Automatic recognition of flow cytometric phytoplankton functional groups using Convolutional Neural Networks

Robin Fuchs[a,b], Melilotus Thyssen[b,1], Véronique Creach[c], Mathilde Dugenne[d], Marie Latimier[e], Arnaud Louchart[f,g], Pierre Marrec[h], Machteld Rijkeboer[i], Gérald Grégori[b], Denys Pommeret[a,j,k,l]

[a]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France; [b]Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France; [c]Cefas, Pakefield Road, NR33 0HT Lowestoft, Suffolk, UK; [d]Department of Oceanography, University of Hawai'i at Mānoa, Honolulu, Hawai'i, USA; [e]IFREMER, DYNECO PELAGOS, F-29280 Plouzane, France; [f]Department of Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Napoli, Italy; [g]IFREMER, Laboratoire Environnement & Ressources, F-62321 Boulogne sur mer, France; [h]Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island, USA; [i]Laboratory for Hydrobiological Analysis, Rijkswaterstaat (RWS), Zuiderwagenplein 2, 8224 AD Lelystad, The Netherlands, [j]Université Claude Bernard Lyon 1, 43 boulevard du 11 Novembre 1918 69622 Villeurbanne cedex, France; [k]ISFA, 50 Avenue Tony Garnier, 69007 Lyon, France; [l]Laboratoire de Sciences Actuarielle et Financière (SAF) EA2429, Lyon France.

[1] Corresponding author: melilotus.thyssen@mio.osupytheas.fr

Abstract

The high variability of phytoplankton distribution has been unraveled by high frequency measurements. Such a resolution can be approached by automated pulse-shape recording flow cytometry (AFCM) operating at hourly sampling resolution. AFCM records morphological and physiological traits as single-cell optical pulse shapes that can be used to classify cells into Phytoplankton Functional Groups (PFG). However, the associated manual post-processing of the data coupled with the increasing size and number of the datasets is time consuming and carries sources of error. Machine learning models are now increasingly used to run automatic classification. Yet, most of the existing methods either present a long training process, need to manually design some features from the raw optical pulse shapes or are dedicated to images only. In this study, we present a Convolutional Neural Network (CNN) to classify PFGs resolved by flow cytometry using the pulse shapes collected by AFCM. The uncertainties of manual classification were first estimated by comparing experts ~~manual gatings on~~ ~~Redpicopro, Orgpicopro, Redpicoeuk, Rednano, Orgnano, Redmicro and Orgmicro phytoplankton cells.~~ Consensual particles in individual PFG were used to train and validate the CNN. The CNN obtained competitive performances compared to ~~the~~ models used in the literature, ~~and presented significant generalization power concerning the sampling area, the AFCM hardware and settings.~~ ~~Finally,~~ we assessed the ability of this classifier to predict phytoplankton counts at a Mediterranean coastal station and from a cruise in the South-West Indian Ocean, providing further comparison with the manual classification of an expert over three months long periods.

Keywords— phytoplankton | pulse-shape recording flow cytometry | automatic classification | deep learning | high frequency

Introduction

Phytoplankton cells are major actors in marine environments and in biogeochemical cycles. The amount of seawater dissolved CO_2 absorbed by phytoplankton cells per unit of time, called primary production, is estimated to be equivalent to all of the primary terrestrial production. This is the case even if they represent less than 1% of the total autotrophic biomass (Field et al. 1998), suggesting a rapid growth capacity and high turnover rates (Fowler et al. 2020). Currently, models estimating primary production in the ocean present a wide uncertainty range (Carr et al. 2006; Saba et al. 2011; Buitenhuis et al. 2012), mainly due to ~~a lack of~~ resolution of the datasets collected (Lévy et al. 2012). Indeed, the heterogeneous distributions of phytoplankton combined with a high structural and functional diversity highlight the need for infra kilometer spatial resolution and infra hour temporal resolution (Kavanaugh et al. 2016).

Phytoplankton biomass and distribution are listed as Essential Ocean Variables (EOV) (Miloslavich et al. 2018), but datasets with resolution inferior to 10km are scarce. Automated pulse-shape recording flow cytometry (AFCM) (Dubelaar et al. 1999; Dubelaar and Gerritzen 2000) enables vast automated data acquisition with hourly sampling strategies on several ~~important size and pigment-related~~ phytoplankton groups. AFCM is now involved in numerous oceanographic field studies and benefits from the growing scientific interest for automated single cell approaches (Boss et al. 2020) in monitoring programs. A dedicated vocabulary with its definition has been published by a wide group of experts to describe the most common groups observed by flow cytometry in natural seawaters, and this nomenclature will be used in this manuscript (`http://vocab.nerc.ac.uk/collection/F02/current/`).

Phytoplankton cells are detected using the emission of fluorescence due to the excitation of chlorophyll (red fluorescence) and accessory pigments (orange fluorescence of phycoerythrin, for instance). AFCM generates a set of pulse shapes or flow cytometric curves (FCCs) which represents the optical profiles of scatter and fluorescences emitted by each particle (cell) when crossing the ~~488 *nm*~~ laser beam. Scatter signals collected at small and large angles (forward scatter (FWS) and sideward scatter (SWS) respectively) are related to the cell size and structure (granularity), while red (FLR) and yellow-orange fluorescence (FLO or FLY) signals are reflecting the ~~pigment nature and content~~ of the cells. From the difference between left angled and right angled FWS pulses, a fifth signal named Curvature is extracted. Instruments can process up to 10 000 cells per second thanks to a frequency acquisition of 4 MHz, with sampled volume up to 5 *mL* routinely. After data collection, AFCM users generally manually ~~gather~~ cells sharing similar optical fingerprints into groups using multiple sets of two dimensional projections (cytograms). Groups recognition and identification are based on seminal papers (Olson et al. 1985; Chisholm et al. 1988; Green et al. 1996; Jacquet et al. 2002; Metfies et al. 2010; Ribeiro et al. 2016; Hamilton et al. 2017; van den Engh et al. 2017; Marrec et al. 2018) ~~describing Redpicopro, Orgpicopro, Redpicoeuk, Rednano, Orgnano characteristics.~~ In addition to these groups, Redmicro and Orgmicro cells can be counted by AFCM and identified to a coarse taxonomic level (typically up to the genus) using recent integration of image-in-flow devices (Dugenne et al. 2014). These size and pigment-related groups belong to several phytoplankton functional groups (PFG), since they fit the initial definition of sets of species sharing similar ecological and biogeochemical functionalities (Le Quere et al. 2005), and will hereafter be identified as cytometric PFG (cPFG).

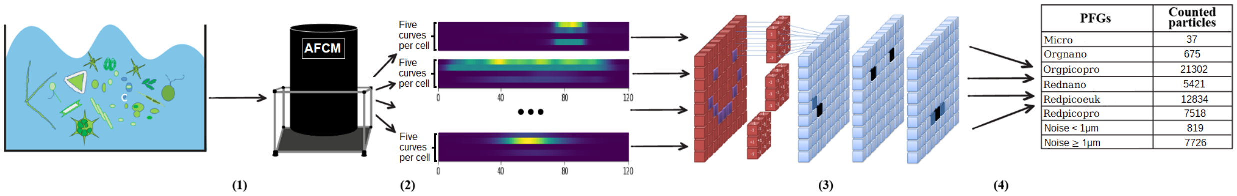
Manual gating is often both time-consuming and error-prone, as it relies on 2D projections and interpretations of simplified descriptors of the complex raw optical profiles (such as pulse maximum height, area under the curve, pulse width) by individual AFCM experts. The spread of this technology will generate datasets too large to be manually processed, constraining the collection of valuable high frequency cPFGs datasets. In order to facilitate the work of an increasing number of AFCM users and decrease the uncertainties linked to manual gating, the classification of cPFGs has to be semi- or fully automated. The automation can be achieved using supervised machine learning methods that assign a label to an observation based on its characteristics, a task named classification.

