain new information? We use a temporal dataset collected in Brittany to train ORDNA; we define an ordination space based on the samples from 1 year and ordinate the data from the samples collected the following year. We expect that the ORDNA structure is mainly stable over time, so that the reprojected points of the second year fall close to those from the first year.

By applying our method to four different eDNA datasets, encompassing marine, freshwater and soil samples, we demonstrate that using ORDNA brings a key improvement in accuracy over traditional methods in various settings, underscoring its potential as a robust tool for eDNA analysis. Moreover, we show the potential of using ORDNA as a predictive method, where it is possible to extract information from raw eDNA samples by reprojecting them in a pre-calibrated ordination space.

2 | MATERIALS AND METHODS

2.1 | Self-supervised ordination of eDNA samples

Similar to traditional ordination methods, our goal with ORDNA is to transform raw eDNA samples into low-dimensional embeddings that retain information about the relationship between ecological communities. We aim to ensure that samples with similar eDNA read compositions have a similar low-dimensional embedding, while dissimilar ones are far apart in that space. To learn meaningful representations, we employ a contrastive SSL objective. Following that objective, we optimize an ANN to find appropriate representations of eDNA samples, resulting in an ordering along the embedding axes.

The contrastive SSL approach requires that we define three sets of raw eDNA reads and employ them as so-called *anchor* (*a*), *positive* (*p*) and *negative* (*n*) examples (detailed below). Our model transforms each set into its embeddings, that is z_a, z_p and z_n. We then use a triplet loss function $\mathscr{L}$ that forces the model to move the embeddings of the *anchor* and *positive* closer together in the embedding space than those of the *anchor* and *negative*, by some margin α:

$$\mathscr{L}(z_a, z_p, z_n) = \max\left(\| z_a - z_p \| - \| z_a - z_n \| + \alpha, 0 \right). \quad (1)$$

Lacking an explicit metric for similarity between eDNA samples, we employ SSL to define suitable *anchor*, *positive* and *negative* examples directly from our unlabelled data. To this end, we assume that two distinct but appropriately large subsets of sequences from the same eDNA sample should result in very similar coordinates in the latent ordination space, as they are expected to represent the same species community. While these two groups of sequences are unlikely to be identical, they are expected to share more information with each other than with samples taken at different locations and, therefore, different read compositions. Under these assumptions, we develop an application of the triplet loss (Schroff et al., 2015), where at each training step the *anchor* and *positive* each consist of 1000 different randomly sampled DNA sequences from the same eDNA sample, while the *negative* is a subset of 1000 sequences sampled from any other eDNA sample. Note that the specific number of 1000 sequences per element is arbitrary and can be adapted based on memory requirements, while keeping in mind that small subsets might not contain enough information for accurate ordinations.

We then implement an ANN architecture capable of transforming groups of sequences into meaningful latent representations. We train it over many examples and iterations, to minimize the triplet loss function $\mathscr{L}$, with the AdamW optimization algorithm (Loshchilov & Hutter, 2017). In addition to the triplet loss function, we add standard L2 regularization to prevent the model from learning representations with arbitrarily large norms, as this could help it overcome the margin penalty. We aim to avoid this scenario and compel the model to address the margin penalty by creating meaningful lower-magnitude latent representations, effectively positioning *negative* examples sufficiently far away. Our final loss function becomes:

$$\mathscr{L}(z_a, z_p, z_n) = \max\left(\| z_a - z_p \| - \| z_a - z_n \| + \alpha, 0 \right) + \lambda \left(\frac{\| z_a \| + \| z_p \| + \| z_n \|}{3} \right), \quad (2)$$

where λ is an additional tuneable parameter. Through the minimization of this final loss function, we anticipate our model to converge to meaningful embeddings for all eDNA samples. The final representation of any eDNA sample corresponds to the average of the representations from all the subsets of 1000 sequences, ensuring that every sequence is used exactly once. Upon convergence, we expect that similar samples will be clustered closely together while very different ones will be far apart in the embedding space. These embeddings can then be used for ordination or further fine-tuning for any downstream

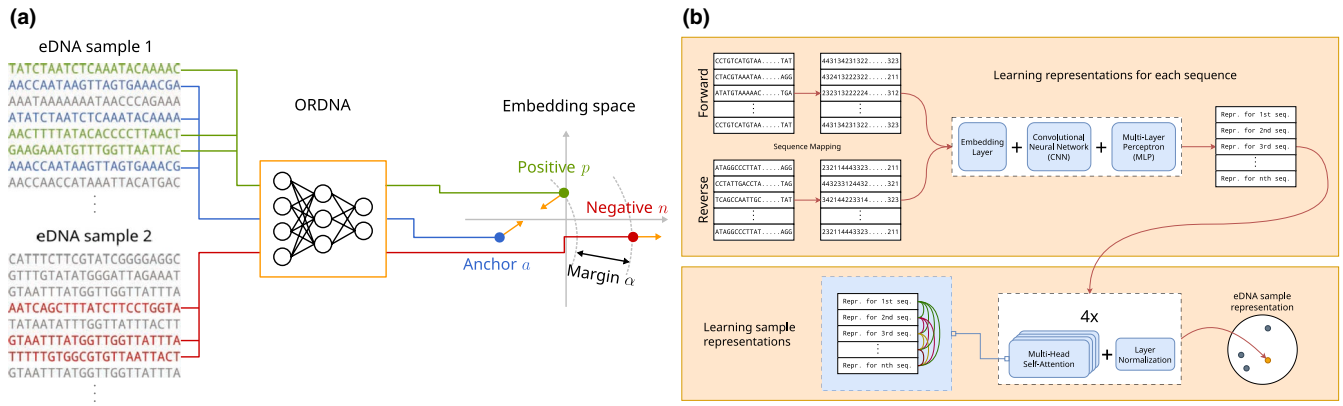

FIGURE 1 Overview of ORDNA architecture and training procedure. (a) Two subsets of sequences are randomly selected from Sample 1, forming the *anchor* and *positive* set. Simultaneously, a third subset is randomly chosen from Sample 2 to serve as the *negative* set. They are then projected into the embedding space by the ORDNA neural network. The final step involves the triplet loss, which the neural network optimizes by drawing the *anchor* and *positive* examples closer in the embedding space while pushing the *negative* set further away. (b) Overview of ORDNA's neural network architecture for transforming groups of sequences into embeddings. Forward and reverse sequences undergo initial encoding before being fed into a sequence of convolutional and fully connected layers. The representations of these sequences are further processed through a permutation-invariant, multi-head self-attention layer to obtain the final representation of the sample (see Supporting Information Section A.2 for more details about the implementation).

supervised learning task on the same dataset. See Figure 1a for an overview and Supporting Information Section A.1 for details about the training procedure.

