New deep learning-based methods for visualizing ecosystem properties using environmental DNA metabarcoding data

Letizia Lamperti, Théophile Sanchez, Sara Si Moussi, David Mouillot, Camille Albouy, Benjamin Flück, Morgane Bruno, Alice Valentini, Loïc Pellissier, Stéphanie Manel

Received: 13 April 2023 | Revised: 29 July 2023 | Accepted: 14 August 2023

DOI: 10.1111/1755-0998.13861

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

New deep learning-based methods for visualizing ecosystem properties using environmental DNA metabarcoding data

Letizia Lamperti[1,2,3] | Théophile Sanchez[2,3] | Sara Si Moussi[4] | David Mouillot[5,6] | Camille Albouy[2,3] | Benjamin Flück[2,3] | Morgane Bruno[1] | Alice Valentini[7] | Loïc Pellissier[2,3] | Stéphanie Manel[1,6]

[1]CEFE, Univ Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France

[2]Ecosystems and Landscape Evolution, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

[3]Ecosystems and Landscape Evolution, Land Change Science Research Unit, Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Switzerland

[4]Laboratoire d'Ecologie Alpine, Univ. Grenoble Alpes, Univ. Savoie MontBlanc, CNRS, Grenoble, France

[5]MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France

[6]Institut Universitaire de France, Paris, France

[7]SPYGEN, Le Bourget-du-Lac, France

Correspondence
Letizia Lamperti, CEFE, Univ Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France.
Email: le.lamperti@gmail.com

Loïc Pellissier, Ecosystems and Landscape Evolution, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland. Ecosystems and Landscape Evolution, Land Change Science Research Unit, Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Switzerland.
Email: loic.pellissier@usys.ethz.ch

Handling Editor: Joanna Kelley

Abstract

Environmental DNA (eDNA) metabarcoding provides an efficient approach for documenting biodiversity patterns in marine and terrestrial ecosystems. The complexity of these data prevents current methods from extracting and analyzing all the relevant ecological information they contain, and new methods may provide better dimensionality reduction and clustering. Here we present two new deep learning-based methods that combine different types of neural networks (NNs) to ordinate eDNA samples and visualize ecosystem properties in a two-dimensional space: the first is based on variational autoencoders and the second on deep metric learning. The strength of our new methods lies in the combination of two inputs: the number of sequences found for each molecular operational taxonomic unit (MOTU) detected and their corresponding nucleotide sequence. Using three different datasets, we show that our methods accurately represent several biodiversity indicators in a two-dimensional latent space: MOTU richness per sample, sequence α-diversity per sample, Jaccard's and sequence β-diversity between samples. We show that our nonlinear methods are better at extracting features from eDNA datasets while avoiding the major biases associated with eDNA. Our methods outperform traditional dimension reduction methods such as Principal Component Analysis, t-distributed Stochastic Neighbour Embedding, Nonmetric Multidimensional Scaling and Uniform Manifold Approximation and Projection for dimension reduction. Our results suggest that NNs provide a more efficient way of extracting structure from eDNA metabarcoding data, thereby improving their ecological interpretation and thus biodiversity monitoring.

KEYWORDS
biodiversity monitoring, data visualization, deep learning, deep metric learning, environmental DNA, machine learning, neural networks, variational autoencoder

Loïc Pellissier and Stéphanie Manel: Co-senior authors.

1 | INTRODUCTION

Human-induced disturbances affect most of the Earth's ecosystems, which are suffering from the accelerating impacts of climate change and overexploitation (Johnston et al., 2022; Jouffray et al., 2020). These threats alter species assemblages and lead to escalating perturbations in ecosystem processes (Frainer et al., 2017; McLean et al., 2019), ultimately altering ecosystem services and thus humanity (Cinner et al., 2020; Tigchelaar et al., 2022). In the context of global change, it is crucial to capture the spatio-temporal dynamics of species assemblages and better understand their responses in order to design appropriate management and mitigation measures (Makiola et al., 2020). Recently, our ability to rapidly generate comprehensive biodiversity inventories has been enhanced by the development of environmental DNA (eDNA) metabarcoding, which allows the retrieval and analysis of DNA naturally shed by organisms in their environment (Deiner et al., 2017; Miya, 2022). eDNA metabarcoding is now operational in many ecosystems for a wide range of micro- and macroorganisms (Cantera et al., 2022; Cordier et al., 2021; Kjær et al., 2022; Mathon et al., 2022), providing information on their taxonomic, functional, but also phylogenetic affiliations (Marques, Castagné, et al., 2021; Marques, Milhau, et al., 2021; Rozanski et al., 2022). Given its limited field effort and ecosystem disturbance (Muff et al., 2023), even in the most remote locations, and the decrease in sequencing cost over the recent years, this approach can be scaled up to monitor many sites at high temporal frequency (Agersnap et al., 2022).

Yet, eDNA metabarcoding produces a huge amount of sequencing data (i.e. a high number of short DNA sequences), that represent complex and high-dimensional information. Typically, these sequences are assigned to known taxonomic units stored in a genetic reference database. The incompleteness of genetic reference databases precludes the identification of many species (Marques et al., 2020), thus working with Molecular Operational Taxonomic Units (MOTUs), representing a cluster of similar sequences, may be required (Deiner et al., 2017; Floyd et al., 2002; Mathon et al., 2022). MOTUs are then defined by a consensus sequence. The attribute attached to an eDNA MOTU is the relative frequency of the sequences in each MOTU and the nucleotide sequence itself. Both attributes can be directly related to ecosystem states and properties (Bakker et al., 2017). Therefore, eDNA data has the potential to reveal ecological patterns that distinguish sampled sites along environmental or human pressure gradients (Marques et al., 2020). Such patterns are expected to emerge from the interaction and nonlinear combination of both abundance and phylogenetic information. However, the dimensionality of the massive amount of sequence information must be reduced to extract relevant features.

Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality (Nissen et al., 2018; van der Maaten & Hinton, 2008). Traditionally, dimensionality reduction is performed using linear techniques such as Principal Component Analysis (PCA; Karl Pearson, 1901), Factor Analysis (Spearman, 1904) and Classical Scaling (Torgerson, 1952). However, due to their underlying hypotheses, these linear techniques cannot adequately deal with complex nonlinear relationships in data such as those provided by eDNA metabarcoding. In the last decade, many nonlinear techniques have been proposed for dimensionality reduction (Facco et al., 2017; Nguyen & Holmes, 2019). Recently, two machine learning techniques – the t-distributed Stochastic Neighbour Embedding (t-SNE; van der Maaten & Hinton, 2008) and the Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) – have shown promising results in generating two-dimensional visualizations of high-dimensional biological data (Diaz-Papkovich et al., 2021). However, the interpretation of t-SNE and UMAP plots remains challenging due to the lack of global structure in the reduced space representation (Battey et al., 2021). Although these methods perform well in clustering similar samples, distances between clusters are not always meaningful (Becht et al., 2019).

On the other hand, neural networks (NNs) have been shown to provide a good representation of learning capacity for various datasets (Sze et al., 2017). NNs are complex mathematical models consisting of many operators called neurons connected as a network.

Within the framework of ordination and dimensionality reduction, there are contrastive methods, such as UMAP and t-SNE, that work on learning features by satisfying distances between observations. Other methods instead use generative latent variable models, where prior distributions are specified for the unobserved structure in the data so that these unknown properties can be inferred by posterior inference. Examples include factor analysis, probabilistic PCA, and variational autoencoders (VAEs). VAEs combine two deep NNs, where the first network (the encoder) encodes input data (e.g. the number of sequences per MOTU detected in each sample) as a probability distribution in a latent space, and the second network (the decoder) attempts to reconstruct the input data given a set of latent coordinates. VAEs have been used extensively in image generation (e.g. Gulrajani et al., 2016; Hou et al., 2018; Larsen et al., 2016), and several recent studies have applied them to dimensionality reduction and classification of single-cell RNAseq data (Becht et al., 2019; Grønbech et al., 2018; Lafarge, Caicedo, et al., 2019; Lafarge, Pluim, et al., 2019; Wang & Gu, 2018). Thanks to the design flexibility of artificial NNs in general, they also have the advantage of being able to encode and mix information from different data types.

In particular, Deep Metric Learning (DML) lies between contrastive and generative latent variable models. DML is an approach based directly on a distance metric that aims to establish similarity or dissimilarity between objects (Kulis, 2013). Although DML aims to reduce the distance between similar objects, it also aims to increase the distance between dissimilar objects (Duffner et al., 2021). Through DML, it is possible to use a distance measure relevant to the case study as a contrastive model but also to encode different inputs via NNs.

