Combining AI and imaging to predict plankton and particulate matter at global scale in the ocean

Description of the task:

The objective of this task is to provide gridded, gap-filled, global maps of ecosystem biological and biogeochemical properties, using the now available wealth of **imaging** data of plankton and particles (e.g. marine snow), regressed on global fields of other variables (provided by satellite imagery, climatologies, possibly models, etc.). The **target variables** will be plankton properties (concentrations, traits, etc.) derived from quantitative imaging data. The references below highlight how we have explored these approaches using classic machine learning tools and subsets of data in the past. The goal here is to shift gears and exploit larger datasets with modern AI approaches, in an operational context, **to be integrated in Digital Twins of the Ocean (DTOs)**.

Practically:

- the **inputs** will be existing **fields** describing the **properties of ocean water** masses (temperature, salinity, oxygen concentration, current speed, vorticity, chlorophyll concentration, etc.), typically those available from the Copernicus Marine Service
- the **outputs** could be:
    - *binary*: taxon **presence**/absence;
    - *continuous*: taxon or particles **concentration**, **biomass**, average **size**, **carbon flux** (derived from images);
    - *multivariate continuous*: taxonomic **composition**, composition of particle types or size classes.

  The different outputs are derived from the same data and can be thought of as **different levels of complexity** in the same problem. The binary case is already well treated so we focus here on the more challenging multivariate case.
- the **task** will be **regression.**

The spatio-temporal **resolution** of the resulting fields could go from climatologies at 1° to daily fields at 4 km; again, this would be considered a gradation in complexity of the same problem. The minimum resolution to go clearly beyond the state of the art would be **monthly-resolved fields at 0.25°**.

Datasets:

The target datasets can be divided into two categories, each managed in a dedicated web-based application:

1) **morphological measurements and taxonomic identification of plankton** (and some particles) in EcoTaxa. Ecotaxa currently hosts 490M individual images, all with comprehensive metadata and sorted according to a reference taxonomy; about 40% of the identifications have been validated by a human operator, the rest are based on machine classifiers. These images were collected from over 300k individual

samples, in over 100k locations. A **third** of the data is **open**, with **60% more accessible** upon request to the data owners. The access policy is being revamped at the time of writing so that the 60% become explicitly available.

2) **inter-calibrated size** (and transparency) measurements **of particles** between ~100µm and a few mm, in EcoPart. The data is already organized in size and depth bins (5m) to provide easy access to size spectra and derive carbon flux and sequestration. The database hosts ~40k profiles, down to an average depth of 1000m; **75%** of them are available **openly**.

Data **coverage** is **worldwide** (Fig 1, Fig 2) with some areas of more intense focus (Atlantic, Mediterranean) and others with fewer data points (Indian, South Pacific). In terms of time, data spans the last 15 years with more intense collection over the last 5 years.

