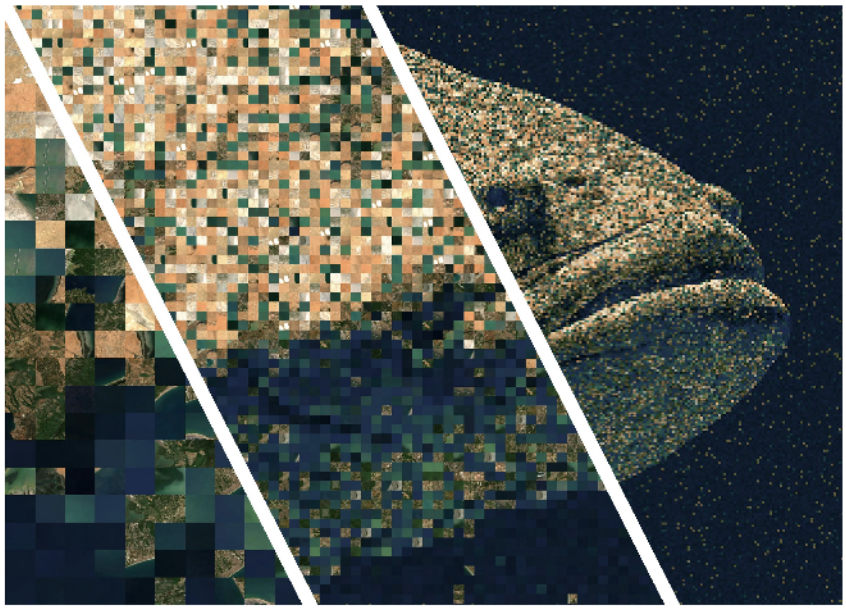
L'Institut Agro Rennes-Angers
□ Site d'Angers X Site de Rennes

| | |
|---|---|
| Année universitaire : 2022-2023<br>Spécialité :<br>Agronomie<br>Spécialisation (et option éventuelle) :<br>Sciences halieutiques et aquacoles, préparée à l'Institut Agro Rennes-Angers (REA) | **Mémoire de fin d'études**<br>X d'ingénieur de l'Institut Agro Dijon (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement)<br>□ de master de l'Institut Agro Rennes-Angers (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement)<br>□ de l'Institut Agro Montpellier (étudiant arrivé en M2)<br>□ d'un autre établissement (étudiant arrivé en M2) |

# Improving the prediction of coastal biodiversity in the Mediterranean Sea using a seascape deep learning approach

Par : Simon BETTINGER



***Soutenu à Rennes*** ***le 14/09/2023***

***Devant le jury composé de :***

| | |
|---|---|
| Président : Etienne Rivot | Autres membres du jury |
| Maître de stage : David Mouillot | *Olivier Le Pape* |
| Enseignant référent : Etienne Rivot | *Verena Trenkel* |

*Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle de l'Institut Agro Rennes-Angers*

## Fiche de confidentialité et de diffusion du mémoire

### Confidentialité

☒ Non ☐ Oui si oui : ☐ 1 an ☐ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible [1].

Date et signature du **maître de stage** [2] :
(ou de l'étudiant-entrepreneur)
13/09/2023

**A la fin de la période de confidentialité,** sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

### Droits d'auteur

L'auteur[3] **---Bettinger Simon----------**
autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☒ Oui ☐ Non

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☒ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☒ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'**auteur** :
13/09/2023

### Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☒ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☒ la diffusion papier et électronique du mémoire

Date et signature de l'**enseignant** :

(1) L'administration, les enseignants et les différents services de documentation de l'Institut Agro Rennes-Angers s'engagent à respecter cette confidentialité.
(2) Signature et cachet de l'organisme
(3).Auteur = étudiant qui réalise son mémoire de fin d'études
(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option)) sera signalée dans les bases de données documentaires sans le résumé

**Remerciements :**

Je tiens tout d'abord à remercier mon maître de stage, David Mouillot, pour m'avoir guidé tout au long de ce projet en me laissant la liberté de l'initiative. J'adresse également toute ma reconnaissance à mon collègue Matthieu, qui a dû répondre cinq mois durant à mes questions techniques, gardant encore de l'énergie pour nos discussions autour des délicieux plats du CROUS. Merci également à Benjamin Bourel, Maximilien Servajean et Alexis Joly pour leur aide précieuse.

Merci à mes cinq colocataires, Clarisse, Erwan, Gaétan, Léna et Hercules pour m'avoir accompagné dans toutes ces aventures, des Cévennes aux bateaux pirates, et m'avoir donné hâte de rentrer chaque soir. Merci à Matteo, Niels, Floriane, et toute l'équipe Montpelliéraine.

Merci également à tous les copains de l'Agro, de Dijon à Rennes, qui ont su s'inviter avec goût à peu près tous les week-ends pour apporter leur bonne humeur.

Enfin, un grand merci à ma famille qui, malgré la distance, m'a offert un soutien constant durant ce stage.

# List of abbreviations

CMEMS : Copernicus Marine Environment Monitoring Service
CNN : Convolutional Neural Network
eDNA : Environmental DNA
EMODNet : European Marine Observation and Data Network
EUNIS : European Nature Information System
GAM : Generalized Additive Model
GBIF : Global Biodiversity Information Facility
GFW : Global Fishing Watch
GLM : General Linear Model
IUCN : International Union for the Conservation of Nature
LIRMM : Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
MARBEC : Marine Biodiversity Exploitation and Conservation
MPA : Marine Protected Area
MSE : Mean Squared Error
NN : Neural Network
NOAA : National Oceanic and Atmospheric Association
RMSE : Root Mean Squared Error
SDM : Species Distribution Model
SGD : Stochastic Gradient Descent
SST : Sea Surface Temperature

# Table of illustrations

# List of tables

# 1. INTRODUCTION

This study, conducted as part of a Master's degree in fishery science at Institut Agro Rennes, with the support of MARBEC and LIRMM laboratories, aims to predict the level of fish biodiversity on Mediterranean coasts through the use of deep learning algorithms and a seascape approach.

Coastal ecosystems are vital to the economy, livelihood and well-being of numerous countries as hotspots of biodiversity and abundance. They are very dynamic systems under biotic and abiotic fluxes which make them interconnected with the open ocean and terrestrial ecosystems but also highly vulnerable to external threats. In the context of growing human density and global warming, coastal ecosystems are exposed to a wide range of stressors such as fishing, direct habitat disruption, heatwaves and pollution (Bevilacqua et al., 2021). The need for the conservation of such vulnerable ecosystems and habitats is urgent and requires a data-driven approach to set-up effective protection measures. Under the condition of careful design and placement, marine protected areas, or MPA, have proven their effectiveness for the conservation of coastal ecosystems and their species (Claudet et al., 2006), notably in the Mediterranean Sea, and are part of the European Biodiversity Strategy for 2030, with a goal of 30% protection and 10% strict protection (Sala et al., 2012). Yet, the Mediterranean is still poorly protected with only 6% of the Mediterranean basin under the MPA status. For 95% of these areas, Claudet et al. (2020) found that protection was not stronger inside than outside, leading to only 0.23% of the total Mediterranean Sea being fully or highly protected.

It is therefore essential to engage in systematic planning processes based on all available scientific sources, and to massively gather data on coastal ecosystems to determine potentially vulnerable areas, as well as to assess the state of ecological communities (Bevilacqua et al., 2021). To reach this objective, data on species richness, abundance and ecological niches are key to better protect marine ecosystems and predict their potential trends under scenarios.

Regrettably, the cost, both in time and resources, to obtain such data, has proven to be a major limitation in the global acquisition of information on coastal ecosystems, (Lundquist and Granek, 2005; Mora et al., 2008). Therefore, the short-term development of conservation planning requires the establishment of new efficient and harmonized data production methods (Edgar et al., 2016). To overcome field limitations, modeling from empirical data has imposed itself as a widespread and powerful approach (Airamé et al., 2003). Over time, a wide range of predictive models were developed, from simple Generalized Linear Models (GLM) based on social, environmental or habitat features of sample locations which linearly correlate to species richness and abundance (Knudby et al., 2010; Mellin et al., 2010) to machine learning methods such as random forests which consider nonlinear multiple interactions and tend to perform better in terms of prediction accuracy. For instance, Knudby et al. (2010) have compared multiple models in their predictive ability for species diversity and abundance values on the coast of Zanzibar using environmental variables as predictors. Machine learning models, namely random forest and bagging, proved to be significantly more accurate than simpler GLM or Generalized Additive Models (GAM) by 11% in predictive power. Most of these models rely on the correlation between the presence of a given species and its immediate environment. Since these models can be used to predict the distribution of species across a given area, they are called Species Distribution Models or SDM.

Yet, SDM and species richness models that consider environmental data as predictors have often been limited in their predictive accuracy : Knudby et al. (2010), reached a minimum RMSE score 6.4 on species richness with a bagging algorithm. This lack of precision partly derives from the exclusive use of local values of these environmental predictors, completely ignoring the effect of the surrounding scene. This classical approach also appears limited in the light of the MacArthur and Wilson's theory of island biogeography (1967), stating that an island's biodiversity increases with its size and distance to the mainland, and setting the basis for landscape ecology and the study of metapopulations (Levins, 1969). In the case of a seascape approach, treating patches of habitats, particularly hard bottom substrates, as islands in a soft bottom matrix appears essential to better understand and model the status and the biodiversity of a given ecosystem. Surprisingly, this seascape approach embedding predictors beyond the local context in predictive models is still rarely implemented owing to the complexity of considering spatial patches of predictor values in classical models. The main breakthrough of my Master Thesis is considering predictors of coastal biodiversity in a seascape context where local values can be treated in their respective locations amidst one another.

This approach considers biotic assemblages as sensitive to both local and regional features, as demonstrated by Belmaker et al. (2011, 2005), as well as Pittman et al. (2004), who showed that species richness in a reef increases with its isolation from a continuous reef patch, as well as with its size. Species richness and abundance are also linked to the shape of a reef (Grober-Dunsmore et al., 2008), as well as to the presence of seagrass beds in its immediate surroundings (Grober-Dunsmore et al., 2007; Mellin et al., 2007; Pittman et al., 2004). The location of a reef in a seascape also influences the response of its biocenosis to local environmental variables (Sievers et al., 2016) as well as to their temporal variability (Gilby et al., 2016). In general, and at the scale of an individual patch or a local ecological community, Kostylev et al. (2005) found that the complexity and structure of the habitat shape the structure of the species assemblage that it shelters. Yet, previous studies did not reach a consensus on the scale of the seascape structure's effect on fish assemblages (Mellin et al., 2009), in part because of the complex relationships between a biocenosis and its biotope and the diversity of seascape features and interactions. For a model to be able to successfully take into account these features across spatial scales to predict species richness, it would arguably be necessary to reach a step beyond traditional methods, into the field of iterative pattern learning via deep neural networks based on seascape information.

The recent emergence of powerful computing devices allowed the development of neural networks, from concepts that had been theorized since the 1980's. Models are now able to handle massive datasets as input variables, and iterate over them in order to extract relevant information for the computing of various tasks, notably predictive models. In deep learning algorithms, data pass through multiple "hidden layers" that interpret abstract information and extract progressively high level relevant features and their subtle interactions. At the end of each pass, an error value is computed and fed backwards, or "back propagated" through the network, which is updated accordingly. This allows for a highly diversified range of cases, making for a better precision in numerous prediction problems (Zhu et al., 2017). Deep learning networks are increasingly used in ecology (589 papers between 2019 and 2022), including in so-called deep-SDM using remote sensing data. They can be seen as an alternative to mechanistic models, mostly used in the past (Borowiec et al., 2022), but are still rarely used on coastal ecosystems.

One of the main challenges and backsides of deep learning models is that they require a large amount of standardized labeled data to train the models, which is particularly critical in the marine environment where each sampling method can provide a different set of species detected (Dalongeville et al., 2022; Valentini et al., 2016). For instance fisheries data miss small and crypto-benthic species while visual surveys miss large and elusive species like sharks. Large datasets on species distribution such as the one provided by the Global Biodiversity Information Facility (GBIF) exist, but traditional sampling methods have a tendency to overlook large portions of cryptic or elusive organisms, thus failing to provide exhaustive inventories. Additionally, the fact that GBIF originates from participatory science programs means that it lacks the homogeneity required for an efficient training.

One way to ensure that biodiversity labels can be standardized across coastal ecosystems is the use of species inventory through novel census methods, such as environmental DNA. Environmental DNA (or eDNA) is based on the fact that all living organisms emit biological debris that contain specific genetic material. In the water, collecting those debris and identifying them through metabarcoding allows to draw up an inventory of species in a narrow time frame and area. The use of eDNA has the advantage to perform quasi-exhaustive biodiversity inventories, at least for fish when the genetic reference database is complete. These inventories are carried out non-invasively and rapidly while avoiding human bias that can lead to the omission of certain species (Dalongeville et al., 2022; Sigsgaard et al., 2017). While eDNA could appear as a perfect candidate for the acquisition of training data to feed deep learning models, two problems arise : firstly, while research is currently being carried out on the estimation of abundance through eDNA samples, it has been found that the quantity of genetic material found in the water does not necessarily correlate with the number of specimen or the amount of biomass that emitted it. While eDNA concentration can provide a rough estimate of abundance for a single species (Rourke et al., 2022; Spear et al., 2021), it remains challenging or even impossible, with current knowledge, to derive precise enough values of species abundance in marine ecosystems through this method. Secondly, eDNA is still a young field, and while numerous campaigns aim to provide biodiversity inventories everywhere around the world (Mathon et al. 2022), the number of samples is still poor (approximately 700 points in the Mediterranean Sea) when it comes to building a dataset for training deep learning models, given that most training sets size at least in the tens of thousands (Benkendorf and Hawkins, 2020). The goal of my Masters Thesis is not to address the first issue since I will focus on fish species richness only but to overcome the second by taking advantage of large datasets on species occurrences provided by the Global Biodiversity Information Facility (GBIF).

One way of addressing the issue of limited sample size is to divide the training into multiple parts, through a process called “transfer learning”. In the first part, the model is trained to recognize patterns in the training data through a pretext task that should be close to the final task. In our case, one solution is to pre-train the neural network with the objective of predicting species occurrences using massive data from Global Biodiversity Information Facility (GBIF). This pre-trained model would then be saved, and the training would be completed on scarce but exhaustive data from eDNA surveys, in a process called “fine-tuning”. If successful, the final model would be able to predict species richness from a large number of seascape predictors composed of satellite imagery and multiple habitat and environmental features stacked as a multi-channel image.

The main goal of my Master Thesis is to set-up a novel deep learning method that would surpass traditional modeling methods on the task of predicting fish species richness on

french Mediterranean coasts from fisheries, protection, environmental and habitat data while taking into account contextual parameters within a seascape approach. Several data sources will be exploited as part of a transfer learning strategy, in order to provide a both accurate and generalizable model for use in the Mediterranean Sea. For comparison, classic random forest models will be tested on the same task and the effect of contextual parameters will be assessed through feature permutation.

# 2. MATERIALS AND METHODS

## a. Mediterranean coasts

The Mediterranean Sea is the largest semi-enclosed sea in the world. Home of 17 000 recorded species (Coll et al., 2010), representing 7% of the world's marine biodiversity with only 0.82% of the global ocean surface (Bianchi and Morri, 2000), it is a hotspot of biodiversity (Bianchi and Morri, 2000; Coll et al., 2010). However, the Mediterranean Sea has become one of the most threatened regions over the span of decades, as ecosystems are undergoing environmental and biological changes. These changes result from a complex combination of anthropogenic pressures, such as overfishing, pollution, and habitat destruction. Mediterranean coastal ecosystems, in particular, are deeply impacted by these pressures as both areas of high diversity and interfaces (Bevilacqua et al., 2021).

As outlined in the introduction, while reserves have been shown to offer benefits in terms of restoration of biodiversity inside and outside their borders, their effectiveness as of today is limited (Claudet et al., 2020), and contingent upon the identification of environments suited for a maximum impact. It was with the aim of improving our current prediction capability and extending the network of effective MPA that French Mediterranean coasts were selected as a framework to this study.