The potential of VAE and DML to ordinate eDNA samples in a low-dimensional space has not yet been demonstrated. In this study, we present two new methods, one based on VAE and the other on DML, to visualize eDNA biodiversity indicators. We tested our methods on three different published eDNA datasets: a fish eDNA dataset collected in the Mediterranean Sea (Boulanger et al., 2021), and two eukaryotic plankton eDNA datasets from the Tara Ocean expedition (de Vargas et al., 2015). We used both the number of sequences found for each molecular operational taxonomic unit (MOTU) detected and their corresponding nucleotide sequences. To validate these two new methods, we compare them with other classical methods: PCA, t-SNE and UMAP for the VAE-based method, and Nonmetric Multidimensional Scaling (NMDS) for the DML-based method. Finally, we show how the proposed methods outperform classical methods in their representation of biodiversity indicators.

2 | MATERIALS AND METHODS

2.1 | VAE-based method applied to eDNA data

The VAE-based method, called VAESeq (Variational AutoEncoder + Sequences) (Figure 1), processes eDNA samples into a two-dimensional latent space. The model consists of one autoencoder (AE) and one VAE. The AE takes as input the nucleotide sequences of the detected MOTUs within each sample and compresses them in the latent encoding z_{AE}. The VAE encoder then receives and processes z_{AE} in combination with the number of sequences found for the MOTUs detected in the samples. By mixing the two inputs, the VAE encodes the samples as points in a 2D latent space called z_{VAE}. In the decoding part, the VAE decoder seeks to recreate the two inputs from z_{VAE}. The decoder measures how much information is lost from the input during the encoding and optimizes the network accordingly. To reduce the running time of the model, we separately trained the AE to encode nucleotide sequences and then trained the VAE.

To encode the DNA sequence information in the AE, the nucleotide sequences are equalized to the same length. We have chosen to keep the maximum length, by padding the sequences with nucleotide code N from the IUPAC nucleotide code. Each canonical base (A, C, T, G) of the sequence and the IUPAC ambiguity codes are translated into an appropriate four-dimensional probability distribution over the four canonical bases (A, T, C, G), including uncertain base sequences (e.g. W and S). For example, A becomes [1, 0, 0, 0] or W becomes [0.5, 0, 0, 0.5] (Flück et al., 2022). Furthermore, N nucleotides added to equalize the sequence length become [0.25, 0.25, 0.25, 0.25]. To maintain the same dimension of the AE input, we combine the nucleotide sequences with the presence/absence of each MOTU in each sample. Therefore, each sample is represented by a tensor containing the translated nucleotide sequences in matrices of the detected MOTUs and, alternatively, a zero matrix if the MOTU was not detected.

The AE component of the network uses the Adam optimizer and the binary cross-entropy loss function to optimize the network. The AE encoder consisted of seven fully connected layers with decreasing widths down to 100, rectified linear unit activations, and dropout regularization (20% dropout). A mirror architecture was used as the decoder.

The VAE component of the network used the Adam optimizer and two loss functions to reconstruct the two inputs: the VAE loss function (Kullback–Leibler divergence + reconstruction error) for the occurrence information and binary cross-entropy for the nucleotide

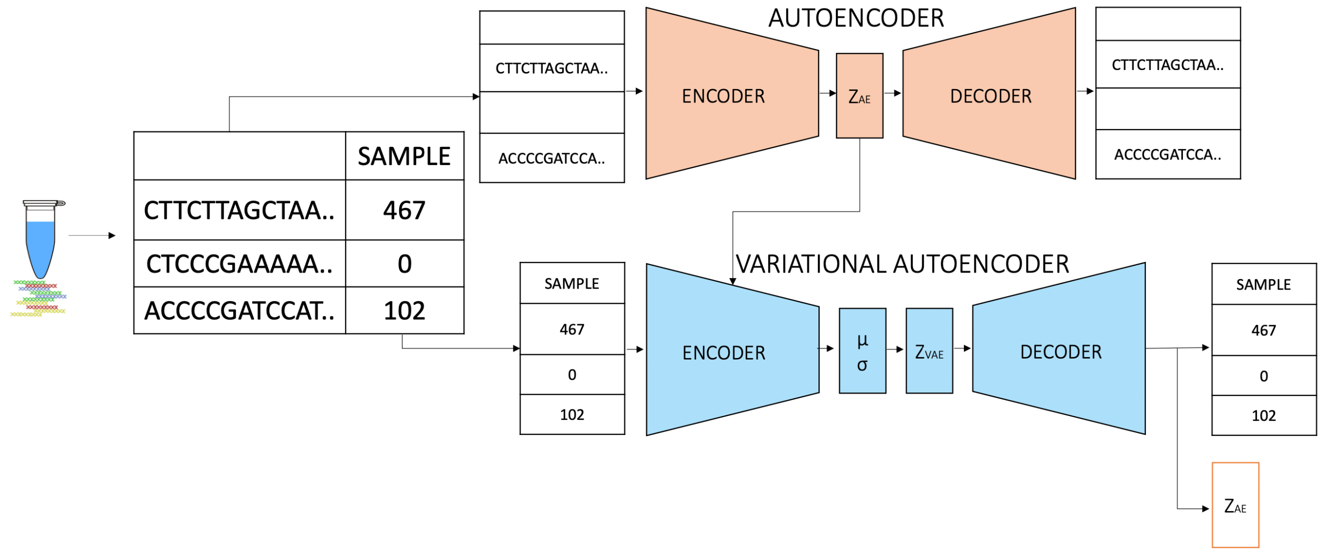

FIGURE 1 Diagram of the VAE-based method applied to eDNA data (VAESeq). The model consists of one autoencoder (AE) and one variational autoencoder (VAE). The AE takes as input the genetic sequence information of each MOTU combined with the presence/absence of each MOTU within each sample to generate the first latent encoding z_{AE}. This information is then passed to the VAE in one encoder layer. Thus, at each iteration, the VAE receives as input the number of sequences detected for each MOTU in one sample and the autoencoder embedding z_{AE}. The VAE processes the two inputs and reduces the dimensionality of the samples to a two-dimensional latent space, z_{VAE}. In z_{VAE}, we find the 2D representation of all data points (Figure S3a,b). In the decoding part, VAE reconstructs the two inputs in order to optimize the network accordingly.

sequences latent encoding different combinations, we set the loss function weights to 1 and 0.2 respectively.

The VAE encoder consisted of three fully connected layers with decreasing widths down to 2, rectified linear unit activations, and dropout regularization (20% dropout). A mirror architecture to the VAE encoder was used as the decoder.

2.2 | DML-based method on β-diversity as a distance

The DML-based method, called ENNBetaDist (Encoder Neural Networks based on Beta diversity Distance matrix), (Figure 2) trains the network according to the pairwise β-diversity calculated between samples. The pairwise β-diversity is used as a distance measure on each pair of samples, to help the network distribute the points in the 2D latent space. We used pairwise Jaccard's dissimilarity as a measure of β-diversity, using the 'betapart' library in R.

The structure of ENNBetaDist consists of two encoder NNs. The encoders of ENNBetaDist are similar to those of VAESeq, with differences in the number of hidden layers, training, and optimization. VAESeq reconstructs the input from the 2D latent space. However, we want the latent space to respect the distances we intend to optimize. Therefore, we implemented a second NN, ENNBetaDist, consisting of two encoder NNs.

At each iteration, each encoder takes as input a sample containing the number of sequences per MOTU and the latent encoding z_{AE} from the AE of the nucleotide sequences detected in the sample. Then, the first encoder projects the first sample in its two-dimensional latent space z_1 and the second encoder projects the second sample in its two-dimensional latent space z_2. To optimize the model, we compute the Euclidean distance between the two points in z_1 and z_2 and compare it with the pairwise β-diversity via a loss function (the mean square error (MSE)). In z_1 and z_2 we find the 2D representations of all data points. The two representations are similar, and for the sake of simplicity, we analyse only the first one as the final latent space produced by our model. Ultimately, ENNBetaDist seeks to represent pairwise the distances related to the species composition of the samples (i.e. the information provided by Jaccard's β-diversity) as distances between points in the 2D space. The encoders consisted of five fully connected layers with decreasing widths down to 2, rectified linear unit activations and dropout regularization (20% dropout).

2.3 | Sensitivity

To perform a cross-validation of our two new methods, we set a global random seed to split 80% of the original dataset in the training set and 20% in the validation set. We repeated the tests until the results were stable, ensuring that we did not overfit by monitoring the loss on both the training and validation set (Figure S1). We implemented the models in R (version 4.1.3, R Core Team, 2022) using TensorFlow (Abadi et al., 2016) and Keras (Chollet & Others, 2015) libraries.