In the case of phytoplankton, automatic classification generally relies on image processing and computer vision. One can for example cite the count of coccoliths using shallow Neural Networks (Beaufort and Dollfus 2004) or more recent works based on Residual Neural Networks and transfer learning (Yosinski et al. 2014) in order to classify images from diverse laboratory cultures and *in situ* monitoring (Dunker 2019; González et al. 2019). However, cameras resolution is relatively low for the identification of pico-nanophytoplankton size classes, which ~~moreover~~ show limited morphological diversity. As such, ~~using the~~ FCCs offers ~~an~~ alternative since it deals also with these small particles. A second main advantage in working on the automatic classification of optical profiles is the shorter training process due to the absence of transfer learning (Pan and Yang 2009) required to fine-tune heavy Neural Networks like Residual Networks (He et al. 2016) for image recognition.

Automatic recognition of cPFGs from the FCCs has received less attention than image-based identification and can be gathered in two main types of approaches. The first family of approaches applies machine learning methods on ~~features~~ computed on the FCCs (for example the mean, the area under the curve, or the length of each FCC). Boddy et al. (1994) started to use neural methods to classify cells ~~into species~~. Wacquet et al. (2013) developed new statistical methods ~~to deal with the features~~ of the FCCs and implemented them along with existing statistical methods in the R package RclusTool. Thomas et al. (2018) and Schmidt et al. (2020) used Random Forests to respectively discriminate between phytoplankton cells of different populations and between phytoplankton and non-phytoplankton particles. Abdelaal et al. (2019) used Linear Discriminant Analysis (LDA) and present performances outperforming Deep Learning approaches.

The second family of approaches relies on the entire FCC signal to perform classification. ~~This is the case of~~ Malkassian et al. (2011) ~~that~~ plunged the FCCs into a Fourier basis and calculated distances to discriminate between populations. (del Barrio et al. 2019) created curves templates to classify AFCM non-marine cells using Wasserstein distance and optimal transport. Finally, (Caillault et al. 2009) relied on the Elastic Matching coupled with standard classifiers. ~~We believe that this second family of approaches can take advantage of the whole signal rather than using some hand-designed descriptors chosen by the user. As a result, our method belongs to this second class of approaches.~~

In this article, we applied for the first time Convolutional Neural Networks (CNN) on pulse shapes recorded by AFCM to automate cPFGs classification as described in Figure 1. CNN have known a fast development in image recognition and computer vision during the last ten years, starting with the seminal work of Krizhevsky et al. (2012). Once in-



PFGs	Counted particles
Micro	37
Orgnano	675
Orgpicopro	21302
Rednano	5421
Redpicoeuk	12834
Redpicopro	7518
Noise < 1µm	819
Noise ≥ 1µm	7726

Figure 1: Explanatory scheme of the predictive pipeline. (1) Particles are sampled from seawater by AFCM. (2) The five flow cytometric curves (FCCs = SWS, FWS, FLR, FLO, Curvature) generated for each particle as they cross a laser beam are interpolated to a fixed length and stacked together into matrices. (3) The CNN predicts the class of each particle using Convolutional layers (red) and Dense layers (blue). (4) The number of particles per group (phytoplankton or background noise) is computed and returned.

terpolated and stacked together as matrices, the FCCs are analogous to images and can be used to train a CNN, rather than computing features on the FCCs. We show the generalization power of the method on two instruments with datasets collected in the South-West Indian and Southern oceans and in the coastal and open Mediterranean sea.

As CNNs rely on robust datasets, individual experts were asked to manually assign a cPFG to particles from samples collected in the different datasets collected. We assessed the heterogeneity between experts classifications and built consensual datasets to evaluate automatic classification models. The performances of four benchmark automatic classification models along with the CNN were compared. Finally, the trained CNN was used to generate predictions spanning three months sampling in a coastal station of the Mediterranean Sea and two months in the South-West Indian Ocean, both at a two hours sampling frequency. The robustness and extremely fast process of the CNN applied open the way to real time cPFG analysis.

Material and procedures

Data collection

Data Origin

In situ AFCM datasets were collected at the SeaWater Sensing Laboratory At MIO Marseille (SSLAMM data), France, a coastal ~~marine~~ Mediterranean station, between September 2019 and December 2019, and onboard the research vessel Marion Dufresne II, from 11 January to 8 March 2021, in the frame of the MAP-IO project (University of la Reunion) during the GEOSCAPE SWINGS cruise (SWINGS data). Two distinct CytoSense flow cytometer (Cytobuoy b.v.), here after identified as SSLAMM-AFCM, and MAP-IO-AFCM were deployed.
For both datasets, seawater was continuously pumped *in situ* and the flow cytometers ran automated acquisitions scheduled every two hours. The SSLAMM coastal seawaters was gently pumped with a VerderFlex40 peristaltic pump at 10 meters away from the coast at a depth of 3 meters, and was delivered unaltered into the laboratory where analyses were conducted. Onboard the Marion Dufresne II, the seawater was collected from the underway clean seawater supply pumped at 7 m depth~~s~~, using a centrifugal pump.

Automated pulse-shape recording flow cytometry

~~The two automated CytoSense flow cytometers (Cytobuoy b.v.) run similarly in both conditions and sampled semi-continuously seawater from the flow-through seawater arrival. The CytoSenses pumped samples from a dedicated external chamber of 200 *ml*. The volume analyzed for each sample was estimated using a calibrated peristaltic pump.~~ Before entering the flow cell, the sample was surrounded by a 0.1 μm filtered seawater sheath fluid and the generated laminar flow aligned each particle prior to cross a 488 nm laser beam (Coherent, 120 mW). Both instruments recorded the optical pulse shapes emitted resulting in forward scatter (FWS), sideward scatter (SWS), and two fluorescences. The SSLAMM-AFCM collected wavebands of > 652 nm (red fluorescence, FLR) and between $552 - 652$ nm (orange fluorescence, FLO). The MAP-IO-AFCM collected wavebands between $668 - 726 nm$ (FLR) and $516 - 650 nm$ (yellow fluorescence, FLY). Particles were recorded in the size range $< 1 - 800$ μm in width and up to a few mm in length for chain forming cells. These optical profiles take the form of a set of curves hereafter called flow cytometric curves (FCC).

Laser scattering at frontal angles (FWS) was collected by two distinct photodiodes to check for the sample core alignment. Difference between left and right photodiodes signatures generates the Curvature curve. To follow the stability of the flow cytometers, 2.0 μm fluorescing polystyrene beads (Polyscience ®) were regularly analyzed. Silica beads (1.01 μm, 2.56 μm, 3.13 μm, 5.02 μm, 7.27 μm in diameter, Bangs Laboratory®) were also used ~~for size retrieving estimates from~~ FWS signals.