## b. Datasets

### i. Response variables

**GBIF**

The Global Biodiversity Information Facility (GBIF) serves as an international network and infrastructure dedicated to providing global data on Earth's biodiversity by leveraging multiple data sources, from scientific campaigns to citizen science initiatives ("GBIF," n.d.). Being open access, GBIF positions itself as a widely used hub to gather data on species distribution on a large scale. For the purpose of this study, the data was narrowed down to marine fishes, specifically chondrichthyes and osteichthyes, using GBIF's filtering tools (Registry-Migration.Gbif.Org, 2022). The occurrences were also limited in time, due to the fact that Sentinel-2 data ranged from 2011 to the present. Geographically, the focus was solely on the Mediterranean Sea. Additionally, a spatial precision filter was applied, retaining only occurrences with a spatial precision within 100 meters. The final dataset comprises

75 391 occurrences of 181 species ranging from 2011 to 2021. Each sample within the final dataset includes the species' complete name, GPS coordinates, survey date, spatial precision and, when available, information on abundance.

**Environmental DNA**

Environmental DNA data was collected through four coastal campaigns conducted by MARBEC from 2018 to 2021 along the French Mediterranean coast (Figure 1), employing various sampling methods such as divers and boats. For each sample, 30 liters of water were filtered along a 2 kilometer, 30 minutes transect through an Athena© peristaltic pump. Water was sent through a VigiLife© 0.2 µM cross flow filtration capsule. Following the transect, the capsule was promptly filled with 80 mL of CL1 conservation buffer and stored at room temperature. To ensure rigorous contamination control, a strict protocol was followed during each sampling operation, which included the use of disposable gloves and filtration equipment. Two pumps were active on most of the sampling points for the purpose of generating replicates. Each replicate was treated as an individual data point. Once acquired, the samples were sent to a dedicated laboratory. Metabarcode sequences were then amplified through PCR using a 12S mitochondrial rRNA primer pair. Extracted DNA was then purified and sequenced. Sequenced material was paired with MARBEC's database comprising 386 sequences from 156 fish species. Points below 50 meters of depth were excluded so as to only keep surface samples. The final dataset comprises 441 data points located for the most part on the French coast, including Corsica, with a few samples around the Balearic islands. Species richness was calculated for each sample as the sum of species detected by the analysis.

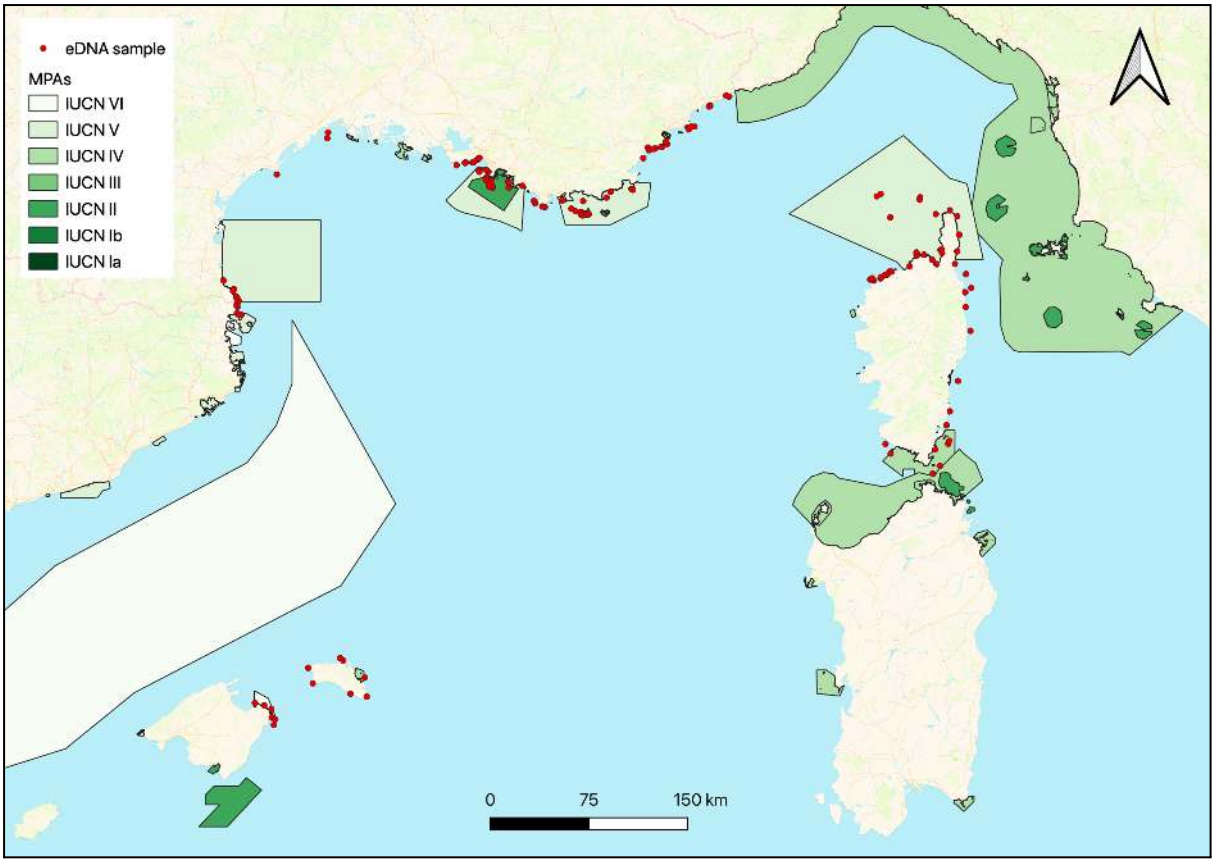



***Figure 1** : Map of sampling points conducted by MARBEC in the Mediterranean Sea, and MPA with corresponding IUCN protection level*

### ii. Explanatory variables

Numerous studies have attempted to explain the richness and abundance of Mediterranean species using different sets of explanatory variables (Bell, 1983; Charton and Ruzafa, 1998; Fanelli et al., 2013). Some crucial abiotic parameters identified in determining the structure of assemblages are presented in Table 1. It is worth noting that most studies do not fully corroborate one another, and that the scale of the effect of each parameter varies greatly with each case. Additionally, the recent development of remote sensing technologies, especially multispectral sensing using satellites and drones, facilitated access to some field data, since various ecological processes can be extracted from certain channels (K. S. He et al., 2015). With the goal of training a deep learning model, environmental variables were selected and extracted as multi-channel tiles for each GBIF and eDNA sample. Satellite imagery was used as a basis for each tile, since it represents raw data from which parameters can be derived without being manually computed, allowing for a greater flexibility in model's interpretation (K. S. He et al., 2015; Kavanaugh et al., 2021). Given the specific conditions of the Mediterranean Sea (opacity, depth etc.), other environmental variables such as bathymetry and substrate had to be extracted from various maps. The sources of explanatory variables and their effects on fish communities are given in Table 1 and developed in the following section.

***Table 1*** *: Main drivers determining fish assemblages as identified in the literature, with their effect and corresponding data sources*

| Variable | Data source and resolution | Tile extent | Effect | Reference |
|---|---|---|---|---|
| Bathymetry | NOAA<br>1.5 km | 30 km | Richness and abundance decrease with depth and increase with slope<br>Richness increases with vertical variability | Rees et al., 2014<br>Rule and Smith, 2007<br>Selfati et al., 2019<br>Stefanoudis et al., 2019 |
| | EMODnet<br>0.095 km | 2.945 km | | |
| Sea Surface Temperature (SST) | CMEMS<br>0.95 km | 30 km | Richness increases with SST | Condal et al., 2012<br>Gibran and Moura, 2012<br>Sigsgaard et al., 2017<br>Vilas et al., 2020 |
| Substrate type | EMODnet<br>~ 50 m | 3 km | Richness increases with rugosity, hotspots on hard-bottom substrate<br>Hard-bottom patches serve as island of high diversity and impact | Monfort et al., 2021<br>Moreno, 2002<br>Planes et al., 2000<br>Sahyoun et al., 2013 |

| | | | | |
|---|---|---|---|---|
| | | | surrounding areas<br>Substrate continuity impacts MPA effects | Ushiama et al., 2016 |
| Chlorophyll | CMEMS<br>1 km | 30 km | Richness decreases and abundance increases with primary production | Awada et al., 2021<br>Chassot et al., 2010<br>van Denderen et al., 2014 |
| Fishing | Global Fishing Watch<br>0.01 degree (approx. 1 km) | 10 km analysis | Higher impact of active bottom gear<br>Impact on abundance ratios | Collie et al., 2000<br>Jennings and Kaiser, 1998<br>Sinclair and Valdimarsson, 2003 |
| Marine Protected Areas | WDPA | 10 km | Increased abundance in protected areas<br>Structural changes of communities<br>Benefits vary among taxa | Claudet et al., 2006<br>Dalongeville et al., 2022<br>Guidetti et al., 2014 |

In order to maximize the efficiency of the training process, a relevant tile size had to be determined. One the one hand, maximum spatial precision is desirable for all data, and would permit more accurate biodiversity analysis. Because of the variety and uncertainty of ranges in the effect of environmental variables, a large spatial extent would be preferable, in order to keep as much information as possible. On the other hand, the image size that can be fed through most deep learning architectures is limited by the capacity of the hardware, often to only a few hundreds of pixels both in width and length. It is also worth noting that extending the tile size too much would, in a Convolutional Neural Network (CNN), result in some confusion for the model since some features that are present in the image, but are too far away from samples to have a real effect on communities, may create noise in the dataset (Benjamin Bourel, personal communications). Data quality is also a limitation, since the resolution of most datasets is above 30 m. Consequently, a compromise had to be established. In terms of precision, multiple studies found that the effect of important benthic features such as depth and substrate cover was variable, ranging from 10 m to a few hundred meters (Grober-Dunsmore et al., 2007; Mellin et al., 2007; Purkis et al., 2008; Rees et al., 2014). However, anthropogenic factors tend to have a higher range : Hackradt et al. (2014) have found that MPA spillover was dependent on the species and could expand over thousands of meters. Additionally, spillover was found to be dependent on the habitat within the MPA, and the connectivity at its boundaries. Green et al. (2015) note that home range varies among and within reef-associated species, scaling from 0.5 km to thousands of kilometers for some predators, with most strictly reef species limited to a few kilometers. Larval spillover is mostly below 5 to 15 km. It is consequently essential to capture this range through a sufficient extent. Considering that the pre-training was carried out in the continuity of trainings that had used ImageNet datasets of input size is 256x256 pixels, tile size was chosen for each variable so as to approach a 256x256 image size given the spatial resolution of the data while preserving the integrity of the data and avoiding interpolation. Groups of different tile extents were created following this rule. Since channel alignment is essential for the analysis of interactions between features, the number of groups was kept as small as possible.

3 groups of variables were consequently created, each with a tile extent of approximately 3 km 10 km, and 63 km. Features that were identified in the literature to interact with each other, such as depth and substrate, were kept in the same group. CNN training using multiple tile sizes as input has already been tested and proven successful during previous projects carried out by the LIRMM laboratory (Benjamin Deneu, personal communications).

Following this decision, square areas were computed around each data point in EPSG 4326 for each variable, in both the GBIF and eDNA datasets.

**Sentinel-2**

Sentinel-2 tiles were gathered through custom requests via Microsoft Planetary Computer (Copernicus, 2011). Since computing power is a limiting factor in the training of CNN, it was decided to limit input channels to the RGB band (band 2, 3 and 4) as well as to the infrared band (band 8).

**Bathymetry**

Bathymetry data was acquired for two resolutions in order to consider both local fine grain structural information and large contextual patches. Large scale depth data is provided by the National Oceanic and Atmospheric Association (NOAA) and was acquired through the geoenrich 0.5.8 package for python. Fine resolution data was acquired through EMODnet via its Digital Bathymetry product based upon a collection of surveys and satellite derived data. The resolution of both rasters is computed into Table 1. Geoenrich's enrichment tool was used to automatically produce 30x30 km georeferenced rasters for the large tiles, while a custom made tool was used for the fine grain data.

**SST**

SST data was provided by the Copernicus Marine Environment Monitoring Service (CMEMS) and acquired through geoenrich. All data comes from the Mediterranean Sea - High Resolution L4 Sea Surface Temperature Reprocessed dataset (CMEMS, 2023a).

**Substrate**

Substrate was acquired based on the high resolution vector EUSeaMap generated by EMODnet in 2021 (Vasquez et al., 2019). This map of the Mediterranean floor provides harmonized information on seabed habitats, based on the European Nature Information System (EUNIS) classification (Davies et al., 2004). Significant discrepancies among classes were noticed between areas of the Mediterranean Sea, in all likelihood caused by inconsistency in mapping campaigns. The similarity between some of the EUNIS classes led to the merging of several substrate types, both in an effort to standardize data and to limit the size of each input : *sandy mud*, *muddy sand*, *fine mud* and *fine mud or muddy sand or sandy mud* were grouped together. *Seabed sediment* was grouped with *lagoon* as both are interchangeably used for both classifications. *Coarse and mixed sediments* and *mixed sediments* were merged. As a consequence, 8 final classes were used. Because a vector map has virtually no resolution, an inspection was performed on coastal areas using QGIS, and revealed the presence of habitat patches smaller than 50 meters along the French coast. Consequently, each habitat type was attributed a single code, and the map was rasterized using QGIS with a resolution of roughly 50 m, in order to keep information on substrate

patches as precise as possible, while keeping a manageable file size for the raster. A custom clipping program was developed and used on the general Mediterranean raster. As CNN take into account the interaction between features, keeping information on habitat type coded as multiple values on a single map would generate artificial closeness between some habitats, and lead the model to interpret false relations. In order to avoid this problem, it was necessary to one-hot encode the data - i.e. coding 1 if the habitat is present, 0 if not, and -1 if there is no data - and produce rasters carrying as many channels as there are habitat types. A custom program was developed to do so.

**Chlorophyll**

Chlorophyll data was acquired through the Mediterranean Sea Ocean Colour Plankton MY L4 daily gapfree observations and climatology and monthly observations dataset provided by CMEMS (CMEMS, 2023b) and accessed through geoenrich.

**Fishing pressure**

Fishing pressure data was provided by Global Fishing Watch (GFW) through their AIS-based commercial fishing dataset. Vessels are identified and classified via CNN, identification registries and expert panels (Kroodsma et al., 2018). Data was available from 2012 to 2020, and consisted of one csv file per day, with each row containing information on a fishing ship, its location at a given time, the duration of fishing and the fishing method employed. A spatial join was performed between the diversity datasets and the GFW dataset, in order to group fishing points by studied tile. Fishing gears were separated, in order to account for the varying impact of different fishing methods on benthic communities, as discussed by Jennings and Kaiser (1998) and Sinclair and Valdimarsson (2003). Seine, trawling and dredging were considered as more destructive methods, and separated from the others. Collie et al. have found that the temporal scale of fishing impact depends on the gear as well as the type of habitat (2000). To incorporate this factor, the number of hours spent fishing was summed within the last 7 days, and last 30 days. To accommodate for the lack of fishing data in 2011 and 2021, it was assumed that no significant change happened between 2011 and 2012, and that 2019 was comparable to 2021. Fishing data from 2012 and 2019 was consequently respectively used for 2011 and 2021. Each occurrence was thus associated with 4 fishing pressure values.

**MPA**

Data on Marine Protected Areas was gathered through the World Database on Protected Areas (UNEP-WCMC and IUCN, 2023). Filters were used to initially narrow down the data on Mediterranean MPA associated with an IUCN classification. A protection value was attributed to each level, 1 being the least protected (level VI), to 7 being the most protected (level Ia) (Day et al., 2019). Each tile was then clipped using the same method as for the substrate data. Level 0 both corresponds to non-protected areas, and to land.

Latitude and longitude were decided to be included as features, in order to account for possible interactions between the geographical situation of samples and features. In order to keep the training process flexible and to facilitate feature ablation procedures, the environmental data tiles were kept separated, and it was decided to selectively merge them as channels at the start of the network. 17 of the 23 channels used for training are presented in Figure 2. They correspond to channels that were not generated from single values.

