Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

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

Letizia Lamperti[1], Théophile Sanchez[2], Sara Si Moussi[3], David Mouillot[4], Camille Albouy[2], Benjamin Flück[2], Morgane Bruno[5], Alice Valentini[6], Loïc Pellissier[7], and Stéphanie Manel[8]

[1]EPHE PSL
[2]ETH Zurich
[3]TIMC-IMAG
[4]MARBEC
[5]CEFE
[6]SPYGEN
[7]ETH Zürich
[8]CNRS

April 15, 2023

Abstract

1. Metabarcoding of environmental DNA (eDNA) has recently improved our understanding of biodiversity patterns in marine and terrestrial ecosystems. However, the complexity of these data prevents current methods to extract and analyze all the relevant ecological information they contain. Therefore, ecological modeling could greatly benefit from new methods providing better dimensionality reduction and clustering. 2. Here we present two new deep learning-based methods that combine different types of neural networks to ordinate eDNA samples and visualize ecosystem properties in a two-dimensional space: the first is based on variational autoencoders (VAEs) and the second on deep metric learning (DML). The strength of our new methods lies in the combination of several inputs: the number of sequences found for each molecular operational taxonomic unit (MOTU), together with the genetic sequence information of each detected MOTU within an eDNA sample. 3. Using three different datasets, we show that our methods represent well three different ecological indicators in a two-dimensional latent space: MOTU richness per sample, sequence α-diversity per sample, and sequence -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, and Uniform Manifold Approximation and Projection for dimension reduction. 4. Our results suggest that neural networks provide a more efficient way of extracting structure from eDNA metabarcoding data, thereby improving their ecological interpretation and thus biodiversity monitoring.

New deep learning-based methods for visualizing ecosystem properties using environmental DNA metabarcoding data Letizia Lamperti[a,b], Théophile Sanchez[b], Sara Si Moussi[e] , David Mouillot[c,d], Camille Albouy[b], Benjamin Flück[b], Morgane Bruno[a], Alice Valentini[f], Loic Pellissier[b]*, Stephanie Manel[a,d] *[a] CEFE, Univ Montpellier, CNRS, **EPHE-PSL University**, IRD, Montpellier, France [b] Landscape Ecology, Inst. of Terrestrial Ecosystems, ETH Zurich, Zurich, Switzerland and Swiss Federal Research Inst. WSL, Birmensdorf, Switzerland[c] MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France [d] Institut Universitaire de France, Paris, France[e] TIMC-IMAG, LECA, Grenoble, France[f] SPYGEN, Savoie Technolac *Co-senior authors**Correspondence:** Letizia Lamperti, e-mail: letizia.lamperti@ephe.psl.eu**Abstract**Metabarcoding of environmental DNA (eDNA) has recently improved