Because of the current memory and computation limitations, optimally sampling the entire size range of the phytoplankton community in natural marine waters require some compromises: to collect small cells ~~such as Orgpicopro and Redpicopro cells~~, the AFCM settings were set on high sensitivity (red fluorescence trigger threshold set on 6 mV (FLR6) for SSLAMM-AFCM and on 5 mV (FLR5) for MAP-IO-AFCM). As a result, the sample was filled by a majority of small and/or dimly fluorescent particles and electronical background noise, hereafter simply called noise. ~~Since the smallest phytoplankton cells are the most abundant in natural samples, they were correctly counted in small volumes between 0.5 ml and 1 ml.~~

Extra white space

In order to collect the largest but less concentrated cells, a second protocol was applied with a red fluorescence trigger threshold (high trigger level) set up to 25 mV (FLR25) for SSLAMM-AFCM, and to 20 mV (FLR20) for MAP-IO-AFCM and a volume analyzed reaching 5 ml. ~~Doing so,~~ the small particles and background noise generating acquisition limitations were not recorded ~~anymore.~~ ~~Except that they use~~ two different thresholds, the two protocols (FLR5/FLR6 and FLR20/FLR25) used the same AFCM settings (same sample pump speed, similar filter mesh sizes, same optical chamber, similar sampling frequency).

~~Finally, the total number of Orgpicopro and Redpicopro cells was computed from the FLR5/FLR6 files and the total number of Orgnano, Redpicoeuk, Rednano and micro cells was computed from the corresponding FLR20/FLR25 files.~~ Raw datafiles were manually gated by experts using the CytoClus4© software (Cytobuoy b.v.).

Manual gating methodology and heterogeneity estimation

The raw data collected by the AFCM are composed of series of five curves exhibiting variable heights, areas and lengths. Experts use a dedicated software, CytoClus4©, to summarize this signal by computing a single value for each curve, typically the area under the curve or the maximal value of the curve. ~~Doing so, one~~ obtains a point of dimension five for each observation and the dataset can be represented by a series of 2D projections. For example, ~~one can~~ plot the Total FLR (the area under the FLR curve) against the Total FLO/FLY (the area under the FLO/FLY curve) to ~~separate Orgpicopro and Orgnano from red only fluorescing particles~~. Total FLR vs Total FWS are commonly used to separate ~~Redpicoeuk, Rednano and Micro~~ size classes, ~~while Total FLR vs Total SWS (or Maximal height of SWS) can help in gating the Redpicopro group.~~

~~Phytoplankton abundance heterogeneity between~~ cPFGs generates imbalanced AFCM datasets. The ratio between the most and the ~~less represented~~ class in our data initially ranged between 10^4 and 10^5. Thus, on the 2D scatter plots used by ~~cytometrists~~ to identify the cPFGs, the less represented particles can be difficult to separate when their distribution overlaps other groups with higher abundances. Furthermore, dealing with large datasets require long periods of assiduity when running manual classification and visual control of groups boundaries, creating frequent errors as these steps are tedious.

This can generate significant biases in the estimated count of some classes. For instance, in the SSLAMM dataset, few dozen of Micro cells are typically observed in a sample, while dozen of thousands of Orgpicopro particles are present. Hence, misclassifying 10 particles of Micro could result in a 30% error rate while misclassifying 10 particles of Orgpicopro would be negligible.

This issue is a type of statistical data contamination and may have significant effects on the patterns learnt by machine learning algorithms. Without any estimation of this contamination, it is impossible to disentangle the errors coming from the data from the error coming from the training process. Furthermore, estimating the variability of functional groups counts is essential to be sure that results coming from different studies are comparable.

The heterogeneity ~~was estimated on classifications performed by a panel of~~ six AFCM experts who were asked to classify SSLAMM and SWINGS data coming from six and twenty acquisitions respectively, acquired at different seasons, geographical zones and times of the day. The list of the cPFGs ~~was given~~, along with two acquisitions of 2.0 μm polystyrene (Polyscience ®) and 3.13 μm silica beads (Bangs Laboratory ®).

The heterogeneity was measured by computing Adjusted Rand Indices (ARI) Steinley (2004) and coefficients of variation (CVs). The ARIs gave an indication about the similarity between two experts overall classifications. The closest the ARI is to 1, the more similar the classifications between two experts are. The ARI have been computed for all pairs of experts and for all files.

~~On the other hand,~~ the coefficient of variation of each cPFG is computed as the standard error divided by the mean of the expert counts for that cPFG. The closest it is to zero, the more the experts agreed on the count of the given cPFG. To summarize, the ARIs assessed the overall agreement between experts' classifications whereas the CVs ~~gave this piece of information~~ at the cPFG level.

Consensual particles, defined as particles for which 2/3 of the experts assigned the same label, were kept to train and evaluate the ~~statistical models~~.

Beyond the initial training samples, one of the experts has manually gated three months of data from the SSLAMM station (from mid-September 2019 to mid-December 2019) and the entire data set from the MAP-IO-SWINGS cruise. The classification obtained from the CNN was then compared with the manual gating.

Data presentation and processing

The datasets composition were fixed to six phytoplankton functional groups determined by their flow cytometry optical properties as they represent the most common groups observed in marine samples . They were identified using the flow cytometry consensual nomenclature (`http://vocab.nerc.ac.uk/collection/F02/current/`): Redpicopro, Orgpicopro, Redpicoeuk, Rednano, Orgnano, Redmicro, Orgmicro. There were however not enough Redmicro and Orgmicro cells *in situ* to distinguish between these two groups and they are treated together in the sequel under the name "Micro" cells.

In addition to these six phytoplankton functional groups, the datasets contained non-phytoplankton particles thereafter called noise particles or events. Noise events were heterogeneous and have been subdivided into $< 1\ \mu m$ and $\geq 1\ \mu m$ groups using silica beads as a size reference (figure 5 in supplementary material). $\geq 1\ \mu m$ noise mainly contained large detrital particles or predators such as ciliates or flagellates cells that have ingested some phytoplankton cells. Conversely, $< 1\ \mu m$ noise often contained optical noise from the sensors, non-fluorescing heterotrophic prokaryotes or decomposing cells.

Due to the acquisition limitations of the two cytometers and because they present dim fluorescence in surface waters, the Redpicopro are hard to distinguish from $< 1\ \mu m$ noise events and a curve shape criterion was used to distinguish between them. Indeed, Redpicopro cells are likely to be spherical cells, and their SWS signal are expected to look as bell curves, whereas $< 1\ \mu m$ noise events can present a significant variety of shapes. Therefore among the consensual Redpicopro cells, only the bell-curved SWS cells were kept in the training, validation and test sets of the CNN.

In order to reach a ~~substantial total dataset size and to reduce the imbalance~~ between groups which ~~disturbs~~ the training process, the over-represented groups were undersampled in the training set~~.~~ Even after undersampling, the relative number of Micro cells in the SSLAMM data remained too low in comparison to the other groups of the training set. Hence, three out of the six FLR25 files were artificially enriched with Micro particles from the FUMSECK campaign (DOI 10.17600/18001155) as if they were part of the original dataset. These FUMSECK Micro cells were collected in the open Mediterranean Sea using the same cytometer with the same settings only four months before the first SSLAMM data acquisition. These additional particles were given for classification to the experts and only the cells identified as Micro cells were kept. The potential batch effect introduced is hence assumed to be negligible.

Before undersampling, the number of particles of the most represented group in the training set was 45 times higher than the less represented one. After undersampling, it was only eight times higher at most for the two datasets.

Conversely, the validation and test sets were not rebalanced. The total size of the training, validation, and test sets were of 33 791, 50 682 and 134 313 particles for the SSLAMM data, and of 57 241, 365 863 and 224 426 particles for the SWINGS data. Table 3 in Supplementary Information describes the number of particles of each group in the training, validation, and test sets.

The length of each AFCM curve is closely linked to the size of the particle (the bigger the particle the longer the sequence). In order to train the CNN, which needs a fixed data format for all observations, the curves have been all set to a fixed length of 120 values interpolated using quadratic interpolation (see Figure 2 in Supplementary Information for an illustration). ~~A length of 120 has been chosen since it corresponds to the third quartile of the curves sizes distribution in our data and as intuitively less information is destroyed when small curves are interpolated to be bigger than the reverse.~~ As the curves were not truncated and the profile shapes were preserved, the choice of this length is not expected to be of prime-importance regarding the performance of the model.