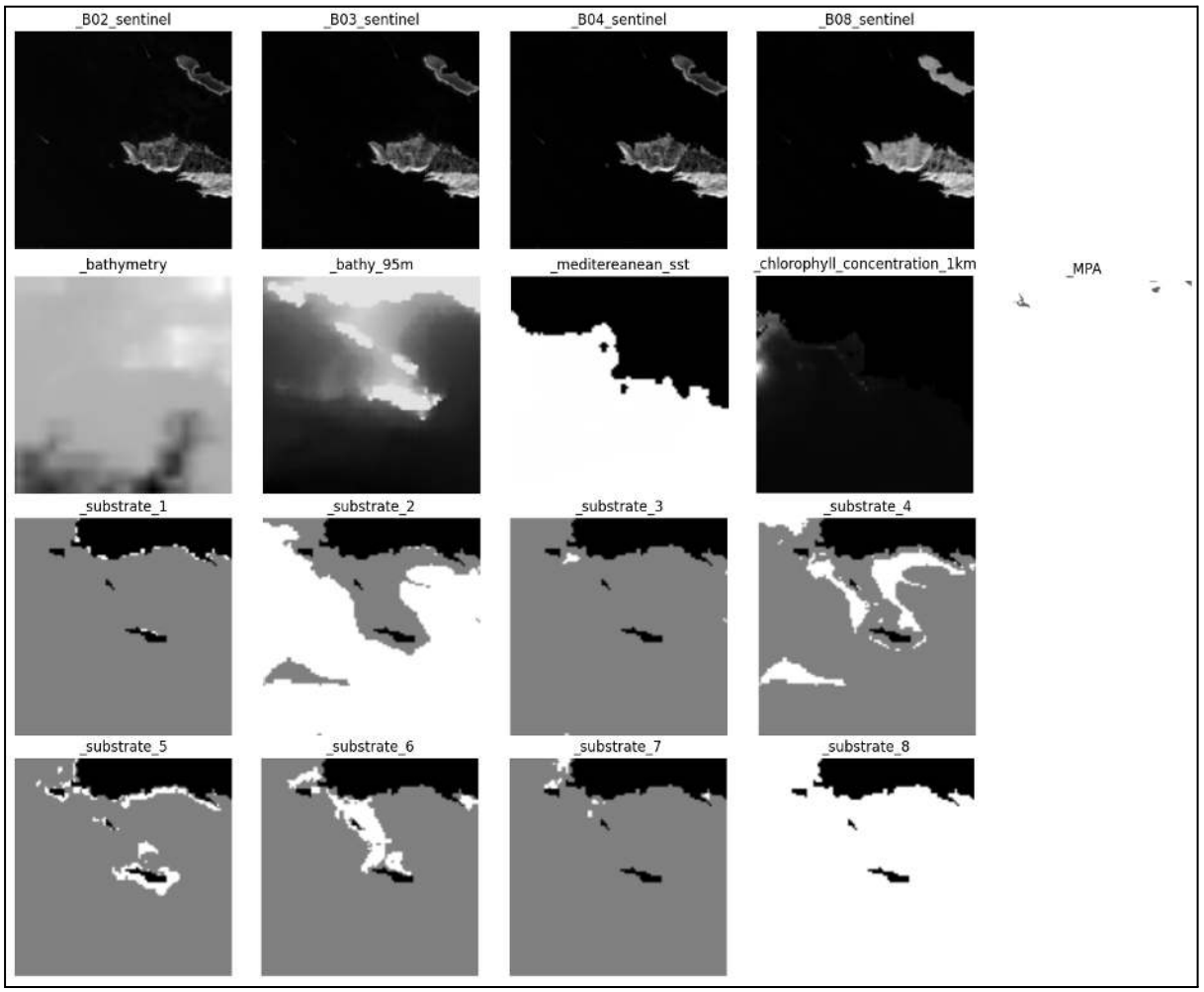



***Figure 2** : 17 channels of the 23 used for the analysis of sample SPY181824, taken near the Riou archipelago*

### iii. Data preprocessing

During the training of a CNN, wide differences in the scale of the input data can cause instability in the process, making it slow or unable to reach convergence. Additionally, inputs of larger magnitude can become overused by the model. In order to avoid these issues, it is essential to normalize the dataset (Maharana et al., 2022). Input layers consisting of continuous variables, such as bathymetry and SST, were scaled to have a mean of 0 and a standard deviation of 1. In order to generalize the weights throughout the pre-training and fine tuning process, standard deviation and mean values of the GBIF pre-training dataset were used for all normalized channels, for both pre-training and fine-tuning. Input consisting of classes, such as substrates and MPA, were not normalized.

For the sake of comparison, data for the random forest models were derived from the CNN input rasters. A center value, corresponding to the value at the sampling location, was extracted for each raster. Additionally, standard deviation was calculated for all continuous variables so as to retain some information about the spatial context and habitat heterogeneity.

For classified substrate data, context was captured in the form of a Shannon entropy index, representing the diversity in substrates for a single tile. Since random forests algorithms are not distance-based models comparing feature values, but rather tree based models making splits in the data, normalization is not required and can impact the results of the regressor. It was consequently not used for the baseline models.

## c. Neural networks

### i. General operations of a neural network

Neural Networks (NN) are models structured around a network of nodes (neurons) that perform mathematical operations on input data, in an attempt to mimic the general functioning of a brain. Each neuron represents a single linear regression model that takes in inputs, treats them, and produces an output that will be sent to the next neuron. It is associated with a bias, a set of weights, and an activation function. Given these parameters, the formula for the output of a neuron is as follows :

$$output = f(b + \sum_i wi \times xi)$$ With :

$f$ : Activation function
$b$ : Bias
$wi$ : Weight for input i
$xi$ : Input i

The activation function is designed to introduce non-linearity into the output by determining whether a neuron should be activated or not, and scale its response to the input. This will, effectively, assess the importance of a single neuron in the final task and make sure that the output cannot be written as a linear combination of the inputs. At the end of the network, in the output layer, an error value will be computed by comparing the prediction with the expected result through the use of a cost, or loss function. In order to train itself, the model will automatically leverage the value of the weights and biases of each neuron through a process called gradient backpropagation : after the calculation of the loss value, it is sent backwards into the network and used to evaluate the participation of each neuron through the use of partial derivatives. Weights and biases of each neuron are updated accordingly. This process is repeated for each sample of the dataset, for a given number of passes, or epochs. Should the hyperparameters of the training be well set, the model will converge.

### ii. Case of the CNN

In our case, the need for the analysis of geospatial structures calls for computer vision through deep learning : in this type of models, images are treated as numerical data upon which gradually specific filters are applied in a process called “convolutions”. This allows, through multiple iterations through the dataset, for the recognition of complex spatial structures and interactions between the channels of the image that is being fed (Zhu et al., 2017). Such models have already been used for SDM using land satellite images acquired through the Sentinel-2 program (Estopinan et al., 2022), as well as to predict rice yields in China through a regression task via satellite imagery (Chu and Yu, 2020). For our task, inputs

are tiles of high resolution satellite images and maps, associated with information on the communities being sampled (richness, abundance etc.). A Convolutional Neural Network (CNN) is a type of neural network designed to treat grid-like data such as images. CNN are typically divided into three types of layers :

**Convolution layer**

In CNN architectures, the weights are in the form of kernels, matrices that slide over the input grid. For each position, an output value is computed through a dot product operation between the kernel and the portion of the grid it is placed over. If the kernel is placed above the structure it was trained to detect, it will output a greater value. Each position of the kernel will produce a single value that will be appended to a new matrix called a feature map, saving the response of the entire grid to the kernel (Figure 3).

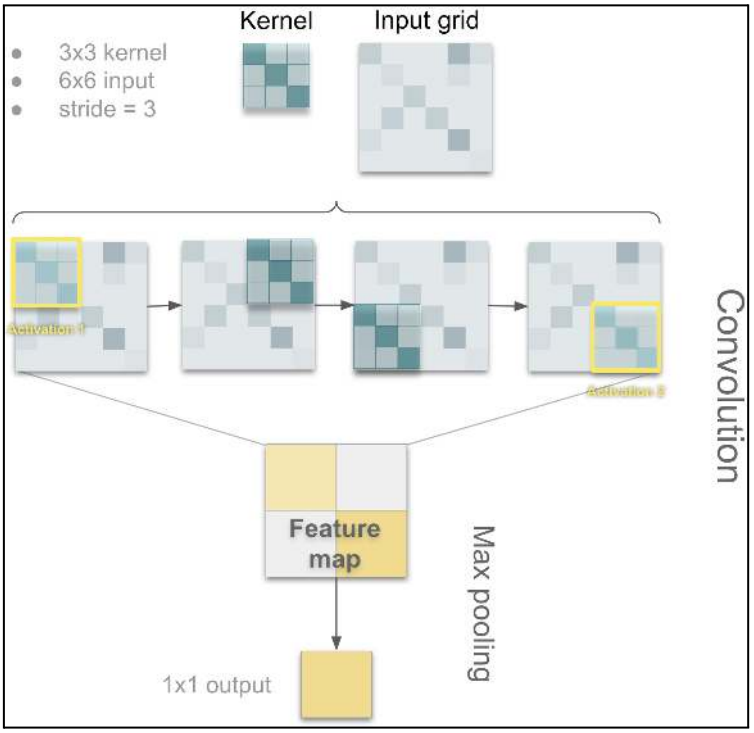



***Figure 3*** *: Schematic convolution and max pooling in a CNN*

**Pooling layer**

Once the activation map is computed, a pooling operation will be applied on it. This will reduce the size of the output and the number of parameters necessary, allowing the network to gain efficiency. Popular types of pooling include max pooling and average pooling : the feature map is partitioned in equal areas, and the maximum value for each area is attributed to it. For average pooling, the average of all values is attributed to each partition. Figure 3 displays a schematic representation of a convolution and max pooling operation of a 3x3 kernel on a 6x6 grid. In this situation, the output corresponds to the probability that the feature was detected at least one time in the input. Here, stride was set to the size of the kernel for better visualization, but it is classically smaller, resulting in higher definition feature maps and outputs.

**Fully connected layer**

After passing through all hidden layers, the information, usually in the form of a 3D tensor, is flattened and fed into a fully connected layer - i.e a layer where all previous neurons are connected with all current neurons - as a 1D tensor. In the case of a regression task, a single final neuron in the fully connected layer will allow the output of a single value (Figure 4).

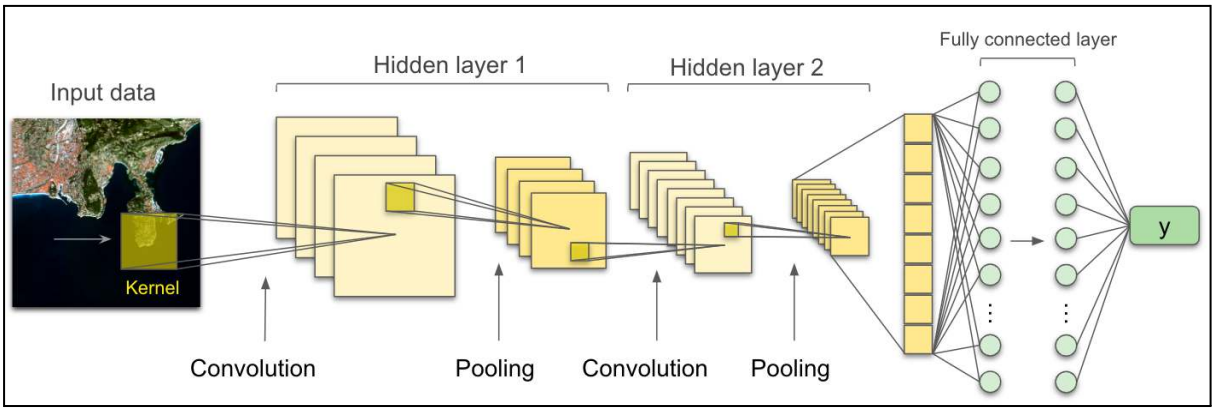



***Figure 4*** *: Schematic operation of a CNN performing a regression analysis on a satellite image to output a predicted value y*

### iii. Choice of architecture

Increasing the number of layers in a CNN can theoretically increase the model's ability to learn complex structures and features. However, it has been observed that using a deeper network decreases performance, notably because the model may encounter issues such as vanishing gradients. This stems from the way weights are updated during training : loss is propagated backwards in the form of a gradient, which is the vector of partial derivatives of the activation function, with respect to the corresponding input variable. In each layer, gradients are multiplied by the weight matrix of each layer they pass through. This may cause them to decrease exponentially, reducing the scale of the weight update for each layer. As a consequence, early layers will receive very small gradient updates leading to slow and ineffective learning. One way of dealing with this issue while keeping the advantages of deep architectures is through the use of residual networks, or ResNet, which were introduced in 2015. ResNet incorporate skip connections, also known as residual connections, that facilitate the flow of information through the network, mitigating the risk of a vanishing gradient (K. He et al., 2015). Multiple types of ResNet exist, and are distinguished from each other by their number of layers. Because of the complexity of identified interactions between environmental predictors and marine communities, it was decided to go beyond a classic ResNet-18, and a ResNet-50 network comprising 50 layers was consequently chosen. The first convolution layer was re-written to accommodate for 23 channels inputs, and the last fully connected layer was set to output a tensor of size one, for the purpose of performing a regression task.

### iv. Hyperparameters

Success in training a CNN is heavily dependent on the tuning of the set of hyperparameters associated with the task. Table 2 presents the three main levers used to tune the training of a

CNN, as well as the effect of their mistuning.

***Table 2** : Main hyperparameters of a CNN and effect of their mistuning*

| Hyperparameter | Definition | Effect if too low | Effect if too high |
|---|---|---|---|
| Learning rate | Amount of parameter change between each batch | Slow training<br>Prone to local minima | Divergence<br>Overshooting<br>Unstable training |
| Number of epochs | Number of times the whole dataset passes through the network | Underfitting<br>Incomplete convergence | Overfitting<br>Wasted resources |
| Batch size | Number of samples treated before network update = one 'batch' | Slow convergence<br>Unstable training | High memory usage<br>Overfitting<br>Prone to local minima |

It will consequently be necessary to take these parameters into account when training the models. The learning rate is often regarded as the most important parameter in the training of a CNN (Smith, 2017). Its optimal value is heavily dependent on the current stage of the training : a high value will be more suited for the beginning of training, since the model is blank and likely to output a very high loss. However, the risk of overshooting the minimum becomes higher with each epoch. One way to address this issue is through the use of a scheduler, which adapts the learning rate to the output of the optimizer function. The schedulers and optimizers used for our CNN are detailed in part 2.d.

## d. CNN Training

### i. Splitting procedure

The training of a CNN is usually divided into 3 steps : training, validating, and testing. Each step requires its own independent dataset. The training set contains the samples that will be used by the model to learn, via backpropagation. The validation set is a smaller set that will be used to evaluate the model's prediction ability at regular steps, allowing the user to fine tune the model by changing the hyperparameters. The CNN knows the validation set, but never learns from it. It is only affected by it indirectly, via the modifications made by the user on the training step. The testing set is never seen by the CNN during training, nor does it affect it in any way. It is only used to provide an unbiased evaluation of the CNN prediction ability at the end of the training and validating steps. The training set usually contains the majority of the samples, and a common split is 80/10/10 (Train/Validation/Test).

While it is common to randomly split a dataset, this basic strategy overlooks the potential spatial dependence or spatial autocorrelation in the data. This can lead to overoptimistic results since the training, validating and testing datasets are not independent,

so the CNN can learn from this spatial dependency regardless of explanatory variables. Ploton et al. (2020), have shown that non spatial split (random) of the dataset to train a model designed to predict forest biomass variation provides a near 50% accuracy while a spatial cross validation, where training and testing datasets are spatially independent gives quasi random results, so no predictive power. One way of overcoming this pitfall is to perform a spatial cross-validation split, that will group samples belonging to a spatial cluster into the same subset, or fold. It will, for instance, prevent two samples that are highly correlated to their spatial situation to be separated into the train set and the test set, which would artificially boost the apparent accuracy of the CNN. While training, one fold is used as a validation set, while the others are used as the training set. This leave-one-out procedure is repeated for each fold.

For the GBIF dataset, samples were segregated following their situation in the Mediterranean Sea, with five groups being created. For the data on eDNA, spatial autocorrelation was tested through Moran's I test using a number of neighbors of 10, in order to capture both local and broader relations. Significant spatial autocorrelation was found, with a Moran I statistic of 0.29 ($p < 2.2e\text{-}16$). Results are illustrated in a Moran scatterplot in Figure 5.