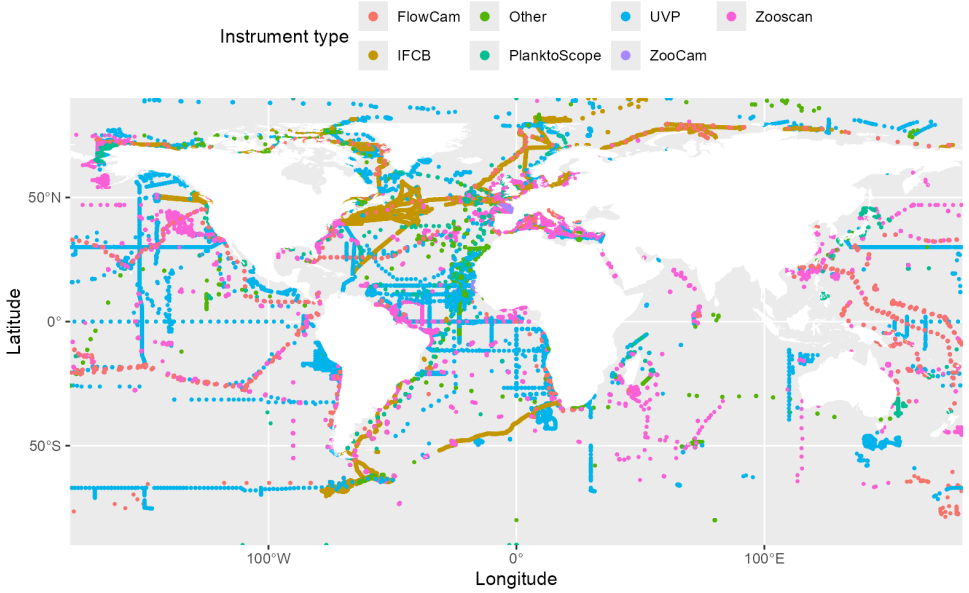

Fig 1: *Map of open (or potentially open within 6 months) data currently present in EcoTaxa (images with taxonomy). The different instruments target organisms of different sizes (IFCB: O10µm; FlowCam, PlanktoScope: O50µm; ZooScan, ZooCam: O500µm; UVP: O1mm).*

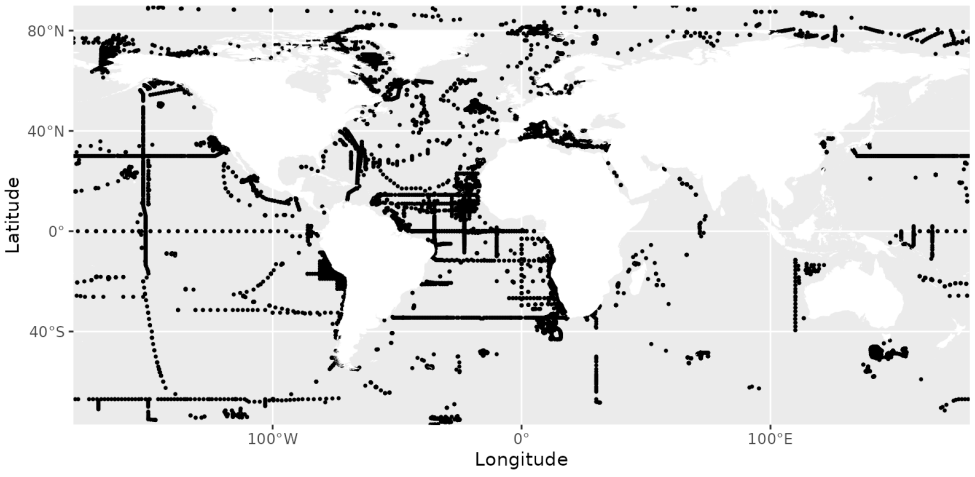

Fig 2: *Map of openly available data in EcoPart (particles size spectra).*

These datasets can be **split** into training, validation, and test sets. Doing so requires care to reduce the effect of **spatio-temporal autocorrelation**; there are community best practices in the matter and we have also explored various techniques to achieve it. An example of common practice is to proceed by **block** cross-validation: input data points are split in geographical blocks, e.g. by ocean basin, and one whole basin is used for evaluation while all the others are used for training.

Evaluation metrics:

Since the objective is regression, the usual metric is the same as the loss: the **mean squared error** between the true value and the prediction, computed here on the validation and test sets. When the response is multivariate, the **multivariate MSE** can be the target metric but is often not ideal when used as a loss when there is correlation among the target variables; this is also a question to be investigated by the future post-doc.

Baselines:

For **phytoplankton** concentration and **broad** community composition, **operational products** (produced without machine learning, from ocean colour sensors) can serve as a baseline. These products are available from the Copernicus Marine Data Store, down to daily and 4 km resolution. For **detailed** phytoplankton taxonomy, **no baseline** exists.