Prediction pipeline presentation

The core of the predictive pipeline is a Convolutional Neural Network initially designed for image recognition. The general idea of such a network is to learn a series of filters that detect some patterns in images and help to discriminate between the classes. More formally, these filters are tables of coefficients iteratively used to compute convolutional operations on the data going through the layers. Compared to Dense layers, the Convolutional ones rely on the assumption that regions in the images convey useful information and that close pixels often carry redundant information. As a result, the total number of parameters of the model is reduced and the training of the model is kept tractable. The Convolutional layers automatically extract features from the signal, which are then used by Dense layers at the end of the network to perform the classification itself. As both images and AFCM data can be represented as tables of coefficients, the same Convolutional Neural Networks can be used to treat both data types with minor adjustments.

~~The CNNs can deal directly with the five FCCs. On the contrary, cytometrists and the machine learning models of the first family of approaches presented above require to compute features on this signal before performing gating.~~ Hence, we expected that the CNN could take advantage of this raw and more complete signal. The CNN architecture is presented in Supplementary Information (see figure 4). The architecture was inspired by the VGG architecture (Simonyan and Zisserman 2014). Features are first extracted by three blocks of convolutional layers separated by "local" average pooling layers in order to reduce the redundant parts of the signal and to automatically design features useful for the classification. These convolutional features are then pooled together using a global average pooling layer so that they can be treated by two dense layers. At the end of the dense layers, a softmax activation function is computing the probabilities that an observation belongs to each class and the loss of the model is computed.

The loss is measuring the gap existing between the class probabilities outputted by the model and the actual class of the observation. This gap represents an error, back-propagated to update the parameters of the network accordingly. white space not needed The negative-likelihood also called the categorical cross-entropy is the most widely used loss for single-label multivariate classification (each observation belongs to one class only) and is the one used here. More refined versions of the categorical cross-entropy such as the weighted version of the categorical cross-entropy, the Focal Loss (FL) (Lin et al. 2017), or the Focal Class-Balanced loss (FCBL) (Cui et al. 2019) have been implemented but brought no additional performances.

Beyond the choice of the loss weights, the significant imbalanceness of the data were also dealt with using undersampling methods. Only a random subset (5000 particles) of the most represented class was kept whereas most of the particles of the less represented classes were sampled. This enabled to reduce the gap between cPFGs in order to have both enough instances per class and a tractable total number of observations in the dataset. Yet, as Figure 2 highlights it, the density of points is not uniform in 2D cytograms. Pure random particles sampling tends to let some of the low density areas of 2D cytograms nearly empty, preventing machine learning models to learn which class to predict for particles in these areas. Hence, additional particles were sampled to fill low density areas. The impact of these zones on the confidence of the CNN cPFG predictions can for instance be seen on figure 6 in Supplementary Information.

Beyond the choice of the loss specification, another important choice is the one of the optimizer which deals with how the network parameters are updated with respect to the loss. We have benchmarked two optimizers: Adam and its extension Ranger. Ranger comes from the combination of two recent publications: RectifiedAdam (or Radam) (Liu et al. 2019) and Lookahead (Zhang et al. 2019).

In order to avoid ~~being stucked in bad~~ local ~~maxima~~, it is a common practice to slowly update the parameters of the models at the beginning of the training, when ~~really~~ promising parameters ~~regions~~ are not identified ~~at the moment~~. This adaptation rate of the parameters with respect to the loss is called the learning rate of the model and is hence often chosen to be small in the early stages of the training process (Popel and Bojar 2018). Radam adapts the learning rate to avoid the learning rate variance to grow too substantially, which is often detrimental to the learning process, according to the authors. On the other hand, Lookahead enables the network to get a better understanding of the loss topology. In order to do so, two sets of weights are used by Lookahead: a faster set of weights that is frequently updated to "explore" the loss surface and a slower set of weights (less frequently updated) to ensure the stability of the learning process. The faster set of weights is updated using not all the data but only a set of several observations batches to get a raw idea of the promising regions to explore. In the Ranger case, these fast weights are updated thanks to the Radam optimizer. ~~It appeared that the Ranger optimizer gave best results than Adam in our case and was therefore preferred in our experiments.~~

The loss, the behaviour of the optimizer and more generally most parts of statistical models are ruled by a set of hyper-parameters chosen by the user. The number of possible combinations is far too high for all the combinations to be tested and then to select the best network specification.

One popular approach relies on Bayesian Hyperoptimisation algorithms (Bergstra et al. 2013), implemented in our case in the Python library Hyperas (Hyperopt for Keras). The idea of Hyperoptimisation methods is to consider hyperparameters as statistical random variables with a prior and to identify posterior regions that present a low loss value. Hence, some draws are taken from the prior distributions, the model is evaluated and low loss regions are identified and focused on. It avoids spending very significant computational efforts on non-promising regions of the hyper-parameters space as it is often the case using standard line search. The hyperparameters spaces used are given in section 1 in Supplementary Information.

Comparison with other classification algorithms

The CNN has been benchmarked against other supervised models in order to ~~illustrate~~ its performance. The benchmark models ~~were models used in the literature mentioned earlier:~~ the k-Nearest Neighbors (kNN) and the Linear Discriminant Analysis (LDA). Tree-based methods such as Random Forest were represented by LGBM which is more recent and takes advantage of gradient-boosting methods.

The data from the ~~inter-gating experiment~~ were used for models evaluation. Once interpolated to a fixed length, the CNN was trained over the five FCCs per particle, while the benchmark models (which ~~cannot deal with the raw curves)~~ were trained on the hand-designed features computed ~~on~~ these FCCs. ~~The list of the features used is given in section 2 of Supplementary Information.~~ The data used to train the models have been randomly separated into a training, a validation and a test set. The models learn how to distinguish between cPFGs on the training set. Once trained, the cPFGs of the validation set are predicted and the hyperparameter optimisation procedure selects the best performing specification of each model on that set. Finally, the best specification of all models are compared on the test set. The benchmark models were trained on features computed over the raw FCCs. The choice of the features created from the signal highly influence the performances of the models and has to be considered when presenting the results. We rely on the ten features per curve created by default by the CytoClus4© software. The feature list is given in Supplementary Information (see section 2).

The performances of the CNN and of the benchmark models were evaluated using the standard per-class precision and recall metrics. The precision is the proportion of particles actually belonging to class k among all those identified as belonging to class k by the algorithm. The recall is the proportion of particles effectively belonging to class k among all the particles of class k existing in the dataset. The closer are both precision and recall to 100%, the closer the classification of a model is to the “true” labels.