2.4 | Case study

2.4.1 | Datasets

We tested our methods on three different published eDNA datasets: a fish eDNA dataset collected in the Western Mediterranean Sea (Boulanger et al., 2021) and two eukaryotic plankton eDNA datasets from the Tara Ocean Campaign (de Vargas et al., 2015). Details are given in Table 1.

eDNA samples from the Western Mediterranean Sea were collected at 77 stations in six marine regions covering the Western Mediterranean, including fished and no-take protected areas (Boulanger et al., 2021). eDNA extraction and amplification were performed at the SPYGEN facility. PCR amplification was performed using the teleo primer pair, targeting a 64 bp fragment of the mitochondrial DNA 12S rRNA gene specific for teleost fishes and elasmobranches (Valentini et al., 2016). Data collection and sample processing are described in detail in Appendix S1.

The Tara Ocean datasets were obtained from the Tara Oceans V9 rDNA metabarcoding dataset (De Vargas et al., 2015) collected across tropical and temperate oceans during the circumglobal Tara Oceans expedition. The analysis was based on metabarcoding data from 129 stations in various oceanic provinces worldwide, using 18S ribosomal DNA sequences across the intermediate plankton-size spectrum. All details on data collection, extraction, and sequencing can be found in the article by de Vargas et al. (2015). We selected the *Dictyochophyceae* and *Telonemia* subsets by taxonomic identification, resulting in two smaller datasets of similar sizes to the Western Mediterranean one whose specifications are shown in Table 1.

2.5 | Comparison and evaluation

We tested the ability of the two new methods to represent biodiversity indicators in a 2D space based on their species and sequence composition. We compared the 2D representation of VAESeq with three other classical dimensionality reduction methods: PCA, which is linear, t-SNE and UMAP, which are nonlinear.

We also analysed the latent encoding z_{AE} generated by the autoencoder using PCA to evaluate the results of the first part of the models, which we call AEgen + PCA (AutoEncoder genetic + PCA). Then, we compared it with a simple VAE where the latent encoding z_{AE} of the nucleotide sequences was not given as input. Likewise, we compared the 2D representation of ENNBetaDist to NMDS, a nonlinear method.

To summarize the inputs employed by each method, PCA, t-SNE, UMAP, VAE and NMDS use the presence/absence information of each MOTU in each sample and the number of sequences found

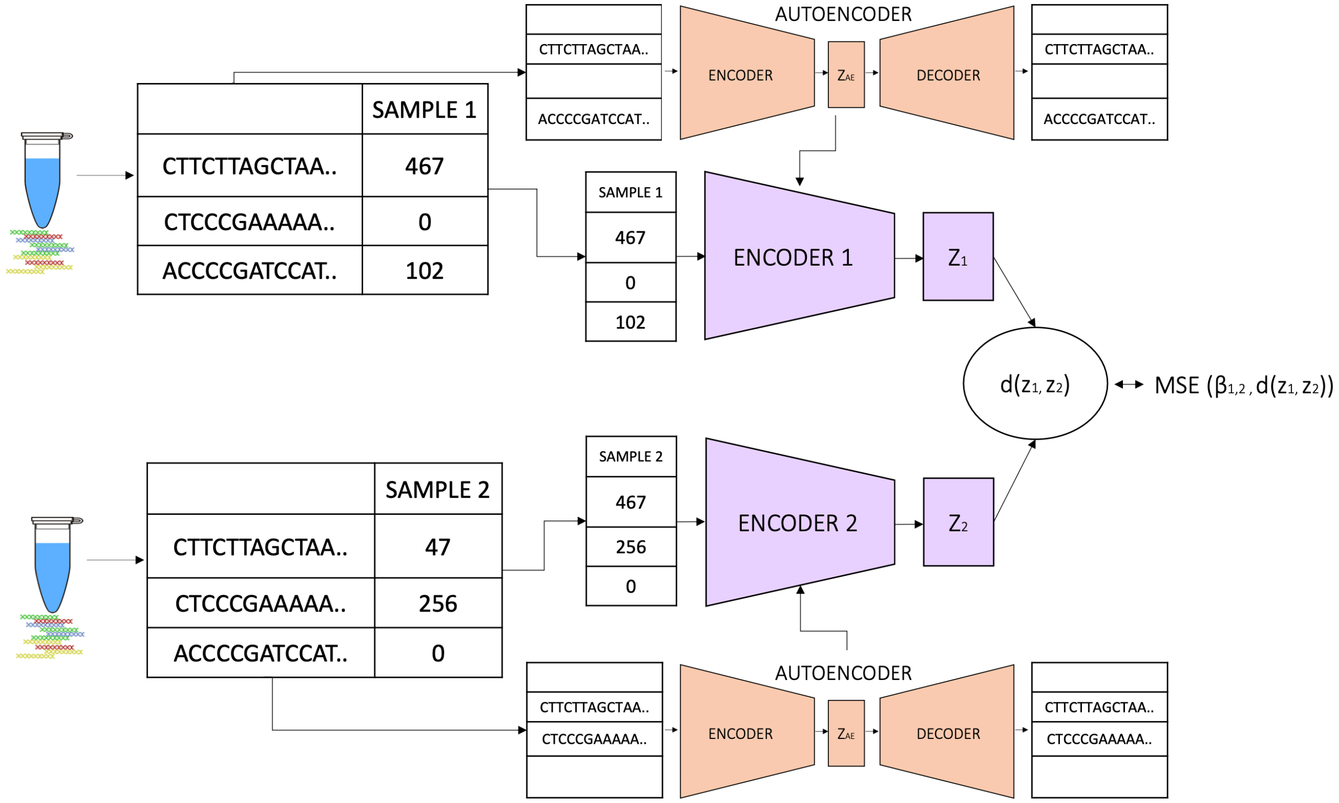

FIGURE 2 Diagram of the DML-based method applied to eDNA data using pairwise Jaccard's β-diversity (ENNBetaDist). The DML-based method ENNBetaDist trains the network according to the pairwise β-diversity calculated for each pair between samples. The pairwise β-diversity is used as a distance measure to help the network distribute the points in the latent space. The two encoders process the two samples, combining the number of sequences found for each MOTU detected and the latent encoding z_{AE} from the AE of the nucleotide sequences detected in each sample, and project them into two points in z_1 and z_2. To optimize the model, we calculate the Euclidean distance between the two points in z_1 and z_2 and compare it with Jaccard's β-diversity via a loss function (the mean square error (MSE)). In z_1 and z_2, we find the 2D representation of all the data points (Figure S3c,d).

TABLE 1 Descriptive table of the three eDNA datasets: Western Mediterranean fish dataset (Boulanger et al., 2021), Tara Ocean *Dictyochophyceae* dataset (de Vargas et al., 2015), and Tara Ocean *Telonemia* dataset (de Vargas et al., 2015). Details on the calculation of sequence α-diversity can be found in the Appendix S1.

	Western Mediterranean Fish Dataset	Tara Ocean Dictyochophyceae Dataset	Tara Ocean Telonemia Dataset
Tot number of samples	394	319	321
Tot number of MOTUs	290	223	237
Min MOTU richness	1	1	1
Max MOTU richness	92	57	49
Min sequence length	53	76	103
Max sequence length	76	151	151
Min sequence α-diversity	1	1	1
Max sequence α-diversity	16.74	10.54	10.41

during eDNA extraction for each detected MOTU. AEgen+PCA uses the MOTU nucleotide sequences in addition to their presence/absence information. Our methods, VAESeq and ENNBetaDist, use all three inputs: the presence/absence information, the number of sequences found for each detected MOTU, and the nucleotide sequences. Moreover, NMDS and ENNBetaDist utilize Jaccard's β-diversity for the optimization process (Table S1).

To evaluate the performance of all methods, we used multiple regression on distance matrices (MRM) representing the sample in the reduced 2D space. For MRM, we implemented two different tests and assessed the statistical significance using permutations. TEST 1 performs a multiple regression between the sample distances in the two-dimensional latent spaces of each method and their Jaccard's β-diversity. TEST 2, instead of using Jaccard's

TABLE 2 Comparison between the 2D representation of VAESeq and AEgen + PCA with three other classical dimension reduction methods: PCA, which is linear, and t-SNE and UMAP, which are nonlinear.