2.2 | Neural network architecture

We develop an ANN that can learn numerical representations of eDNA samples by transforming a subset of raw reads into numerical embeddings that minimize the previously defined triplet loss. The raw nucleotide base characters are first converted into specific numerical values, making them suitable inputs for our neural network (see Figure 1b for a simplified overview of our network). The first layer then learns short vector representations for each nucleotide base. Note that using fixed representations, for example one-hot encodings, for the nucleotide bases would also be possible, but making them learnable potentially facilitates training and increases the expressivity of the network. Then, each sequence is fed to convolutional layers that accumulate information over the entire sequence, and a multi-layer perceptron (also known as fully connected layers) combines information from the forward and reverse reads into a single representation for each sequence. Finally, these sequence representations are passed through a stack of multi-head self-attention layers until we retrieve a final representation for our full input. These self-attention layers allow the model to put more or less weight on different sequences or combinations of sequences (e.g. the combined presence of two species that only co-occur in a specific eDNA sample) by processing the sequence embeddings pairwise. Moreover, they are invariant to permutations of the input sequences, an advantageous feature considering our focus is solely on the overall species composition, rather than the specific order of sequences, which is arbitrary. More details about the implementation are provided in Supporting Information Section A.2.

2.3 | Soil and water eDNA datasets

To evaluate the performance of our approach with different types of eDNA data, we compared ordination approaches with four different empirical eDNA datasets from diverse locations around the globe (Figure 2) and with different targeted species. The first dataset consisted of 85 freshwater samples from French Guiana and targeted fishes through amplification of the 12S rRNA 'teleo' gene fragment. The second dataset comprised 98 marine samples collected in the Atlantic Ocean around Pointe du Raz (Brittany, France) from 2020 to 2022 (34 samples in 2020, 30 samples in 2021 and 34 samples in 2022). The 12S rRNA 'teleo' gene fragment was also amplified for this dataset. The third dataset consisted of 180 forest soils collected from nine locations in Switzerland. At each location, soil samples were obtained from forest reserves (no wood exploitation) and adjacent managed forests (where wood is commercially exploited). The prokaryotic community was targeted using 16S rRNA V3–V4 primers. The fourth dataset comprised 200 soil samples collected from the Rhone valley around the town of Visp (Switzerland), where soils have been exposed to different levels of mercury pollution. The

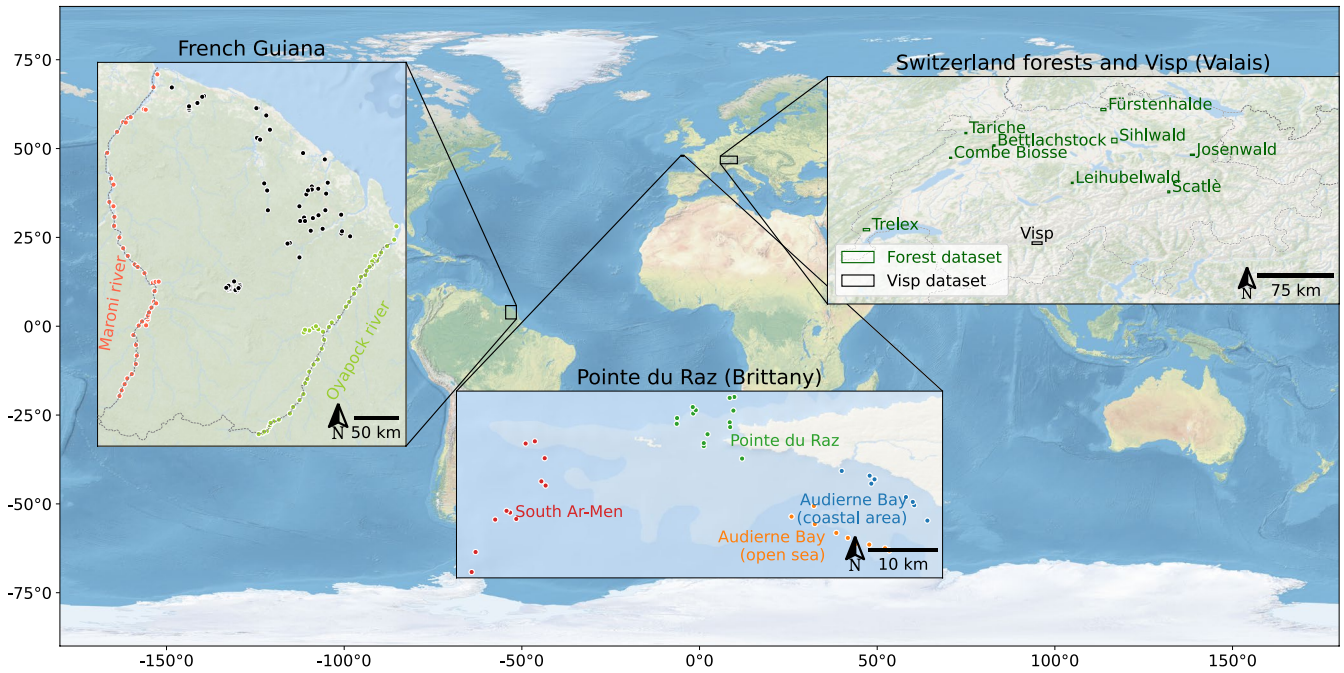

FIGURE 2 Sampling locations of the four eDNA datasets used in this study: Marine samples from Pointe du Raz in Brittany (98 samples), freshwater samples from French Guiana (85 samples), forest soils across Switzerland (180 samples from nine locations) and mercury-polluted soils from Visp in Switzerland (200 samples). See Figures A5 and A6 in Supporting Information for the exact locations of the Swiss forest and Visp samples.

same 16S rRNA V3–V4 primers were used to assess the prokaryotic community. In Supporting Information A, we explain in detail how the samples were collected and sequenced, and we describe the data processing phases of each dataset.

2.4 | Principal coordinates analysis baseline

To compare ORDNA with established ordination techniques, we applied PCoA, a multidimensional scaling method, to each of the datasets. Unlike ORDNA, which processes raw sequence data, PCoA necessitates first computing a distance matrix that describes the compositional differences between eDNA samples and then representing them in a multidimensional space. The approach applied in this study to characterize eDNA sample composition varied across the datasets, contingent on the available information about the taxa and the environment under study. For the freshwater dataset from French Guiana rivers and the marine dataset from Pointe du Raz in Brittany, we performed MOTU detection and taxonomic assignment using obitools (Boyer et al., 2016). In the case of both soil datasets—forest soils across Switzerland and mercury-polluted soils in Visp, we quantified the ASV counts in each sample using DADA2 (Callahan et al., 2016). Subsequently, we employed the Bray–Curtis dissimilarity, as implemented in the *ecodist* package (Goslee & Urban, 2007) in R V4.2.2 (R Core Team, 2024), to generate a distance matrix reflecting the relative abundance of taxa or ASVs in each sample. PCoAs were computed using the wcmdscale function from the *vegan* package in R (Dixon, 2003), and the first two axes were retained for visualization. We used these PCoAs as a benchmark to evaluate ORDNA's performance in extracting ecological information from the raw eDNA data.