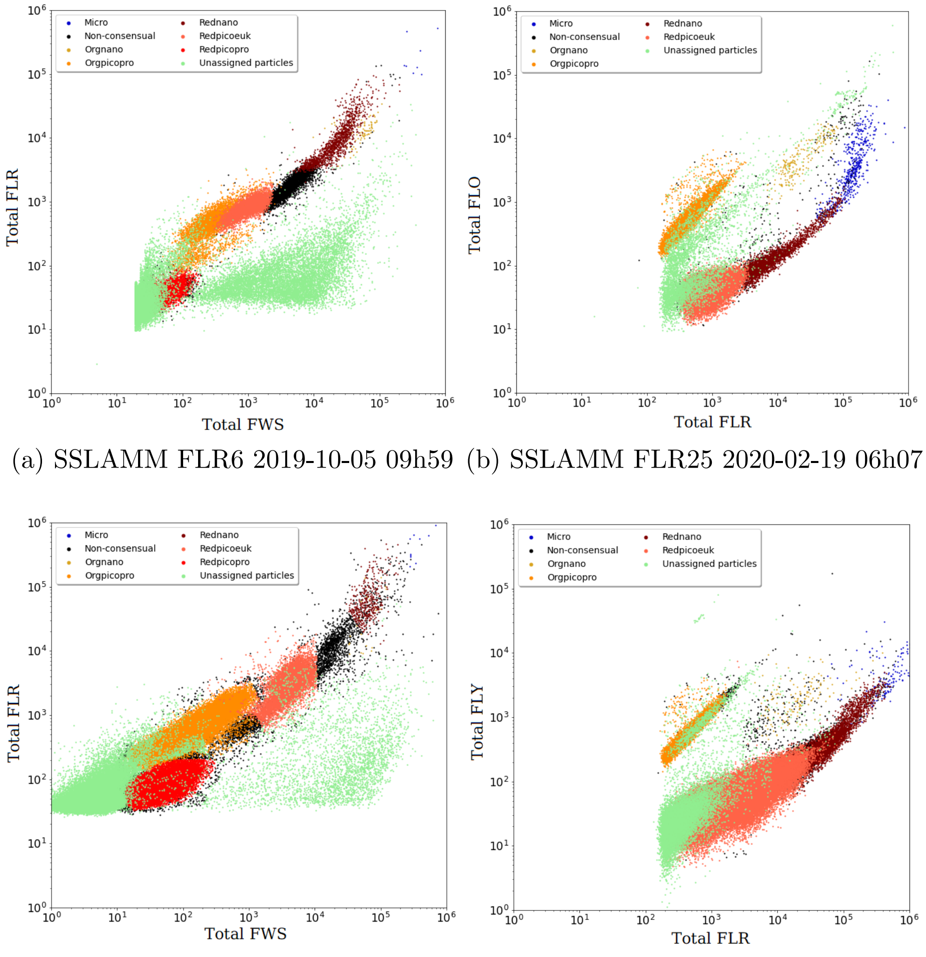
The Python code used to produce the results of this work is available as a Github repository named phyto_curves_reco.

Results

Manual gating uncertainty estimation

The main groups observed by AFCM are represented on Figure 2. It presents descriptive 2D cytograms associated with two files for each data source. The 2D cytograms are the main tools used for manual gating and evidence here the disparities existing between experts. The non-consensual particles - on which less than 2/3 of the experts agreed - were located mainly at the frontiers between groups. The less consensual demarcation lines were between Rednano and Redpicoeuk and between Redpicopro and the background noise events.

The uncertainties of manual classification for individual cPFGs are reported in Supplementary Information (Figure 1 and 2). The patterns observed in terms of ARIs and CVs were similar between SSLAMM and SWINGS data. For both data sources, 75% of the pairwise ARIs were higher than 0.78, which underlined that the experts shared a com-



(a) SSLAMM FLR6 2019-10-05 09h59 (b) SSLAMM FLR25 2020-02-19 06h07

(c) SWINGS FLR5 2021-03-02 18h11 (d) SWINGS FLR20 2021-01-30 03h59

Figure 2: 2D cytograms showing the particles contained in two files from the SSLAMM data (a and b) and two files from the SWINGS data (c and d). Cytograms (a) and (c) present the Total Red Fluorescence (a.u., Total FLR) as a function of the Total Forward Scatter (a.u., Total FWS) and cytograms (b) and (d) show the Total Orange/Yellow Fluorescence (a.u., Total FLO, Total FLY) as a function of the Total Red Fluorescence (a.u., Total FLR). Total refers to the area under the curve of the optical variable. Each dot represents a particle. A particle is considered as consensual if 2/3 of the experts have voted for the same cPFG for this particle. Non-consensual particles are represented in black.

mon way to perform the overall classification. However, these high ARI were driven by several over-represented cPFGs which were also well identified. This was the case of Orgpicopro cells that obtained CVs between 0.01 and 0.14 for the SSLAMM data and between 0.04 and 0.50 for the SWINGS data and the case of Redpicoeuk (SSLAMM $CV \in [0.05, 0.50]$ and SWINGS $CV \in [0.10, 0.45]$). Conversely, Micro cells (SSLAMM $CV \in [0.26, 1.60]$ and SWINGS $CV \in [0.20, 1.30]$), Orgnano (SSLAMM $CV \in [0.48, 0.90]$ and SWINGS $CV \in [0.30, 1.70]$), Rednano (SSLAMM $CV \in [0.48, 0.90]$ on and SWINGS $CV \in [0.30, 1.70]$), and Redpicopro (SSLAM $CV \in [0.16, 2.50]$ and SWINGS $CV \in [0.5, 1.20]$) were far less identified.

Model benchmark on SSLAMM data

Tables 1 and 2 report the precision and the recall obtained by the four models for each data class.

Based on the specific precision and recall values, the CNN and the LGBM obtained the best performances on the quasi-totality of cPFGs. The performance spread between the two methods was often inferior to 1%. The kNN presented the worst performances for both datasets. The LDA results are more mixed as it well distinguished noise events from phytoplankton particles classified but got ~~for instance~~ the worst precision on three cPFGs on the SWINGS data.

The cPFGs that were the best identified manually were also the ones that were the best classified by machine learning models. This is the case of Orgpicopro, Redpicoeuk particles. Similarly, the Redpicropro and Orgnano cells were weakly identified manually and were less well gated by machine learning models. ~~On the contrary,~~ Micro and Rednano cells which experienced poor manual identifiability presented good precisions and recalls for near all methods.

The generalization ~~power~~ of the models was also tested by training them on one data source (SSLAMM or SWINGS) and by making predictions on the other data source. Results are given in Tables 4 and 5 in Supplementary Information.
When the models were trained on the SWINGS data, the CNN obtained the best performances, with precisions higher than 90% for five out of the eight classes and kNN the worst performances. Concerning the cPFGs, noise events and Orgpicopro were the best classified and Redpicopro and Micro cells were the less well gated.
When trained on the SSLAMM data and predict on SWINGS data, the LGBM obtains the best performances and LDA the worst. Redpicopro cells and noise events $\geq 1\mu m$ were the worst identified by the models. Rednano cells obtained precisions lower than 34% but recalls higher than 87%. The opposite pattern was observed for the Redpicoeuk class, denoting that a significant number of manually identified Redpicoeuk cells were predicted as Rednano cells by the models.

The running time of the models is given in Supplementary Information.

Prediction of the SSLAMM Time Series

Figure 3 presents the ~~automatically and manually classified~~ time series for all cPFGs counted particles from the SSLAMM files and the SWINGS files. As accurate cPFG predictions imply accurate predictions of the total noise events, the background noise events-

	Precision				Recall			
Model	kNN ~~(prec)~~	LDA ~~(prec)~~	lgbm ~~(prec)~~	~~cnn~~ ~~(prec)~~	kNN ~~(rec)~~	LDA ~~(rec)~~	lgbm ~~(rec)~~	~~cnn~~ ~~(rec)~~
Micro	73.68	96.54	97.13	98.00	72.20	93.95	98.65	98.88
Orgnano	27.80	50.30	89.74	96.59	35.43	94.86	100.00	97.14
Orgpicopro	97.41	98.74	99.91	99.84	76.36	98.97	99.35	99.31
Rednano	79.00	94.18	98.04	97.33	90.78	85.58	99.32	99.08
Redpicoeuk	71.45	83.80	99.02	99.32	83.26	99.45	98.33	97.60
Redpicopro	4.67	28.72	73.73	79.51	54.08	96.65	98.62	95.34
Noise $< 1\mu m$	91.95	99.41	99.97	99.67	85.66	96.11	99.47	99.50
Noise $\geq 1\mu m$	91.06	97.59	97.23	96.22	71.17	78.38	98.22	97.39

Table 1: Precision ~~(prec)~~ and recall ~~(rec)~~ of the benchmarked models on SSLAMM data

kNN: k-nearest ....