	Western Mediterranean Dataset				Tara Ocean Dictyochophyceae Dataset				Tara Ocean Telonemia Dataset			
	TEST 1		TEST 2		TEST 1		TEST 2		TEST 1		TEST 2	
	D_{1s}–$D_{\beta\text{-jac}}$		D_{1s}–$D_{\beta\text{-gen}}$		D_{1s}–$D_{\beta\text{-jac}}$		D_{1s}–$D_{\beta\text{-gen}}$		D_{1s}–$D_{\beta\text{-jac}}$		D_{1s}–$D_{\beta\text{-gen}}$	
	R^2	*p*-value	R^2	*p*-value	R^2	*p*-value	R^2	*p*-value	R^2	*p*-value	R^2	*p*-value
PCA	.0038	.129	.0027	.192	.0290	.001	.0077	.016	.0111	.004	.0045	.072
VAE	.0479	.001	.0245	.002	.0613	.001	.0254	.001	.0129	.001	.0053	.061
UMAP	.1164	.001	.0068	.014	.0931	.001	.0082	.001	.1271	.001	.1245	.001
t-SNE	.1501	.001	.0190	.001	.0673	.001	.0141	.001	.0988	.001	.0216	.001
AEgen + PCA	.2821	.001	.4645	.001	.3629	.001	**.4596**	.001	.2795	.001	.3619	.001
VAESeq	**.3103**	.001	**.5041**	.001	**.3648**	.001	.4567	.001	**.3609**	.001	**.4834**	.001

Note: For AEgen + PCA, we performed a PCA on the embedding generated by the autoencoder in order to evaluate the results of the first part of the models. VAE designates a simpler variational autoencoder, i.e., to which the autoencoder embedding of the nucleotide sequence has not been given as input. See Table S1 for more details on the inputs for the different methods. We explore the use of multiple regression on distance matrices (MRM), and tests of statistical significance are performed using permutations. TEST 1 performs the multiple regression between the distance matrix between the sample point distances of the two-dimensional latent spaces of each method and Jaccard's β-diversity between the samples. In TEST 2, instead of using the distance matrix based on Jaccard's β-diversity, we use the distance matrix calculated on the β-diversity between sequences within the Hill number framework. The best results for each test are shown in bold.

β-diversity matrix, uses the distance matrix calculated on the β-diversity between sequences within the Hill number framework. In addition, we examined the ability of our methods to represent the MOTU richness and sequence α-diversity indicators through linear correlations with the two dimensional space. For simplicity, we only reported the correlations for the axis with the highest correlation denoted as V.

The baseline methods were performed using R version 4.1.3 with the packages 'stats' for PCA, 'umap' (Konopka, 2019) for UMAP, 'Rtsne' (Kjær et al., 2022) for t-SNE, 'TensorFlow' with 'Keras' for VAE. The NMDS method was performed using the function *metaMDS* from the package 'vegan' specifying 'Jaccard's' as distance. We used the *MRM* function of the 'ecodist' package to perform the tests. We computed the Euclidean distance matrix between each pair of samples in the 2D latent spaces using the function *dist* from the package 'stats' and the Jaccard's β-diversity using the library 'betapart'. We calculated the distance matrix of sequence β-diversity between each pair of samples using the Hill number framework (Abadi et al., 2016). The genetic distance between each pair of sequences was computed with the function *dist.gene* from the package 'ape'. Sequence β-diversity was calculated with the function *beta.fd.hill* from package 'mFD', with parameters $q = 1$ and τ = 'mean' (Chao et al., 2020; Magneville et al., 2022), using Sørensen's β-diversity.

We have visualized the 2D space results of ENNBetaDist in the Western Mediterranean fish dataset on a geographic map (Figure 4). To visualize these data, we used the HSV (hue-saturation-value) approach to transform the coordinates of the points into colours. This visualization technique allows us to gain valuable insights into the 2D space results of the ENNBetaDist analysis. For each point, the HSV hue component was determined from the distance of the point from zero (i.e. the origin) of the 2D latent space. This resulted in creating a colour gradient. Furthermore, the HSV value component was derived from the slope of the vector defined by the point and the origin of the 2D latent space, introducing additional variations in the colour representation. The HSV saturation component was set to one. Points from full reserve (i.e. no-take areas) are represented by circles, while points from other locations are represented by squares. We used the *hsv* function from R's grDevices.

3 | RESULTS

3.1 | Comparison with other methods

We first tested the representation of sequence α-diversity, MOTU richness, Jaccard's β-diversity, and sequence β-diversity in the 2D spaces of several classical dimension reduction methods. We then compared the results of VAESeq and the autoencoder embedding followed by a PCA (AEgen + PCA) with the classical PCA, t-SNE, UMAP, and a simple VAE (Table 2). Likewise, we compared the results of ENNBetaDist with NMDS (Table 3) since both work with Jaccard's β-diversity information a priori.

TABLE 3 Comparison between the 2D representation of ENNBetaDist and the NMDS method.

	MOTU richness		Sequence α-diversity		β – Diversity			
					TEST 1		TEST 2	
	MOTU richness – V		Sequence α-diversity – V		D_{1s}–$D_{\beta\text{-jac}}$		D_{1s}–$D_{\beta\text{-gen}}$	
	R^2	*p*-value	R^2	*p*-value	R^2	*p*-value	R^2	*p*-value
Western Mediterranean Fish Dataset								
NMDS	0.0821	.001	0.0124	.001	0.0117	.001	0.0369	.006
ENNBetaDist	**0.9018**	.001	**0.7116**	.001	**0.4106**	.001	**0.7415**	.001
Tara Ocean Dictyochophyceae Dataset								
NMDS	0.7527	.001	0.6907	.001	0.3802	.001	0.5337	.001
ENNBetaDist	**0.9612**	.001	**0.8639**	.001	**0.4440**	.001	**0.6288**	.001
Tara Ocean Telonemia Dataset								
NMDS	0.5238	.001	0.4918	.001	**0.6428**	.001	0.5640	.001
ENNBetaDist	**0.9245**	.001	**0.8034**	.001	0.4411	.001	**0.5802**	.001

Note: The table shows the correlation test values for the two methods in representing the three different ecological indicators of MOTU richness, sequence α-diversity and β-diversity. For the MOTU richness and the sequence α-diversity, we study the linear regression with the two axes of the 2D space and we reported the best, represented by V. For the β-diversity, we explore the use of multiple regression on distance matrices (MRM), and tests of statistical significance are performed using permutations. TEST 1 performs the multiple regression between the distance matrix between the sample point distances of the two-dimensional latent spaces of each method and Jaccard's β-diversity between the samples. In TEST 2, instead of using the distance matrix based on Jaccard's β-diversity, we use the distance matrix calculated on the β-diversity between sequences within the Hill number framework. The best results for each test are shown in bold.

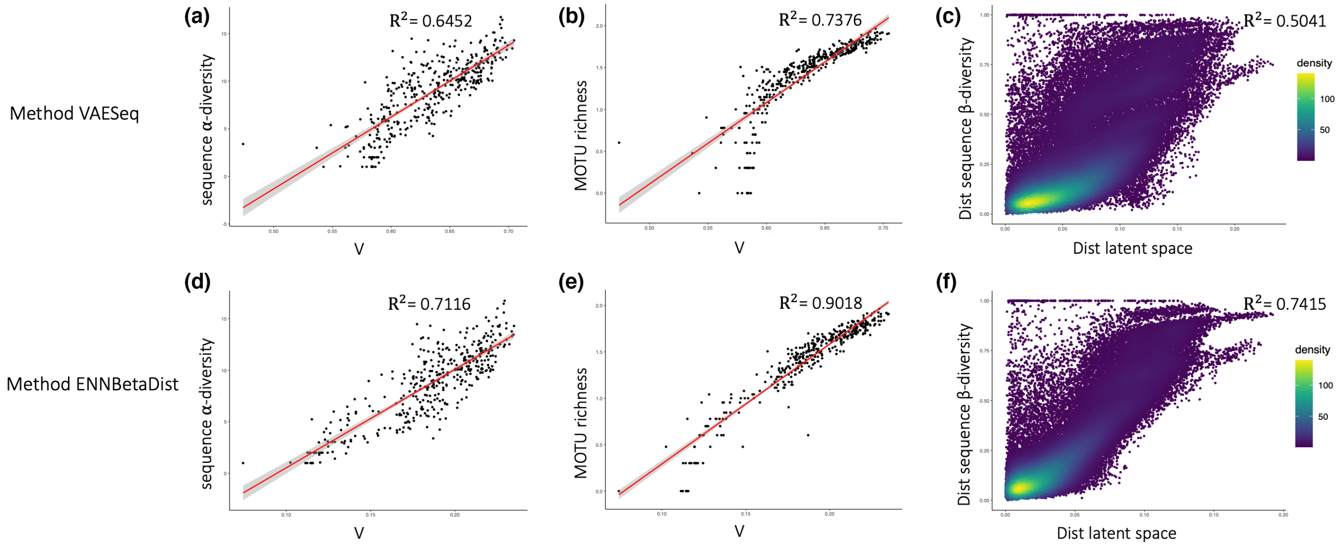

FIGURE 3 Correlation plots between the ordination of data samples in the 2D spaces of the VAESeq and ENNBetaDist with MOTU richness (a, d), sequence α-diversity (b, e) with one axe of the 2D latent spaces (V), and between the Euclidean distance between the samples in the 2D latent spaces of the two methods and the sequence β-diversity matrix (c, f) on the Western Mediterranean fish dataset. The R^2 values for each correlation test performed are shown in each figure ($p < .001$).