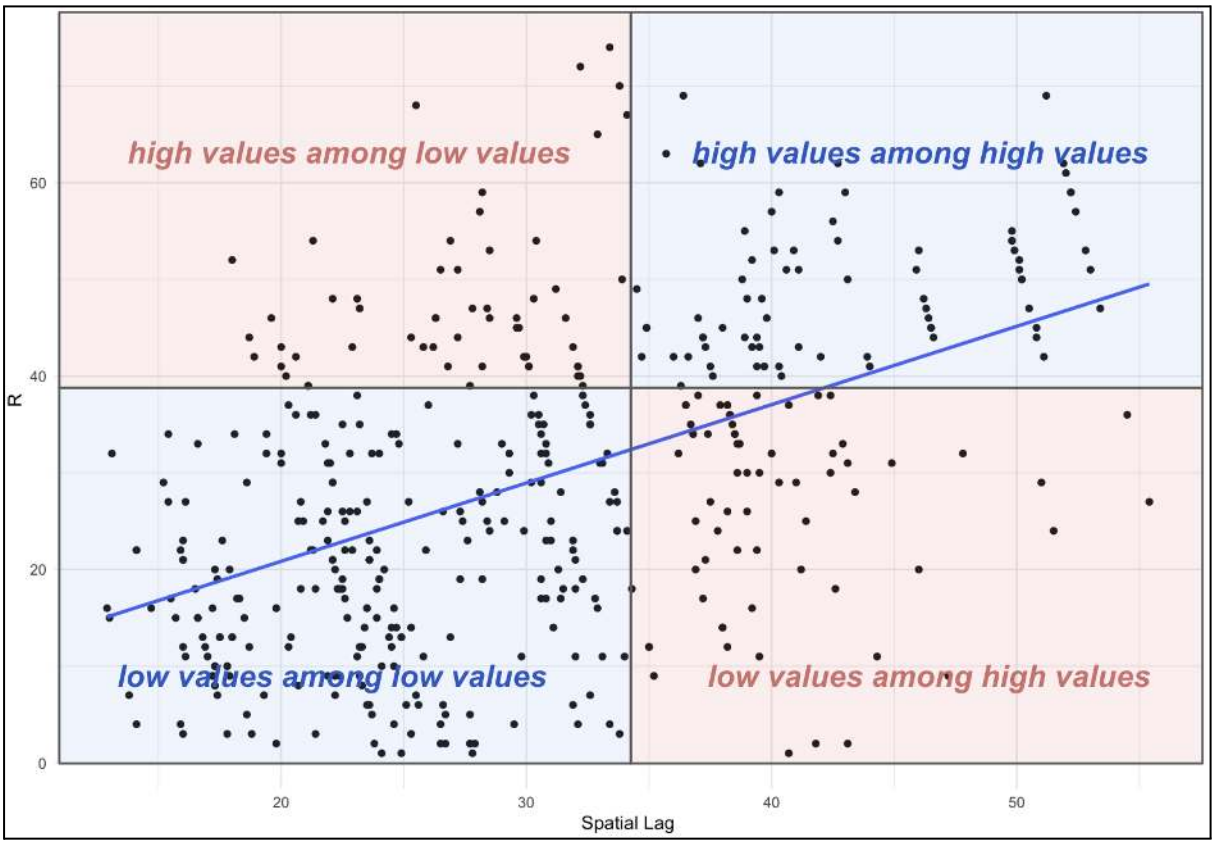



***Figure 5*** *: Moran scatterplot showing spatial autocorrelation of species richness (R) in the eDNA dataset (k=10)*

As a consequence, sample points were first grouped by distance into 20 clusters using a k-means clustering algorithm on QGIS. Clusters were then randomly attributed to 6 folds so that each fold contains approximately 1/6th of the total eDNA data. For each training

procedure, one fold was chosen for the test phase of the model, while the other five were split into a train set (4 folds) and a validation set (one fold). This split was repeated for each possible permutation of the folds, so that each fold was used as a test set. All 6 folds as well as a permutation example are presented in Figure 6.

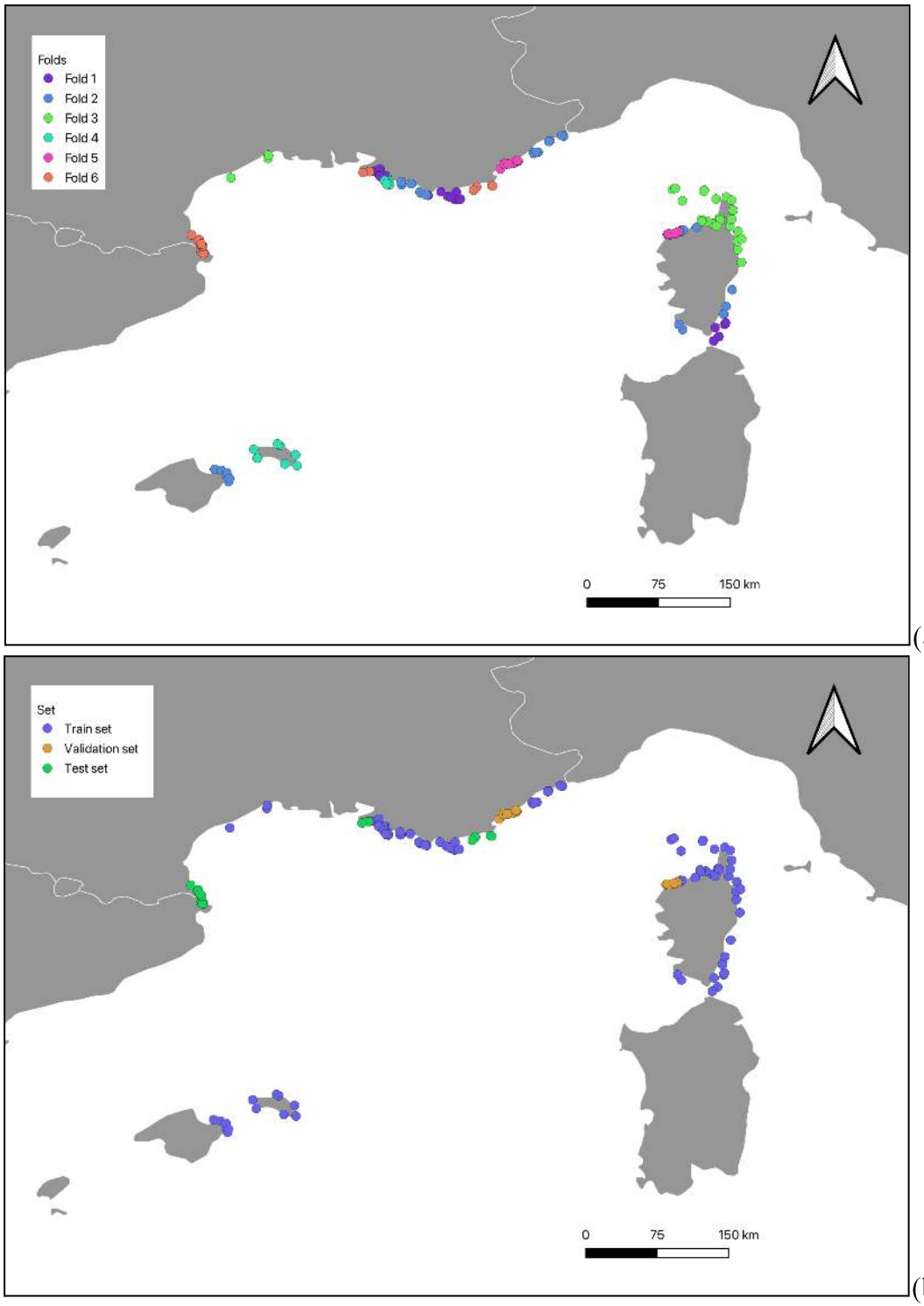



***Figure 6*** *: Classifications of folds for the eDNA dataset (a), example of a 4/1/1 split between the folds (b).*

### ii. Multimodal classification task on GBIF

Transfer learning takes advantage of the fact that feature extraction performed by a CNN on a given task can be reused for another CNN on a similar task, given similar inputs (Figure 7). A single CNN was trained with the GBIF dataset, on a multimodal classification task that was originally designed for the IA-Biodiv challenge as part of the FishPredict project. The goal was to accurately predict probabilities of presence of fish species in the Mediterranean Sea. Learning rate was initialized at 0.0112 and gradually decreased following a plateau scheduler, which is the most common scheduling method and consists in gradually reducing the learning rate when the loss output hits a plateau (Al-Kababji et al., 2022). Cross Entropy Loss was used to compute loss, and the optimizer function itself was set to Stochastic Gradient Descent (SGD), as it has been observed to offer better generalization performance than default Adam optimization, at the cost of longer training (Gupta et al., 2021). The training of the classification model was carried out by LIRMM personnel. Parameters of the classification CNN can be found in Table 3.

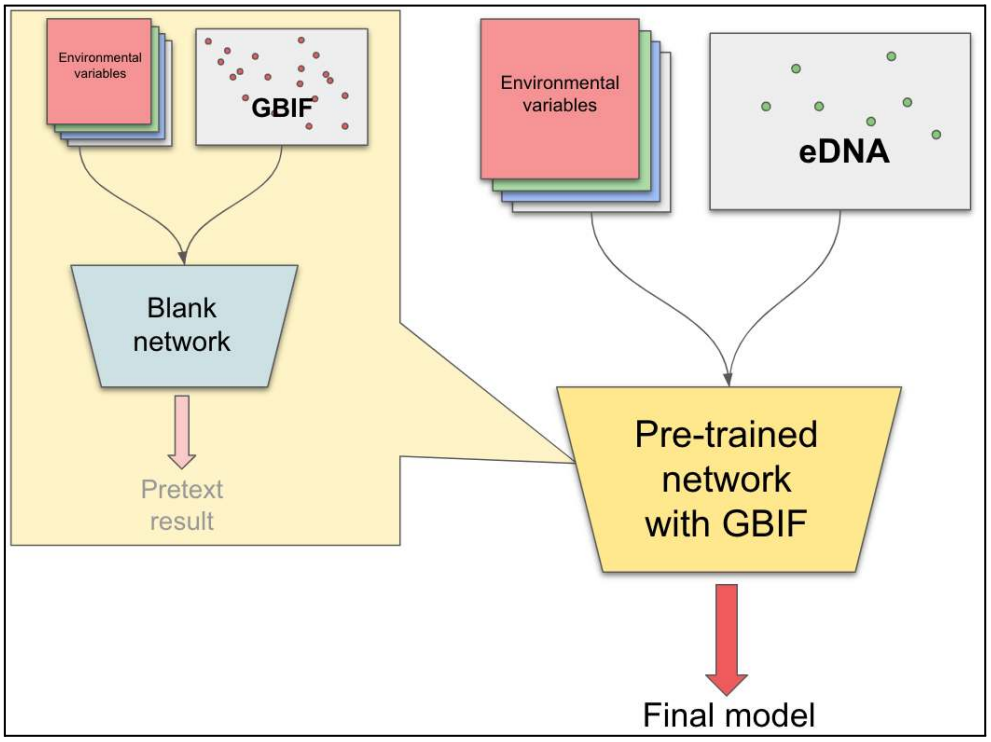



***Figure 7** : Conceptual workflow for transfer learning on GBIF and eDNA data*

### iii. Regression task on eDNA

Two types of CNN were trained on the eDNA dataset. The first one was set to measure the accuracy of a model that would be trained from scratch, using randomly initialized weights for all 23 bands. Training was divided between all 6 possible permutations among the folds, and 6 models were output for both modalities. For each modality, several runs were carried-out in order to find the best possible tuning for the hyperparameters, through trial and error.

In the case of the untrained CNN, multiple batch sizes and learning rates were experimented, as it was observed that training was unstable at typically used values. Multiple methods of learning rate scheduling were tested. While the plateau scheduler works well in the context of a smooth drop in loss, preliminary tests indicated that the network was prone to encounter local minima, which can be mistaken by the scheduler with an actual optimum plateau. Learning rate consequently decreased too quickly and training stalled with a high loss value. In order to avoid this issue, a cyclic scheduler, as developed by Smith (2017) was used for both the untrained and the pre-trained eDNA CNN. Cyclic schedulers make the learning rate oscillate periodically between two threshold values, allowing the network to escape local minima with higher learning rates while avoiding very fast loss divergence known as gradient explosions. This method has proven to result in fast and efficient training, including on ResNet architectures, bypassing the need to tune a global rate. Loss decreased much more smoothly once a cyclic triangular scheduler was applied to the learning rate, while keeping its base and maximum values low. As transfer learning implies the use of the same architecture as the CNN that was used for pre-training, the pre-trained CNN was initialized with the same architecture and in the same way as the untrained CNN. Weights were loaded as a dictionary into the network, and the last fully connected layer, which was designed for multimodal classification with 181 output classes, was rewritten to only output one class, making it fit for regression. All layers were frozen, i.e. their weights and bias were set to be insensitive to gradient backpropagation. This allows the preservation of the features that were extracted during pre-training.

Tests were then performed with multiple training methods and hyperparameters. Following common transfer learning practice, layers were unfrozen one by one starting from the last (fully connected layer). Following this order, the CNN keeps in memory the lowermost level features learnt during pre-training, and learns to detect new ones specific to the transfer dataset and task. Additionally, since large discrepancies between the training and validation accuracies suggested high overfitting during the training of the untrained CNN - i.e. the model memorized the noise of the dataset and fit too closely to the training data - it was decided to incorporate an L2 regularization method in the form of a weight decay of 0.1 in the optimizer. L2 regularization limits the scale of a CNN's weights during backpropagation, and helps prevent overfitting.

Since folds were slightly unequal and the dataset was small, an adaptive batch size was chosen for each model, mostly to keep the CNN from training on extremely small batches. Batch size was computed as a third of the size of the validation dataset for each training. Both regression CNN used MSE as a loss value. The parameters of each CNN can be found in Table 3. While training was carried out for 50 epochs, early checkpoints corresponding to peaks of high validation R2 for each model were kept in an attempt to evaluate the model's generalization power before being subject to overfitting. This modality will thereafter be referred to as "early stopping".

***Table 3** : Architecture and hyperparameters chosen for the GBIF CNN, and both training modalities on the eDNA dataset*

| Modality | GBIF pre-training | Untrained CNN | Pre-trained CNN |
|---|---|---|---|
| Architecture | ResNet-50 | ResNet-50 | ResNet-50<br>Layer 4 / fc unfrozen |
| Optimizer | Stochastic Gradient Descent (SGD)<br>Weight decay 0.001 | Stochastic Gradient Descent (SGD)<br>Weight decay 0.1 | Stochastic Gradient Descent (SGD)<br>Weight decay 0.1 |
| Scheduler / Learning rate | ReduceLROnPlateau | Cyclic (1e-3 - 1e-4) | Cyclic (1e-3 - 1e-4) |
| Loss | Cross Entropy Loss | MSE Loss | MSE Loss |
| Batch size | 256 | Adaptive<br>(Training dataset/10) | Adaptive<br>(Training dataset/10) |
| Training epochs | 10 | 80 | 50 |

### e. Random forests

Two random forests were carried out on both the punctual and contextual dataset in order to assess the contribution of context features in the accuracy of a model. Each model was fitted using 500 decision trees, with the maximum amount of predictors available. As a consequence, 23 predictors were used for the contextual data, and 15 predictors were used for the punctual data. Models were trained following a leave-on-out procedure, i.e. each model was trained on five of the folds, with the remaining fold being used as a test set. This process was repeated so that each fold was used as a test set, and a predicted value was output for each sample of the total dataset. 6 models were trained. Richness predicted on the test sets were then compared to actual values used to produce a general accuracy value. Feature importance was measured by shuffling features one by one on the test sets, before calculating the $R^2$ as an accuracy metric. The accuracy values were then compared. General importance for a feature was calculated as the mean of its importance on all models. The randomForest package for R was used.

### f. Metrics

Accuracy was used as the metric for the multimodal classification task on GBIF. $R^2$ was used as the final accuracy score for both the regression CNN and the random forest models and calculated as :

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

With $y_i$ the true value and $\widehat{y}_i$ the predicted value for all n samples.