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

our understanding of biodiversity patterns in marine and terrestrial ecosystems. However, the complexity of these data prevents current methods to extract and analyze all the relevant ecological information they contain. Therefore, ecological modeling could greatly benefit from new methods providing better dimensionality reduction and clustering. Here we present two new deep learning-based methods that combine different types of neural networks to ordinate eDNA samples and visualize ecosystem properties in a two-dimensional space: the first is based on variational autoencoders (VAEs) and the second on deep metric learning (DML). The strength of our new methods lies in the combination of several inputs: the number of sequences found for each molecular operational taxonomic unit (MOTU), together with the genetic sequence information of each detected MOTU within an eDNA sample. Using three different datasets, we show that our methods represent well three different ecological indicators in a two-dimensional latent space: MOTU richness per sample, sequence α-diversity per sample, 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, and Uniform Manifold Approximation and Projection for dimension reduction. Our results suggest that neural networks provide a more efficient way of extracting structure from eDNA metabarcoding data, thereby improving their ecological interpretation and thus biodiversity monitoring.**Keywords:** biodiversity monitoring; deep learning; deep metric learning; data visualization; environmental DNA; machine learning; neural networks; variational autoencoder **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 (Miya 2022; Deiner et al., 2017). eDNA metabarcoding is now operational in many ecosystems for a wide range of micro- and macroorganisms (Cantera et al., 2022; Kjær et al., 2022; Mathon et al., 2022; Cordier et al., 2021), providing information on their taxonomic, functional, but also phylogenetic affiliations (Marques et al., 2021; Rozanski et al., 2022). Given the low field effort and disturbance (Muff et al., 2022), even in the most remote locations, and the decrease in sequencing costs in recent years, this approach can be scaled up to monitor many sites at high temporal frequency (Agersnap et al., 2022). eDNA metabarcoding produces massive 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 (Marques et al., 2020) precludes the identification of many species, thus working with Molecular Operational Taxonomic Units (MOTUs) representing a cluster of similar sequences may be required (Deiner et al., 2017; 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 genetic sequence itself. Both attributes can be directly related to ecosystem states and properties (Shelton et al., 2019; Bakker et al., 2017). Therefore, eDNA data are potentially relevant for revealing 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 (Van Der Maaten et al., 2009; Nguyen and Holmes, 2019). Traditionally, dimensionality reduction is performed using linear techniques such as Principal Component Analysis (PCA; Pearson, 1901), Factor Analysis (Spearman, 1904), and classical scaling (Torgerson, 1952). However, due to their hypotheses, these linear techniques cannot adequately deal with complex non-linear relationships in data such as eDNA. In the last decade, many nonlinear techniques have been proposed for dimensionality reduc-

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

tion (Nguyen et al., 2019; Facco et al., 2017). Recently, two machine learning techniques - the t-distributed stochastic neighbour embedding (t-SNE; van der Maaten and Hinton, 2008) and the Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) - have shown promising results in generating two-dimensional visualisations of high-dimensional biological data (Diaz-Papkovich et al., 2019). 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., 2019). Although these methods perform well in clustering similar samples, distances between clusters are not always meaningful (Becht et al., 2019). However, neural networks (NN) have been shown to have a good representation of learning capacity (Sze et al., 2017). NN are complex mathematical models consisting of many operators called neurons, organized in a network of interconnected nodes. UMAP and t-SNE learn features by satisfying distances between observations, i.e., in contrast. 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 neural networks, 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., Larsen et al., 2015; Gulrajani et al., 2016; Hou et al., 2016), and several recent studies have applied them to dimensionality reduction and classification of single-cell RNAseq data (Grønbech et al., 2018; Lafarge et al., 2019; Wang and Gu 2018; Battey et al., 2019). Thanks to the design flexibility of artificial neural networks in general, they also have the advantage of being able to encode and mix information from different data types. Deep metric learning (DML) lies between contrastive and generative latent variable models. Metric learning is an approach directly based on a distance metric that aims to establish similarity or dissimilarity between objects (Kulis, 2013). Although metric learning 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 neural networks. VAE and DML have not yet demonstrated their potential to ordinate eDNA samples in a low-dimensional space. In this study, we present two new methods, one based on VAE and the other on DML, to perform data visualisation. In our methods, we assemble, modify, and adapt these neural networks with others to work best with eDNA data. We tested our methods on three different published eDNA datasets: a fish eDNA dataset collected in the Mediterranean Sea (Boulanger et al. 2020), and two eukaryotic plankton eDNA datasets from the Tara Ocean expedition (de Vargas et al. 2015). We used both the number of sequences per detected MOTU and the genetic sequence information of each MOTU detected in each sample as input data. To validate these two new methods, we compare them with three classical methods: PCA, t-SNE, and UMAP. Finally, we show how the proposed methods outperform classical methods in their representation of ecological indicators.**2 MATERIALS and METHODS 2.1 VAE-based method applied to eDNA Data**The VAE-based method, called VAESeq (Fig. 1), processes eDNA samples into a two-dimensional latent space. The model consists of an autoencoder (AE) and a variational autoencoder (VAE). The AE takes as input the genetic sequence information and the presence/absence of each MOTU within each sample to generate the first latent encoding*zAE*. The VAE encoder then receives the embedding generated by the genetic autoencoder and analyses it in combination with the number of sequences found for each MOTU detected in the sample under consideration. By mixing the two inputs, it encodes the samples as points in a 2D latent space called *zVAE*. In the decoding part, the VAE decoder seeks to recreate the two inputs from *zVAE*. 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 embed genetic sequences. Because we were interested in the compression of the genetic information rather than its representation, we decided to use an AE rather than a VAE. To encode the DNA sequence information in the AE, the sequences are equalised to the same length. We have chosen to keep the maximum length, adding nucleotides X to equalise. 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,