2.5 | Evaluation of ORDNA and comparison with PCoA

The contrastive SSL approach used to learn ORDNA's network weights ordinates the samples based on their similarity in raw read composition, which depends not only on environmental factors but also on the potential biases of the data collection and the intrinsic stochasticity of the species captured by eDNA sequencing. Therefore, we evaluated ORDNA performances by assessing them empirically with the four datasets described previously. We then visually and numerically compared ORDNA's ordinations obtained from these datasets with the ordinations generated by the PCoA baseline method. We assigned colours to each sample based on a continuous colour wheel centred at the origins of the two methods' ordination spaces, where the angle corresponded to the hue, and we reported the colours on the geographic map. Next, we calculated the Pearson correlation coefficient r between the Euclidean pairwise distances of samples in the ORDNA embedding and the pairwise distances in the PCoA ordination space.

2.6 | Comparison of the embeddings and associations with spatial and environmental distances

Ordination methods involving dimensional reductions should retain relevant ecological information, and thus embedding axes are expected to correlate with underlying abiotic environmental drivers of assemblage composition. For each dataset, we related the composition distance from ORDNA and the baseline PCoA to geographic distances, assuming that geographic closeness approximates similarities in ecological constraints. First, for each dataset, we computed the geographic distance between sampled points in terms of Euclidean distance, except for the French Guiana dataset of freshwater fish, where we used distances following river channels. We correlated the distances between points in the embedding space with those in the geographic space. Second, we tested the association between embedding distances and key environmental parameters related to each dataset. For the dataset of freshwater fish, we considered both temperature and the distance to the sea as environmental parameters, since the environmental parameters of rivers primarily shift from upstream to downstream. For the soil datasets from Switzerland, we used soil pH. In addition, for the contaminated area around Visp, we used soil mercury concentrations. Details on the soil acidity and pollution measurements at each sample location are presented in Supporting Information A. For the Brittany marine fish dataset, we extracted the annual sea surface temperature for each of the sampled sites. For both geographic and environmental parameters, we computed the Pearson's correlation r between the Euclidean pairwise distances in the embedding and either geographic or environmental distances. To further evaluate these associations, we performed Mantel tests to assess the statistical significance of the correlation, and reported the values on the figures when they were not significant (i.e. >0.01).

2.7 | Re-projection of sampled points across years

We applied ORDNA to unseen data to evaluate the method's ability to relate new samples to previous ones based on a pre-trained embedding. We used the Brittany marine dataset, where eDNA samples were collected at the same locations during the years 2020, 2021 and 2022. We first selected the data for the first year (2020) to learn the embedding space and then used the trained model to project the samples from the following years (2021 and 2022), which were not used for calibration. Since the locations of the sampling points did not strictly match across years, we grouped them into four areas for the analysis (Figure 2: Pointe du Raz, Audierne Bay [open sea], Audierne Bay [coastal area] and South Ar-Men). We investigated whether the model trained on the data from 2020 could accurately associate the samples collected in 2021 and 2022 with their area of origin. We computed the percentage of attributions to the correct area by the embedding. This allowed us to evaluate ORDNA's robustness when ordinating new samples in a pre-configured space in order to compare them with earlier samples.

3 | RESULTS

3.1 | Comparison of the structure of ordinations obtained with PCoA and ORDNA

We compared the organization of the samples in the ordination and geographic spaces between the baseline method (PCoA) and ORDNA. The Pearson correlations between the embedding pairwise distances of PCoA and ORDNA vary from 0.296 to 0.761, considering the first two axes (Table 1). For all datasets except the Brittany dataset, ORDNA arranges the samples in a circular pattern. Similarly, PCoA reveals a horseshoe pattern in all datasets except the French Guiana one (Figure 3; Figures A2, A5 and A6 in Supporting Information). Although the structures of the clusters in the embedding space match globally, there are discrepancies between the two methods in each dataset. For the French Guiana dataset, the ordination generated by ORDNA in Figure 3 shows a more gradual transition from the source to the sea for the Maroni River, which translates to a lower correlation between the two methods' ordinations for this river (0.296 in Table 1). Furthermore, for the Maroni River, ORDNA exhibits a distinct separation between inland samples and those located nearer to the coastal area, in contrast to the PCoA ordination. For the Brittany dataset (Figure A2 in Supporting Information), the two methods give similar results, clustering together samples located in the open sea (i.e. south-west of

TABLE 1 Pearson correlation coefficient *r* between pairwise embedding distances from principal coordinates analysis (PCoA) and ORDNA, and the pairwise geographic distances of the sampling locations.

	Swiss forests	Visp	Brittany (2021–2022)	French Guiana	
				Oyapock	Maroni
PCoA	0.211	0.027	0.421	**0.202**	0.397
ORDNA	**0.265**	**0.189**	**0.509**	0.178	**0.446**
PCoA vs. ORDNA	0.718	0.356	0.533	0.761	0.296

Note: Bold values indicate the approach that yields the highest correlation between embedding and geographic pairwise distances. The last row contains the correlation coefficients between the embedding distances of the PCoA and the embedding distances of ORDNA. For the Brittany dataset, ORDNA trained with data from 2021 and 2022 is compared with PCoA on data from the same years. For the French Guiana dataset, we retained only the data sampled from the Oyapock River, Maroni River and their tributaries for comparative analysis and calculated geographic distances based on the river network.