For **zooplankton** biomass, we used gradient boosted trees to extrapolate biomass of a few groups globally, by regressing it on yearly climatologies of environmental variables (Drago et al., 2022). The resulting fields are available together with the code to generate them, on GitHub.
RandomForests were used to upscale *in-situ* **particle** data using monthly climatologies (Clements et al., 2022). The resulting fields are distributed on BCO-DMO.
A similar but more comprehensive approach is being developed as part of the EU BlueCloud2026 project, to provide the community with standardised tools to do such upscaling from monthly climatologies, using a combination of non-AI and non-deep AI tools[1]. This approach should be used in the coming months to generate fields of various plankton-related variables.
These first results demonstrate the potential of such datasets (which have grown since 2022) and of traditional machine learning approaches. But, for now, **no baseline** in the form of a continuously updated, operational product exists for **zooplankton** and **particle** distribution. This is what is needed for Digital Twins,

[1] Schickele et al "Standardizing marine habitat modelling practices to enhance inter-comparability across biological observations" https://www.biorxiv.org/content/10.1101/2024.09.02.610745v1

and current efforts in this direction are also being pushed at EU level(e.g. project DTO-BioFlow).

Relevance of the proposal in terms of AI-native solutions for the DTO:

While there are significant advances in the use of AI for the physics of the ocean, using state of the art **AI** approaches based on **biological** data is still at its **infancy**. Yet, as shown above, some biological data, related to plankton and particles, is collected at scale and in a consistent manner. Current **biogeochemical models** represent these biological compartments very **coarsely** and somewhat **inaccurately**. As the "baselines" section above highlights, some **AI-based solutions** are emerging and they are likely the **best avenue forward** to derive global scale products and understanding. These AI-based products will likely be used in the future to **validate mechanistic models**, rather than considered as (faster) approximations of those models.

Relevance of the proposal in terms of topical demonstrations:

Such products are more relevant than ever to understand the impact of **climate**, the structure of open ocean **ecosystems** and potential consequences on **carbon sequestration**. In addition, products based on these data have the potential to be integrated into the DTO for **decision making** by key stakeholders such as fisheries management, MPA planners, or any structure related to Ocean "Health". This is relevant for **Défi 3** of the PPR (protection et résilience des milieux marins); partly for **Défi 2** (oceans polaires, since some data and predictions are in these regions); and will be fed by and inform the **Défi 6** (observations programs).

Scientific and technical staff involved:

Jean Olivier Irisson (MCF, Sorbonne Université). He is a computational ecologist and uses numerical techniques, including machine learning, to accelerate the processing of biological data and to analyze the resulting large datasets. This allows him to study the distribution of plankton from global scale to submesoscale; he is particularly interested in the interplay between plankton behavior and small scale physical forcing and how it affects global properties of communities. He coordinates the development of EcoTaxa.

Lionel Guidi (DR CNRS - INSU). His research interests range from plankton diversity to the ocean carbon cycle, with a focus on the biological carbon pump linking the two subjects. His research is based on the use of standard biogeochemical approaches as well as the development of innovative imaging and genomic instruments and methods (including machine learning) to study the dynamics of plankton and marine particles at different spatial and temporal scales. He coordinates the development of EcoPart.

Olivier Bernard (DR Inria). He is a specialist of biological system modelling and control, developing approaches in the domain of nonlinear automatic control, in combination with machine learning techniques. He has mainly applied his

developments to planktonic ecosystems (and especially to bacteria and phytoplankton) with metabolic viewpoints, integrating data from artificial ecosystems to the open ocean, to predict diversity and carbon fluxes.

Each member will dedicate **1 month per year** to the project, initially to provide the data, expertise on it, and existing code; later on to assess results.

In addition to these core members, the work of other technical staff will be at least indirectly involved in the DC:
Julie Coustenoble (IE Sorbonne Université, CDI): development of EcoPart
Béatrice Caraveo (IE Sorbonne Université, CDD): development of EcoTaxa
Victoria Bancel (IE Sorbonne Université, CDD): data management
Although non-permanent, the current contracts of these personnel extend at least to the end of 2026.

Data challenge versioning plan:

The following is one **proposition**, among many that could be interesting (see the various outputs in the first section). We would be happy to discuss modifications of this plan if they are deemed necessary.