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

X nucleotides added to equalise the sequence length become [0.25, 0.25, 0.25, 0.25]. We combine the genetic information with the presence/absence of each MOTU in each sample. Therefore, each sample is represented by a tensor containing the translated binary 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 optimiser and the binary cross-entropy loss function to optimise the network. The AE encoder consisted of seven fully connected layers with decreasing widths down to 100, rectified linear unit activations, and dropout regularization. A mirror architecture was used as the decoder. The VAE component of the network uses the Adam optimiser 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 genetic sequence information. After testing 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. A mirror architecture to the encoder was used as the decoder. **2.2 ΔΜΛ-βασεδ μετηοδ ον β-διερσιτψ ας α δισταν ςε**The ENNBetaDist DML-based method (Fig. 2) trains the network accordingly to the pairwise β-diversity calculated between the 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 latent space. We used pairwise Jaccard dissimilarity as a measure of β-diversity, using the 'betapart' library in R. The structure of ENNBetaDist consists of two encoder neural networks. The encoders of ENNBetaDist are similar to those of VAESeq, with differences in the number of hidden layers, training, and optimisation. VAESeq reconstructs input from the latent space, however, we want the latent space to respect the distances we want to optimise. At each iteration, each encoder takes as input a sample containing the number of sequences per MOTU and the genetic latent encoding from the AE of that sample.Then, the two encoders process the two samples, combining the number of sequences found for each MOTU detected and the autoencoder embedding of the sequences into a point in the two-dimensional latent space *z1* and *z2*. To optimise the model, we calculate the Euclidean distance between the two points in *z1* and *z2* and compare it with the pairwise β-diversity via a loss function (the mean square error (MSE)). In *z1* and *z2* we find the visual representation of all data points. Indeed, ENNBetaDist represents 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 5 fully connected layers with decreasing widths down to 2, rectified linear unit activations, and dropout regularization.**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 the validation set. We implemented the models in R (version 4.1.3 R Core Team, 2020) using TensorFlow (Abadi et al., 2015) and Keras (Chollet, F. et al., 2015) libraries.**2.4 Case study**2.4.1 Data sets We tested our methods on three different published eDNA datasets: a fish eDNA dataset collected in the western Mediterranean Sea (Boulanger et al. 2020) 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 dataset were collected at 77 stations in six marine regions covering the western Mediterranean, including fished and no-take protected areas (Fig. S1) (Boulanger et al. 2020). 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 elasmobranchs, according to the protocol described in (Valentini et al. 2016). Data collection and sample processing are described in detail in Appendix 1 of the supplementary material. 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 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 and whose specifications are shown in Table 1.**2.5 Comparison and evaluation**We tested the ability of the two new methods to better represent the distance between samples based on their species and sequence composition. We compared the 2D representations of