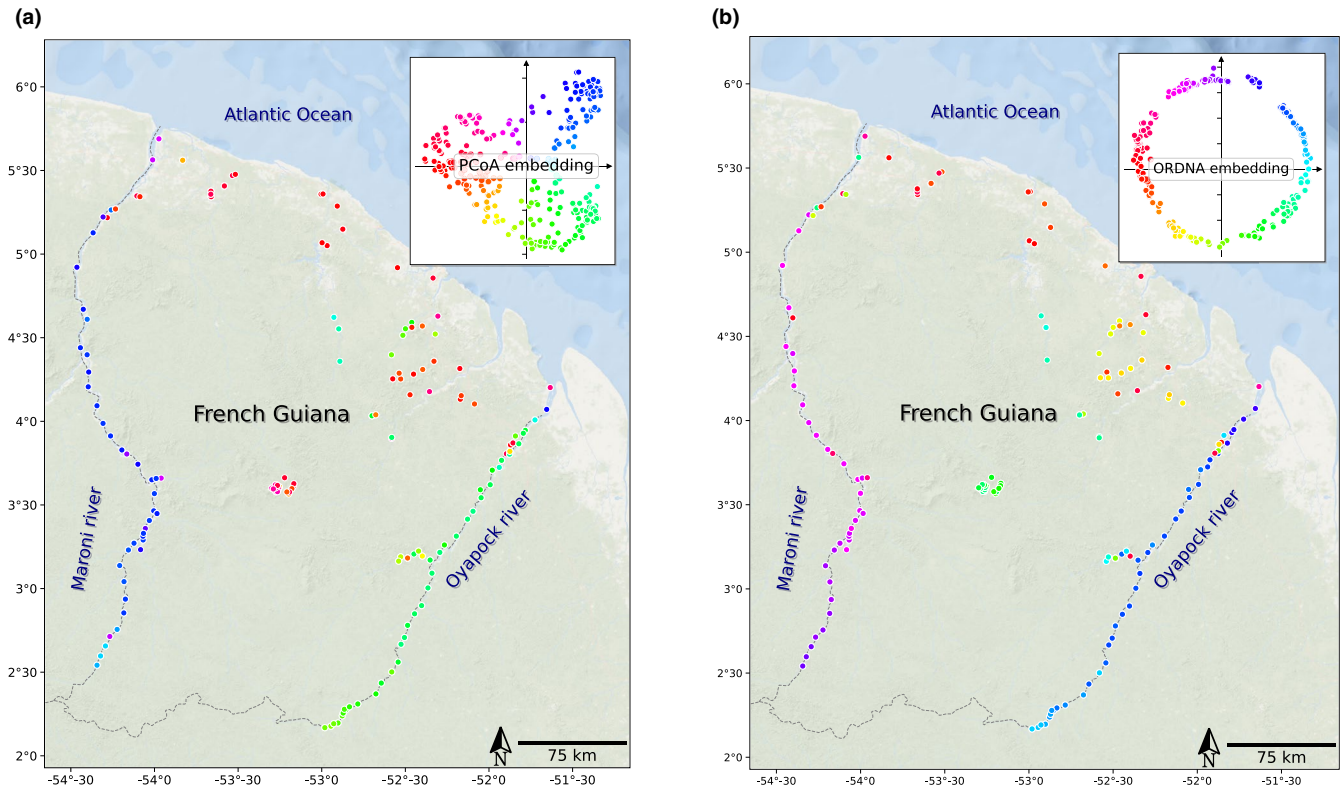

FIGURE 3 Ordination obtained with principal coordinates analysis (PCoA) on Bray–Curtis distances computed from detected species (a) and ORDNA using raw reads (b) for the French Guiana freshwater dataset. The colours of the sample points on the map correspond to their positions in the embedding space.

Audierne Bay and far south of the Ar-Men area), even though ORDNA was trained only with data from 2020 while the PCoA used data from all the years. For the soil samples from Swiss forests (Figure A5 in Supporting Information), both methods distinguish between managed and reserve forests as two different clusters for the Sihlwald, Fürstenhalde and Combe Biosse locations. Only PCoA distinguishes between the managed and reserve forests in the Scatlè location, and only ORDNA distinguishes between them in the Josenwald location. Finally, for the dataset from Visp (Figure A6 in Supporting Information), only ORDNA strongly clusters the samples on the west side of the map, downstream of multiple industrial facilities.

3.2 | Geographic and environmental distances

We evaluated the relationship between the distances among sample embeddings and their corresponding spatial and environmental distances. Our analysis reveals a systematically stronger correlation between the pairwise distances of the embeddings generated by ORDNA and the geographic distances (Table 1) compared with PCoA, except for one test case involving data from the Oyapock River in French Guiana (r value of 0.202 for PCoA and 0.178 for ORDNA). For example, the correlation for the Visp dataset is much lower for PCoA (0.027) than for ORDNA (0.189). The correlation between the distances among sample embeddings and environmental distances shows more varied signals. In the Visp dataset, ORDNA has lower correlations with mercury concentration (ORDNA $r=0.029$ and non-significant Mantel test, PCoA $r=0.179$; Figure 4). In the Brittany marine fish dataset, ORDNA shows higher correlations than PCoA with temperature (ORDNA $r=0.124$, PCoA $r=0.067$ and non-significant Mantel test; Figure 4). For the freshwater fish dataset from French Guiana, ORDNA also shows higher correlations than PCoA with temperature (ORDNA $r=0.219$, PCoA $r=0.182$; Figure 4), but lower correlations with distance to the sea (ORDNA $r=0.382$, PCoA $r=0.549$; Figure A4 in Supporting Information). For the Swiss forest soils dataset, PCoA shows higher correlations than ORDNA with soil pH (ORDNA $r=0.660$, PCoA $r=0.737$; Figure 4). To summarize, ORDNA embeddings overall have stronger associations with geographic distances across the four datasets tested, while both ORDNA and PCoA embeddings exhibit varying degrees of environmental correlations across the different locations.

3.3 | Robustness of ORDNA

To evaluate the capability of ORDNA, trained on 2020 data, to project 2021–2022 samples, we compared the embeddings of the 2021–2022 samples within the ORDNA space, as illustrated in Figure 5. The minimum Euclidean distance between points from the 2021–2022 samples and points from 2020 for the same area is, on average, 0.118 for PCoA and 0.139 for ORDNA (Table A1 in Supporting Information). PCoA places samples from 2021 to 2022 in the convex hull corresponding to the same area in 34.8% of cases (Figure A3 in Supporting Information), compared with 39.4% for ORDNA. In the case of ORDNA, attribution is the highest for the Audierne Bay (open sea) and South Ar-Men areas, both at 50%, while the Audierne Bay (coastal area) showed the lowest level of attribution, at 10%.

4 | DISCUSSION

Sequence data from eDNA is rich in information on species composition, possibly relative abundance and even intraspecific genetic variation (Andres et al., 2023). Therefore, reducing this information into species or MOTU composition (Marques et al., 2020; Mathon et al., 2021) may be a suboptimal approach for handling this source of biodiversity data. Here, we presented a deep learning method leveraging the triplet loss (Ge et al., 2018) to ordinate eDNA metabarcoding datasets directly from their raw sequence composition. We demonstrated the flexibility of our method by applying it to four distinct datasets, which included various types of environmental samples from soil, freshwater and marine ecosystems at both regional and national scales. Our results indicated that ORDNA produced robust embeddings comparable to those obtained with PCoA but without the need for any bioinformatic pre-processing. Moreover, the organization of the samples in the embedding space adequately represented the geography and the environment of these assemblages. Hence, our method offers a robust alternative approach to previous bioinformatic pipelines for processing the increasing number of eDNA metabarcoding datasets (Blackman et al., 2024), allowing the extraction of ecological information, which in turn can support biodiversity quality assessments and guide management strategies (Li et al., 2023).

Bioinformatic pipelines are usually employed to process metabarcoding data from eDNA, by performing quality control measures, correcting errors, removing contaminants, assembling and aligning sequences through dereplication, merging paired reads and assigning taxonomic classifications using reference databases (Mathon et al., 2021). Consequently, most previous analyses of assemblage from eDNA metabarcoding have relied on such processed tables (Burian et al., 2021). These pipelines define MOTUs as clusters of sequences or ASVs from which ecological and statistical analyses are computed, including ordination techniques (e.g. NMDS or PCoA) that support the interpretation of community composition differences (Polanco Fernández et al., 2021; West et al., 2020). Our results indicate that ANNs offer an opportunity to simplify the analysis and reduce the number of steps while improving the output. They excel at handling high-dimensional data and can process large datasets more efficiently than traditional methods. In addition, as eDNA metabarcoding focuses on small markers, which tend to perform better than longer ones (Zhang et al., 2020), paired-end merging is not required. This allows direct exploitation of the raw sequences, eliminating the need for intermediate steps and manual intervention. Our analyses show that ORDNA enables the automatic extraction of relevant features from raw sequence data to constrain sample differences in the embedding space, advancing beyond