Year 1

The most readily usable data are the **particle** spectra; this also would connect better with other data challenges which are likely to be focused on physics and biogeochemistry.
The target would be the **composition** of particles by **size** and **type** (defined based on morphology): the absolute concentration in each size bin/type. Those are relevant variables to compute carbon export fluxes.
The **output** would be the predicted **composition at different depths** (e.g. [0-100m], [100-500m], [500-1000m]) of samples in the **test** set; evaluation will be done on this. The model should then be applied to produce **global fields**, with monthly resolution each year; no objective assessment of their quality can be made but their coherence can be evaluated by experts.

To achieve this we would need, from the PPR:
1PM: prepare the train/val/test split and the system to compute quality metrics on the test set.
1.5PM: preparing input fields in easily queryable formats (e.g. zarr) or directly extracting the values around points in the train/val/test sets or writing a generic code to do this
1.5PM: finalising a system for the objective categorisation of particles from their morphology; the code is public and functional but needs to be scaled to 1-2 million input images for "training" and several millions in inference.

Year 2

The second step would be focused on images that provide **plankton taxonomy**. The target would be the absolute **concentration** of a few dominant groups of phyto- and zooplankton in the first 200m of the water column, predicted in isolation or together.
The rest will be **similar to year 1** (extraction of predictors, evaluation on test set, inference of global fields, etc.).

The required resources would therefore be:
0.5PM: use the same method as in year 1 to split train/val/test on new datasets, compute the same/similar metrics.
0.5PM: adjust the preparation of input fields if it appears necessary after year 1.
2PMs: train image classifiers from available training sets (containing 500k to 1M labelled images); we can provide architectures and implementations that are known to perform well. Run the trained classifier(s) on all target images (millions) to get consistent machine-generated labels. Those labels will be used to assess the consistency of the human sorting and quality control the input dataset, or even directly as the substrate for the model predicting concentrations, which is the focus of this challenge.

Links with other proposals:

Within this PPR action, insights could be gained from discussions with **ReeFishBench** due to the biological nature of both proposals and possible commonality of some tools. The proposal oriented towards reanalyses (**iGOR**) could provide input fields to the models. Finally, the ones oriented towards ocean interior physics (**DC10, 12**) would be relevant as input to predict deep particles concentrations fields (although it may not work schedule wise as this is planned in the first year, it could be a further development).

Related scientific references:

(1) Clements, D. J.; Yang, S.; Weber, T.; McDonnell, A. M. P.; Kiko, R.; Stemmann, L.; Bianchi, D. Constraining the Particle Size Distribution of Large Marine Particles in the Global Ocean With *In Situ* Optical Observations and Supervised Learning. *Global Biogeochemical Cycles* **2022**, *36* (5), e2021GB007276. https://doi.org/10.1029/2021GB007276.

(2) Dugenne, M.; Corrales-Ugalde, M.; Luo, J. Y.; Kiko, R.; O'Brien, T. D.; Irisson, J.-O.; Lombard, F.; Stemmann, L.; Stock, C.; Anderson, C. R.; Babin, M.; Bhairy, N.; Bonnet, S.; Carlotti, F.; Cornils, A.; Crockford, E. T.; Daniel, P.; Desnos, C.; Drago, L.; Elineau, A.; Fischer, A.; Grandrémy, N.; Grondin, P.-L.; Guidi, L.; Guieu, C.; Hauss, H.; Hayashi, K.; Huggett, J. A.; Jalabert, L.; Karp-Boss, L.; Kenitz, K. M.; Kudela, R. M.; Lescot, M.; Marec, C.; McDonnell, A.; Mériguet, Z.; Niehoff, B.; Noyon, M.; Panaïotis, T.; Peacock, E.; Picheral, M.; Riquier, E.; Roesler, C.; Romagnan, J.-B.; Sosik, H. M.; Spencer, G.; Taucher, J.; Tilliette, C.; Vilain, M. First Release of the Pelagic Size Structure Database: Global Datasets of Marine Size Spectra Obtained from Plankton Imaging Devices. *Earth Syst. Sci. Data* **2024**, *16* (6), 2971–2999. https://doi.org/10.5194/essd-16-2971-2024.