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

VAESeq and ENNBetaDist with three other classical dimension reduction methods: PCA, which is linear, and t-SNE, UMAP, which are non-linear. Furthermore, we analysed the genetic embedding generated by the autoencoder using the PCA to evaluate the results of the first part of the models. We also compared it with a simpler VAE, i.e., to which the embedding of the sequence autoencoder has not been given as input. The inputs to PCA, t-SNE and UMAP are the presence/absence information of each MOTU in each sample and the number of sequences found during eDNA extraction for each detected MOTU (Table 3). Our neural network-based methods also receive genetic information on the sequences of the latent embedding of the autoencoder (Table 2). To evaluate the performance of all the methods, we used both multiple regressions on distance matrices (MRM) representing the sample in space (Table 3). 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 (Table 3). TEST 2, instead of using Jaccard's β-diversity matrix, uses the distance matrix calculated on the β-diversity between sequences within the Hill number framework (Table 3). The baseline methods were performed using R version 4.1.3 with the packages "stats" for PCA, "umap" (Konopka 2019) for UMAP, "Rtsne" (Jesse Krijthe 2022) for t-SNE and "TensorFlow" with "Keras" for VAE. 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 latent spaces using the function *dist* from the package "stats" and the Jaccard 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 (Alberdi 2019). 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 = "mean" (Chao et al. 2019), using Sørensen's β-diversity.**3 RESULTS3.3 Comparison with other methods**We decide to test the representations of the points in the 2D latent spaces of the three different data sets with the methods shown in Table 3. We test the 2D representations with the Jaccard β-diversity matrix and the sequence β-diversity matrix (TEST 1 and 2). Out of the three datasets considered, the highest R^2 values are achieved by the neural network-based methods. For our methods, the results of TEST 2 are better than those of TEST 1. In TEST 1, using the matrix based on Jaccard's β-diversity, the R^2 values in the latent space of the autoencoder (AEgen+PCA) are 10^2 times higher than those of PCA, t-SNE, and UMAP. Furthermore, for all three datasets, the R^2 values increase in the case of VAEseq and ENNBetaDist, that is, when the number of sequences is also given as input. The best results for all three datasets are obtained with ENNBetaDist, i.e., when the model is optimised in pairwise β-diversity. The worst-performing method is PCA, which is the only linear method used. Furthermore, we have shown that VAE fails to extract information for the three datasets when genetic sequence information is not added. We describe below in more detail the results in the case of the Mediterranean eDNA fish data set. The same results obtained on the Dictyochophyceae and Telonemia data sets are shown in Figure S2 in the supplementary material.**3.2 VA Latent spaces representations and ecological interpretation on western Mediterranean eDNA fish data set**The 2D latent space representations of the eDNA fish data samples using two new methods (Fig. S3) reveal gradients both in terms of MOTU richness (Figs. 3 a-c) and sequence α-diversity (Figs. 3 b-d). The two gradients were visible along both directions of the 2D latent space. For simplicity, we report only the studies along the vertical direction, where the correlations are the strongest. We performed correlation tests to validate the relationship between the vertical direction of our 2D latent space and the diversity metrics. For both methods, we found that the latent variable along the second axis is significantly correlated with the sequence α-diversity (Pearson's $r = 0.80$, $p < 0.001$ for VAESeq Fig. 3a; $r = 0.84$, $p < 0.001$for ENNBetaDist Fig. 3d). The same latent variable axis 2 is also significantly correlated with the MOTU richness ($r = 0.86$, $p < 0.001$ for VAESeq Fig. 3b; $r = 0.95$, $p < 0.001$ for ENNBetaDist Fig. 3e). We tested the correlation between the 2D spatial representation of the two new methods with the Jaccard's β-diversity matrix and with the sequence β-diversity matrix (TEST 1 and 2; Table 3). We found that the two new methods outperform traditional methods (i.e., PCA, t-SNE, UMAP, Fig. S4). The R^2 values of the VAESeq and ENNBetaDist are 10^2 times higher than PCA, and more than twice that of t-SNE and UMAP. The first principal component of the PC axes in the PCA on the eDNA dataset explains only 2.9% of the variance of the data, and there is a nonsignificant correlation in both tests (Fig. S4 a, Table 3). Among the

