Automatic recognition of flow cytometric phytoplankton functional groups using convolutional neural networks

Robin Fuchs, Melilotus Thyssen, Véronique Creach, Mathilde Dugenne, Lloyd Izard, Marie Latimier, Arnaud Louchart, Pierre Marrec, Machteld Rijkeboer, Gérald Grégori, et al.

LIMNOLOGY and OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 20, 2022, 387–399

doi: 10.1002/lom3.10493

Automatic recognition of flow cytometric phytoplankton functional groups using convolutional neural networks

Robin Fuchs,[1,2] **Melilotus Thyssen**,[2]* **Véronique Creach**,[3] **Mathilde Dugenne**,[4] **Lloyd Izard**,[5] **Marie Latimier**,[6] **Arnaud Louchart**,[7,8] **Pierre Marrec**,[9] **Machteld Rijkeboer**,[10] **Gérald Grégori**,[2] **Denys Pommeret**[1,11,12,13]

[1]Aix Marseille Univ, CNRS, I2M, Marseille, France
[2]Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France
[3]Cefas, Suffolk, UK
[4]Department of Oceanography, University of Hawai'i at Manoa, Honolulu, Hawai'i
[5]Sorbonne Université, CNRS, IRD, MNHN, Laboratoire d'Océanographie et du Climat: Expérimentations et Approches Numériques (LOCEAN-IPSL), Paris, France
[6]IFREMER, DYNECO PELAGOS, Plouzane, France
[7]Department of Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Villa Comunale, Naples, Italy
[8]IFREMER, Laboratoire Environnement and Ressources, Boulogne-sur-Mer, France
[9]Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island
[10]Laboratory for Hydrobiological Analysis, Rijkswaterstaat (RWS), Lelystad, The Netherlands
[11]Université Claude Bernard Lyon 1, Villeurbanne, France
[12]ISFA, Lyon, France
[13]Laboratoire de Sciences Actuarielle et Financière (SAF), Lyon, France

Abstract

The 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 (PFGs). However, the associated manual post-processing of the data coupled with the increasing size and number of datasets is time-consuming and error-prone. Machine learning models are increasingly used to run automatic classification. Yet, most of the existing methods either present a long training process, need to manually design 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 several PFGs using AFCM pulse shapes. The uncertainties of manual classification were first estimated by comparing experts' recognition of six PFGs. Consensual particles from the manual PFG classification were used to train and validate the CNN. The CNN obtained competitive performances compared to other models used in the literature and remained robust across several sampling areas, and instrumental 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 a comparison with the manual classification over 3-month periods and a 2h frequency. These promising results strengthen the near real-time observation of PFGs, especially required with the increasing use of AFCM in monitoring research programs.

*Correspondence: melilotus.thyssen@mio.osupytheas.fr

Additional Supporting Information may be found in the online version of this article.

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 autotrophic carbon fixation, 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 the coarse 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 functional diversity, biomass, and distribution are listed as essential ocean variables (EOV) (Miloslavich et al. 2018), but datasets with resolutions inferior to 10 km are scarce. Automated pulse-shape recording flow cytometry (AFCM) such as the CytoSense instrument (Cytobuoy, b.v.; Dubelaar et al. 1999; Dubelaar and Gerritzen 2000) enables vast automated data acquisition with hourly sampling strategies of several phytoplankton groups at a single-cell-level resolution. AFCM is now involved in numerous oceanographic field studies and benefits from the growing scientific interest in automated single-cell approaches (Boss et al. 2020) in monitoring programs.

The CytoSense AFCMs generate a set of pulse shapes or flow cytometric curves (FCCs), which represent the optical profiles of scatter and fluorescences emitted by each particle (detritus, cell, or colony) when crossing a laser beam. Scatter signals collected at small and large angles (forward scatter [FWS] and sideward scatter [SWS], respectively) are related to the particle size and structure (granularity), while red fluorescence (FLR) and yellow-orange fluorescence (FLY or FLO) signals are reflecting pigment contents of the photosynthetic cells (such as chlorophyll *a* or phycoerythrin). 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 particles per second thanks to a frequency acquisition of 4 MHz, with sampled volume up to 5 mL routinely.

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 the most common groups observed by flow cytometry in natural seawater. In addition to these groups of pico-nanophytoplankton, AFCM resolves microphytoplankton size classes with a coarse taxonomic-level identification (typically up to the genus) using the recent integration of image-in-flow devices (Dugenne et al. 2014). A dedicated vocabulary, relying on these papers, has been recently suggested by a wide group of flow cytometry experts (http://vocab.nerc.ac.uk/collection/F02/current/). These size and pigment-related groups belong to several phytoplankton functional groups (PFGs), 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).

Raw data recorded by AFCM has to be manually processed. This processing, called manual gating of cPFG, is performed on 2D projections of reduced statistics of the FCCs (such as pulse maximum height, area under the curve, pulse width). The long periods of assiduity required, coupled with experts' diversity of practices and the significant differences in cPFG abundances can be substantial sources of errors. Furthermore, the spread of the AFCM technology generates datasets too numerous 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-automated 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, camera resolution is relatively low for the identification of pico-nanophytoplankton size classes, which show limited morphological diversity. As such, using the FCCs offers an alternative since it deals also with these small particles that can represent up to 90% of the total phytoplankton biomass (Li et al. 1983; Detmer and Bathmann 1997; Ribeiro et al. 2016). 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 a set of reduced statistics derived from the FCCs. Boddy et al. (1994) started to use neural methods to classify cells at the species level. Wacquet et al. (2013) developed original statistical methods 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, to which this study belongs, relies on the entire FCC signal to perform classification. For example, Malkassian et al. (2011) plunged the FCCs into a Fourier basis and calculated distances to discriminate between populations. Del Barrio et al. (2019) created curve templates to classify AFCM nonmarine cells using Wasserstein distance and optimal transport. Finally, Caillault et al. (2009) relied on the elastic matching coupled with standard classifiers. While these two families of approaches attempt to classify cPFGs in an objective and reproducible manner, they all

present unique advantages and trade-offs. A comparison of all these approaches has yet to be reported.