Model	kNN (prec)	LDA (prec)	lgbm (prec)	cnn (prec)	kNN (rec)	LDA (rec)	lgbm (rec)	cnn (rec)
Micro	24.20	67.66	95.22	75.26	93.15	93.61	100.00	100.00
Orgnano	10.74	31.68	86.18	96.30	45.38	80.67	89.08	65.55
Orgpicopro	67.93	48.54	99.58	99.24	49.04	90.78	99.30	99.16
Rednano	62.02	83.02	75.56	85.04	82.82	92.58	99.05	96.08
Redpicoeuk	97.19	97.11	99.77	99.65	79.99	91.74	96.93	98.23
Redpicopro	12.04	34.13	98.24	94.53	53.75	65.70	95.88	95.80
Noise $< 1\mu m$	87.01	97.11	99.63	99.59	75.32	83.60	99.79	99.38
Noise $\geq 1\mu m$	53.55	98.88	93.65	92.02	77.75	61.04	98.10	97.26

Table 2: Precision (prec) and recall (rec) of the benchmarked models on SWINGS data

related curves are not reported for concision purposes. The R^2 for the noise particles was of 1.0 for both data sources (data not shown). The CNN and the manual expert hence discriminated similarly between phytoplankton and non-phytoplankton cells (the counts only differed by 2.5%).

The R^2 and the slope coefficients on Figure 3 are close to 1.0 for the quasi-totality of the cFPGs of both data sources. The counts resulting from the manual and CNN gatings are in adequation. The two main exceptions are the Micro and Rednano cells from the SSLAMM data. In the SSLAMM data, Micro cells were rare (less than 300 cells per file) which made the identification of this population difficult. Concerning the Rednano cells, the R^2 of 0.61 is partly explained by a ~~different~~ Redpicoeuk / Rednano frontier between the CNN and the expert. This is confirmed by the 0.84 slope coefficients of the SSLAMM Redpicoeuk cells: the largest manually gated Redpicoeuk cells were regarded as Rednano cells by the CNN.

The CNN average prediction time for each file of the series was of 90 seconds (7 seconds for the prediction itself and more than a minute for the pre-processing steps). We ran the pipeline on two machines in parallel and the total prediction time was of 15 CPU usage hours for the 1639 files of the SSLAMM time series and 10 hours for the 1184 files of the SWINGS time series.

Discussion

The use of automated systems is often mandatory to get resolutive datasets, common in the field of physical oceanography, but still limited in marine microbial ecology. ~~Microbiological entities~~ in marine environments are influenced by physics, chemistry, and biological interactions that shape their distribution. Yet, they also have internal clocks and specific physiological-morphological characteristics that affect their fitness and require studies integrating biodiversity and dynamic processes (Dutkiewicz et al. 2020). The measurements of cell abundances and morphological traits extracted from *in situ* samples collected with AFCM have already provided numerous insights into the complex distribution of phytoplankton and its interaction with environmental factors (Ribalet et al. 2015; Hyun et al. 2020), such as physical conditions (Partensky et al. 1999; Vaulot et al. 2008; Marrec et al. 2018; Louchart et al. 2020) and trophic network interactions (Christaki et al. 2011).

Automatic classification of AFCM data is built upon referenced cPFGs used for training purpose. Manual gating is prone to subjectivity and assessments of the heterogeneity between experts classifications are rarely performed in flow cytometric studies. Garcia et al. (2014) evidenced up to 20% variability between two experts on two groups of bacterioplankton. In the present study, a consensus between six experts from different laboratories was evaluated on six cPFGs and noise events. The most abundant cPFGs, Orgpicopro and Redpicoeuk, were identified by all experts with small error margins. This can be attributed to the high number of cells, combined to the very characteristic orange fluorescence of Orgpicopro particles. On the contrary, there was a lack of consensus concerning the boundaries between Redpicoeuk and Rednano, with counts variations of more than 100% between experts. The origin of this discrepancy came from the non consensual criteria used to differentiate these groups using 2D projections. Some experts used the 3.13 μm silica beads provided to them for the experiment, while other experts used a threshold between the 2 and 3.13 μm beads. The choice of a criterion to distinguish Redpicoeuk from Rednano is an issue already reported in Buitenhuis et al. (2012). In ad-