(3) Drago, L.; Panaïotis, T.; Irisson, J.-O.; Babin, M.; Biard, T.; Carlotti, F.; Coppola, L.; Guidi, L.; Hauss, H.; Karp-Boss, L.; Lombard, F.; McDonnell, A. M. P.; Picheral, M.; Rogge, A.; Waite, A. M.; Stemmann, L.; Kiko, R. Global Distribution of Zooplankton Biomass Estimated by In Situ Imaging and Machine Learning. *Front. Mar. Sci.* **2022**, *9*, 894372. https://doi.org/10.3389/fmars.2022.894372.

(4) Kaneko, H.; Endo, H.; Henry, N.; Berney, C.; Mahé, F.; Poulain, J.; Labadie, K.; Beluche, O.; El Hourany, R.; Tara Oceans Coordinators; Acinas, S. G.; Babin, M.; Bork, P.; Bowler, C.; Cochrane, G.; De Vargas, C.; Gorsky, G.; Guidi, L.; Grimsley, N.; Hingamp, P.; Iudicone, D.; Jaillon, O.; Kandels, S.; Karsenti, E.; Not, F.; Poulton, N.; Pesant, S.; Sardet, C.; Speich, S.; Stemmann, L.; Sullivan, M. B.; Sunagawa, S.; Chaffron, S.; Wincker, P.; Nakamura, R.; Karp-Boss, L.; Boss, E.; Bowler, C.; De Vargas, C.; Tomii, K.; Ogata, H. Predicting Global Distributions of Eukaryotic Plankton Communities from Satellite Data. *ISME COMMUN.* **2023**, *3* (1), 101. https://doi.org/10.1038/s43705-023-00308-7.

(5) Kiko, R.; Picheral, M.; Antoine, D.; Babin, M.; Berline, L.; Biard, T.; Boss, E.; Brandt, P.; Carlotti, F.; Christiansen, S.; Coppola, L.; de la Cruz, L.; Diamond-Riquier, E.; Durrieu de Madron, X.; Elineau, A.; Gorsky, G.; Guidi, L.; Hauss, H.; Irisson, J.-O.; Karp-Boss, L.; Karstensen, J.; Kim, D.; Lekanoff, R. M.; Lombard, F.; Lopes, R. M.; Marec, C.; McDonnell, A. M. P.; Niemeyer, D.; Noyon, M.; O'Daly, S. H.; Ohman, M.; Pretty, J. L.; Rogge, A.; Searson, S.; Shibata, M.; Tanaka, Y.; Tanhua, T.; Taucher, J.; Trudnowska, E.; Turner, J. S.; Waite, A.; Stemmann, L. A Global Marine Particle Size Distribution Dataset Obtained with the Underwater Vision Profiler 5. *Earth Syst. Sci. Data Discuss.* **2022**. https://doi.org/10.5194/essd-2022-51.

(6) Panaïotis, T.; Babin, M.; Biard, T.; Carlotti, F.; Coppola, L.; Guidi, L.; Hauss, H.; Karp-Boss, L.; Kiko, R.; Lombard, F.; McDonnell, A. M. P.; Picheral, M.; Rogge, A.; Waite, A. M.; Stemmann, L.; Irisson, J. Three Major Mesoplanktonic Communities Resolved by in Situ Imaging in the Upper 500 m of the Global Ocean. *Global Ecol Biogeogr* **2023**, *32* (11), 1991–2005. https://doi.org/10.1111/geb.13741.

(7) Ricour, F.; Guidi, L.; Gehlen, M.; DeVries, T.; Legendre, L. Century-Scale Carbon Sequestration Flux throughout the Ocean by the Biological Pump. *Nat. Geosci.* **2023**, *16* (12), 1105–1113. https://doi.org/10.1038/s41561-023-01318-9.

(8) Schickele, A.; Debeljak, P.; Ayata, S.-D.; Bittner, L.; Pelletier, E.; Guidi, L.; Irisson, J.-O. The Genomic Potential of Photosynthesis in Piconanoplankton Is Functionally Redundant but Taxonomically Structured at a Global Scale. *Sci. Adv.* **2024**, *10* (33), eadl0534. https://doi.org/10.1126/sciadv.adl0534.