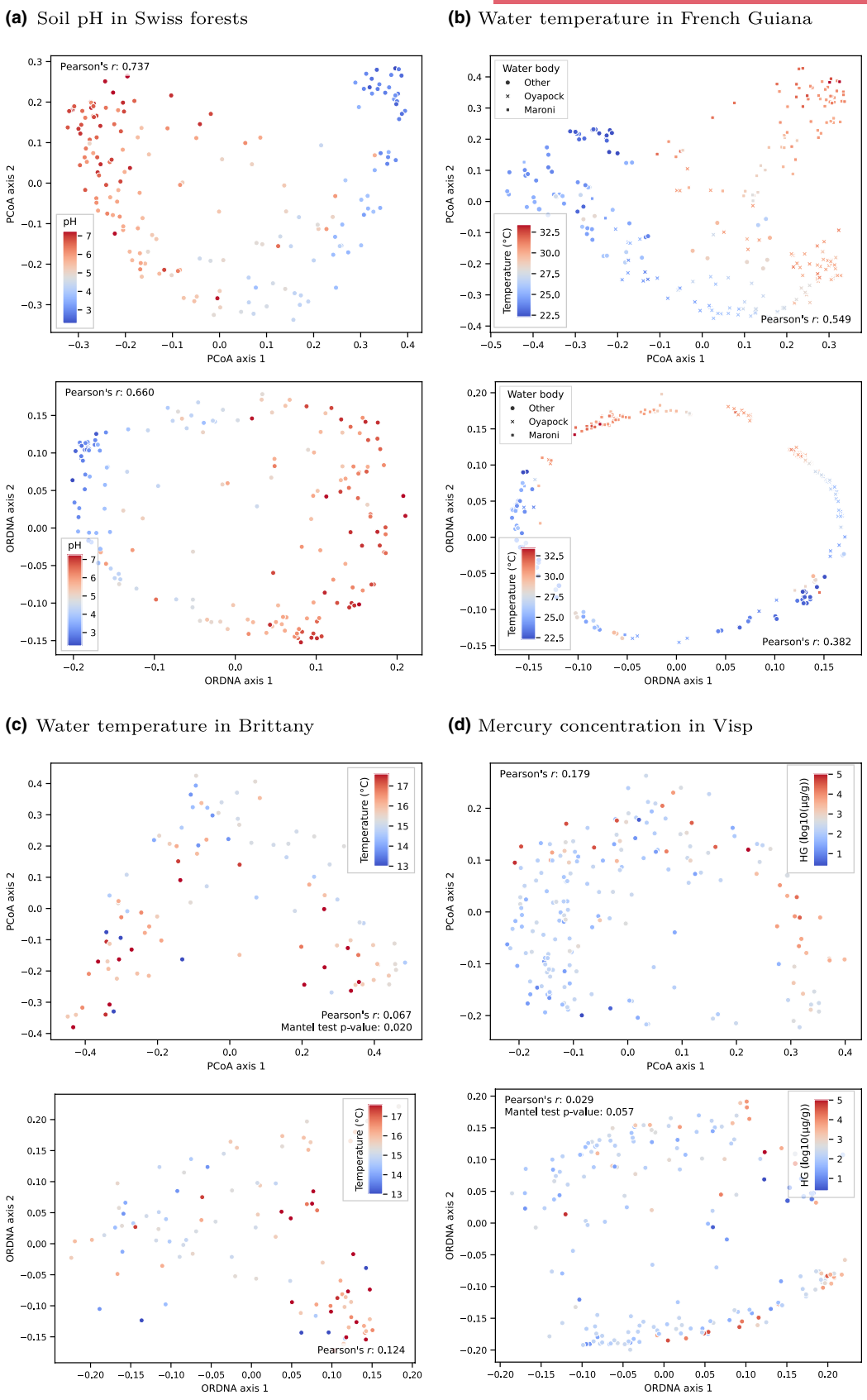

FIGURE 4 Comparison between principal coordinates analysis (PCoA) (top row) and ORDination via Deep Neural Algorithm (ORDNA) (lower row) ordinations for various measurements. Pearson's correlations are computed between the pairwise Euclidean distance of the embeddings and the absolute pairwise difference of the measurement values. Mantel test *p*-values are indicated when they are greater than 0.01.

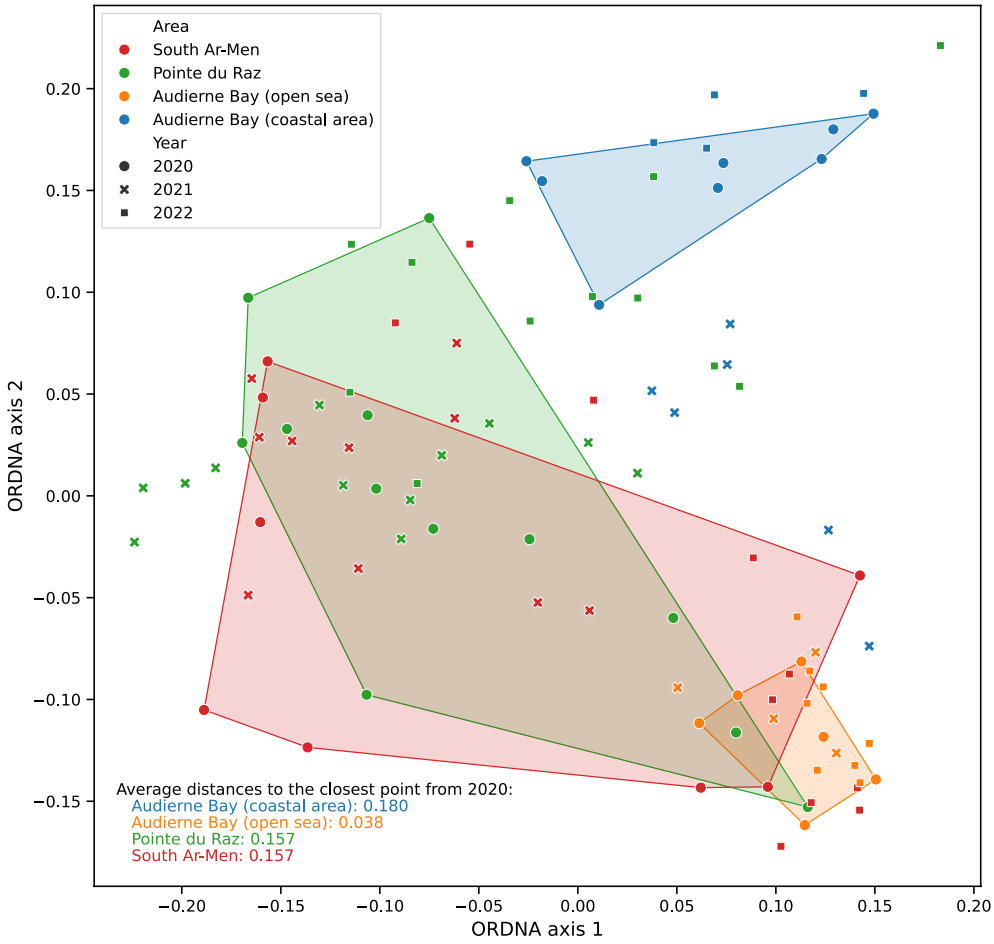

FIGURE 5 Ordination of the Brittany marine dataset (2020, 2021 and 2022) obtained with ORDination via Deep Neural Algorithm (ORDNA) trained with data from 2020 (b). Points are grouped into four subregions (South Ar-Men, Pointe du Raz and two in Audierne Bay), and their year of sampling is indicated. The reported distances represent the average Euclidean distance between samples from 2021 and 2022 and the closest sample from 2020.

previous applications of deep learning with processed eDNA data (Frühe et al., 2021; Lamperti et al., 2023). Here, we focused on a two-dimensional ordination of the samples. However, ORDNA can be optimized to produce representations of any dimensionality. By comparison, ordination methods such as PCoA and NMDS initially generate $N-1$ dimensions, ordered by their importance, which can then be reduced for analysis. Moreover, once trained, ORDNA can rapidly process new data from the same ecosystem, making it suitable for nearly real-time analysis. The advantage of using raw data directly could lead to the discovery of new ecological patterns among the samples, providing deeper insights into the underlying biodiversity.