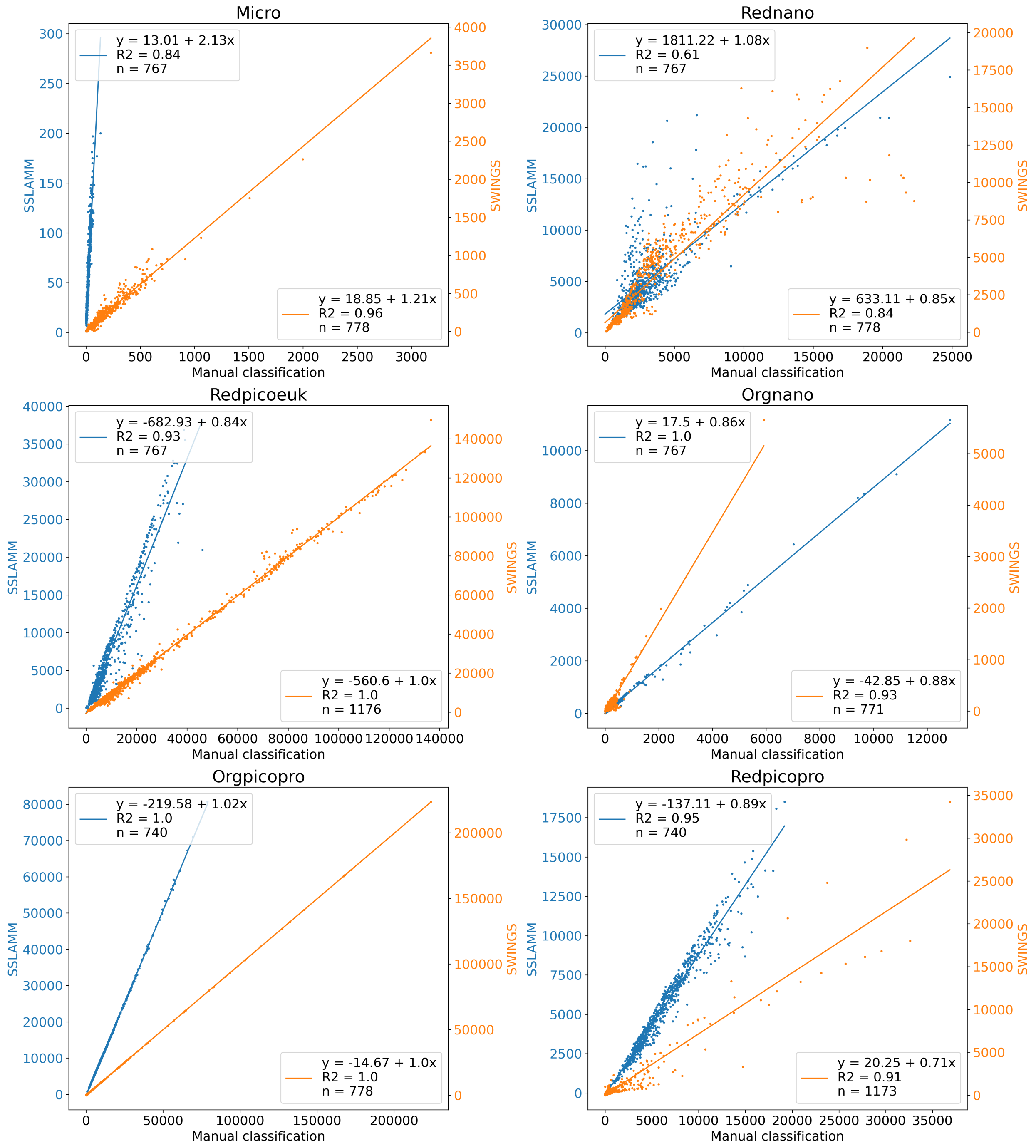


Figure 3: Automatic classification count (number of particles) as a function of the manual gating count (number of particles) for each cPFG: the Orgnano (a), the Micro (b), the Rednano (c), the Redpicpeuk (d), the Redpicopro (e), the Orgpicopro (f). For each cPFG a linear regression has been fitted and the resulting line coefficients and the R^2 coefficient are given.

dition, the Redpicopro / noise $< 1\mu m$ frontier differed significantly between experts. Finally, the differences in cPFG relative abundances made the manual classification of rare cPFGs hard and entailed divergences in Micro, Rednano and Orgnano counts.

As such, the intercomparison highlighted the necessity of consensual rules and criteria to distinguish groups and the need for peer-reviewed data in order to obtain reliable cPFG observations for automation purposes.

Such multi-reviewed datasets are increasing in popularity in the machine learning community, the best example being the ImageNet repository (Fei-Fei 2010).

Despite the heterogeneity in manual gating, a robust and reliable dataset has been built by keeping the particles that were consensual between experts. Using the consensual observations, three statistical models were trained and their performances compared with the ones of the Convolutional Neural Network presented here.

On the SSLAMM and SWINGS test sets, the CNN model proposed in this study achieved precision and recall values competitive with the ones of the LGBM and higher than the ones of the kNN and of the LDA. It exhibited performances higher than 90% in a vast majority of cases. When compared to a manual expert gating the CNN has given proofs that it was a reliable method to track the cPFG abundance in near-real time in two very different contexts. Furthermore, it exhibited significant generalization properties when trained on the SWINGS dataset and used for prediction on the SSLAMM dataset. When trained on the SSLAMM data to predict SWINGS data, the generalization power of the CNN was still solid but lower. This may be due to the lower diversity of SSLAMM data that were sampled in a unique geographical point compared to the SWINGS data collected in very contrasted areas of the South-West Indian and Southern oceans. This could also be due to the lower size of the SSLAMM dataset to which neural methods are particularly sensitive.

~~As a conclusion,~~ this preliminary and highly promising work applies a CNN on interpolated raw pulse shapes acquired on an hourly basis by pulse-shape recording flow cytometry. It opens the way to the integration of cPFGs into forecasting biogeochemical models, depending on near real time data inputs. High frequency sampling of phytoplankton and determination of the communities structure and abundances in near real time will permit a better integration of pulsed events and responses capacities of some functional groups in these models. It will also enable to adjust near real time spatial sampling strategies where influences of physical structures such as fronts and eddies directly affect the distribution of phytoplankton groups (d'Ovidio et al. 2019).

Acknowledgments

We thank Cytobuoy b.v. for their assistance to design special CytoClus4© features mandatory to conduct this work. We thank Olivier Grosso and Michel Durand for technical assistance at the SeaWater Sensing Laboratory At MIO Marseille (SSLAMM), and the support of the MIO Service Atmosphere Mer (Deny Malengros and Fabrice Garcia) and UMS OSU Pytheas divers, Laurent VanBostal, Christian Marshal and Dorian Guillemain for installing and maintaining the pumping inlet. Supports for the SSLAMM were provided by Aix Marseille Université, MIO, and OSU PYTHEAS. We thank Stéphanie Barrillon and the participants of the FUMSECK cruise. We thank the captain and crew of the Marion Dufresne II. We are deeply grateful to Lea Gest and

Lloyd Izard for assistance on the MAP-IO flow cytometer. The MAP-IO project has been funded by the CPER, the FEDER, the FOF and INSU and conducted by the LACY-University of la Reunion. We are also very thankful to Lloyd Izard, Maiwenn Hascoët and Chloé Caille for their participation to the manual gating process. Funding of R.F. PhD thesis was provided by the Ministry of Higher Education, Research and Innovation. The project leading to this publication has received funding from European FEDER Fund under project 1166-39417. The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" program.

References

Abdelaal, T., V. van Unen, T. Höllt, F. Koning, M. J. Reinders, and A. Mahfouz 2019. Predicting cell populations in single cell mass cytometry data. *Cytometry Part A 95*(7), 769–781.

Beaufort, L. and D. Dollfus 2004. Automatic recognition of coccoliths by dynamical neural networks. *Marine Micropaleontology 51*(1-2), 57–73.

Bergstra, J., D. Yamins, and D. D. Cox 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proceedings of the 30th International Conference on Machine Learning 28(1)*, 115–123.

Boddy, L., C. Morris, M. Wilkins, G. Tarran, and P. Burkill 1994. Neural network analysis of flow cytometric data for 40 marine phytoplankton species. *Cytometry: The Journal of the International Society for Analytical Cytology 15*(4), 283–293.

Boss, E., A. M. Waite, J. Uitz, and others 2020. Recommendations for plankton measurements on the go-ship program with relevance to other sea-going expeditions. *SCOR Working Group GO-SHIP Report 154*, 1–70.

Buitenhuis, E. T., W. K. Li, D. Vaulot, and others 2012. Picophytoplankton biomass distribution in the global ocean. *Earth System Science Data 4*(1), 37–46.

Caillault, É., P.-A. Hébert, and G. Wacquet 2009. Dissimilarity-based classification of multidimensional signals by conjoint elastic matching: application to phytoplanktonic species recognition. In *International Conference on Engineering Applications of Neural Networks*, pp. 153–164. Springer.

Carr, M.-E., M. A. Friedrichs, M. Schmeltz, and others 2006. A comparison of global estimates of marine primary production from ocean color. *Deep Sea Research Part II: Topical Studies in Oceanography 53*(5-7), 741–770.

Chisholm, S. W., R. J. Olson, E. R. Zettler, R. Goericke, J. B. Waterbury, and N. A. Welschmeyer 1988. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature 334*(6180), 340–343.

Christaki, U., C. Courties, R. Massana, P. Catala, P. Lebaron, J. M. Gasol, and M. V. Zubkov 2011. Optimized routine flow cytometric enumeration of heterotrophic flagellates using sybr green i. *Limnology and Oceanography: Methods 9*(8), 329–339.

Cui, Y., M. Jia, T.-Y. Lin, Y. Song, and S. Belongie 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277.

del Barrio, E., H. Inouzhe, J.-M. Loubes, C. Matrán, and A. Mayo-Íscar 2019. optimalflow: Optimal-transport approach to flow cytometry gating and population matching. arXiv preprint arXiv:1907.08006.

Dubelaar, G. and P. Gerritzen 2000. Cytobuoy: a step forward towards using flow cytometry in operational oceanography. *Scientia Marina 64*(2), 255–265.

Dubelaar, G. B., P. L. Gerritzen, A. E. Beeker, R. R. Jonker, and K. Tangen 1999. Design and first results of cytobuoy: A wireless flow cytometer for in situ analysis of marine and fresh waters. *Cytometry: The Journal of the International Society for Analytical Cytology 37*(4), 247–254.