As the predicted value can significantly differ from the true value, the R2 score may be negative, meaning that the model performs worse than if it systematically output the average value of $y$. Each $R^2$ value was computed on the whole dataset for each modality by combining the predictions of all models. Training of the CNN was monitored through $R^2$ and loss, which were output at each epoch end for the validation set. RMSE was also output for the random forest and regression CNN models.

# 3. RESULTS

## a. Random forests

$R^2$ was calculated from the actual and predicted species richness values. Accuracy for the contextual model was found to be non-significant with a $R^2$ of 0.08. The difference in accuracy between the contextual and punctual models was found to be significant. Results for both random forest models can be found in Table 4.

***Table 4*** *: $R^2$ score output for the random forest model based on contextual and punctual data*

| Dataset | $R^2$ score on test set | RMSE on test set |
|---|---|---|
| Punctual data | -0.02 | 16.35 |
| Punctual + contextual data | 0.08 | 15.55 |

Feature importance analysis was only carried out for the contextual model in view of the poor R2 score of the punctual model. Importance was very variable among features. (Figure 7). Substrate and SST were found to be globally important in the prediction of species richness for both types of models. RGB Sentinel-2 bands were determined to be inequally important, with only bands 3 and 4 (respectively green and red) contributing to the prediction. Fishing was globally observed to have a negative impact on the prediction. Contextual data in the form of map standard deviation was mostly found to be important in the Sentinel-2 infrared band.

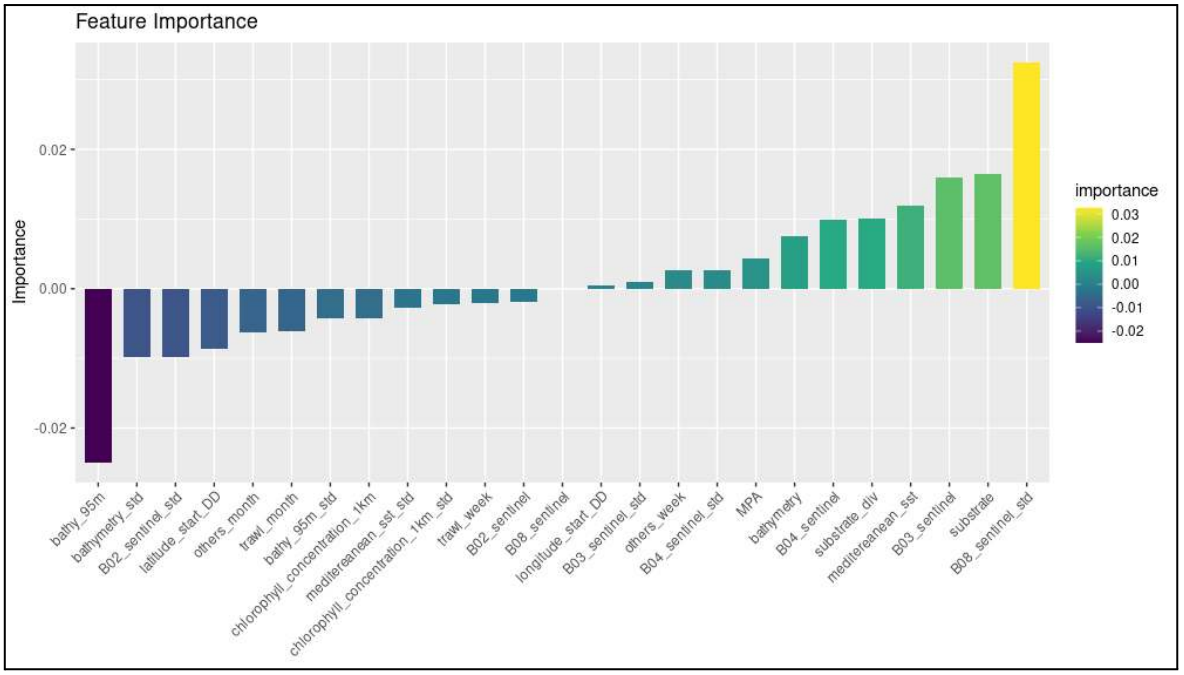



***Figure 8*** *: Mean feature importance plot for the contextual random forest model. Importance is the contribution of a feature to the test $R^2$ score*

## b. Multimodal classification task

Multimodal classification output provided mitigated results. Loss was observed to decrease slightly for validation, while it experienced a sharp decrease for training, resulting in a discrepancy between the training and validation loss. Maximum validation accuracy was measured at 0.078. Figure 8 displays the evolution of both the training and validation loss value.

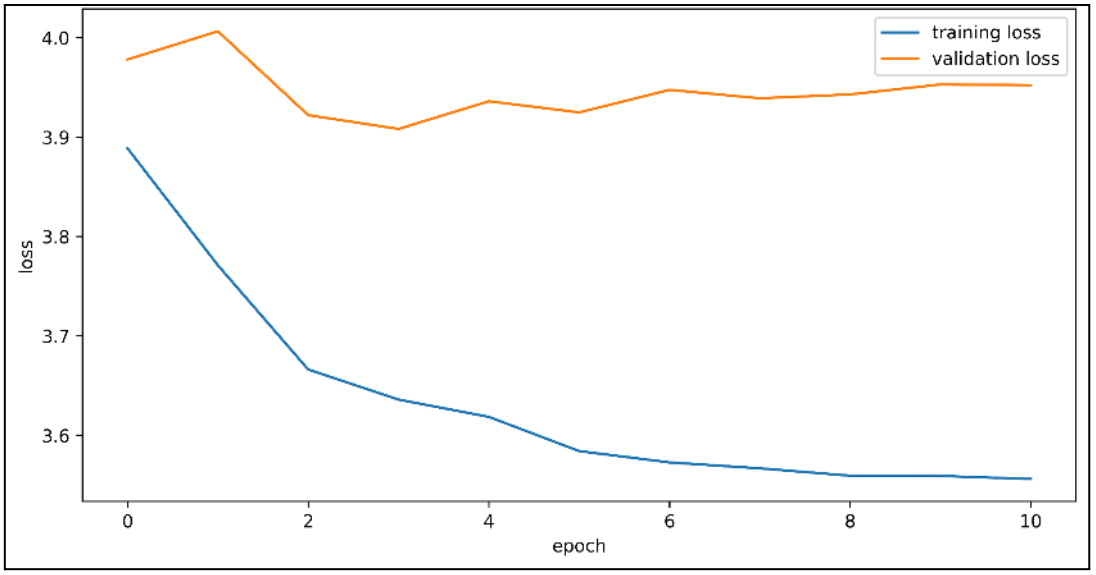



***Figure 9*** *: Evolution of the training and validation loss for the multimodal classification pre-training CNN*

### c. Untrained CNN

Across all untrained models, a general trend of convergence was observed during training. However, it was found that all models were subject to a large discrepancy between the training $R^2$ and the validation $R^2$. While training $R^2$ plateaued at approximately 0.5 for all models, validation $R^2$ stabilized well below 0, with some of the models displaying extremely local positive surges. Large fluctuations were also observed in the validation $R^2$ after convergence. Figure 9 displays the evolution of the training and validation $R^2$ scores for all six untrained models.

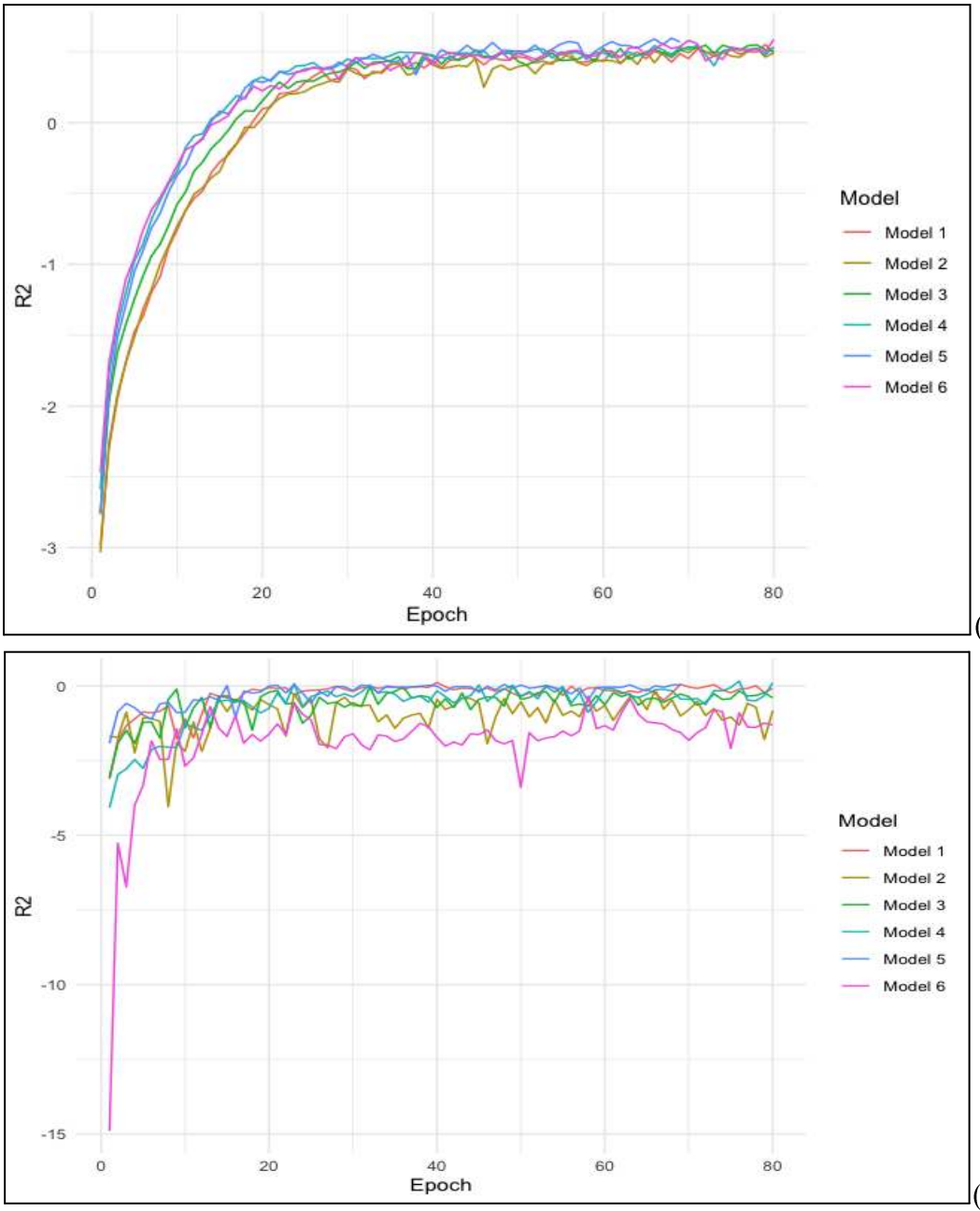
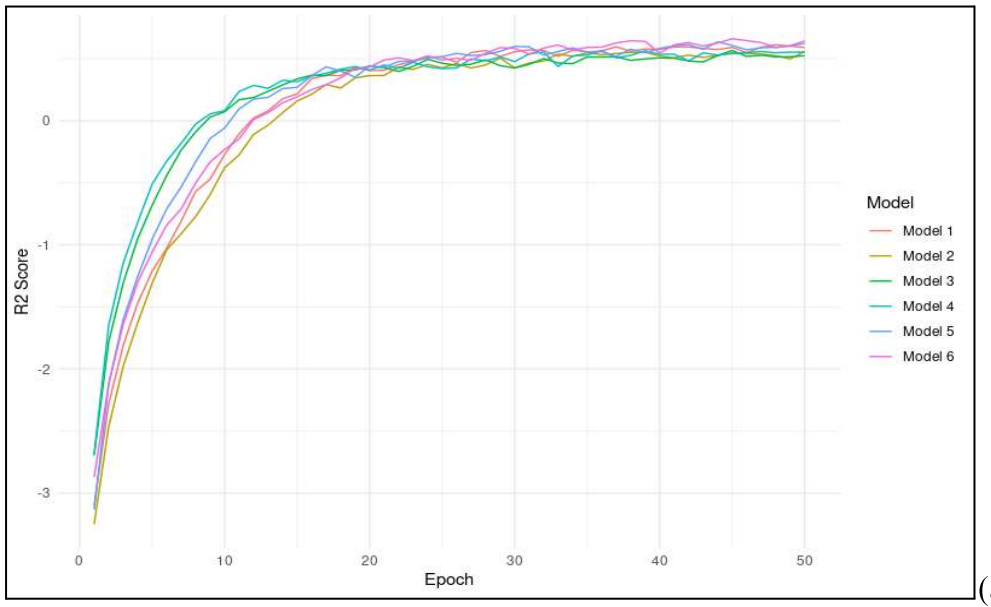


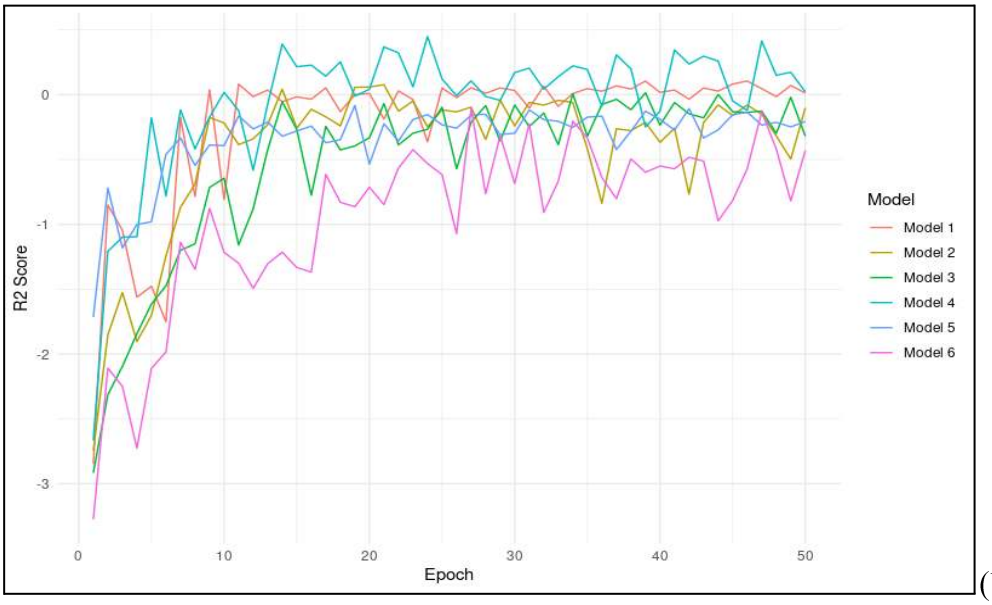

***Figure 10*** *: Evolution of the train $R^2$ (a) and validation $R^2$ (b) for all six models during training with untrained weights*

### d. Pre-trained CNN

In the case of pre-trained models, convergence was reached for all six CNN. However, similar to untrained models, the CNN displayed high instability in their validation $R^2$ after convergence was reached. Training $R^2$ displayed a similar trend as without pre-training. However, validation $R^2$ scores, while negative for most models after convergence, tended to plateau above 0 for model 1 and model 4, reaching peaks of approximately 0.1 and 0.4 respectively. Additionally, models 1 and 4 plateaued above 0. Figure 10 displays the trends of the training and validation $R^2$ scores for all six pre-trained models. Global accuracy scores for the combined predictions of the models are variable, with an $R^2$ of -0.24 for the early stopping, and 0.14 for the full training, which is superior to the accuracy of both random forest models. Results are displayed in Table 5.




(a)




(b)

***Figure 11*** *: Evolution of the train $R^2$ (a) and validation $R^2$ (b) for all six pre-trained models during training*

***Table 5*** *: $R^2$ score outputs for the totality of the CNN-predicted values, for both variable early stopping, and end-of-training weights*

| Stopping method | Global $R^2$ Score | Global RMSE score |
|---|---|---|
| Variable stop | -0.24 | 17.71 |
| Full 50 epochs | 0.14 | 15.01 |

# 4. DISCUSSION

## a. Random forests reveal contextual importance

While both random forest models show very poor performance in the prediction of species richness given our dataset, the contextual model is superior to the punctual model, with an $R^2$ score of 0.08 against -0.02. This finding is consistent with a theory of landscapes, and highlights the importance of studying features within their context, also in the marine realm.