In this article, we provide a comparison of expert manual classifications of cPFGs detected by AFCM. We used the consensual particles to develop, for the first time, a convolutional neural network (CNN) trained on pulse shapes recorded by AFCM as described in Fig. 1. We compared the performance of our CNN, along with other automatic approaches, and tested its robustness across two instruments and multiple study areas. Finally, the CNN was used to generate predictions spanning 3 months in a coastal station of the Mediterranean Sea and 2 months in the South-West Indian Ocean, both at a 2h sampling frequency. The robustness and extremely fast process of the applied CNN open the way to near real-time cPFG analysis.

Material and procedures

Data origin and collection

Two datasets collected using different approaches were used in this study. The first one, referred to as SSLAMM data, was acquired in different Mediterranean areas using the same flow cytometer and settings: at a coastal marine Mediterranean station (the SSLAMM, SeaWater Sensing Laboratory At MIO Marseille, France), between September 2019 and December 2019 and in an open Mediterranean sea area, during the FUMSECK cruise (DOI 10.17600/18001155) in the Gulf of Genoa from 30 April 2019 to 05 May 2019. The second dataset, named hereafter SWINGS data, originated from the South-West Indian Ocean and the Southern Ocean and was collected onboard the R/V Marion Dufresne II, from 11 January 2021 to 08 March 2021, in the frame of the MAP-IO project (Marion Dufresne Atmospheric Program—Indian Ocean, University of la Reunion) during the GEOTRACES SWINGS cruise (South-West Indian Geotraces Section, DOI 10.13155/83989, SWINGS data). Two distinct CytoSense flow cytometers (Cytobuoy b.v.), hereafter identified as SSLAMM-AFCM, and MAP-IO-AFCM, were deployed. A map indicating the location of the different sampling areas is given in Supporting Information Fig. S1.

For both datasets, seawater was continuously pumped in situ and the flow cytometers ran automated acquisitions scheduled every 2 h. The SSLAMM coastal seawater was gently pumped with a VerderFlex40 peristaltic pump at 10 m away from the coast at a depth of 3 m, and was delivered unaltered into the laboratory where analyses were conducted. The FUMSECK data were collected onboard the R/V le Tethys II from the underway clean seawater supply pumped at 2 m depth. Onboard the Marion Dufresne II, the seawater was collected from the underway clean seawater supply pumped at 7 m depth, using a centrifugal pump.

The two automated CytoSense flow cytometers (Cytobuoy b.v.) were operated similarly in the three conditions. They pumped samples from a dedicated external chamber of 300 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 before crossing a 488-nm laser beam (Coherent, 120 mW). Both instruments recorded the optical pulse shapes emitted resulting in FWS, SWS, and two fluorescences. The SSLAMM-AFCM collected wavebands of > 652 nm (red fluorescence, FLR) and between 552 and 652 nm (orange fluorescence, FLO). The MAP-IO-AFCM collected wavebands between 668 and 726 nm (FLR) and 516 and 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.

Laser scattering at frontal angles (FWS) was collected by two distinct photodiodes to check for the sample core alignment. The difference between left and right photodiodes signatures generated the Curvature curve. SWS, FLR, and FLY were collected with photomultiplier tubes. To follow the stability of the flow cytometers, 2.0-μm fluorescing polystyrene beads (Polyscience®) were regularly analyzed. Silica beads (1.01, 2.56, 3.13, 5.02, and 7.27 μm in diameter, Bangs Laboratory®) were also used to calibrate FWS into particle size.

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

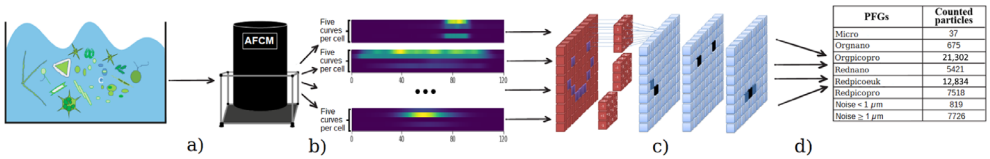


Fig. 1. Explanatory scheme of the predictive pipeline. (**a**) Particles are sampled from seawater by AFCM. (**b**) The five flow cytometric curves (FCCs = SWS, FWS, FLR, FLO, Curvature) generated for each particle as they cross a laser beam are interpolated to a fixed length and stacked together into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

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. With this setting, the small particles and background noise generating acquisition limitations were not recorded. Except for their use of 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, similar gains).

Flow cytometry groups nomenclature

A set of six phytoplankton functional groups determined by their optical properties were selected in this study. They were identified and labeled using the flow cytometry consensual nomenclature (http://vocab.nerc.ac.uk/collection/F02/current/): Redpicopro, Orgpicopro, Redpicoeuk, Rednano, Orgnano, Redmicro, Orgmicro. A correspondence table between this new nomenclature and previous denominations observed in the literature is given in Supplemental Information Table S1. There were not enough Redmicro and Orgmicro cells in situ to distinguish between these two groups and they will be gathered together in the sequel under the name "Micro" cells. The HSnano, Redredpico, Redrednano, and Orgpico groups defined in the nomenclature were not abundant enough to be resolved or not found in our case.

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 and ≥1 μm groups using silica beads as a size reference (Supporting Information Fig. S2). The ≥1 μm noise group mainly contained large detrital particles or predators such as ciliates or flagellates cells that have ingested some phytoplankton cells. Conversely, <1 μm noise group often contained optical noise from the sensors, non-fluorescing heterotrophic prokaryotes, or decaying cells.

The total number of Orgpicopro and Redpicopro cells was obtained from the FLR5/FLR6 files and the total number of Orgnano, Redpicoeuk, Rednano, and Micro cells was obtained from the corresponding FLR20/FLR25 files.