The physical characteristics of an area directly influence its climate, soil, natural resources and ecosystems, all of which shape the local environmental conditions (Garibaldi et al., 2014). In this context, ORDNA embeddings showed systematically similar or higher correlations with the geographic distances compared with those from PCoA. Therefore, ORDNA may be more efficient in extracting dissimilarities among assemblages associated with environmental gradients when those are linked to environmental distances. For example, it generated smoother dissimilarity gradients for the freshwater fish dataset from French Guiana than previously mapped ones (Flück et al., 2022), and a more spatially organized community structure in the Visp dataset with mercury-polluted soils. These results indicate that the non-linear embeddings of ORDNA enable the extraction of compositional axes that represent geographic dissimilarities well and improve the visualization on a map. Nevertheless, when we tested correlations with the environmental variables expected to drive composition dissimilarities in each dataset, correlations were often stronger with PCoA than with ORDNA. For example, with the Swiss forest soils dataset, PCoA showed higher correlations than ORDNA with soil pH. This could be explained by ORDNA capturing a change in composition associated with environmental variables unmeasured in this study. However, even when directly processing highly complex input data without bioinformatic pre-processing, we find that ORDNA produces outputs that are at least as informative as those from classical approaches.

Ecologists regularly use ecological similarities between samples to interpret global change factors that could impact communities through comparisons with pristine locations (Philippi et al., 1998).

Communities diverging from a state of reference would quickly indicate that a disturbance factor is causing a shift in the ecosystem. In this context, we demonstrated that ORDNA is effective for rapidly projecting new eDNA samples when the ordination model has been pre-trained. We demonstrated that the representation produced by ORDNA is robust, enabling us to reproject data points from later time points into this space. Previous eDNA studies have involved using ordination to identify changes in communities. However, due to their linear structure and the bioinformatic reduction, such approaches might have been more limited in detecting signs of change appropriately. In our marine dataset from Brittany, we found only limited change between the 2 years. This example illustrates the ability of our approach to identify trends in the data. In particular, the trained embeddings could be directly used to extract relevant information from newly collected samples, enabling applications in areas such as forensic investigations (Frøslev et al., 2023), ecological assessments and pollution monitoring (Hinz et al., 2022).

While the present approach enhances the toolkit for processing eDNA metabarcoding data, it has some limitations that could be overcome in future studies leveraging advances in neural networks. First, the model training time can be substantial, in particular when GPUs are not available, which might limit the application of the method for large eDNA datasets. As the dataset size increases, the number of possible triplets grows polynomially, making the training process more complex. Additionally, ORDNA frequently generated a representation that tended towards a circle in the two embedding dimensions that emerged. Although providing a definitive explanation would require further experiments, the emergence of a circular representation rather than a linear gradient in both dimensions suggests that a single dimension may explain most of the dataset. This pattern could, therefore, be an artefact of the network regularization. However, it may also facilitate easier detection of the dataset's underlying properties compared with PCoA, which exhibits more complex behaviour. Indeed, our results may relate to the horseshoe pattern or arch effect, observed with PCoA, which is a common artefact that arises due to a gradient-like structure in the data and non-linear relationships inherent in the dataset (Podani & Miklós, 2002). This pattern typically occurs when there is a strong underlying gradient or continuous variable influencing the data points, such as in ecological data where species composition changes gradually along environmental gradients. The aim of ordinations is to preserve pairwise distances in a lower-dimensional space, and when the original data points lie along a curve in a high-dimensional space, the best two-dimensional representation often may result in a curved or bent shape. The horseshoe pattern highlights the limitation of PCoA and other dimensionality reduction techniques in capturing complex, high-dimensional relationships in fewer dimensions, but in this regard, ORDNA only suffers from the same constraints as other ordination methods. Finally, evaluating the quality of embeddings in comparison with other approaches is non-trivial, and future studies should develop approaches to benchmark datasets, as done in other machine learning tasks to optimize algorithms, for example in genomic studies (Luecken et al., 2022).

In the short term, training ORDNA's network with larger datasets encompassing diverse locations, taxa, eDNA primers and sequencing technologies could unify different eDNA data types into a single framework. Such a generalized model could ordinate new datasets and seamlessly relate them to existing ones, providing a universal tool for eDNA data integration and comparison. The flexibility of deep learning could also be leveraged by integrating additional regularization techniques and loss functions to better align the model with specific ecological research objectives, such as detecting subtle environmental gradients or focusing on specific taxa. Moreover, the embeddings produced by ORDNA could serve as valuable inputs for other analytical methods, enabling complementary analyses that harness the strengths of both direct sequence-based embeddings and derived biological interpretations. For instance, these embeddings could be used alongside other data sources as inputs for deep learning models or classical statistical analyses, extending ORDNA's utility beyond ordination to tasks, such as predicting biodiversity indicators, species distribution modelling and ecological network analysis. Additionally, explainability methods in deep learning could be employed to elucidate the similarities and differences between classical bioinformatics pipelines and ORDNA. These methods could also be leveraged to identify which aspects of the input data have the greatest influence on the ordination, offering deeper insights into the key markers of biodiversity.

To conclude, ORDNA is able to ordinate eDNA samples based on their read compositions by sub-sampling directly from the demultiplexed raw sequence data. We observed that, by enabling the learning of non-linear embeddings, ORDNA's ordination recovers contrasts between sites more effectively and leads to better visualization compared with traditional ordination methods. These findings are supported by overall higher correlations between spatial and known ecological distances with ORDNA's sample distances compared with those retrieved by PCoA. We further demonstrated the ability of ORDNA to project points from a newly sampled year, that is eDNA samples not seen during training, into the space calibrated by independently acquired data from a previous year. This functionality paves the way for methods to easily detect changes in composition over time without the need to continuously fit a model, as the inference time of our neural network is negligible, even when run on a single CPU. All things considered, its high speed of processing from the sequencer to representations amenable to ecological interpretation gives ORDNA the potential to significantly accelerate accurate ecological discovery in order to support decision-making in conservation and ecosystem management.