Dugenne, M., M. Thyssen, D. Nerini, C. Mante, J.-C. Poggiale, N. Garcia, F. Garcia, and G. J. Grégori 2014. Consequence of a sudden wind event on the dynamics of a coastal phytoplankton community: an insight into specific population growth rates using a single cell high frequency approach. *Frontiers in microbiology 5*, 485.

Dunker, S. 2019. Hidden secrets behind dots: Improved phytoplankton taxonomic resolution using high-throughput imaging flow cytometry. *Cytometry Part A 95*(8), 854–868.

Dutkiewicz, S., P. Cermeno, O. Jahn, M. J. Follows, A. E. Hickman, D. A. Taniguchi, and B. A. Ward 2020. Dimensions of marine phytoplankton diversity. *Biogeosciences 17*(3), 609–634.

d'Ovidio, F., A. Pascual, J. Wang, and others 2019. Frontiers in fine-scale in situ studies: Opportunities during the swot fast sampling phase. *Frontiers in Marine Science 6*, 168.

Fei-Fei, L. 2010. Imagenet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*, Volume 16, pp. 18–25.

Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *science 281*(5374), 237–240.

Fowler, B. L., M. G. Neubert, K. R. Hunter-Cevera, R. J. Olson, A. Shalapyonok, A. R. Solow, and H. M. Sosik 2020. Dynamics and functional diversity of the smallest phytoplankton on the northeast us shelf. *Proceedings of the National Academy of Sciences 117*(22), 12215–12221.

Garcia, F. C., A. Lopez-Urrutia, and X. A. G. Moran 2014. Automated clustering of heterotrophic bacterioplankton in flow cytometry data. *Aquatic Microbial Ecology 72*(2), 175–185.

González, P., A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik 2019. Automatic plankton quantification using deep features. *Journal of Plankton Research 41*(4), 449–463.

Green, J., P. Course, and G. Tarran 1996. The life-cycle of emiliania huxleyi: A brief review and a study of relative ploidy levels analysed by flow cytometry. *Journal of marine systems 9*(1-2), 33–44.

Hamilton, M., G. M. Hennon, R. Morales, and others 2017. Dynamics of teleaulax-like cryptophytes during the decline of a red water bloom in the columbia river estuary. *Journal of Plankton Research 39*(4), 589–599.

He, K., X. Zhang, S. Ren, and J. Sun 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hyun, S., M. R. Cape, F. Ribalet, and J. Bien 2020. Modeling cell populations measured by flow cytometry with covariates using sparse mixture of regressions. arXiv preprint arXiv:2008.11251.

Jacquet, S., M. Heldal, D. Iglesias-Rodriguez, A. Larsen, W. Wilson, and G. Bratbak 2002. Flow cytometric analysis of an emiliana huxleyi bloom terminated by viral infection. *Aquatic Microbial Ecology 27*(2), 111–124.

Kavanaugh, M. T., M. J. Oliver, F. P. Chavez, R. M. Letelier, F. E. Muller-Karger, and S. C. Doney 2016. Seascapes as a new vernacular for pelagic ocean monitoring, management and conservation. *ICES Journal of Marine Science 73*(7), 1839–1850.

Krizhevsky, A., I. Sutskever, and G. E. Hinton 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.

Le Quere, C., S. P. Harrison, I. Colin Prentice, and others 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology 11*(11), 2016–2040.

Lévy, M., R. Ferrari, P. J. Franks, A. P. Martin, and P. Rivière 2012. Bringing physics to life at the submesoscale. *Geophysical Research Letters 39*(14).

Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liu, L., H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han 2019. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265.

Louchart, A., F. Lizon, A. Lefebvre, M. Didry, F. G. Schmitt, and L. F. Artigas 2020. Phytoplankton distribution from western to central english channel, revealed by automated flow cytometry during the summer-fall transition. *Continental Shelf Research 195*, 104056.

Malkassian, A., D. Nerini, M. A. van Dijk, M. Thyssen, C. Mante, and G. Gregori 2011. Functional analysis and classification of phytoplankton based on data from an automated flow cytometer. *Cytometry part A 79*(4), 263–275.

Marrec, P., G. Grégori, A. M. Doglioli, and others 2018. Coupling physics and biogeochemistry thanks to high-resolution observations of the phytoplankton community structure in the northwestern mediterranean sea. HAL preprint. HAL Id: hal-01735426.

Metfies, K., C. Gescher, S. Frickenhaus, and others 2010. Contribution of the class cryptophyceae to phytoplankton structure in the german bight 1. *Journal of Phycology 46*(6), 1152–1160.

Miloslavich, P., N. J. Bax, S. E. Simmons, and others 2018. Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. *Global change biology 24*(6), 2416–2433.

Olson, R., D. Vaulot, and S. Chisholm 1985. Marine phytoplankton distributions measured using shipboard flow cytometry. *Deep Sea Research Part A. Oceanographic Research Papers 32*(10), 1273–1280.

Pan, S. J. and Q. Yang 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*(10), 1345–1359.

Partensky, F., J. Blanchot, and D. Vaulot 1999. Differential distribution and ecology of prochlorococcus and synechococcus in oceanic waters: a review. *Bulletin-Institut Oceanographique Monaco Special Number 19*, 457–476.

Popel, M. and O. Bojar 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics 110*(1), 43–70.

Ribalet, F., J. Swalwell, S. Clayton, and others 2015. Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre. *Proceedings of the National Academy of Sciences 112*(26), 8008–8012.

Ribeiro, C. G., A. L. dos Santos, D. Marie, V. H. Pellizari, F. P. Brandini, and D. Vaulot 2016. Pico and nanoplankton abundance and carbon stocks along the brazilian bight. *PeerJ 4*, e2587.

Saba, V. S., M. A. Friedrichs, D. Antoine, and others 2011. An evaluation of ocean color model estimates of marine primary productivity in coastal and pelagic regions across the globe. *Biogeosciences 8*(2), 489–503.

Schmidt, K. C., S. L. Jackrel, D. J. Smith, G. J. Dick, and V. J. Denef 2020. Genotype and host microbiome alter competitive interactions between microcystis aeruginosa and chlorella sorokiniana. *Harmful Algae 99*, 101939.

Simonyan, K. and A. Zisserman 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Steinley, D. 2004. Properties of the hubert-arable adjusted rand index. *Psychological methods 9*(3), 386.

Thomas, M. K., S. Fontana, M. Reyes, and F. Pomati 2018. Quantifying cell densities and biovolumes of phytoplankton communities and functional groups using scanning flow cytometry, machine learning and unsupervised clustering. *PloS one 13*(5), e0196225.

van den Engh, G. J., J. K. Doggett, A. W. Thompson, M. A. Doblin, C. N. Gimpel, and D. M. Karl 2017. Dynamics of prochlorococcus and synechococcus at station aloha revealed through flow cytometry and high-resolution vertical sampling. *Frontiers in Marine Science 4*, 359.

Vaulot, D., W. Eikrem, M. Viprey, and H. Moreau 2008. The diversity of small eukaryotic phytoplankton ($\leq 3\mu$m) in marine ecosystems. *FEMS microbiology reviews 32*(5), 795–820.

Wacquet, G., É. P. Caillault, D. Hamad, and P.-A. Hébert 2013. Constrained spectral embedding for k-way data clustering. *Pattern Recognition Letters 34*(9), 1009–1017.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson 2014. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328.

Zhang, M., J. Lucas, J. Ba, and G. E. Hinton 2019. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pp. 9593–9604.