Manual gating methodology and heterogeneity estimation

The raw data collected by the AFCM are composed of a series of five curves exhibiting variable heights, areas, and lengths. Experts use a dedicated software, CytoClus4©, and single values for each curve, typically the area under the curve or the maximal value of the curve, to perform their gating. With the summary statistics, experts obtain a point of dimension five for each observation and the dataset can be represented by a series of 2D projections. For example, experts commonly project the Total FLR (the area under the FLR curve) against the Total FLO or FLY (the area under the FLO or 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. The manual gating procedure is illustrated in Supporting Information Fig. S3.

The heterogeneity among 6 AFCM manual classifications was assessed on multiple SSLAMM and SWINGS acquisitions (6 and 20, respectively), spanning multiple seasons, study areas, and times of the day. The list of the cPFGs was given, along with two acquisitions of 2.0-μm polystyrene beads (Polyscience®) and 3.13-μm silica beads (Bangs Laboratory®).

The heterogeneity was measured by computing the Adjusted Rand Indices (ARIs) Steinley (2004) on the experts' overall classification and the coefficients of variation (CVs) of each cPFG count. The ARIs indicate the similarity between two experts' overall classifications. The closest the ARI is to 1, the more similar the classifications between two experts are. The ARIs have been computed for all pairs of experts and all files.

In addition, 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. As a result, the ARIs assessed the overall agreement between experts' classifications whereas the CVs summarized the similarities of manual classifications at the cPFG level.

Beyond the initial training samples, one of the experts has manually gated 3 months of data from the SSLAMM station (from mid-September 2019 to mid-December 2019) and the entire dataset from the MAP-IO-SWINGS cruise. The classification obtained from the CNN was then compared with the manual gating.

Data processing for automatic classification

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

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 μm 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 signals are expected to look like bell curves, whereas <1 μm noise events can present a significant variety of shapes. Therefore among the consensual Redpicopro cells, only the bell-curved SWS cells were kept to train and validate the models.

The consensual particles were split into three sets: the training set, the validation set, and the test set. The training set is used by the models to learn how to distinguish between cPFGs, the validation to compare several specifications of a given model, and the test set to compare the best specifications of different models. In order to reach a substantial total dataset size and to reduce the imbalance between groups that affect the training process, the over-represented groups were undersampled in the training set.

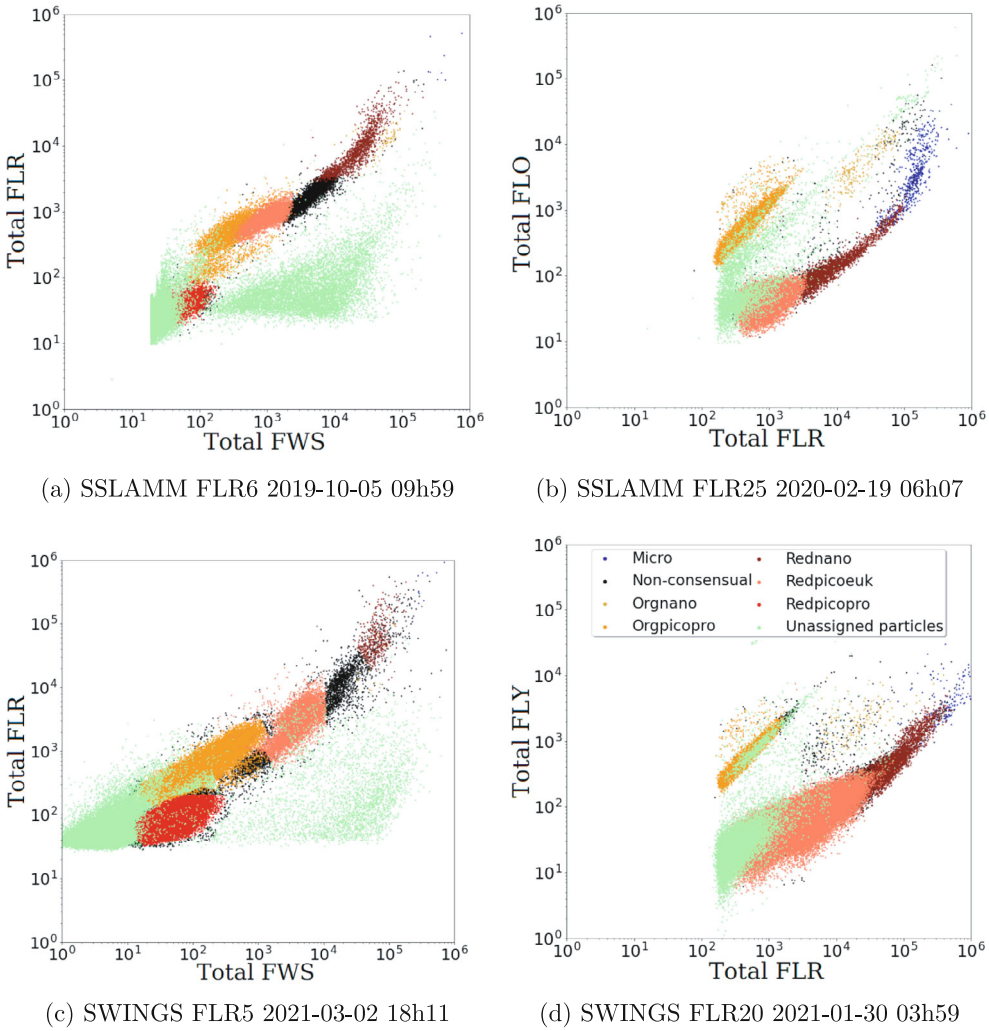
Yet, as Fig. 2 highlights it, the density of points is not uniform in 2D cytograms. Pure random particle sampling tends to let some of the low-density areas of 2D cytograms nearly empty, preventing machine learning models to learn which class to predict for particles in these areas. Hence, additional particles were sampled to fill low-density areas in the limit of 5% of the dataset size. The impact of these zones on the confidence of the CNN cPFG predictions can for instance be seen in Supporting Information Fig. S4.