Although the analysis of the feature importance results should be handled in respect to the poor performance of the random forests, observable trends tend to confirm complex relations between predictors and biodiversity values. Feature importance results mostly align with existing literature, highlighting the significance of substrate type and SST (Gibran and Moura, 2012). Though of lesser significance, the presence of substrate diversity - initially considered, with the standard deviation of bathymetry, to be one of the ideal proxies for depicting seascape complexity - exerts a positive influence on the prediction of fish species richness by the contextual model, which underscores the importance of habitat structure. Despite the modest importance of these contextual features, these results, when seen in the light of the improvement provided by the contextual model, suggest that interactions with other local predictors at the seascape scale could play a more dominant role in enriching the models' predictive capacity. This observation also aligns with existing literature, highlighting how the impact of local characteristics depends on their broader contextual surroundings, as discussed in the introduction (Gilby et al., 2016; Sievers et al., 2016).

However, the discrepancy between the better predictive ability of the contextual model over the local one and the low importance of most contextual features is an indicator of the necessity to bypass the creation of such limited proxies using seascape images and CNN. Surprisingly, and while exerting a positive influence on the predicting ability, depth does not emerge as one of the most important features contrary to literature findings (Rule

and Smith, 2007; Selfati et al., 2019; Stefanoudis et al., 2019). This low individual importance might be due to the fact that while depth alone is known to have a significant effect on fish assemblages, complex interactions with other features may be at play and dilute its importance among other predictors, especially given the low range of bathymetric data in seas surface eDNA samples. Additionally, the importance of Sentinel-2 imagery varies among bands, with the green band (Band 3) being the third most important identified feature. While the infrared band (Band 8) is not of a significant importance in the model, its standard deviation is among the most important features for the contextual random forest model. The contribution of these two bands to the predictions provided by the contextual random forest model may be linked to their potential role as proxies for primary production driven by phytoplankton through an equivalent of the Normalized Difference Vegetation Index (NDVI). The presence of features of highly negative importance raises the question of the quality of the data, as well as the selection of the predictors.

### b. Successful training of the CNN, with mitigated results

The multimodal classification CNN's results are mitigated. While the network displayed fast training, and reached a positive validation accuracy, this value remains low, and the large discrepancy between the training and validation loss suggests overfitting, which is a problem to solve. This low performance questions the choice of predictors, but might also be linked to the fact that GBIF data does not derive from standardized campaigns.

Similarly, the untrained network's training has yielded poor results. Both the $R^2$ scores for training and validation increased to reach a plateau, but the validation $R^2$ remained notably low, even after convergence of the network. This suggests high overfitting despite the L2 regularization, which is expected given the small size of the dataset. The instability observed in the trend of the validation $R^2$ is indicative of poor generalization performance, which was also expected since the network did not benefit from any pre-training.

Likewise, results for the fine-tuning of the pre-trained CNN were highly variable among models. While all models show similar relative performances, model 1 and 4 converged towards a positive value, meaning that the models perform better than an average output. Although this may be attributed to the pre-training since the training conditions and hyperparameters between the untrained and the pre-trained CNN were identical, $R^2$ for the validation of model 4 is surprisingly high and might be due to spatial auto-correlation. Model 4 was validated on fold 4, which can be observed to be a densely packed cluster in the proximity of fold 1. Fold 1 is among those used for training model 4. $R^2$ starting points for most models were surprisingly similar between the untrained and the pre-trained CNN, which suggest that pre-training on GBIF data had a limited benefit. As fine-tuning was performed on layer 4 and the fully connected layer of our CNN, which are by definition supposed to extract higher-level features, this result questions the similarity of the pre-training (GBIF) and fine-tuning task (eDNA), notably in the significance of high definition seascape features. While pre-training involved structural predictions of fish communities, fine-tuning was set to

predict species richness. Though this choice stemmed from a will to test our approach on a simpler task, this could conversely be considered as an oversimplification of assemblages, which may be detrimental to the predictive power of the CNN. While local species richness, as a measure of alpha diversity (Whittaker, 1960), is widely used in ecological studies of marine environments, it was for instance found to be insensitive to the presence of MPA by Dalongeville et al. (2022). However, other indicators such as functional diversity, phylogenic diversity and the ratio between pelagic and demersal fish differ significantly inside and outside of marine reserves, suggesting they are more sensitive to human impacts. While time was insufficient in the frame of this project, CNN accuracy might benefit from the use of beta diversity indicators focusing on taxonomic dissimilarity between samples (Ricotta and Szeidl, 2009), if not from species trait-based metrics, which have been observed to reflect the state of ecosystems in an efficient and consistent way (Mouillot et al., 2013). Tools such as the mFD package for R (Magneville et al., 2022) may be used for such implementations.

Despite the poor performances of the model in regard to the validation, the CNN performed surprisingly well when used to predict species richness on the test sets. While early stopping output mostly provided negative $R^2$, and a global $R^2$ of -0.24 with an RMSE of 17.71, models that were trained for 50 epochs collectively showed an $R^2$ score of 0.14 with an RMSE of 15.01, which is above the accuracy of both random forests. This is in accordance with our hypothesis and goes in the way of demonstrating the superiority of a seascape approach when compared with standard methods. However, while this result is promising , it must be evaluated in the light of the contrasting poor performance on the validation sets, and of the limitations of the dataset. Clustering, notably, might have failed to completely avoid spatial auto-correlation, and could induce a performance boost for some of the CNN random forests. While superior to the random forest, the CNN display an overall poor predicting power. Additionally, the fact that accuracy is significantly better at the end of training than for early stopping, despite similar validation $R^2$, suggests that increased test accuracy might be linked to overfitting on the training set.

Throughout training, the $R^2$ consistently increases, demonstrating that the architecture is working and that the CNN is learning as intended. However, instability in the validation $R^2$ suggests that the CNN unlearns features, indicating that some samples in the training set are likely to carry contradictory information. Globally, these poor results likely stem from a combination of problems concerning the biodiversity indicators that were used and limitations of the dataset that will be detailed in part 3.c.

All other things being equal, our results suggest that CNN can provide better predictions of biodiversity indicators than standard regression models by efficiently taking into account contextual data.

### c. Lack of data and poor predictors of diversity

The main limitation we faced during training was the overall lack of data, as well an apparent lack of diversity and relevance in the predictors. The number of eDNA samples constituting the final set is very low when compared to the usual size of datasets used to train CNN. Additionally, the range of species richness values is high and their frequencies homogenous (Figure 11), which can arguably be seen as the sign of a robust dataset, fit for training on a good span of situations. However, when compared to the distribution of modalities in some of the predictors that were identified as most important in the literature, it becomes apparent that sampling mostly occurred in low-depth zones (Figure 13), with an over-representation of posidonia meadows and rocky substrates (Figure 12), which are known to be high-potential environments in terms of biodiversity (Moreno, 2002). This may lead to the CNN being fed a diversity of response values corresponding to a rather homogenous set of inputs, leading to the confusion of the network from one batch to the other. Habib et al. (2019) have shown that training accuracy on medical imagery increased with sample diversity rather than with the size of the dataset. Ideally, the CNN should have access to a large dataset consisting of diverse examples of predictors associated with a good range of diversity values. In our case, both of these limitations arise from the necessarily biased sampling plan that was carried out before this project : eDNA data was mostly collected in areas that were known to be rich, with the main goal being to assess the efficiency of MPA in the French Mediterranean Sea by comparing inventories performed both inside and outside of protected areas (Dalongeville et al., 2022).

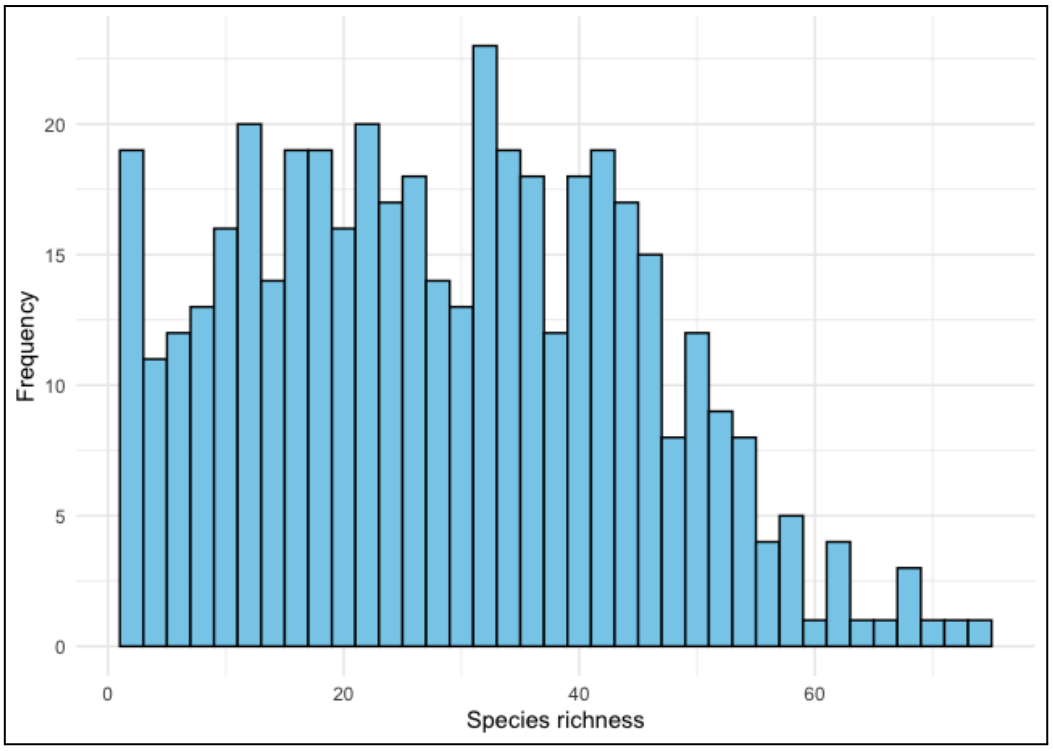



***Figure 12*** *: Frequency distribution of species richness in the eDNA dataset*

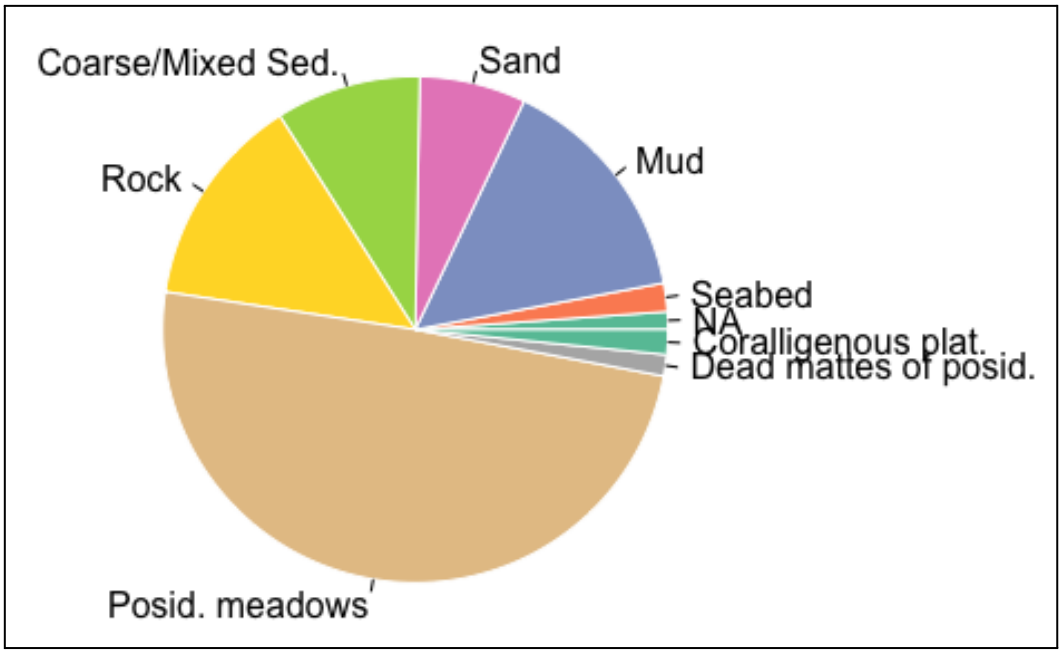



***Figure 13*** *: Frequency distribution of substrates in the eDNA dataset*

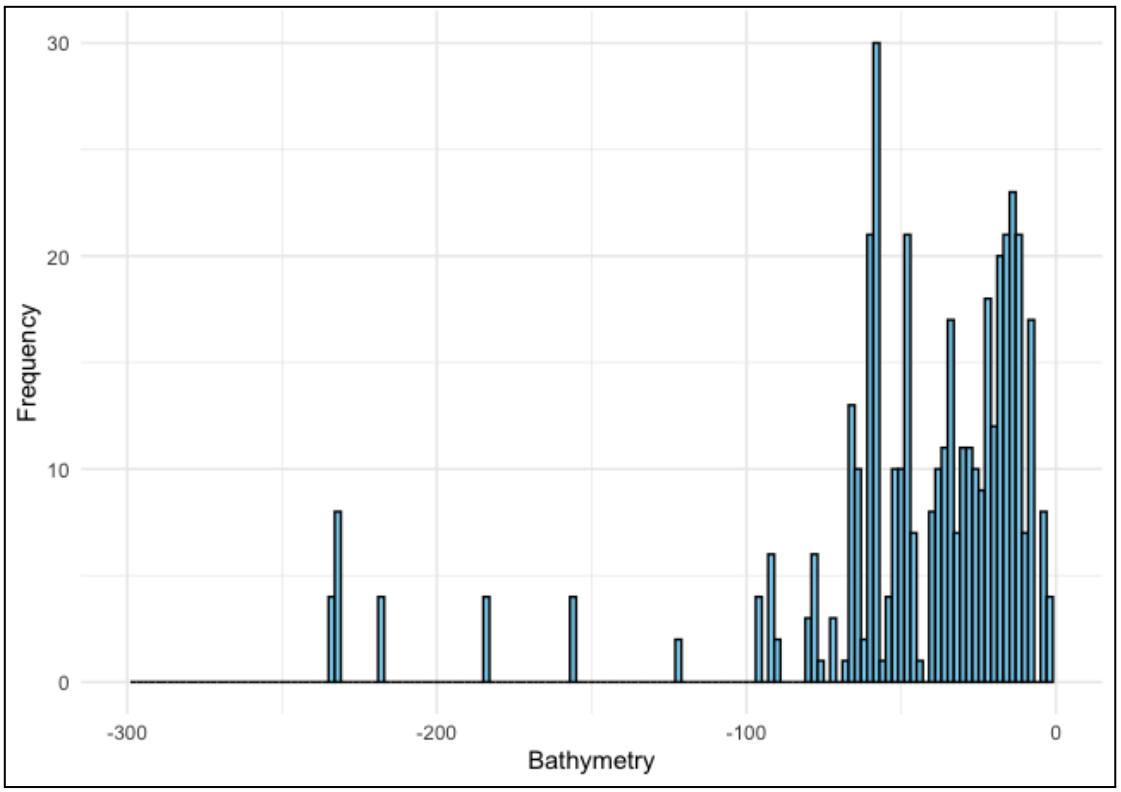



***Figure 14*** *: Frequency distribution of bathymetry in the eDNA dataset*

Furthermore, the distribution of sampling points was observed to be highly clustered, which greatly complicated the creation of the folds used for cross-validation. Because one of the main objectives was to build folds of a similar size, k-mean clustering led to the generation of non-homogenous groups in terms of distance between the samples. For this

issue to be resolved, all sampling points should be homogeneously distributed. Should a new dataset be similarly clustered, other cross-validation methods, such as density-weighted fold generation (de Bruin et al., 2022), are to be explored. Another solution may be to use a weighted sampler, both for the GBIF and the eDNA CNN, to generate the batches while keeping an even spread of predictor classes.

The first step in increasing the predicting power of our CNN should therefore be to rebuild a robust dataset containing a large diversity of seascapes. The BioDivMed 2023 campaign could take part in the building of such data, as 700 new eDNA samples will be gathered at locations spanning over 2000 km (Université de Montpellier, 2023).