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

non-linear methods, ENNBetaDist shows the best results (TEST 2: $R^2 = 0.5041$; $p < 0.001$ for VAESeq and $R^2 = 0.7415$; $p < 0.001$ for ENNBetaDist; Table 3).**4 DISCUSSION**The massive arrival of big data in ecology, facilitated by new technologies (Farley et al. 2018; Besson et al. 2022), makes dimensionality reduction, as well as data visualization, an important analytical tool. In this study, we introduce two new deep learning-based methods that combine different types of neural networks 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 for eDNA the number of sequences found for each molecular operational taxonomic unit (MOTU), together with the genetic sequence information for each detected MOTU. Using three different datasets - a fish eDNA dataset collected in the Mediterranean Sea (Boulanger et al. 2020), and two eukaryotic plankton eDNA datasets from the Tara Ocean expedition (de Vargas et al. 2015) - we show that our methods well represent three different ecological indicators in the two-dimensional latent space: (i) MOTU richness per sample, (ii) sequence α-diversity per sample, and (iii) sequence β-diversity between samples along a gradient in the latent space (Fig. 3 and Fig. S2, Table 3). Thus, the 2D representation obtained reveals the ecological information underlying the study of communities. Additionally, we have shown how our two new methods outperform the representation of eDNA data compared to other dimensionality reduction techniques such as PCA, t-SNE, UMAP, and even a simple VAE to which no sequence information is added (Table 3). In contrast, linear methods such as PCA result in poor dimensionality reduction to ordinate eDNA samples (Table 3; Fig. S4 a). This is due to the complexity of eDNA data (Miya 2022, Xiong et al., 2022). eDNA metabarcoding is defined as the use of general or universal polymerase chain reaction (PCR) primers on mixed DNA samples of any origin, followed by high-throughput next-generation sequencing (NGS) to determine the species composition in a given sample (see, e.g., Deiner et al., 2017). Despite its potential in biodiversity monitoring (Pawlowski et al., 2022, V.d. Heyde et al., 2022, Mathon et al., 2022), it can be limited by false reads due to contamination, errors that can occur during the extraction, PCR, or sequencing process (Bohmann et al., 2014; Ficetola et al., 2016; Creer et al., 2016; Hering et al., 2018, Calderón-Sanou, et al., 2020). 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 high-dimensional, with a single observation from the experimental system containing measurements describing multiple traits (Hallam et al., 2021). For this reason, application neural architectures such as VAESeq and ENNBetaDist provide a better solution for clustering or understanding eDNA data. Neural networks allow the integration of multiple inputs into a single model (Cichy et al., 2019; LeCun et al., 2015; Schmidhuber, J. 2015). This is particularly relevant for the analysis of eDNA metabarcoding data, which are combinations of different types of information (Table 2). Our two new methods combine the number of sequences found for each MOTU and the genetic 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 (Fig. 3, Fig. S2). 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 particular management constraints 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 genetic 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 dataset approximately 60 pb, Table 1), our results indicate that such genetic information can inform species ecology. Here, the manipulation of the genetic information of the sequences highlights the proximity of the 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 diversity and its indicators. In addition, we show that a simple VAE, using only the information on the number of sequences present in each sample, gives a poor representation of the data (Table 3). The sequences turn out to be crucial for good

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

information extraction. Instead of relying solely on the number of sequences identified per MOTU or the genetic information within the sequences as in Cordier et al., (2018) and Flück et al. (2022), our methods combine both information (Table 2). VAESeq is based on VAE that is optimised to reconstruct the input. 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 is not relying on any pairwise distance between samples. In this case, the model is free to discover connections and highlight possible new ones in a fully unsupervised learning process. Alternately, ENNBetaDist helps to represent samples in a latent space according to an input metric. In addition, two new methods allow users to define the output dimensionality while preserving the global geometry (i.e., relative positions in the latent space) better than competing methods. Our results demonstrate that neural networks provide a more efficient way of extracting structure from eDNA metabarcoding data than traditional dimension reduction methods, thereby improving future ecological interpretation. The resulting diversity indices can 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 complexity can improve our understanding of biodiversity and ecosystems, and thus help stakeholders manage ecosystems.**Acknowledgments**This work is part of the Horizon 2020-Marie Skodowska-Curie Actions-COFUND project Artificial Intelligence for the Sciences (AI4theSciences) and the ANR FISH-PREDICT project. This project was partly funded by the WSL internal grant "eDNA Proof" and the Swiss Data Science Centre "DNai".**Conflict of Interest** The authors declare no conflict of interest.**Authors' 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.**Data Availability**If the paper is accepted for publication, all the eDNA data will be available in a Dryad Digital Repository. Codes and scripts for reproducing analyses in this manuscript are available at*https://github.com/letizialamperti/DLeDNA*.**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*. https://doi.org/10.1111/ele.14123ï 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/https://doi.org/10.1016/j.tree.2014.04.003 Boulanger, E., Dalongeville, A., Andrello, M., Mouillot, D., & Manel, S. (2020). Spatial graphs highlight how multi-generational dispersal shapes landscape genetic patterns. *Ecography*, *43*(8), 1167–1179. https://doi.org/https://doi.org/10.1111/ecog.05024 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/https://doi.org/10.1111/1755-0998.13367 Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2020). From environmental DNA sequences