Before undersampling, the number of particles of the most represented group in the training set was 130 times higher than the less represented one. After undersampling, it was only eight times higher at most for the two datasets.

Conversely, the validation set was undersampled in a stratified manner, that is, non-rebalanced. Finally, the test set was constituted of three genuine files to give the best representation possible of in situ conditions at different seasons and times of the day. The total size of the training, validation, and test sets were 33,791, 50,682, and 134,313 particles for the SSLAMM data, and 57,241, 365,863, and 224,426 particles for the SWINGS data. Supporting Information Tables S2 and S3 describe 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). The size distribution of the FCCs suggested that 75% of our observations were recorded with 120 or fewer values.

In order to train the CNN, which needs a fixed data format for all observations, the curves have been all interpolated to the fixed length of 120 values using quadratic interpolation



(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

Fig. 2. 2D cytograms showing the particles contained in two files from the SSLAMM data (**a**,**b**) and two files from the SWINGS data (**c**,**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 consensual if 2/3 of the experts have voted for the same cPFG for this particle. Non-consensual particles are represented in black.

(*see* Supporting Information Fig. S5 for an illustration). The choice of the 3rd quartile was motivated by the fact that, intuitively, less information is destroyed when small curves are interpolated to be bigger than the reverse. Besides, 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.

CNN specification

The core of the predictive pipeline is a CNN 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 CNNs can be used to treat both data types with minor adjustments. The CNN architecture is presented in Supporting Information (see Fig. S6). The architecture was inspired by the VGG architecture (Simonyan and Zisserman 2014). Other architectures such as the Inception Architecture (Szegedy et al. 2015) have been implemented but brought no additional performance (result not shown). The number of observations was not sufficient to implement deeper architectures such as residual networks (He et al. 2016).

In our network, features are first extracted by three blocks of convolutional layers separated by "local" average pooling layers 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 computes the probabilities that an observation belongs to each class and the loss of the model is evaluated.

The loss measures 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. 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 (Lin et al. 2017) or the focal class-balanced loss (Cui et al. 2019) have been implemented but brought no additional performances.

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. Ranger, a generalization of the widely used Adam optimizer (Kingma and Ba 2014), was here used. Ranger comes from the combination of two recent publications: RectifiedAdam (or Radam) (Liu et al. 2019) and Lookahead (Zhang et al. 2019).

In order for the optimization process not to remain stuck in very local minima, it is a common practice to slowly update the parameters of the models at the beginning of the training, when promising parameter regions are not yet identified. 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. 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 observation 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.

Comparison with other classification algorithms

The CNN has been benchmarked against other supervised models to compare the performance of individual machine learning algorithms. The benchmark models have been published in the literature: the k-nearest neighbors (kNN) and the LDA (Abdelaal et al. 2019). Tree-based methods such as Random Forest were represented by the light gradient boosting machine (LGBM) (Ke et al. 2017), which is more recent and takes advantage of gradient-boosting methods.

The data from the manual classifications comparison experiment were used for model 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 from these FCCs (commonly referred to as "Listmode features"). The choice of the features created from the signal highly influences the performances of the models and has to be considered when presenting the results. We rely on the 13 features per curve created by default by the CytoClus4© software. The feature list is given in Supporting Information (see Section S1).

Most parts of statistical models are ruled by a set of hyperparameters chosen by the user (e.g., number of neurons and layers, number of neighbors, learning rate, batch size). The number of possible combinations is far too high for all the combinations to be tested and then to select the best model specifications.

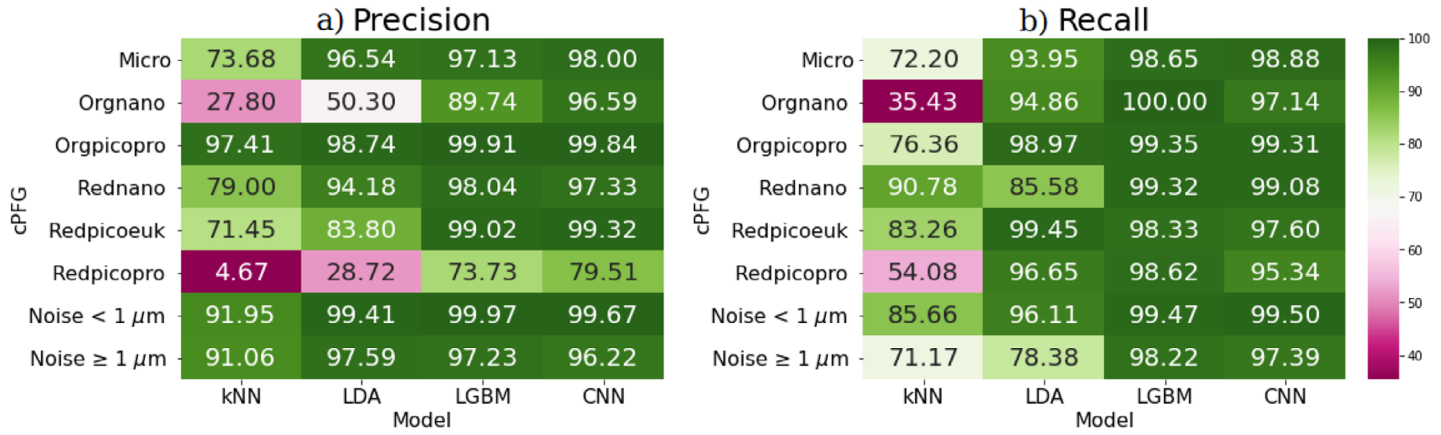


Fig. 3. Precision (**a**) and recall (**b**) (%) of the benchmarked models on SSLAMM data.

One popular approach relies on Bayesian Hyperoptimization algorithms (Bergstra et al. 2013), implemented in our case in the Python libraries Hyperopt and Hyperas (Hyperopt for Keras). The idea of Hyperoptimization methods is to consider hyper-parameters 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 substantial 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 Supporting Information Section S2.