### d. Time and resources limitations

Researching, acquiring and treating data in order to produce a dataset that was suitable for training has accounted for a large portion of this project. Because no pre-made dataset was available for the majority of our predictors, it became necessary to develop custom tools to treat available data. These procedures tended to be very time-consuming as data was in the form of large rasters, vector maps, or, in the case of fishing data, very large data frames. Similarly, deep learning networks, in particular CNN, are rather opaque fields for untrained operators, and some tasks would benefit from more user-friendly procedures. As CNN become increasingly used in the field of ecology, the development of large multi-purpose datasets and tools will become essential in disseminating deep-learning techniques. Such tools already exist : TorchGeo (Stewart et al., 2022), a python package designed to help build geospatial datasets used in CNN, served as the basis for some of the tools developed here.

Additionally, training a CNN is very labor intensive and costly in both time and computing resources. Using high end graphics cards such as a Tesla A100 GPU, a single training procedure can take up to an hour and half. Following the trial and error process that was used here to try and find an optimal set of hyperparameters, this number is to be multiplied by the tens of runs that were carried out for each of the models and modalities. Considering the resources needed to successfully train a CNN this way, it is possible that some of the selected hyperparameters and/or methods were sub-optimal given the dataset and task that were developed. While more time might lead to the selection of a better set of hyperparameters, it is important to highlight that one set is fit for a given task and that it appears essential, in our situation, to prioritize the development of a better dataset rather than to spend more resources trying to optimize training on current data.

### e. Alternative methods and architectures

While it is essential to construct a better biodiversity dataset for training, some alternative methods might be of use in preparation for a variety of tasks. In this project, pre-training was carried-out on a task that was designed to be similar to the end objective.

While this is theoretically an efficient method, building a preliminary dataset on a similar task may reveal itself to be very resource-consuming, and can lead, as shown here, to mitigated results. Moreover, this strategy relies on the existence of a large dataset of similar response variables. One way to maximize the size of the pre-training dataset while removing the need for actual biodiversity data would be through the use of self-supervised training. Self-supervised training is based on the idea that feature extraction can be performed without necessarily having access to actual values to predict (Yuan et al., 2021). During this procedure, CNN are fed regular inputs, but pretext tasks are used to output a loss value : this includes completing a part of the input, predicting relative positions of image patches etc. Using self-supervised training, one can pre-train a CNN on a virtually infinite dataset. Preliminary tests were carried out using 20 000 random bathymetry patches in the Mediterranean Sea, and produced promising results. Data augmentation methods may also be used to artificially increase the size of the final training dataset, by modifying existing data (Maharana et al., 2022). Random cropping and rotations have already shown promising results on the GBIF dataset, and might be tested on the eDNA data. Others tests also showed that the joint use of the loss and of the recent Koleo loss, which is part of the DINOv2 architecture (Oquab et al., 2023), increases the performance of the GBIF CNN.

Finally, recent network architectures can be used to bypass the compromise between high resolution imagery and geographically large tiles. Notably, transformer networks allow to take into account both fine grain inputs and their respective geographical position inside of a large scale context. Contrary to standard CNN, transformers divide images into multiple tiles that are then linearized and treated as ‘words’ in a ‘sentence’. This ‘sentence’ is embedded with a positional vector that contains information about each tile’s position within the original image. Geographical position then becomes an input, and attention maps can be computed to explore the impact of features on the predicted value (Vaswani et al., 2017).

# CONCLUSION

The CNN that was developed in this study yielded a significant improvement over standard random forest methods, but remains limited in terms of predictive ability. However, the poor performances of the random forest model highlight the probable responsibility of the dataset in the CNN's ineffectiveness. Contextual data was successfully identified as a pathway to improve model accuracy, and underscores the need for the study of coastal environments in the context of their seascape. With the implementation of new datasets built using upcoming eDNA sampling campaigns, deep learning networks are likely to be central towards a new generation of biodiversity models.

While our CNN were not successfully trained to accurately accomplish the task that they were designed for, it is important to highlight that they were nonetheless successfully trained. It is highly unlikely that significantly better results are in reach given the limitations of our dataset. However, this project has contributed to the development of tools and workflows that will be reused, and constitutes a theoretical and practical groundwork for further biodiversity modeling tasks.

# BIBLIOGRAPHY


Airamé, S., Dugan, J.E., Lafferty, K.D., Leslie, H., McArdle, D.A., Warner, R.R., 2003. APPLYING ECOLOGICAL CRITERIA TO MARINE RESERVE DESIGN: A CASE STUDY FROM THE CALIFORNIA CHANNEL ISLANDS. Ecol. Appl. 13, 170–184. https://doi.org/10.1890/1051-0761(2003)013[0170:AECTMR]2.0.CO;2

Al-Kababji, A., Bensaali, F., Dakua, S.P., 2022. Scheduling Techniques for Liver Segmentation: ReduceLROnPlateau Vs OneCycleLR.

Awada, H., Aronica, S., Bonanno, A., Basilone, G., Zgozi, S.W., Giacalone, G., Fontana, I., Genovese, S., Ferreri, R., Mazzola, S., Spagnolo, B., Valenti, D., Denaro, G., 2021. A novel method to simulate the 3D chlorophyll distribution in marine oligotrophic waters. Commun. Nonlinear Sci. Numer. Simul. 103, 106000. https://doi.org/10.1016/j.cnsns.2021.106000

Bell, J.D., 1983. Effects of Depth and Marine Reserve Fishing Restrictions on the Structure of a Rocky Reef Fish Assemblage in the North-Western Mediterranean Sea. J. Appl. Ecol. 20, 357. https://doi.org/10.2307/2403513

Belmaker, J., Shashar, N., Ziv, Y., 2005. Effects of small-scale isolation and predation on fish diversity on experimental reefs. Mar. Ecol. Prog. Ser. 289, 273–283. https://doi.org/10.3354/meps289273

Belmaker, J., Ziv, Y., Shashar, N., 2011. The influence of connectivity on richness and temporal variation of reef fishes. Landsc. Ecol. 26, 587–597. https://doi.org/10.1007/s10980-011-9588-0

Benkendorf, D.J., Hawkins, C.P., 2020. Effects of sample size and network depth on a deep learning approach to species distribution modeling. Ecol. Inform. 60, 101137. https://doi.org/10.1016/j.ecoinf.2020.101137

Bevilacqua, S., Airoldi, L., Ballesteros, E., Benedetti-Cecchi, L., Boero, F., Bulleri, F., Cebrian, E., Cerrano, C., Claudet, J., Colloca, F., Coppari, M., Di Franco, A., Fraschetti, S., Garrabou, J., Guarnieri, G., Guerranti, C., Guidetti, P., Halpern, B.S., Katsanevakis, S., Mangano, M.C., Micheli, F., Milazzo, M., Pusceddu, A., Renzi, M., Rilov, G., Sarà, G., Terlizzi, A., 2021. Mediterranean rocky reefs in the Anthropocene: Present status and future concerns, in: Advances in Marine Biology. Elsevier, pp. 1–51. https://doi.org/10.1016/bs.amb.2021.08.001

Bianchi, C.N., Morri, C., 2000. RMESUaLTrSine Biodiversity of the Mediterranean Sea: Situation, Problems and Prospects for Future Research. Mar. Pollut. Bull. 40.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E., 2022. Deep learning as a tool for ecology and evolution. Methods Ecol. Evol. 13, 1640–1660. https://doi.org/10.1111/2041-210X.13901

Charton, J.A.G., Ruzafa, A.P., 1998. Correlation Between Habitat Structure and a Rocky Reef Fish Assemblage in the Southwest Mediterranean. Mar. Ecol. 19, 111–128. https://doi.org/10.1111/j.1439-0485.1998.tb00457.x

Chassot, E., Bonhommeau, S., Dulvy, N.K., Mélin, F., Watson, R., Gascuel, D., Le Pape, O., 2010. Global marine primary production constrains fisheries catches. Ecol. Lett. 13, 495–505. https://doi.org/10.1111/j.1461-0248.2010.01443.x

Chu, Z., Yu, J., 2020. An end-to-end model for rice yield prediction using deep learning fusion. Comput. Electron. Agric. 174, 105471. https://doi.org/10.1016/j.compag.2020.105471

Claudet, J., Loiseau, C., Sostres, M., Zupan, M., 2020. Underprotected Marine Protected Areas in a Global Biodiversity Hotspot. One Earth 2, 380–384. https://doi.org/10.1016/j.oneear.2020.03.008

Claudet, J., Pelletier, D., Jouvenel, J.-Y., Bachet, F., Galzin, R., 2006. Assessing the effects of marine protected area (MPA) on a reef fish assemblage in a northwestern Mediterranean marine reserve: Identifying community-based indicators. Biol. Conserv. 130, 349–369. https://doi.org/10.1016/j.biocon.2005.12.030

CMEMS, 2023a. Mediterranean Sea - High Resolution L4 Sea Surface Temperature

Reprocessed [WWW Document]. URL https://data.marine.copernicus.eu/product/SST_MED_SST_L4_REP_OBSERVATIONS_010_021/description (accessed 8.17.23).
CMEMS, 2023b. Mediterranean Sea Ocean Colour Plankton MY L4 daily gapfree observations and climatology and monthly observations [WWW Document]. URL https://data.marine.copernicus.eu/product/OCEANCOLOUR_MED_BGC_L4_MY_009_144/description (accessed 8.17.23).
Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Ben Rais Lasram, F., Aguzzi, J., Ballesteros, E., Bianchi, C.N., Corbera, J., Dailianis, T., Danovaro, R., Estrada, M., Froglia, C., Galil, B.S., Gasol, J.M., Gertwagen, R., Gil, J., Guilhaumon, F., Kesner-Reyes, K., Kitsos, M.-S., Koukouras, A., Lampadariou, N., Laxamana, E., López-Fé de la Cuadra, C.M., Lotze, H.K., Martin, D., Mouillot, D., Oro, D., Raicevich, S., Rius-Barile, J., Saiz-Salinas, J.I., San Vicente, C., Somot, S., Templado, J., Turon, X., Vafidis, D., Villanueva, R., Voultsiadou, E., 2010. The Biodiversity of the Mediterranean Sea: Estimates, Patterns, and Threats. PLoS ONE 5, e11842. https://doi.org/10.1371/journal.pone.0011842
Collie, J.S., Hall, S.J., Kaiser, M.J., Poiner, I.R., 2000. A quantitative analysis of fishing impacts on shelf-sea benthos: Effects of fishing on benthos. J. Anim. Ecol. 69, 785–798. https://doi.org/10.1046/j.1365-2656.2000.00434.x
Condal, F., Aguzzi, J., Sardà, F., Nogueras, M., Cadena, J., Costa, C., Del Río, J., Mànuel, A., 2012. Seasonal rhythm in a Mediterranean coastal fish community as monitored by a cabled observatory. Mar. Biol. 159, 2809–2817. https://doi.org/10.1007/s00227-012-2041-3
Copernicus, 2011. Copernicus Sentinel-2 data. Retrieved from Planetary Computer, processed by ESA [WWW Document]. URL (accessed 6.1.23).
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., Romant, L., Manel, S., Mouillot, D., 2022. Benchmarking eleven biodiversity indicators based on environmental DNA surveys: More diverse functional traits and evolutionary lineages inside marine reserves. J. Appl. Ecol. 59, 2803–2813. https://doi.org/10.1111/1365-2664.14276
Davies, C.E., Moss, D., Hill, M.O., 2004. EUNIS HABITAT CLASSIFICATION REVISED 2004.
Day, J., Dudley, N., Hockings, M., Holmes, G., Laffoley, D., Stolton, S., Wells, S., Wenzel, L., 2019. Guidelines for applying the IUCN protected area management categories to marine protected areas. Prot. Area Manag.
de Bruin, S., Brus, D.J., Heuvelink, G.B.M., van Ebbenhorst Tengbergen, T., Wadoux, A.M.J.-C., 2022. Dealing with clustered samples for assessing map accuracy by cross-validation. Ecol. Inform. 69, 101665. https://doi.org/10.1016/j.ecoinf.2022.101665
Edgar, G.J., Bates, A.E., Bird, T.J., Jones, A.H., Kininmonth, S., Stuart-Smith, R.D., Webb, T.J., 2016. New Approaches to Marine Conservation Through the Scaling Up of Ecological Data. Annu. Rev. Mar. Sci. 8, 435–461. https://doi.org/10.1146/annurev-marine-122414-033921
Estopinan, J., Servajean, M., Bonnet, P., Munoz, F., Joly, A., 2022. Deep Species Distribution Modeling From Sentinel-2 Image Time-Series: A Global Scale Analysis on the Orchid Family. Front. Plant Sci. 13, 839327. https://doi.org/10.3389/fpls.2022.839327
Fanelli, E., Cartes, J.E., Papiol, V., López-Pérez, C., 2013. Environmental drivers of megafaunal assemblage composition and biomass distribution over mainland and insular slopes of the Balearic Basin (Western Mediterranean). Deep Sea Res. Part Oceanogr. Res. Pap. 78, 79–94. https://doi.org/10.1016/j.dsr.2013.04.009
GBIF [WWW Document], n.d. URL https://www.gbif.org/ (accessed 6.26.23).
Gibran, F.Z., Moura, R.L.D., 2012. The structure of rocky reef fish assemblages across a

nearshore to coastal islands' gradient in Southeastern Brazil. Neotropical Ichthyol. 10, 369–382. https://doi.org/10.1590/S1679-62252012005000013
Gilby, B.L., Tibbetts, I.R., Olds, A.D., Maxwell, P.S., Stevens, T., 2016. Seascape context and predators override water quality effects on inshore coral reef fish communities. Coral Reefs 35, 979–990. https://doi.org/10.1007/s00338-016-1449-5
Green, A.L., Maypa, A.P., Almany, G.R., Rhodes, K.L., Weeks, R., Abesamis, R.A., Gleason, M.G., Mumby, P.J., White, A.T., 2015. Larval dispersal and movement patterns of coral reef fishes, and implications for marine reserve network design. Biol. Rev. 90, 1215–1247. https://doi.org/10.1111/brv.12155
Grober-Dunsmore, R., Frazer, T.K., Beets, J.P., Lindberg, W.J., Zwick, P., Funicelli, N.A., 2008. Influence of landscape structure on reef fish assemblages. Landsc. Ecol. 23, 37–53. https://doi.org/10.1007/s10980-007-9147-x
Grober-Dunsmore, R., Frazer, T.K., Lindberg, W.J., Beets, J., 2007. Reef fish and habitat relationships in a Caribbean seascape: the importance of reef context. Coral Reefs 26, 201–216. https://doi.org/10.1007/s00338-006-0180-z
Guidetti, P., Baiata, P., Ballesteros, E., Di Franco, A., Hereu, B., Macpherson, E., Micheli, F., Pais, A., Panzalis, P., Rosenberg, A.A., Zabala, M., Sala, E., 2014. Large-Scale Assessment of Mediterranean Marine Protected Areas Effects on Fish Assemblages. PLoS ONE 9, e91841. https://doi.org/10.1371/journal.pone.0091841
Gupta, A., Ramanath, R., Shi, J., Keerthi, S.S., 2021. Adam vs. SGD: Closing the generalization gap on image classification.
Habib, A., Karmakar, C., Yearwood, J., 2019. Impact of ECG Dataset Diversity on Generalization of CNN Model for Detecting QRS Complex. IEEE Access 7, 93275–93285. https://doi.org/10.1109/ACCESS.2019.2927726
Hackradt, C.W., García-Charton, J.A., Harmelin-Vivien, M., Pérez-Ruzafa, Á., Le Diréach, L., Bayle-Sempere, J., Charbonnel, E., Ody, D., Reñones, O., Sanchez-Jerez, P., Valle, C., 2014. Response of Rocky Reef Top Predators (Serranidae: Epinephelinae) in and Around Marine Protected Areas in the Western Mediterranean Sea. PLoS ONE 9, e98206. https://doi.org/10.1371/journal.pone.0098206
He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition.
He, K.S., Bradley, B.A., Cord, A.F., Rocchini, D., Tuanmu, M., Schmidtlein, S., Turner, W., Wegmann, M., Pettorelli, N., 2015. Will remote sensing shape the next generation of species distribution models? Remote Sens. Ecol. Conserv. 1, 4–18. https://doi.org/10.1002/rse2.7
Jennings, S., Kaiser, M.J., 1998. The Effects of Fishing on Marine Ecosystems, in: Advances in Marine Biology. Elsevier, pp. 201–352. https://doi.org/10.1016/S0065-2881(08)60212-6
Kavanaugh, M., Bell, T., Catlett, D., Cimino, M., Doney, S., Klajbor, W., Messié, M., Montes, E., Muller Karger, F., Otis, D., Santora, J., Schroeder, I., Triñanes, J., Siegel, D., 2021. Satellite Remote Sensing and the Marine Biodiversity Observation Network: Current Science and Future Steps. Oceanography 34. https://doi.org/10.5670/oceanog.2021.215
Knudby, A., LeDrew, E., Brenning, A., 2010. Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. Remote Sens. Environ. 114, 1230–1241. https://doi.org/10.1016/j.rse.2010.01.007
Kostylev, V.E., Erlandsson, J., Ming, M.Y., Williams, G.A., 2005. The relative importance of habitat complexity and surface area in assessing biodiversity: Fractal application on rocky shores. Ecol. Complex. 2, 272–286. https://doi.org/10.1016/j.ecocom.2005.04.002
Kroodsma, D.A., Mayorga, J., Hochberg, T., Miller, N.A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T.D., Block, B.A., Woods, P., Sullivan, B., Costello, C., Worm, B., 2018. Tracking the global footprint of fisheries. Science 359, 904–908. https://doi.org/10.1126/science.aao5646
Levins, R., 1969. Some Demographic and Genetic Consequences of Environmental