3.1.1 | VAESeq

Out of the three datasets considered, the highest R^2 values were achieved by VAESeq and AEgen + PCA (Table 2, Table S2). Indeed, when we tested the correlation between the distance matrix of samples in 2D latent spaces and the Jaccard's distance matrix (TEST 1), the R^2 values of AEgen + PCA and VAESeq (VAESeq: $R^2 = .34$; AEgen + PCA: $R^2 = .31$) are almost 10 times higher than those of PCA ($R^2 = .04$), and 3 times higher than those of *t*-SNE ($R^2 = .11$), and UMAP ($R^2 = .11$). The worst-performing method was PCA, which is the only linear method used. For example, in the Western Mediterranean fish dataset, the first principal component of PCA explained only 2.9% of the variance in the data, and the correlation was not significant in any test (Figure S2a, Table 2). VAE failed to extract information for the three datasets when the nucleotide sequence was not used ($R^2 = .04$). Furthermore, for VAESeq and AEgen + PCA, the R^2 values increased when we tested the correlation between the distance matrix of samples in 2D latent space and the distance matrix

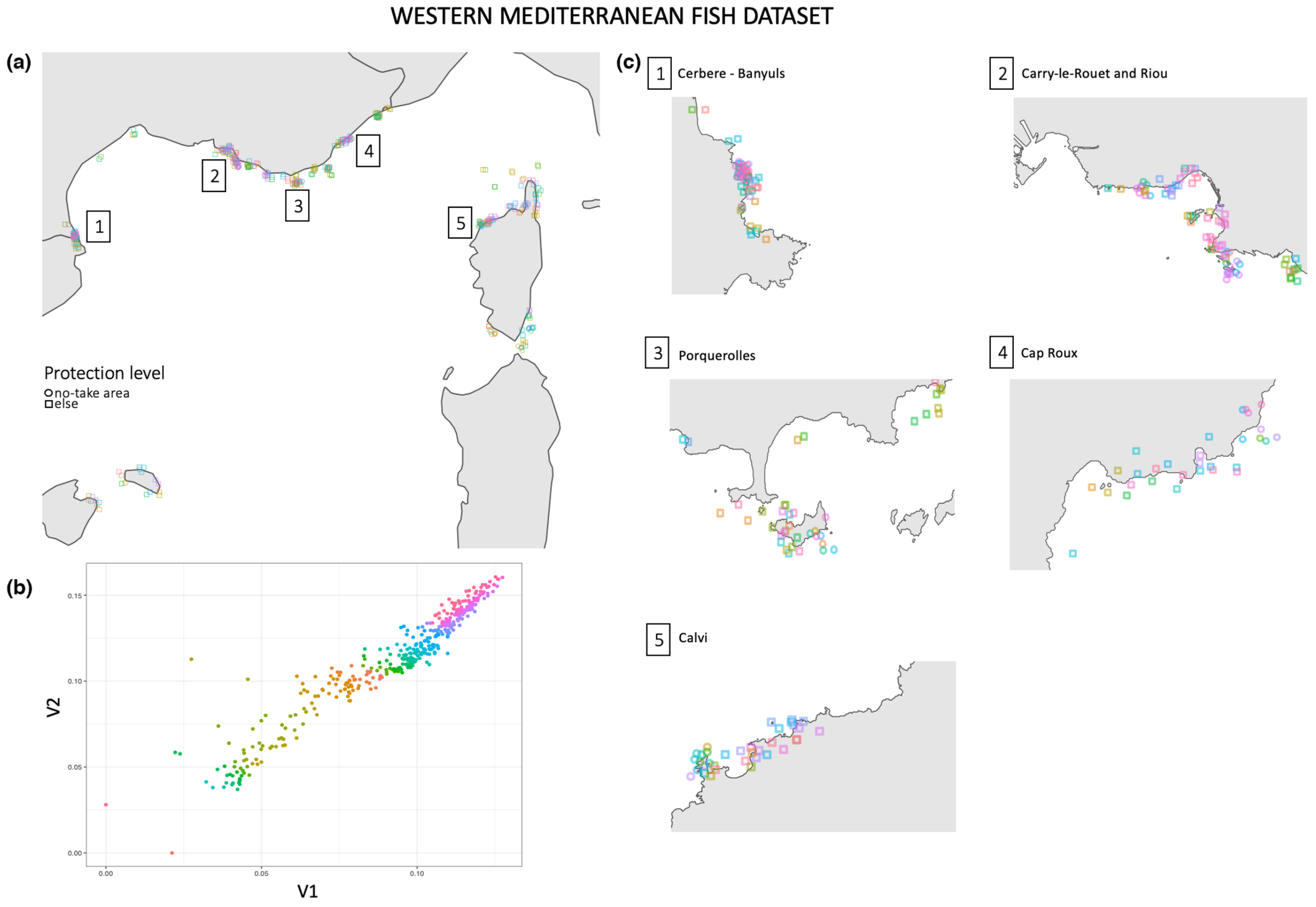

FIGURE 4 Geographical map of the Western Mediterranean fish dataset (a), illustrating the resulting 2D latent space of ENNBetaDist. To achieve this visualization, we employed the HSV (hue-saturation-value) approach, which allowed us to transform the coordinates of eDNA samples into colors (b). Specifically, the HSV hue component of each sample was determined based on its distance from zero in the 2D latent space, resulting in a colour gradient. The HSV value component was derived from the slope of the vector defined by the sample and the origin, introducing further colour variations. The HSV saturation component was set to one, ensuring a consistent saturation for all samples. Circles represent samples from reserves (i.e. no-take areas), while squares represent other regions. The visualization revealed interesting insights, particularly in identifying the pink cluster, corresponding to protected areas in the Mediterranean Sea and its nearby regions. Moving from west to east, these areas were identified as Cerbere-Banyuls, Carry-le-Rouet and Riou, Porquerolles, Cap Roux, and Calvi (c).
Note: To better visualize the data in the maps (c), we applied a small amount of noise to the coordinates of the samples, resulting in artificial distribution points on land.

based on genetic β-diversity (for VAESeq: R^2 = .48; for AEgen + PCA R^2 = .43). VAESeq and AEgen+PCA were able to accurately represent the gradient of MOTU richness and sequence α-diversity for all three datasets, where the other classical methods failed (Table S2).

3.1.2 | ENNBetaDist

For Tara Ocean datasets, both models succeeded in representing the Jaccard's β-diversity (for ENNBetaDist: R^2 = .44; For NMDS: R^2 = .51), and the sequence β-diversity (for ENNBetaDist: R^2 = .61; For NMDS: R^2 = .55) (Table 3). For the Western Mediterranean fish dataset ENNBetaDist succeeded in representing the Jaccard's β-diversity (R^2 = .41) and the sequence β-diversity (R^2 = .74), while NMDS failed due to the presence of one outlier (Jaccard's β-diversity R^2 = .01; sequence β-diversity R^2 = .04; Figure S3). By removing the outlier, NMDS and ENNBetaDist achieved similar β-diversity estimates. Moreover, for the representation of MOTU richness and sequence α-diversity, ENNBetaDist proved to be better than NMDS in representing both indicators, achieving values of R^2 > .92 for MOTU richness, and R^2 > .71 for sequence α-diversity in the three datasets (Table 3).

We describe below in more detail only the results of the Western Mediterranean eDNA fish dataset, while the results for the other datasets are provided in Figure S4.