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

to ecological conclusions: How strong is the influence of methodological choices? *Journal of Biogeography*, *47*(1), 193–206. https://doi.org/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/https://doi.org/10.1111/1440-1703.12102 Chao, A., & Ricotta, C. (2019). Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*,*100*(12), e02852. https://doi.org/https://doi.org/10.1002/ecy.2852 Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/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/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, Issue 9, pp. 1008–1018). British Ecological Society. https://doi.org/10.1111/2041-210X.12574 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/https://doi.org/10.1111/ddi.13534 Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure.*IEEE Transactions on Pattern Analysis and Machine Intelligence*,*PAMI-1*(2), 224–227.*https://doi.org/10.1109/TPAMI.1979.4766909*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/https://doi.org/10.1111/mec.14350*de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Bescot, N. le, 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 Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2021). A review of UMAP in population genetics. In *Journal of Human Genetics*(Vol. 66, Issue 1, pp. 85–91). Springer Nature. https://doi.org/10.1038/s10038-020-00851-4 Duffner, S., Garcia, C., Idrissi, K., & Baskurt, A. (n.d.).*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). 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/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). https://doi.org/10.1038/s41598-022-13412-w 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 Frøslev, T. G.,

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates.*Nature Communications*, *8*(1), 1188. https://doi.org/10.1038/s41467-017-01312-x 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/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/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/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, Issue 1, pp. 43–54). Cell Press. https://doi.org/10.1016/j.oneear.2019.12.016 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 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 MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M. S., Ducar, M. D., Meyerson, M., & Thorner, A. R. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing.*BMC Genomics*, *19*(1), 30. https://doi.org/10.1186/s12864-017-4428-5 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*(1). https://doi.org/https://doi.org/10.1111/ecog.05904 Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies.*PeerJ*, *2*, e593. https://doi.org/10.7717/peerj.593 Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., Brennan, G., Bush, A., Canard, E., & Cordier, T. (n.d.). *Key questions for next-generation biomonitoring*. https://doi.org/10.3389/fenvs.2019.00197ï Marques, V., Castagné, P., Polanco, A., Borrero-Pérez, G. H., Hocdé, R., Guérin, P. É., Juhel, J. B., Velez, L., Loiseau, N., Le-

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

tessier, 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*. 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/https://doi.org/10.1002/edn3.358 Nguyen Lan Huong AND 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/https://doi.org/10.1016/j.scitotenv.2021.151783 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. https://doi.org/https://doi.org/10.1002/edn3.140 Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*,*4*, e2584. https://doi.org/10.7717/peerj.2584 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/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/https://doi.org/10.1016/j.neunet.2014.09.003 Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303. https://doi.org/https://doi.org/10.1111/1755-0998.12402 Shelton, A. O., Kelly, R. P., O'Donnell, J. L., Park, L., Schwenke, P., Greene, C., Henderson, R. A., & Beamer, E. M. (2019). Environmental DNA provides quantitative estimates of a threatened salmon species.*Biological Conservation*, *237*, 383–391. https://doi.org/https://doi.org/10.1016/j.biocon.2019.07.003 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.

Posted on 15 Apr 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168156381.10025180/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

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/https://doi.org/10.1016/j.gfs.2022.100637 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/https://doi.org/10.1016/j.scitotenv.2022.157617 van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. In *Journal of Machine Learning Research* (Vol. 9). 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/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/https://doi.org/10.1016/j.watbs.2022.100007

Hosted file

`Figures_Manuscript_Lamperti_et_al.docx` available at https://authorea.com/users/606745/articles/635670-new-deep-learning-based-methods-for-visualizing-ecosystem-properties-using-environmental-dna-metabarcoding-data