Heterogeneity for Biological Control. Bull. Entomol. Soc. Am. 15, 237–240. https://doi.org/10.1093/besa/15.3.237
Lundquist, C.J., Granek, E.F., 2005. Strategies for Successful Marine Conservation: Integrating Socioeconomic, Political, and Scientific Factors. Conserv. Biol. 19, 1771–1778. https://doi.org/10.1111/j.1523-1739.2005.00279.x
MacArthur, R., Wilson, E.O., 1967. The theory of island biogeography, Princeton University Press. ed. https://doi.org/10.1515/9781400881376
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
Maharana, K., Mondal, S., Nemade, B., 2022. A review: Data pre-processing and data augmentation techniques. Glob. Transit. Proc., International Conference on Intelligent Engineering Approach(ICIEA-2022) 3, 91–99. https://doi.org/10.1016/j.gltp.2022.04.020
Mellin, C., Andréfouët, S., Kulbicki, M., Dalleau, M., Vigliola, L., 2009. Remote sensing and fish–habitat relationships in coral reef ecosystems: Review and pathways for multi-scale hierarchical research. Mar. Pollut. Bull. 58, 11–19. https://doi.org/10.1016/j.marpolbul.2008.10.010
Mellin, C., Andréfouët, S., Ponton, D., 2007. Spatial predictability of juvenile fish species richness and abundance in a coral reef environment. Coral Reefs 26, 895–907. https://doi.org/10.1007/s00338-007-0281-3
Mellin, C., Bradshaw, C.J.A., Meekan, M.G., Caley, M.J., 2010. Environmental and spatial predictors of species richness and abundance in coral reef fishes: Predicting coral reef fish species richness and abundance. Glob. Ecol. Biogeogr. 19, 212–222. https://doi.org/10.1111/j.1466-8238.2009.00513.x
Monfort, T., Cheminée, A., Bianchimani, O., Drap, P., Puzenat, A., Thibaut, T., 2021. The Three-Dimensional Structure of Mediterranean Shallow Rocky Reefs: Use of Photogrammetry-Based Descriptors to Assess Its Influence on Associated Teleost Assemblages. Front. Mar. Sci. 8, 639309. https://doi.org/10.3389/fmars.2021.639309
Mora, C., Tittensor, D.P., Myers, R.A., 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. Proc. R. Soc. B Biol. Sci. 275, 149–155. https://doi.org/10.1098/rspb.2007.1315
Moreno, I., 2002. Effects of substrate on the artificial reef fish assemblage in Santa Eulalia Bay (Ibiza, western Mediterranean). ICES J. Mar. Sci. 59, S144–S149. https://doi.org/10.1006/jmsc.2002.1271
Mouillot, D., Graham, N.A.J., Villéger, S., Mason, N.W.H., Bellwood, D.R., 2013. A functional approach reveals community responses to disturbances. Trends Ecol. Evol. 28, 167–177. https://doi.org/10.1016/j.tree.2012.10.004
Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. DINOv2: Learning Robust Visual Features without Supervision.
Pittman, S., McAlpine, C., Pittman, K., 2004. Linking fish and prawns to their environment: a hierarchical landscape approach. Mar. Ecol. Prog. Ser. 283, 233–254. https://doi.org/10.3354/meps283233
Planes, S., Galzin, R., Rubies, A.G., Goñi, R., Harmelin, J.-G., Diréach, L.L., Lenfant, P., Quetglas, A., 2000. Effects of marine protected areas on recruitment processes with special reference to Mediterranean littoral ecosystems. Environ. Conserv. 27, 126–143. https://doi.org/10.1017/S0376892900000175
Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11, 4540.

https://doi.org/10.1038/s41467-020-18321-y
Purkis, S.J., Graham, N.A.J., Riegl, B.M., 2008. Predictability of reef fish diversity and abundance using remote sensing data in Diego Garcia (Chagos Archipelago). Coral Reefs 27, 167–178. https://doi.org/10.1007/s00338-007-0306-y
Rees, M.J., Jordan, A., Price, O.F., Coleman, M.A., Davis, A.R., 2014. Abiotic surrogates for temperate rocky reef biodiversity: implications for marine protected areas. Divers. Distrib. 20, 284–296. https://doi.org/10.1111/ddi.12134
Registry-Migration.Gbif.Org, 2022. GBIF Backbone Taxonomy. https://doi.org/10.15468/39OMEI
Ricotta, C., Szeidl, L., 2009. Diversity partitioning of Rao's quadratic entropy. Theor. Popul. Biol. 76, 299–302. https://doi.org/10.1016/j.tpb.2009.10.001
Rourke, M.L., Fowler, A.M., Hughes, J.M., Broadhurst, M.K., DiBattista, J.D., Fielder, S., Wilkes Walburn, J., Furlan, E.M., 2022. Environmental DNA (eDNA) as a tool for assessing fish biomass: A review of approaches and future considerations for resource surveys. Environ. DNA 4, 9–33. https://doi.org/10.1002/edn3.185
Rule, M.J., Smith, S.D.A., 2007. Depth-associated patterns in the development of benthic assemblages on artificial substrata deployed on shallow, subtropical reefs. J. Exp. Mar. Biol. Ecol. 345, 38–51. https://doi.org/10.1016/j.jembe.2007.01.006
Sahyoun, R., Bussotti, S., Di Franco, A., Navone, A., Panzalis, P., Guidetti, P., 2013. Protection effects on Mediterranean fish assemblages associated with different rocky habitats. J. Mar. Biol. Assoc. U. K. 93, 425–435. https://doi.org/10.1017/S0025315412000975
Sala, E., Ballesteros, E., Dendrinos, P., Di Franco, A., Ferretti, F., Foley, D., Fraschetti, S., Friedlander, A., Garrabou, J., Güçlüsoy, H., Guidetti, P., Halpern, B.S., Hereu, B., Karamanlidis, A.A., Kizilkaya, Z., Macpherson, E., Mangialajo, L., Mariani, S., Micheli, F., Pais, A., Riser, K., Rosenberg, A.A., Sales, M., Selkoe, K.A., Starr, R., Tomas, F., Zabala, M., 2012. The Structure of Mediterranean Rocky Reef Ecosystems across Environmental and Human Gradients, and Conservation Implications. PLoS ONE 7, e32742. https://doi.org/10.1371/journal.pone.0032742
Selfati, M., El Ouamari, N., Franco, A., Lenfant, P., Lecaillon, G., Mesfioui, A., Boissery, P., Bazairi, H., 2019. Fish assemblages of the Marchica lagoon (Mediterranean, Morocco): Spatial patterns and environmental drivers. Reg. Stud. Mar. Sci. 32, 100896. https://doi.org/10.1016/j.rsma.2019.100896
Sievers, K., Barr, R., Maloney, J., Driscoll, N., Anderson, T., 2016. Impact of habitat structure on fish populations in kelp forests at a seascape scale. Mar. Ecol. Prog. Ser. 557, 51–63. https://doi.org/10.3354/meps11885
Sigsgaard, E.E., Nielsen, I.B., Carl, H., Krag, M.A., Knudsen, S.W., Xing, Y., Holm-Hansen, T.H., Møller, P.R., Thomsen, P.F., 2017. Seawater environmental DNA reflects seasonality of a coastal fish community. Mar. Biol. 164, 128. https://doi.org/10.1007/s00227-017-3147-4
Sinclair, M., Valdimarsson, G. (Eds.), 2003. Responsible fisheries in the marine ecosystem, 1st ed. CABI Publishing, UK. https://doi.org/10.1079/9780851996332.0000
Smith, L.N., 2017. Cyclical Learning Rates for Training Neural Networks.
Spear, M.J., Embke, H.S., Krysan, P.J., Vander Zanden, M.J., 2021. Application of eDNA as a tool for assessing fish population abundance. Environ. DNA 3, 83–91. https://doi.org/10.1002/edn3.94
Stefanoudis, P.V., Gress, E., Pitt, J.M., Smith, S.R., Kincaid, T., Rivers, M., Andradi-Brown, D.A., Rowlands, G., Woodall, L.C., Rogers, A.D., 2019. Depth-Dependent Structuring of Reef Fish Assemblages From the Shallows to the Rariphotic Zone. Front. Mar. Sci. 6, 307. https://doi.org/10.3389/fmars.2019.00307
Stewart, A.J., Robinson, C., Corley, I.A., Ortiz, A., Ferres, J.M.L., Banerjee, A., 2022. TorchGeo: Deep Learning With Geospatial Data.
UNEP-WCMC, IUCN, 2023. Protected Planet: The World Database on Protected Areas (WDPA) [Online]. URL www.protectedplanet.net
Université de Montpellier, 2023. Mission BioDivMed 2023 : l'ADN environnemental pour une

cartographie inédite de la biodiversité marine méditerranéenne [WWW Document]. https://www.umontpellier.fr. URL https://www.umontpellier.fr/articles/mission-biodivmed-2023-ladn-environnemental-pour-une-cartographieinedite-de-la-biodiversite-marine-mediterraneenne (accessed 8.28.23).
Ushiama, S., Smith, J.A., Suthers, I.M., Lowry, M., Johnston, E.L., 2016. The effects of substratum material and surface orientation on the developing epibenthic community on a designed artificial reef. Biofouling 32, 1049–1060. https://doi.org/10.1080/08927014.2016.1224860
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., Peroux, T., Crivelli, A.J., Olivier, A., Acqueberge, M., Le Brun, M., Møller, P.R., Willerslev, E., Dejean, T., 2016. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Mol. Ecol. 25, 929–942. https://doi.org/10.1111/mec.13428
van Denderen, P.D., Hintzen, N.T., Rijnsdorp, A.D., Ruardij, P., van Kooten, T., 2014. Habitat-Specific Effects of Fishing Disturbance on Benthic Species Richness in Marine Soft Sediments. Ecosystems 17, 1216–1226. https://doi.org/10.1007/s10021-014-9789-x
Vasquez, M., Albrecht, J., Manca, E., Agnesi, S., Al Hamdani, Z., Andersen, J., Annunziatellis, A., Bekkby, T., Bruschi, A., Doncheva, V., Drakopoulou, V., Duncan, G., Inghilesi, R., Kyriakidou, C., Lalli, F., Lillis, H., Mo, G., Muresan, M., Salomidi, M., Sakellariou, D., Simboura, M., Teaca, A., Tezcan, D., Todorova, V., Tunesi, L., 2019. EUSeaMap. A European broad-scale seabed habitat map. Ifremer. https://doi.org/10.13155/49975
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All you Need.
Vilas, D., Pennino, M.G., Bellido, J.M., Navarro, J., Palomera, I., Coll, M., 2020. Seasonality of spatial patterns of abundance, biomass, and biodiversity in a demersal community of the NW Mediterranean Sea. ICES J. Mar. Sci. 77, 567–580. https://doi.org/10.1093/icesjms/fsz197
Whittaker, R.H., 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecol. Monogr. 30, 279–338. https://doi.org/10.2307/1943563
Yuan, Y., Borrmann, D., Hou, J., Ma, Y., Nüchter, A., Schwertfeger, S., 2021. Self-Supervised Point Set Local Descriptors for Point Cloud Registration. Sensors 21, 486. https://doi.org/10.3390/s21020486
Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a review. IEEE Geosci. Remote Sens. Mag. 5, 8–36. https://doi.org/10.1109/MGRS.2017.2762307

| | |
|---|---|
| L'INSTITUT agro Rennes Angers | Diplôme : Ingénieur agronome<br>Spécialité : **Sciences Halieutiques et Aquacoles**<br>Spécialisation / option : Ressources et Environnements Aquatiques<br>Enseignant référent : Etienne Rivot |


| | |
|---|---|
| Auteur(s) : Simon Bettinger<br><br>Date de naissance* : 02/06/1999 | Organisme d'accueil : UMR MARBEC<br>Adresse : 093 Pl. Eugène Bataillon, 34090 Montpellier |
| Nb pages : 31 Annexe(s) : 0 | |
| Année de soutenance : 2023 | Maître de stage : David Mouillot |


Titre français : Amélioration de la prédiction de la biodiversité côtière en mer Méditerranée à travers une approche d'apprentissage profond des paysages marins

Titre anglais : Improving the prediction of coastal biodiversity in the Mediterranean Sea using a seascape deep learning approach


Résumé (1600 caractères maximum) :

La mer Méditerranée est l'un des plus grands réservoirs de biodiversité marine au monde, mais c'est aussi l'une des zones les plus menacées. Étant donné la protection limitée dont elle bénéficie actuellement, il devient urgent de développer des stratégies de conservation basées sur une cartographie de sa biodiversité afin de mieux identifier les espèces et habitats vulnérables. Pour y parvenir, nous devons améliorer notre capacité à prédire la biodiversité côtière dans cet environnement hautement hétérogène, dynamique et morcelé. Dans cette étude, nous avons développé une nouvelle approche basée sur un paysage de variables socio-environnementales associées à des algorithmes d'apprentissage profond afin d'améliorer notre capacité prédictive par rapport aux méthodes classiques fondées sur des variables explicatives locales. Des modèles d'apprentissage profond utilisant des réseaux neuronaux convolutifs (CNN) préalablement entraînés sur l'ensemble de la mer Méditerranée ont été développés et testés sur des données de richesse spécifique de poissons provenant d'échantillonnages d'ADN environnemental le long de la côte française. Ces tests de validation croisée spatiale démontrent la supériorité de cette approche par rapport aux méthodes d'apprentissage automatique standard, mais la faible capacité prédictive des modèles met en évidence le besoin de quantités plus importantes de variables normalisées et pertinentes.

Abstract (1600 caractères maximum) :

The Mediterranean Sea is one of the largest reservoirs of marine biodiversity in the world, but also one of the most threatened areas. Given the limited protection it currently benefits, it becomes urgent to develop conservation strategies based on a mapping of its biodiversity to better identify vulnerable species and habitats. To achieve this, we need to improve our ability to predict coastal biodiversity in this highly heterogeneous, dynamic and patchy environment. In this study, we developed a new approach based on the seascape of socio-environmental variables coupled with deep learning algorithms to improve our predictive capacity compared to conventional methods based on local explanatory variables. Deep learning models using convolutional neural networks (CNN) pre-trained over the entire Mediterranean Sea were developed and tested on fish species richness data from environmental DNA sampling along the French coast. These spatial cross-validation tests demonstrate the superiority of this approach against standard machine learning methods, but the globally poor predictive capacity of the models highlights the need for larger amounts of standardized and relevant variables.

Mots-clés : Modélisation, Écosystème, Deep Learning, ADN environnemental, Biodiversité

Key Words: Modelization, Ecosystem, Deep Learning, Environmental DNA, Biodiversity

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*