3.2 | Latent spaces representations and ecological interpretation on Western Mediterranean eDNA dataset

The 2D latent space representations of the eDNA fish samples using two new methods (Figure S5) revealed marked gradients both in terms of MOTU richness (Figures 3a–c) and sequence α-diversity (Figures 3b–d). The two gradients were visible along both directions of the 2D latent space. For simplicity, we only reported the axis with the highest correlation denoted as V. For both methods, we found that V was significantly and positively correlated with the sequence α-diversity (Pearson's $r=.80$, $p<.001$ for VAESeq Figure 3a; $r=.84$, $p<.001$ for ENNBetaDist Figure 3d). V was also significantly and positively correlated with the MOTU richness ($r=.86$, $p<.001$ for VAESeq Figure 3b; $r=.95$, $p<.001$ for ENNBetaDist Figure 3e). We tested the correlation between the 2D spatial representation of VAESeq and ENNBetaDist with Jaccard's and sequence β-diversity matrices (TEST 1 and 2; Tables 2 and 3). We found that the distance between samples in the 2D spaces of the different methods was strongly correlated with the distance between samples expressed as Jaccard's and sequence β-diversity. Furthermore, we reported the resulting 2D latent space of ENNBetaDist in the geographical map of the Western Mediterranean Sea using an HSV approach (Figures 4a,b). The geographical map of the dataset showed a clustering of protection levels (Figure 4a). This was supported by a significant association tested by a generalized linear model (GLM) between the level of protection and the distance of the samples from zero in the 2D latent space (HSV hue value) (likelihood ratio test on the GLM, p-value=.001). Remarkably pink clusters on the map corresponded to the protected areas located in the Mediterranean Sea and its immediate surroundings. Moving from west to east, these areas were identified as Cerbere-Banyuls, Carry-le-Rouet and Riou, Porquerolles, Cap Roux and Calvi (Figure 4c). Additionally, the pink cluster also corresponded to eDNA samples with the highest values of MOTU richness and the highest values of sequence α-diversity (Figure S5c,d).

4 | DISCUSSION

The arrival of big data in ecology, facilitated by new technologies (Besson et al., 2022; Farley et al., 2018), makes dimensionality reduction, as well as data visualization, important analytical tools for ecological interpretations. In this study, we introduce two new deep learning-based methods that combine different types of NNs to ordinate eDNA samples and visualize ecosystem properties in a two-dimensional space: the first is based on variational autoencoder (VAE), and the second on deep metric learning (DML). The strength of our new methods lies in their ability to combine multiple inputs simultaneously, namely the number of sequences found for each molecular operational taxonomic unit (MOTU) detected and their corresponding nucleotide sequence. Using three different datasets - a fish eDNA dataset collected in the Mediterranean Sea (Boulanger et al., 2021), and two eukaryotic plankton eDNA datasets from the Tara Ocean expedition (de Vargas et al., 2015) – we show that our methods are able to represent three different biodiversity indicators in the two-dimensional latent space: (i) MOTU richness per sample, (ii) sequence α-diversity per sample and (iii) Jaccard's and sequence β-diversity between samples along a gradient in the latent space (Figure 3 and Figure S4, Tables 2 and 3). Thus, the 2D representation obtained reveals the ecological information underlying the structure of fish communities. In addition, we highlight how VAESeq outperforms other dimensionality reduction techniques, such as PCA, t-SNE, UMAP and even a simple VAE without sequence information added, in the visualization of ecosystem properties (Table 2). Likewise, ENNBetaDist outperforms NMDS in 11 of the 12 tests performed and manages to cope with the presence of an outlier in the Western Mediterranean fish dataset (Table 3).

In contrast, linear methods such as PCA result in poor dimensionality reduction to ordinate eDNA samples (Table 2; Figure S2a). This is due to the complexity of eDNA data (Miya, 2022; Xiong et al., 2022). Indeed, despite its potential in biodiversity monitoring (Mathon et al., 2022; Pawlowski et al., 2022; van der Heyde et al., 2022), eDNA metabarcoding can be limited by false reads due to contamination, errors that can occur during the extraction, PCR or sequencing process (Bohmann et al., 2014; Calderón-Sanou et al., 2020; Creer et al., 2016; Ficetola et al., 2016; Hering et al., 2018). Although field and laboratory practices can mitigate some of this, the risk of error cannot be eliminated and must be considered (Burian et al., 2021). Furthermore, eDNA metabarcoding sampling produces large, high-resolution datasets that are complex and highly dimensional, with a single observation from the experimental system containing measurements describing multiple traits (Hallam et al., 2021). For this reason, the application of neural architectures such as VAESeq and ENNBetaDist provides a better solution for understanding and representing eDNA data.

Neural networks allow for the integration of multiple inputs into a single model (Cichy & Kaiser, 2019; LeCun et al., 2015; Schmidhuber, 2015). This is particularly relevant for the analysis of eDNA metabarcoding data, which combines different types of information (Table S1). Our two new methods combine the number of sequences found for each MOTU and the nucleotide sequence of the detected MOTUs, which provide complementary information about the rarity and dissimilarity of the sequences, respectively. Our methods can then represent eDNA samples in 2D space, placing samples in relation to each other according to their composition (Figure 3, Figure S4). Due to the process of phylogenetic niche conservatism (Wiens et al., 2010) and environmental filtering (Guimarães, 2020), species present in a particular habitat or under a particular management (e.g. reserve) may show some phylogenetic and trait clustering (Jarzyna et al., 2021). In the context of eDNA, it is therefore expected that if two MOTUs are present in the same habitat, their nucleotide sequence similarity, even based on a short sequence, will be higher than for MOTUs from different habitats. Therefore, this genetic 'proximity' information, taken into account in our two methods, contributes to the ordination of eDNA samples in a lower-dimensional space along ecological, environmental or management gradients. Furthermore, despite the short length of the recovered sequence in metabarcoding (teleo fish marker is approximately 60 pb, Table 1), our results indicate that such nucleotide sequence information can inform species ecology and biogeography. Here, the manipulation of the nucleotide sequence highlights the proximity of sequences where the respective MOTUs are present in the different samples. Therefore, the composition of each MOTU together with its DNA sequence improves the representation of fish biodiversity and its indicators. In addition, we show that a simple VAE, using only the information on the number of sequences present in each sample, provides a poor representation of the data (Table 2). The sequences turn out to be crucial for good information extraction.

Instead of relying solely on the number of sequences identified per MOTU or the nucleotide sequences as in Cordier et al. (2021) and Flück et al. (2022), our methods combine both information (Table S1). VAESeq is based on VAE that is optimized to reconstruct the input data. ENNBetaDist is a DML method that also uses the diversity information (here the β-diversity) as a distance metric between samples. Using VAESeq for data extraction has the advantage of treating each data independently because it does not rely on any pairwise distance between samples. In this case, the model is free to discover connections and highlights possible new ones in a fully unsupervised learning process. Alternatively, ENNBetaDist helps to represent samples in a 2D latent space according to an input metric. ENNBetaDist is also able to represent the data and the three biodiversity indicators in the 2D space despite the strong perturbation due to an outlier, in contrast to the NMDS method (Table 3, Figures S3–S5). In addition, two new methods allow users to define the output dimensionality while preserving the global geometry (i.e. relative positions in the 2D latent space) better than classical methods. Furthermore, in the case of the Western Mediterranean eDNA dataset, ENNBetaDist revealed a clustering of protected sites (pink cluster on the map, Figure 4a, c). This result of an effect of protection on β-diversity supports previous eDNA studies analyzing the same dataset (Boulanger et al., 2021; Dalongeville, Boulanger, et al., 2022; Dalongeville, Nielsen, et al., 2022) and non-eDNA studies in the Mediterranean (Giakoumi et al., 2017).

Our results demonstrate that NNs provide a more efficient way of extracting structure from eDNA metabarcoding data than traditional dimension reduction methods, thereby improving future ecological interpretation. The resulting biodiversity indices can thus be used in future applications to improve our understanding of the processes behind spatial patterns coming from other types of monitoring approaches and in any other fields. Visualizing ecosystem eDNA sequences can improve our understanding of biodiversity and ecosystem properties, and thus help stakeholders in their decisions.

AUTHOR CONTRIBUTIONS

L. Lamperti, S. Si Moussi, B. Flück, L. Pellissier, and S. Manel conceived the ideas and designed the methodology; M. Bruno and A. Valentini analyzed the eDNA fish data samples; L. Lamperti, C. Albouy, D. Mouillot, T. Sanchez, S. Manel and L. Pellissier led the writing of the manuscript. All authors critically contributed to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

This work is part of the Horizon 2020-Marie Skodowska-Curie Actions-COFUND project Artificial Intelligence for the Sciences (AI4theSciences) and the IA-Biodiv ANR project FISH-PREDICT (ANR-21-AAFI-0001-01). Funding was provided by the ANR Lab-Com DiagADNe (ANR-20-LCV1-0008) and Agence de l'Eau RMC to perform Mediterranean eDNA sampling and metabacoding. This project was partly funded by the WSL internal grant 'eDNA Proof' and the Swiss Data Science Centre 'DNai'. We would also like to thank Olivier François and Giovanni Poggiato for their help in carrying out this study. Open access funding provided by Eidgenossische Technische Hochschule Zurich.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The Western Mediterranean fish dataset is archived in the Dryad Digital Repository at 10.5061/dryad.18931zcx1 and 10.5061/dryad.j9kd51cbr (Boulanger et al., 2021). The Tara Ocean datasets are archived in Pangaea at http://doi.pangaea.de/10.1594/PANGAEA.843022 (de Vargas et al., 2015). Codes and scripts for reproducing the analyses in this manuscript are available at https://github.com/letizialamperti/DLeDNA.