AUTHOR CONTRIBUTIONS

Théophile Sanchez, Steven Stalder, Loïc Pellissier, Michele Volpi, Camille Albouy and Stéphanie Manel conceived the ideas and designed the methodology. Sébastien Brosse, Aline Frossard, Flurin Leugger, Romane Rozanski, Laura Medicis, Fabienne Kuhn, Xingguo Han, Adrien Mestrot and Camille Albouy collected the data. Théophile Sanchez, Steven Stalder, Letizia Lamperti and Shuo Zong developed the code and performed the analysis. Théophile Sanchez, Steven Stalder, Loïc Pellissier and Michele Volpi led the writing of the

manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

This research was supported by the Swiss Data Science Center (SDSC) collaborative grant C21-05: DNAi and by the ANR SNF project 205556: ShifteDNA awarded to LP and SM.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70033.

DATA AVAILABILITY STATEMENT

The code used to develop ORDNA is available at https://gitlab.renkulab.io/dnai/ordna. The repository includes simple instructions on how to run a docker image to directly train and apply ORDNA on new datasets and retrieve ordinations. The code used to generate the plots, along with alternative versions of the ordination plots featuring colorblind-friendly palettes, is available at https://gitlab.renkulab.io/dnai/ORDNA-plots. Both repositories are archived in https://doi.org/10.5281/zenodo.15011432 (Sanchez et al., 2025). The raw French Guiana dataset is available via https://doi.org/10.3897/BDJ.7.e37518 (Murienne et al., 2019). The raw Swiss forests dataset is available on GenBank under the accession number PRJNA1121203. The raw Visp dataset is available on GenBank under the accession number PRJNA1121073.

ORCID

Théophile Sanchez https://orcid.org/0000-0001-8571-0578
Steven Stalder https://orcid.org/0009-0000-4568-8652
Letizia Lamperti https://orcid.org/0000-0001-8059-1354
Sébastien Brosse https://orcid.org/0000-0002-3659-8177
Aline Frossard https://orcid.org/0000-0003-1699-6220
Flurin Leugger https://orcid.org/0000-0001-9027-6892
Romane Rozanski https://orcid.org/0000-0002-5558-6622
Shuo Zong https://orcid.org/0000-0002-7458-3291
Stéphanie Manel https://orcid.org/0000-0001-8902-6052
Xingguo Han https://orcid.org/0000-0002-5753-5255
Adrien Mestrot https://orcid.org/0000-0002-4387-3886
Camille Albouy https://orcid.org/0000-0003-1629-2389
Michele Volpi https://orcid.org/0000-0003-2771-0750
Loïc Pellissier https://orcid.org/0000-0002-2289-8259

REFERENCES

Andres, K. J., Lodge, D. M., Sethi, S. A., & Andrés, J. (2023). Detecting and analysing intraspecific genetic variation with eDNA: From population genetics to species abundance. *Molecular Ecology*, *32*(15), 4118–4132.

Arranz, I., Fournier, B., Lester, N. P., Shuter, B. J., & Peres-Neto, P. R. (2022). Species compositions mediate biomass conservation: The case of lake fish communities. *Ecology*, *103*(3), e3608.

Battey, C., Ralph, P. L., & Kern, A. D. (2020). Predicting geographic location from genetic variation with deep neural networks. *eLife*, *9*, e54507. https://doi.org/10.7554/eLife.54507

Blackman, R., Couton, M., Keck, F., Kirschner, D., Carraro, L., Cereghetti, E., Perrelet, K., Bossart, R., Brantschen, J., Zhang, Y., & Altermatt, F. (2024). Environmental DNA: The next chapter. *Molecular Ecology*, *33*(11), e17355. https://doi.org/10.1111/mec.17355

Borcard, D., Gillet, F., Legendre, P., Borcard, D., Gillet, F., & Legendre, P. (2018). Spatial analysis of ecological data. In *Numerical ecology with R* (pp. 299–367). Springer International Publishing.

Borcard, D., Legendre, P., Avois-Jacquet, C., & Tuomisto, H. (2004). Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, *85*(7), 1826–1832.

Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., & White, A. E. (2022). Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, *13*(8), 1640–1660.

Boyer, F., Mercier, C., Bonin, A., le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182.

Burian, A., Mauvisseau, Q., Bulling, M., Domisch, S., Qian, S., & Sweet, M. (2021). Improving the reliability of eDNA data interpretation. *Molecular Ecology Resources*, *21*(5), 1422–1433.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). Dada2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., Iribar, A., Taberlet, P., Vigouroux, R., & Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Molecular Ecology Resources*, *19*(1), 27–46.

Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T., & Pawlowski, J. (2019). Embracing environmental genomics and machine learning for routine biomonitoring. *Trends in Microbiology*, *27*(5), 387–397.

Coutant, O., Jézéquel, C., Mokany, K., Cantera, I., Covain, R., Valentini, A., Dejean, T., Brosse, S., & Murienne, J. (2023). Environmental DNA reveals a mismatch between diversity facets of Amazonian fishes in response to contrasting geographical, environmental and anthropogenic effects. *Global Change Biology*, *29*(7), 1741–1758.

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.

Dixon, P. (2003). Vegan, a package of r functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930.

Dornelas, M., Magurran, A. E., Buckland, S. T., Chao, A., Chazdon, R. L., Colwell, R. K., Curtis, T., Gaston, K. J., Gotelli, N. J., Kosnik, M. A., McGill, B., McCune, J. L., Morlon, H., Mumby, P. J., Øvreås, L., Studeny, A., & Vellend, M. (2013). Quantifying temporal change in biodiversity: Challenges and opportunities. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1750), 20121931.

Dray, S., Pélissier, R., Couteron, P., Fortin, M. J., Legendre, P., Peres-Neto, P. R., Bellier, E., Bivand, R., Blanchet, F. G., de Cáceres, M., Dufour, A. B., Heegaard, E., Jombart, T., Munoz, F., Oksanen, J., Thioulouse, J., & Wagner, H. H. (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs*, *82*(3), 257–275.

Flück, B., Mathon, L., Manel, S., Valentini, A., Dejean, T., Albouy, C., Mouillot, D., Thuiller, W., Murienne, J., Brosse, S., & Pellissier, L. (2022). Applying convolutional neural networks to speed up

environmental DNA annotation in a highly diverse ecosystem. *Scientific Reports*, *12*(1), 10247.

Frøslev, T. G., Ejrnæs, R., Hansen, A. J., Bruun, H. H., Nielsen, I. B., Ekelund, F., Vestergård, M., & Kjøller, R. (2023). Treated like dirt: Robust forensic and ecological inferences from soil eDNA after challenging sample storage. *Environmental DNA*, *5*(1), 158–174. https://doi.org/10.1002/edn3.367

Frossard, A., Donhauser, J., Mestrot, A., Gygax, S., Bååth, E., & Frey, B. (2018). Long-and short-term effects of mercury pollution on the soil microbiome. *Soil Biology and Biochemistry*, *120*, 191–199. https://doi.org/10.1016/j.soilbio.2018.01.028

Frühe, L., Cordier, T., Dully, V., Breiner, H. W., Lentendu, G., Pawlowski, J., Martins, C., Wilding, T. A., & Stoeck, T. (2021). Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. *Molecular Ecology*, *30*(13), 2988–3006.

Garibaldi, C., Nieto-Ariza, B., Macía, M. J., & Cayuela, L. (2014). Soil and geographic distance as determinants of floristic composition in the Azuero Peninsula (Panama). *Biotropica*, *46*(6), 687–695.