The performances of the CNN and of 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 both precision and recall are 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 freely available as a Github repository named phyto_-curves_reco (https://github.com/RobeeF/phyto_curves_reco) with the following DOI: 10.5281/zenodo.5681642.

Results

Manual gating uncertainty estimation

The main groups observed by AFCM are represented in Fig. 2. It presents descriptive 2D cytograms associated with two files for each data source. 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 Supporting Information (Figs. S7, S8). 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. However,

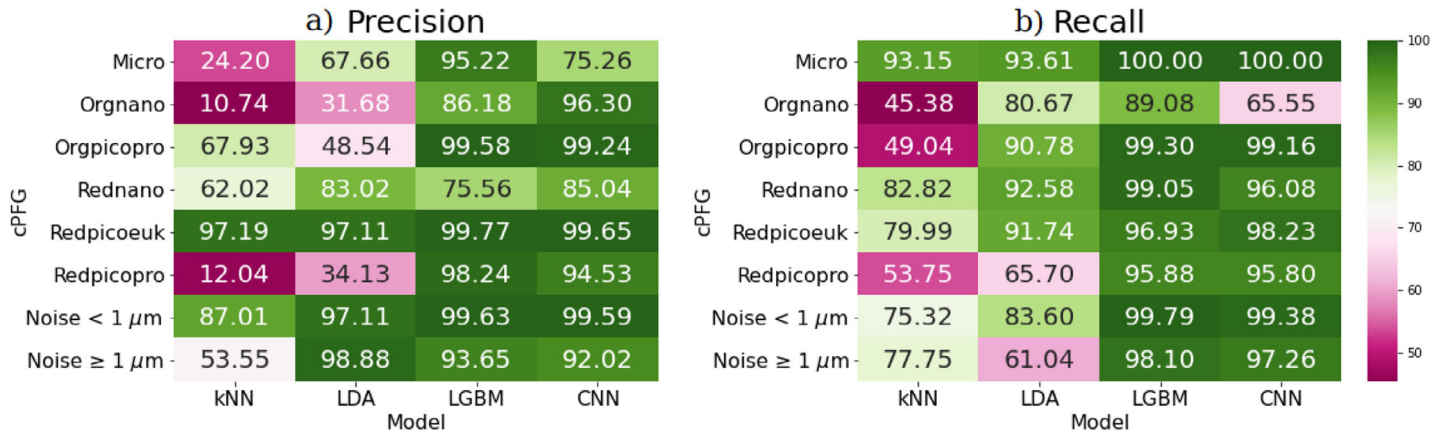


Fig. 4. Precision (**a**) and recall (**b**) (%) of the benchmarked models on SWINGS data.