ORCID

Letizia Lamperti https://orcid.org/0000-0001-8059-1354
Benjamin Flück https://orcid.org/0000-0002-0396-6383
Alice Valentini https://orcid.org/0000-0001-5829-5479
Loïc Pellissier https://orcid.org/0000-0002-2289-8259
Stéphanie Manel https://orcid.org/0000-0001-8902-6052

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A system for Large-Scale Machine Learning. *Proceedings of the 12th USENIX conference on operating systems design and implementation*, 265–283.

Agersnap, S., Sigsgaard, E. E., Jensen, M. R., Avila, M. D. P., Carl, H., Møller, P. R., Krøs, S. L., Knudsen, S. W., Wisz, M. S., & Thomsen, P. F. (2022). A National Scale "BioBlitz" using citizen science and eDNA Metabarcoding for monitoring coastal marine fish. *Frontiers in Marine Science*, *9*. https://doi.org/10.3389/fmars.2022.824100

Bakker, J., Wangensteen, O. S., Chapman, D. D., Boussarie, G., Buddo, D., Guttridge, T. L., Hertler, H., Mouillot, D., Vigliola, L., & Mariani, S. (2017). Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. *Scientific Reports*, *7*(1), 16886. https://doi.org/10.1038/s41598-017-17150-2

Battey, C. J., Coffing, G. C., & Kern, A. D. (2021). Visualizing population structure with variational autoencoders. *G3: Genes*, *Genomes*, *Genetics*, *11*(1). https://doi.org/10.1093/G3JOURNAL/JKAA036

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

Besson, M., Alison, J., Bjerge, K., Gorochowski, T. E., Høye, T. T., Jucker, T., Mann, H. M. R., & Clements, C. F. (2022). Towards the fully automated monitoring of ecological communities. *Ecology Letters*, *25*, 2753–2775.

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, *29*(6), 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Boulanger, E., Loiseau, N., Valentini, A., Arnal, V., Boissery, P., Dejean, T., Deter, J., Guellati, N., Holon, F., Juhel, J. B., Lenfant, P., Manel, S., & Mouillot, D. (2021). Environmental DNA metabarcoding reveals and unpacks a biodiversity conservation paradox in Mediterranean marine reserves. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1949), 20210112. https://doi.org/10.1098/rspb.2021.0112

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. https://doi.org/10.1111/1755-0998.13367

Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2020). From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? *Journal of Biogeography*, *47*(1), 193–206. https://doi.org/10.1111/jbi.13681

Cantera, I., Coutant, O., Jézéquel, C., Decotte, J.-B., Dejean, T., Iribar, A., Vigouroux, R., Valentini, A., Murienne, J., & Brosse, S. (2022). Low level of anthropization linked to harsh vertebrate biodiversity declines in Amazonia. *Nature Communications*, *13*(1), 3290. https://doi.org/10.1038/s41467-022-30842-2

Chao, A., Kubota, Y., Zelený, D., Chiu, C.-H., Li, C.-F., Kusumoto, B., Yasuhara, M., Thorn, S., Wei, C.-L., Costello, M. J., & Colwell, R. K. (2020). Quantifying sample completeness and comparing diversities among assemblages. *Ecological Research*, *35*(2), 292–314. https://doi.org/10.1111/1440-1703.12102

Chollet, F., & Others. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cinner, J. E., Zamborain-Mason, J., Gurney, G. G., Graham, N. A. J., MacNeil, M. A., Hoey, A. S., Mora, C., Villéger, S., Maire, E., McClanahan, T. R., Maina, J. M., Kittinger, J. N., Hicks, C. C., D'agata, S., Huchery, C., Barnes, M. L., Feary, D. A., Williams, I. D., Kulbicki, M., … Mouillot, D. (2020). Meeting fisheries, ecosystem function, and biodiversity goals in a human-dominated world. *Science*, *368*(6488), 307–311. https://doi.org/10.1126/science.aax9412

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, *30*(13), 2937–2958. https://doi.org/10.1111/mec.15472

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. In *Methods in ecology and evolution* (Vol. 7, pp. 1008–1018). British Ecological Society. https://doi.org/10.1111/2041-210X.12574

Dalongeville, A., Boulanger, E., Marques, V., Charbonnel, E., Hartmann, V., Santoni, M. C., Deter, J., Valentini, A., Lenfant, P., Boissery, P., Dejean, T., Velez, L., Pichot, F., Sanchez, L., Arnal, V., Bockel, T., Delaruelle, G., Holon, F., Milhau, T., … Mouillot, D. (2022). Benchmarking eleven biodiversity indicators based on environmental DNA surveys: More diverse functional traits and evolutionary lineages inside marine reserves. *Journal of Applied Ecology*, *59*(11), 2803–2813. https://doi.org/10.1111/1365-2664.14276

Dalongeville, A., Nielsen, E. S., Teske, P. R., & von der Heyden, S. (2022). Comparative phylogeography in a marine biodiversity hotspot provides novel insights into evolutionary processes across the Atlantic-Indian Ocean transition. *Diversity and Distributions*, *28*(12), 2622–2636. https://doi.org/10.1111/ddi.13534

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., … Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605. https://doi.org/10.1126/science.1261605

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

Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2021). A review of UMAP in population genetics. *Journal of Human Genetics*, *66*(1), 85–91. https://doi.org/10.1038/s10038-020-00851-4

Duffner, S., Garcia, C., Idrissi, K., & Baskurt, A. (2021). *Similarity Metric Learning*. https://hal.archives-ouvertes.fr/hal-03465119

Facco, E., D'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, *7*(1), 12140. https://doi.org/10.1038/s41598-017-11873-y

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *Bioscience*, *68*(8), 563–576. https://doi.org/10.1093/biosci/biy068

Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, *16*(3), 604–607. https://doi.org/10.1111/1755-0998.12508

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. https://doi.org/10.1038/s41598-022-13412-w

Floyd, R., Abebe, E., Papert, A., & Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology*, *11*(4), 839–850. https://doi.org/10.1046/j.1365-294X.2002.01485.x

Frainer, A., Primicerio, R., Kortsch, S., Aune, M., Dolgov, A. v., Fossheim, M., & Aschan, M. M. (2017). Climate-driven changes in functional biogeography of Arctic marine fish communities. *Proceedings of the National Academy of Sciences*, *114*(46), 12202–12207. https://doi.org/10.1073/pnas.1706080114

Giakoumi, S., Scianna, C., Plass-Johnson, J., Micheli, F., Grorud-Colvert, K., Thiriet, P., Claudet, J., Di Carlo, G., Di Franco, A., Gaines, S. D., García-Charton, J. A., Lubchenco, J., Reimer, J., Sala, E., & Guidetti, P. (2017). Ecological effects of full and partial protection in the crowded Mediterranean Sea: A regional meta-analysis. *Scientific Reports*, *7*(1), 8940. https://doi.org/10.1038/s41598-017-08850-w

Grønbech, C., Vording, M., Timshel, P., Sønderby, C., Pers, T., & Winther, O. (2018). *scVAE: Variational auto-encoders for single-cell gene expression data*. https://doi.org/10.1101/318295

Guimarães, P. R. (2020). The structure of ecological networks across levels of organization. *Annual Review of Ecology, Evolution, and Systematics*, *51*(1), 433–460. https://doi.org/10.1146/annurev-ecolsys-012220-120819

Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., & Courville, A. (2016). *PixelVAE: A Latent Variable Model for Natural Images*. http://arxiv.org/abs/1611.05013

Hallam, J., Clare, E. L., Jones, J. I., & Day, J. J. (2021). Biodiversity assessment across a dynamic riverine system: A comparison of eDNA metabarcoding versus traditional fish surveying methods.