Ge, W., Huang, W., Dong, D., & Scott, M. R. (2018). Deep metric learning with hierarchical triplet loss. *CoRR abs/1810.06951*. http://arxiv.org/abs/1810.06951

Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, *22*, 1–19.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*(3–4), 325–338.

Hinz, S., Coston-Guarini, J., Marnane, M., & Guarini, J. M. (2022). Evaluating eDNA for use within marine environmental impact assessments. *Journal of Marine Science and Engineering*, *10*(3), 375.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution*, *3*(4), 539–551. https://doi.org/10.1038/s41559-019-0826-1

Kaya, M., & Bilge, H. Ş. (2019). Deep metric learning: A survey. *Symmetry*, *11*(9), 1066.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Lamperti, L., Sanchez, T., Si Moussi, S., Mouillot, D., Albouy, C., Flück, B., Bruno, M., Valentini, A., Pellissier, L., & Manel, S. (2023). New deep learning-based methods for visualizing ecosystem properties using environmental DNA metabarcoding data. *Molecular Ecology Resources*, *23*(8), 1946–1958.

Lang, N., Jetz, W., Schindler, K., & Wegner, J. D. (2023). A high-resolution canopy height model of the earth. *Nature Ecology & Evolution*, *7*(11), 1778–1789.

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., & Aide, T. M. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, *59*, 101113.

Leprieur, F., Descombes, P., Gaboriau, T., Cowman, P. F., Parravicini, V., Kulbicki, M., Melián, C. J., de Santana, C. N., Heine, C., Mouillot, D., Bellwood, D. R., & Pellissier, L. (2016). Plate tectonics drive tropical reef biodiversity dynamics. *Nature Communications*, *7*(1), 11461.

Li, F., Qin, S., Wang, Z., Zhang, Y., & Yang, Z. (2023). Environmental DNA metabarcoding reveals the impact of different land use on multitrophic biodiversity in riverine systems. *Science of the Total Environment*, *855*, 158958.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint* arXiv:171105101. https://doi.org/10.48550/arXiv.1711.05101

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., & Theis, F. J. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, *19*(1), 41–50.

Mace, G. M., Norris, K., & Fitter, A. H. (2012). Biodiversity and ecosystem services: A multilayered relationship. *Trends in Ecology & Evolution*, *27*(1), 19–26.

Marques, V., Guérin, P. É., Rocle, M., Valentini, A., Manel, S., Mouillot, D., & Dejean, T. (2020). Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. *Ecography*, *43*(12), 1779–1790.

Marshall, E., Wintle, B. A., Southwell, D., & Kujala, H. (2020). What are we measuring? A review of metrics used to describe biodiversity in offsets exchanges. *Biological Conservation*, *241*, 108250.

Mathon, L., Valentini, A., Guérin, P. E., Normandeau, E., Noel, C., Lionnet, C., Boulanger, E., Thuiller, W., Bernatchez, L., Mouillot, D., Dejean, T., & Manel, S. (2021). Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Molecular Ecology Resources*, *21*(7), 2565–2579.

Murienne, J., Cantera, I., Cerdan, A., Cilleros, K., Decotte, J.-B., Dejean, T., Vigouroux, R., & Brosse, S. (2019). Aquatic eDNA for monitoring French Guiana biodiversity. *Biodiversity Data Journal*, *7*, e37518. https://doi.org/10.3897/BDJ.7.e37518

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., Heip, C., Höft, R., Hurtt, G., Jetz, W., … Wegmann, M. (2013). Essential biodiversity variables. *Science*, *339*(6117), 277–278.

Philippi, T. E., Dixon, P. M., & Taylor, B. E. (1998). Detecting trends in species composition. *Ecological Applications*, *8*(2), 300–308.

Podani, J., & Miklós, I. (2002). Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*, *83*(12), 3331–3343.

Polanco Fernández, A., Marques, V., Fopp, F., Juhel, J. B., Borrero-Pérez, G. H., Cheutin, M. C., Dejean, T., González Corredor, J. D., Acosta-Chaparro, A., Hocdé, R., Eme, D., Maire, E., Spescha, M., Valentini, A., Manel, S., Mouillot, D., Albouy, C., & Pellissier, L. (2021). Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environmental DNA*, *3*(1), 142–156.

Purvis, A., & Hector, A. (2000). Getting the measure of biodiversity. *Nature*, *405*(6783), 212–219.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rozanski, R., Trenkel, V. M., Lorance, P., Valentini, A., Dejean, T., Pellissier, L., Eme, D., & Albouy, C. (2022). Disentangling the components of coastal fish biodiversity in southern Brittany by applying an environmental DNA approach. *Environmental DNA*, *4*(4), 920–939. https://doi.org/10.1002/edn3.305

Sanchez, T., Cury, J., Charpiat, G., & Jay, F. (2021). Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, *21*(8), 2645–2660.

Sanchez, T., Stalder, S., Lamperti, L., Brosse, S., Frossard, A., Leugger, F., Rozanski, R., Zong, S., Manel, S., Medici, L., Kuhn, F., Han, X., Mestrot, A., Albouy, C., Volpi, M., & Pellissier, L. (2025). ORDNA: Deep learning ordination for raw environmental DNA samples. *Zenodo*, https://doi.org/10.5281/zenodo.15011432

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *CoRR* abs/1503.03832. http://arxiv.org/abs/1503.03832

Strange, N., zu Ermgassen, S., Marshall, E., Ermgassen, S. z., Bull, J. W., & Jacobsen, J. B. (2024). Why it matters how biodiversity is measured in environmental valuation studies compared to conservation science. *Biological Conservation*, *292*, 110546. https://doi.org/10.1016/j.biocon.2024.110546

Vélez, J., McShea, W., Shamon, H., Castiblanco-Camacho, P. J., Tabak, M. A., Chalmers, C., Fergus, P., & Fieberg, J. (2023). An evaluation

of platforms for processing camera-trap data using artificial intelligence. *Methods in Ecology and Evolution*, *14*(2), 459–477.

West, K. M., Stat, M., Harvey, E. S., Skepper, C. L., DiBattista, J. D., Richards, Z. T., Travers, M. J., Newman, S. J., & Bunce, M. (2020). eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical Island ecosystem. *Molecular Ecology*, *29*(6), 1069–1086.

Yan, J., & Wang, X. (2023). Machine learning bridges omics sciences and plant breeding. *Trends in Plant Science*, *28*(2), 199–210.

Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution*, *11*(12), 1609–1625.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1: Supporting information for ORDNA: Deep-learning-based ordination for raw environmental DNA samples.

How to cite this article: Sanchez, T., Stalder, S., Lamperti, L., Brosse, S., Frossard, A., Leugger, F., Rozanski, R., Zong, S., Manel, S., Medici, L., Kuhn, F., Han, X., Mestrot, A., Albouy, C., Volpi, M., & Pellissier, L. (2025). ORDNA: Deep-learning-based ordination for raw environmental DNA samples. *Methods in Ecology and Evolution*, *00*, 1–14. https://doi.org/10.1111/2041-210X.70033