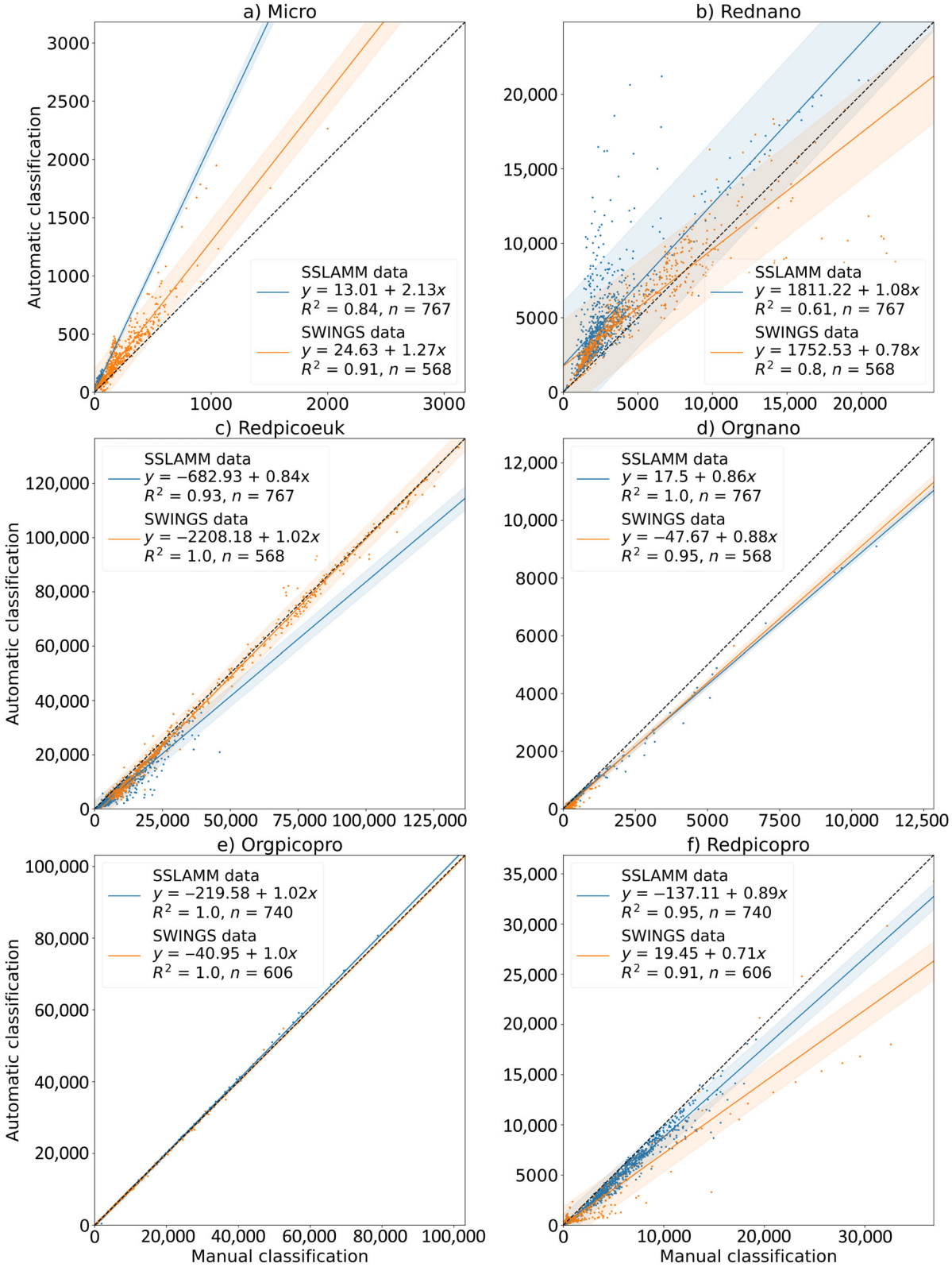


Fig. 5. Automatic classification count (number of particles) as a function of the manual gating count (number of particles) for each cPFG: the Micro (**a**), the Rednano (**b**), the Redpicoeuk (**c**), the Orgnano (**d**), the Orgpicopro (**e**), the Redpicopro (**f**). Blue dots are for SSLAMM data, Orange dots are for SWINGS data. For each cPFG a linear regression has been fitted and the corresponding regression coefficients and R^2 are reported. The resulting 95% confidence intervals are illustrated as light orange and blue bands. The black dashed line indicates a 1 : 1 ratio between the manual and automatic classifications.

these high ARIs 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 0.02 and 0.50 for the SWINGS data and the case of Redpicoeuk (SSLAMM CV $\in$ [0.05, 0.44] and SWINGS CV $\in$ [0.03, 0.28]). Conversely, Micro cells (SSLAMM CV $\in$ [0.27, 1.55] and SWINGS CV $\in$ [0.12, 1.26]), Orgnano (SSLAMM CV $\in$ [0.50, 0.85] and SWINGS CV $\in$ [0.21, 1.75]), Rednano (SSLAMM CV $\in$ [0.25, 0.92] and SWINGS CV $\in$ [0.10, 1.34]), and Redpicopro (SSLAMM CV $\in$ [0.13, 2.45] and SWINGS CV $\in$ [0.56, 1.07]) were far less identified (Supporting Information Fig. S8).

Model benchmark on the test set

Figures 3 and 4 report the precision and the recall obtained by the four models for each cPFG and noise classes.

Based on the specific precision and recall values, the CNN and the LGBM obtained the best performances on the quasi-totality of cPFGs. The kNN presented the worst performances for both datasets. The LDA results are mixed as it distinguished noise events from phytoplankton particles classified but got the worst precision on three cPFGs on the SWINGS data.

The best manually identified cPFGs were also the best classified by machine learning models, i.e. Orgpicopro and Redpicoeuk cells. Similarly, the Redpicopro and Orgnano cells were weakly manually identified and less well gated by machine learning models. Finally, Micro and Rednano cells that experienced poor manual identifiability presented good precision and recall values for near all methods.

The generalization capacity of the models was tested by training them on one data source (SSLAMM or SWINGS) and by making predictions on the other data source. Results are given in Supporting Information Figs. S9 and S10.

When the models were trained on the SWINGS data and used to predict SSLAMM 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 used to predict SWINGS data, the LGBM obtained the best performances and LDA the worst. Redpicopro cells and noise events ≥1 μm were the worst identified by the models. Rednano cells obtained precisions lower than 34% but recall values 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 Supporting Information (Table S4).

Automatic classification on the full datasets

Figure 5 presents the regression between the automatically and manually counted cPFGs particles from the SSLAMM files and the SWINGS files.

The R^2 and the slope coefficients in Fig. 5 are close to 1.0 for the majority of the cPFGs of both data sources: The counts resulting from the manual and CNN gatings are in adequation. The main exceptions are the Micro and Rednano cells from the SSLAMM data and the Redpicopro cells from the SWINGS data. In the SSLAMM data, Micro cells were scarce (less than 300 cells per file) which made the identification of this population difficult. The CNN counted twice as many Micro cells as the manual expert, but the counts seemed to be proportional ($R^2 = 0.84$). 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 automatic Redpicopro count from SWINGS data presented a strong correlation with the manual count ($R^2 = 0.91$). However, the CNN was more conservative and considered some of the manually gated Redpicopro cells as *noise* < 1 μm cells. Finally, the R^2 for the noise particles was equal to 1.0 for both data sources (data not shown). The CNN and the manual expert hence discriminated similarly between phytoplankton and non-phytoplankton particles (the counts only differed by 2.5%).

The CNN average prediction time for each file of the series was 66 s (7 s 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 h for the 1184 files of the SWINGS time series.

Discussion

The use of automated sensors is often mandatory to get resolutive datasets, common in the field of physical oceanography, but still limited in marine microbial ecology. Microbial populations 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 sensors integrating biodiversity and dynamic processes (Dutkiewicz et al. 2020). Flow cytometry measurements of phytoplankton cell abundances and single-cell morphological traits have already provided numerous insights into their interaction with environmental factors (Ribalet et al. 2015; Hyun et al. 2020), such as physical conditions (Partensky et al. 1999; Marrec et al. 2018; Louchart et al. 2020) and trophic network interactions (Christaki et al. 2011). The collected morphological traits have also enabled hourly growth rates and primary

production assessments per phytoplankton group (Sosik et al. 2003; Dugenne et al. 2014; Hunter-Cevera et al. 2014).

Although AFCM is a powerful tool for the study of phytoplankton functional groups and benefits from recent technological advances, AFCM data post-processing is often performed manually. Yet, this post-processing (also named 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 overall classification methodology was shared by the experts as confirmed by the high pairwise ARIs. On the contrary, the uncertainties existing in the exact manual gates frontiers coupled with the underrepresentation of several cPFGs led to significant differences in cPFG counts.

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 with 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 for Rednano cells. The origin of this discrepancy came from the nonconsensual 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 addition, the observation of Redpicopro cells by AFCM has been enabled only recently thanks to advances in filtration of the sheath fluid or more powerful lasers Marrec et al. (2018). Yet, these particles still remain close to the flow cytometer detection limits and Redpicopro cells were hardly distinguished from the noise < 1 μm by the experts. Finally, the differences in cPFG relative abundances made the manual classification of rare cPFGs equivocal 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 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 (Deng et al. 2009).