Environmental DNA, *3*(6), 1247–1266. https://doi.org/10.1002/edn3.241

Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B., Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., & Kelly, M. (2018). Implementation options for DNA-based identification into ecological status assessment under the European water framework directive. *Water Research*, *138*, 192–205. https://doi.org/10.1016/j.watres.2018.03.003

Hou, Y., Li, Z., Wang, P., & Li, W. (2018). Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(3), 807–811. https://doi.org/10.1109/TCSVT.2016.2628339

Jarzyna, M. A., Quintero, I., & Jetz, W. (2021). Global functional and phylogenetic structure of avian assemblages across elevation and latitude. *Ecology Letters*, *24*(2), 196–207. https://doi.org/10.1111/ele.13631

Johnston, E. L., Clark, G. F., & Bruno, J. F. (2022). The speeding up of marine ecosystems. *Climate Change Ecology*, *3*, 100055. https://doi.org/10.1016/j.ecochg.2022.100055

Jouffray, J. B., Blasiak, R., Norström, A. v., Österblom, H., & Nyström, M. (2020). The blue acceleration: The trajectory of human expansion into the ocean. In *One earth* (Vol. *2*, pp. 43–54). Cell Press. https://doi.org/10.1016/j.oneear.2019.12.016

Karl Pearson, F. R. S. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Kjær, K. H., Winther Pedersen, M., de Sanctis, B., de Cahsan, B., Korneliussen, T. S., Michelsen, C. S., Sand, K. K., Jelavić, S., Ruter, A. H., Schmidt, A. M. A., Kjeldsen, K. K., Tesakov, A. S., Snowball, I., Gosse, J. C., Alsos, I. G., Wang, Y., Dockter, C., Rasmussen, M., Jørgensen, M. E., … Consortium, P. (2022). A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature*, *612*(7939), 283–291. https://doi.org/10.1038/s41586-022-05453-y

Konopka, T. (2019). *umap: Uniform manifold approximation and projection. R package version 0.2.3.1*. https://CRAN.R-project.org/package=umap

Kulis, B. (2013). Metric learning: A survey. *Foundations and Trends® in Machine Learning*, *5*(4), 287–364. https://doi.org/10.1561/2200000019

Lafarge, M. W., Caicedo, J. C., Carpenter, A. E., Pluim, J. P. W., Singh, S., & Veta, M. (2019). Capturing single-cell phenotypic variation via unsupervised representation learning. In M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. Unal, & T. Vercauteren (Eds.), *Proceedings of the 2nd international conference on medical imaging with deep learning* (Vol. *102*, pp. 315–325). PMLR. https://proceedings.mlr.press/v102/lafarge19a.html

Lafarge, M. W., Pluim, J. P. W., Eppenhof, K. A. J., & Veta, M. (2019). Learning domain-invariant representations of histological images. *Frontiers in Medicine*, *6*. https://doi.org/10.3389/fmed.2019.00162

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (Vol. *48*, pp. 1558–1566). PMLR. https://proceedings.mlr.press/v48/larsen16.html

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Magneville, C., Loiseau, N., Albouy, C., Casajus, N., Claverie, T., Escalas, A., Leprieur, F., Maire, E., Mouillot, D., & Villéger, S. (2022). mFD: an R package to compute and illustrate the multiple facets of functional diversity. *Ecography*, *2022*. https://doi.org/10.1111/ecog.05904

Makiola, A., Dickie, I. A., Holdaway, R. J., Wood, J. R., Orwin, K. H., & Glare, T. R. (2019). Land use is a determinant of plant pathogen alpha- but not beta-diversity. *Molecular Ecology*, *28*, 3786–3798. https://doi.org/10.1111/mec.15177

Marques, V., Castagné, P., Polanco, A., Borrero-Pérez, G. H., Hocdé, R., Guérin, P. É., Juhel, J. B., Velez, L., Loiseau, N., Letessier, T. B., Bessudo, S., Valentini, A., Dejean, T., Mouillot, D., Pellissier, L., & Villéger, S. (2021). Use of environmental DNA in assessment of fish functional and phylogenetic diversity. *Conservation Biology*, *35*(6), 1944–1956. https://doi.org/10.1111/cobi.13802

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. https://doi.org/10.1111/ecog.05049

Marques, V., Milhau, T., Albouy, C., Dejean, T., Manel, S., Mouillot, D., & Juhel, J. B. (2021). GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Diversity and Distributions*, *27*(10), 1880–1892. https://doi.org/10.1111/ddi.13142

Mathon, L., Marques, V., Mouillot, D., Albouy, C., Baletaud, F., Borrero-Pérez, G. H., Dejean, T., Edgar, G. J., Grondin, J., Guerin, P.-E., Hocdé, R., Juhel, J.-B., Maire, E., Mariani, G., McLean, M., Polanco, A. F., Pouyaud, L., Stuart-Smith, D., Yulia Sugeha, H., … dan Domestikasi, K. (2022). Cross-ocean patterns and processes in fish biodiversity on coral reefs through the lens of eDNA metabarcoding. *Proceedings of the Royal Society B: Biological Sciences*, *289*, 20220162.

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. http://arxiv.org/abs/1802.03426

McLean, M., Auber, A., Graham, N. A. J., Houk, P., Villéger, S., Violle, C., Thuiller, W., Wilson, S. K., & Mouillot, D. (2019). Trait structure and redundancy determine sensitivity to disturbance in marine fish communities. *Global Change Biology*, *25*(10), 3424–3437. https://doi.org/10.1111/gcb.14662

Miya, M. (2022). Environmental DNA Metabarcoding: A novel method for biodiversity monitoring of marine fish communities. *Annual Review of Marine Science*, *14*(1), 161–185. https://doi.org/10.1146/annurev-marine-041421-082251

Muff, M., Jaquier, M., Marques, V., Ballesta, L., Deter, J., Bockel, T., Hocdé, R., Juhel, J.-B., Boulanger, E., Guellati, N., Fernández, A. P., Valentini, A., Dejean, T., Manel, S., Albouy, C., Durville, P., Mouillot, D., Holon, F., & Pellissier, L. (2023). Environmental DNA highlights fish biodiversity in mesophotic ecosystems. *Environmental DNA*, *5*(1), 56–72. https://doi.org/10.1002/edn3.358

Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLoS Computational Biology*, *15*(6), 1–19. https://doi.org/10.1371/journal.pcbi.1006907

Nissen, J. N., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Bjørn Nielsen, H., Petersen, T. N., Winther, O., & Rasmussen, S. (2018). Binning microbial genomes using deep learning. *BioRxiv*. https://doi.org/10.1101/490078

Pawlowski, J., Bruce, K., Panksep, K., Aguirre, F. I., Amalfitano, S., Apothéloz-Perret-Gentil, L., Baussant, T., Bouchez, A., Carugati, L., Cermakova, K., Cordier, T., Corinaldesi, C., Costa, F. O., Danovaro, R., Dell'Anno, A., Duarte, S., Eisendle, U., Ferrari, B. J. D., Frontalini, F., … Fazi, S. (2022). Environmental DNA metabarcoding for benthic monitoring: A review of sediment sampling and DNA extraction methods. *Science of the Total Environment*, *818*, 151783. https://doi.org/10.1016/j.scitotenv.2021.151783

R Core Team. (2022). *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

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, *15*, 201–292. https://doi.org/10.2307/1412107

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, *105*(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740

Tigchelaar, M., Leape, J., Micheli, F., Allison, E. H., Basurto, X., Bennett, A., Bush, S. R., Cao, L., Cheung, W. W. L., Crona, B., DeClerck, F., Fanzo, J., Gelcich, S., Gephart, J. A., Golden, C. D., Halpern, B. S., Hicks, C. C., Jonell, M., Kishore, A., … Wabnitz, C. C. C. (2022). The vital roles of blue foods in the global food system. *Global Food Security*, *33*, 100637. https://doi.org/10.1016/j.gfs.2022.100637

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*, 401–419. https://doi.org/10.1007/BF02288916

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.-M., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*, 929–942. https://doi.org/10.1111/mec.13428

van der Heyde, M., Bunce, M., & Nevill, P. (2022). Key factors to consider in the use of environmental DNA metabarcoding to monitor terrestrial ecological restoration. *Science of the Total Environment*, *848*, 157617. https://doi.org/10.1016/j.scitotenv.2022.157617

van der Maaten, L., & Hinton, G. (2008). Visualizing data using *t*-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wang, D., & Gu, J. (2018). VASC: Dimension reduction and visualization of single-cell RNA-seq data by deep Variational autoencoder. *Genomics, Proteomics and Bioinformatics*, *16*(5), 320–331. https://doi.org/10.1016/j.gpb.2018.08.003

Wiens, J. J., Ackerly, D. D., Allen, A. P., Anacker, B. L., Buckley, L. B., Cornell, H. v., Damschen, E. I., Jonathan Davies, T., Grytnes, J.-A., Harrison, S. P., Hawkins, B. A., Holt, R. D., McCain, C. M., & Stephens, P. R. (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*, *13*(10), 1310–1324. https://doi.org/10.1111/j.1461-0248.2010.01515.x

Xiong, F., Shu, L., Zeng, H., Gan, X., He, S., & Peng, Z. (2022). Methodology for fish biodiversity monitoring with environmental DNA metabarcoding: The primers, databases and bioinformatic pipelines. *Water Biology and Security*, *1*(1), 100007. https://doi.org/10.1016/j.watbs.2022.100007

SUPPORTING INFORMATION

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

How to cite this article: 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*, 1946–1958. https://doi.org/10.1111/1755-0998.13861