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 CNN 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 the LDA. It exhibited performances higher than 90% in a vast majority of cases. When compared to a manual expert gating the CNN has evidenced its reliability to track the cPFG abundance in near real time in two very different contexts. The small discrepancies between manual and automatic classifications can be considered marginal when compared to the length and the high temporal and functional diversity resolution of the predicted time series. Furthermore, the CNN exhibited significant generalization properties when trained on the SWINGS data and used for prediction on the SSLAMM data. 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 and number of observations of SSLAMM data, where pico-nanophytoplankton cells dominated all over the year, compared to the SWINGS data collected in very contrasted areas of the South-West Indian and Southern oceans, the latter being considered as dominated by nano-microphytoplankton cells (Rembauville et al. 2017).

More generally, the training sets used in this study are of moderate sizes ($\sim 10^4$ observations compared to $\sim 10^6$ observations generally encountered in CNN image classification as in Simonyan and Zisserman (2014)). Yet, deep learning methods seem to take a bigger advantage of dataset sizes than traditional machine learning methods (Ng 2017), at least when the dataset size grows from a moderate to substantial size (several millions of observations) (Hestness et al. 2017; Neyshabur et al. 2017; Sun et al. 2017). Thus, the current increase in AFCM dataset sizes and dataset number should give an additional edge to the CNN over the LGBM which currently present comparable performances.

In summary, 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 will permit a better integration of pulsed events and response 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).

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 A **95**: 769–781. doi: 10.1002/cyto.a.23738

Beaufort, L., and D. Dollfus. 2004. Automatic recognition of coccoliths by dynamical neural networks. Mar. Micropaleontol. **51**: 57–73. doi:10.1016/j.marmicro.2003.09.003

Bergstra, J., D. Yamins, and D. D. Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. PMLR **28**: 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 **15**: 283–293. doi:10.1002/cyto.990150403

Boss, E., 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. SCOR, p. 1–70. doi:10.25607/OBP-718

Buitenhuis, E. T., and others. 2012. Picophytoplankton biomass distribution in the global ocean. Earth Syst. Sci. Data **4**: 37–46. doi:10.5194/essd-4-37-2012

Caillault, É., P.-A. Hébert, and G. Wacquet. 2009. Dissimilarity-based classification of multidimensional signals by conjoint elastic matching: Application to phytoplanktonic species recognition. Communications in Computer and Information Science, 153–164. doi:10.1007/978-3-642-03969-0_15

Carr, M.-E., and others. 2006. A comparison of global estimates of marine primary production from ocean color. Deep-Sea Res. II Top. Stud. Oceanogr. **53**: 741–770. doi:10.1016/j.dsr2.2006.01.028

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**: 340–343. doi:10.1038/334340a0

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. Limnol. Oceanogr. Methods **9**: 329–339. doi:10.4319/lom.2011.9.329

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

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. BMC Bioinformatics 21(1): doi:10.1186/s12859-020-03795-w

Deng, J., W. Dong, R. Socher, L.-J. Li, Li Kai, and Fei-Fei Li. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2009.5206848

Detmer, A., and U. Bathmann. 1997. Distribution patterns of autotrophic pico- and nanoplankton and their relative contribution to algal biomass during spring in the Atlantic sector of the southern ocean. Deep-Sea Res. II Top. Stud. Oceanogr. **44**: 299–320. doi:10.1016/s0967-0645(96)00068-9

d'Ovidio, F., and others. 2019. Frontiers in fine-scale in situ studies: Opportunities during the SWOT fast sampling phase. Front. Mar. Sci. **6**: 168. doi:10.3389/fmars.2019.00168

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 **37**: 247–254. doi:10.1002/(sici)1097-0320(19991201)37:4<247::aid-cyto1>3.0.co;2-9

Dubelaar, G., and P. Gerritzen. 2000. Cytobuoy: A step forward towards using flow cytometry in operational oceanography. Sci. Mar. **64**: 255–265. doi:10.3989/scimar.2000.64n2255

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. Front. Microbiol. **5**: 485. doi:10.3389/fmicb.2014.00485

Dunker, S. 2019. Hidden secrets behind dots: Improved phytoplankton taxonomic resolution using high-throughput imaging flow cytometry. Cytometry A **95**: 854–868. doi:10.1002/cyto.a.23870

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**: 609–634. doi:10.5194/bg-17-609-2020

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**: 237–240. doi:10.1126/science.281.5374.237

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. Proc. Natl. Acad. Sci. **117**: 12215–12221. doi:10.1073/pnas.1918439117

Garcia, F. C., A. Lopez-Urrutia, and X. A. G. Moran. 2014. Automated clustering of heterotrophic bacterioplankton in flow cytometry data. Aquat. Microb. Ecol. **72**: 175–185. doi:10.3354/ame01691

González, P., A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik. 2019. Automatic plankton quantification using deep features. J. Plankton Res. **41**: 449–463. doi:10.1093/plankt/fbz023

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. J. Mar. Syst. **9**: 33–44. doi:10.1016/0924-7963(96)00014-0

Hamilton, M., and others. 2017. Dynamics of teleaulax-like cryptophytes during the decline of a red water bloom in the Columbia river estuary. J. Plankton Res. **39**: 589–599. doi:10.1093/plankt/fbx029

He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition, p. 770–778. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. doi:10.1109/cvpr.2016.90

Hestness, J., and others (2017). Deep learning scaling is predictable, empirically (Version 1). arXiv. doi:10.48550/ARXIV.1712.00409

Hunter-Cevera, K. R., M. G. Neubert, A. R. Solow, R. J. Olson, A. Shalapyonok, and H. M. Sosik. 2014. Diel size distributions reveal seasonal growth dynamics of a coastal phytoplankter. Proc. Natl. Acad. Sci. **111**: 9852–9857. doi:10.1073/pnas.1321421111

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. doi:10.48550/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. Aquat. Microb. Ecol. **27**: 111–124. doi:10.3354/ame027111

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 J. Mar. Sci. **73**: 1839–1850. doi:10.1093/icesjms/fsw086

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree, p. 3146–3154. *In* Advances in neural information processing systems.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. arXiv. doi:10.48550/ARXIV.1412.6980

Le Quere, C., and others. 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Glob. Chang. Biol. **11**: 2016–2040. doi:10.1111/j.1365-2486.2005.1004.x

Lévy, M., R. Ferrari, P. J. Franks, A. P. Martin, and P. Rivière. 2012. Bringing physics to life at the submesoscale. Geophys. Res. Lett. **39**: L14602. doi:10.1029/2012gl052756

Li, W., D. S. Rao, W. Harrison, J. Smith, J. Cullen, B. Irwin, and T. Platt. 1983. Autotrophic picoplankton in the tropical ocean. Science **219**: 292–295. doi:10.1126/science.219.4582.292

Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal loss for dense object detection, p. 2980–2988. *In* Proceedings of the IEEE International Conference on Computer Vision. IEEE. doi:10.1109/iccv.2017.324

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. doi:10.48550/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. Cont. Shelf Res. **195**: 104056. doi:10.1016/j.csr.2020.104056

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 A **79**: 263–275. doi:10.1002/cyto.a.21035

Marrec, P., and others (2018). Coupling physics and biogeochemistry thanks to high-resolution observations of the phytoplankton community structure in the northwestern Mediterranean Sea. HAL preprint. Biogeosciences **15**: 1579–1606. doi:10.5194/bg-15-1579-2018

Metfies, K., and others. 2010. Contribution of the class cryptophyceae to phytoplankton structure in the german bight 1. J. Phycol. **46**: 1152–1160. doi:10.1111/j.1529-8817.2010.00902.x

Miloslavich, P., and others. 2018. Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. Glob. Chang. Biol. **24**: 2416–2433. doi:10.1111/gcb.14108

Neyshabur, B., S. Bhojanapalli, D. McAllester, and N. Srebro (2017). Exploring generalization in deep learning. Curran Associates, Inc. Advances in neural information processing systems, 30.

Ng, A. 2017. Machine learning yearning. Available from https://www.deeplearning.ai

Olson, R., D. Vaulot, and S. Chisholm. 1985. Marine phytoplankton distributions measured using shipboard flow cytometry. Deep Sea Res. Part A Oceanogr. Res. Pap. **32**: 1273–1280. doi:10.1016/0198-0149(85)90009-3

Pan, S. J., and Q. Yang. 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**: 1345–1359. doi:10.1109/tkde.2009.191

Partensky, F., J. Blanchot, and D. Vaulot. 1999. Differential distribution and ecology of prochlorococcus and synechococcus in oceanic waters: A review. Bull. Inst. Oceanogra. Monaco Spec. Num. **19**: 457–476.

Popel, M., and O. Bojar. 2018. Training tips for the transformer model. Prague Bull. Math. Linguistics **110**: 43–70. doi:10.48550/ARXIV.1804.00247

Rembauville, M., and others. 2017. Plankton assemblage estimated with bgc-argo floats in the southern ocean: Implications for seasonal successions and particle export. J. Geophys. Res. Oceans **122**: 8278–8292. doi:10.1002/2017jc013067

Ribalet, F., and others. 2015. Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre. Proc. Natl. Acad. Sci. **112**: 8008–8012. doi:10.1073/pnas.1424279112

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. doi:10.7717/peerj.2587

Saba, V. S., and others. 2011. An evaluation of ocean color model estimates of marine primary productivity in coastal and pelagic regions across the globe. Biogeosciences **8**: 489–503. doi:10.5194/bg-8-489-2011

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. doi:10.1016/j.hal.2020.101939

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. doi:10.48550/ARXIV.1409.1556

Sosik, H. M., R. J. Olson, M. G. Neubert, A. Shalapyonok, and A. R. Solow. 2003. Growth rates of coastal phytoplankton from time-series measurements with a submersible flow cytometer. Limnol. Oceanogr. **48**: 1756–1765. doi:10.4319/lo.2003.48.5.1756

Steinley, D. 2004. Properties of the hubert-arable adjusted rand index. Psychol. Methods **9**: 386–396. doi:10.1037/1082-989x.9.3.386

Sun, C., A. Shrivastava, S. Singh, and A. Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era, p. 843–852. *In* Proceedings of the IEEE International Conference on Computer Vision. IEEE.

Szegedy, C., and others. 2015. Going deeper with convolutions, p. 1–9. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE.

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**: e0196225. doi:10.1371/journal.pone.0196225

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. Front. Mar. Sci. **4**: 359. doi:10.3389/fmars.2017.00359

Wacquet, G., É. P. Caillault, D. Hamad, and P.-A. Hébert. 2013. Constrained spectral embedding for k-way data clustering. Pattern Recogn. Lett. **34**: 1009–1017. doi:10.1016/j.patrec.2013.02.003

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

Zhang, M., J. Lucas, J. Ba, and G. E. Hinton. 2019. Lookahead optimizer: K steps forward, 1 step back, p. 9593–9604. *In* Advances in neural information processing systems.

Acknowledgments

The authors thank Cytobuoy b.v. for their assistance to design special CytoClus4© features mandatory to conduct this work. The authors 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. The authors thank Stéphanie Barrillon and the participants of the FUMSECK cruise, and the captain and crew of the R/V Tethys II. MAP-IO is a scientific program led by the LACy/La Réunion University and was funded by the European Union through the ERDF program, the University of Reunion, the SGAR-Réunion, the région Réunion, the CNRS, the TAAF, the IFREMER and the Flotte Océanographique Française. The authors thank the technical team of the LACy engaged in the data acquisition and the maintenance of the instruments of the MAP-IO program. The authors are also very thankful to Maiwenn Hascoët, Gaëtan Viardot, and Chloé Caille for their participation in the manual gating process and in the data post-processing. 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 the ERDF 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.

Submitted 17 November 2021
Revised 25 February 2022
Accepted 04 May 2022

Associate editor: Tammi Richardson