General answer

Dear Editors and Referees,

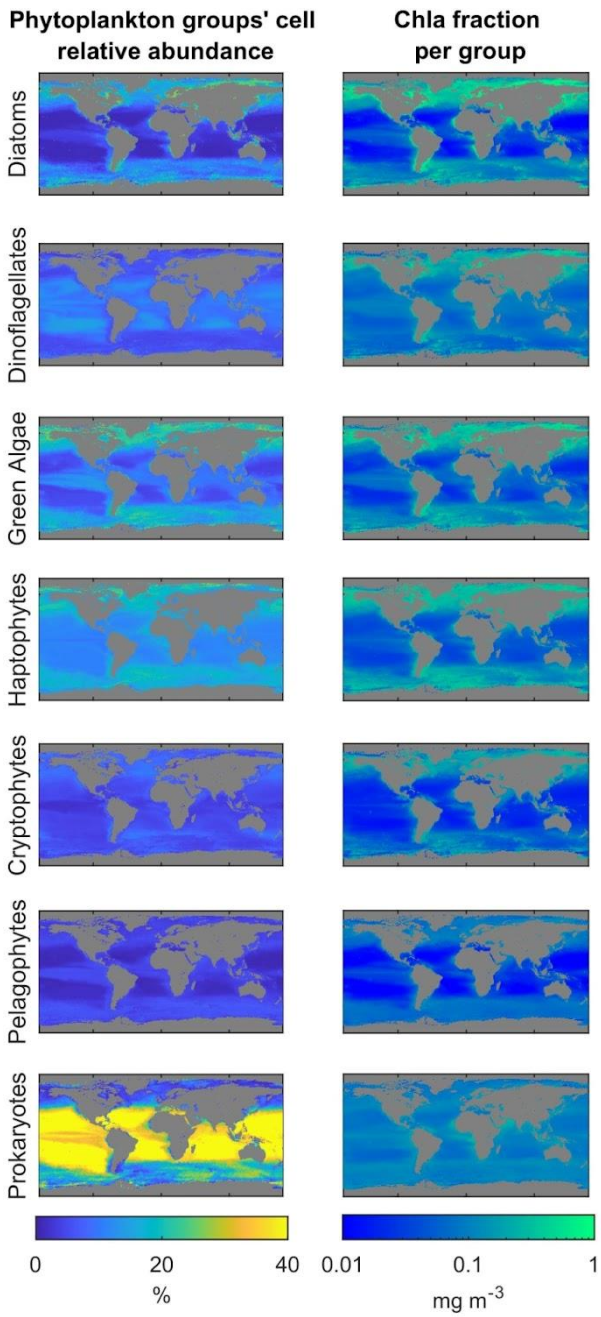
We would like to express our gratitude for the valuable comments and suggestions provided for improving our manuscript. We acknowledge the referee's observations regarding communication ambiguities and technical issues in the initial version, and we have prepared this revised manuscript to address these concerns and clarify the highlighted aspects.

This response aims to address the common major issues raised by both referees. We acknowledge that the development steps of the omic-based satellite algorithm in the paper were unclear, and the inclusion of a pigment-based approach for validation was misleading.

To clarify, our method is based on the link between omic and satellite data. The pigment approach only played a role in the post-training process to compare the outputs of both approaches. We incorporated the pigment HPLC data in this study due to its widespread use in ocean color remote sensing techniques for estimating phytoplankton groups, primarily because of its high data availability. However, it is important to note that numerous studies have demonstrated significant uncertainties between the pigment approach and phytoplankton abundance observed through other methods (Chase et al., 2020). These uncertainties arise from factors such as the overlapping presence of pigments across phytoplankton classes, photoacclimation, and physiological processes. Therefore, it is crucial to recognize that our study addresses two types and levels of information: Omics and Pigments. The use of pigments in this work is more for comparison purposes rather than validation, and we acknowledge that our previous message regarding this matter was misleading.

Additionally, the referees found it unconvincing to introduce physiological uncertainties when transforming omics data into Chla fraction per phytoplankton class. We introduced this aspect to compare the omic and pigment-based approaches.

Based on the comments from both referees, we have chosen to thoroughly revise the methodology section. We have added flowcharts to simplify the process and enhance its applicability for readers. The entire methodology has been revised in light of the suggestions provided by the referees. To address the concerns regarding chlorophyll-a fractionation and enable the emergence of different levels of information as outputs, we trained two algorithms using the same satellite data and SOM methodology. One algorithm provides the relative cell abundance of phytoplankton; SOMRCA (including the estimation of direct psbO relative abundance values), while the other algorithm estimates the phytoplankton Chla fraction per group; SOMChlF. Importantly, this revised methodology now considers the psbO occurrence per size fraction, which was not taken into account in the initial version of the manuscript. In both algorithms, uncertainties on the outputs were evaluated and therefore are presented with the outputs.



The outputs from both algorithms will allow us to address questions regarding phytoplankton diversity from an ecological perspective (through relative cell abundance) and a biogeochemical perspective (through Chla fraction per group), while considering physiological uncertainties.

We sincerely hope that this response is convincing and meets your expectations. We appreciate the thorough review process and are confident that the revisions have significantly improved the manuscript.

Figure 1: Different levels of information on phytoplankton groups. Noting that cell relative abundance (SOMRCA) and Chla fraction per group (SOMChlF) are two outputs of two different algorithms based on the same SOM methodology

Comments Referee #1

The work by El Hourany et al. describes machine learning techniques for application to (blue) ocean color data to determine the global distribution of phytoplankton functional types. A special focus is on the description of ML techniques with the identification of crucial features based on parameters of the merged GlobColour dataset. The details of the methods used are often cryptically written and difficult to follow, and reproduction of the methods and results is not possible. The methods section should be revised accordingly. Besides the application of ML methods in the context, the advantage of the method remains unclear and is not further specified; it could well be higher accuracy or computing speed. I recommend a thorough revision of the paper to describe the methods in a more understandable way and to prove the added value (also of future ML methods).

We thank the referee for the valuable comments and suggestions provided for improving our manuscript. We acknowledge the referee's observations regarding communication ambiguities and technical issues in the initial version, and we have prepared this revised manuscript to address these concerns and clarify the highlighted aspects.

In this paper, we approach the estimation of phytoplankton groups as a unified community structure to preserve inter-group coherence. Our objective was to develop a method capable of estimating all seven groups using a single set of satellite predictors. The challenge we faced was twofold: the problem was multivariate in nature, and the dataset was relatively small, with missing values in the satellite matchups.

To address these challenges and ensure that no valuable psbO measurements were lost, we turned to the technique of Self-Organizing Maps (SOM) and topology conservation. SOM is a powerful unsupervised learning algorithm that allows for the establishment and reproduction of relationships between variables. By utilizing SOM, we were able to fill in the gaps in the dataset and exploit the preserved topology to estimate the phytoplankton groups.

The advantage of using SOM in this context is its ability to handle multivariate data and preserve the underlying structure of the variables. It enables us to capture the complex relationships between the predictors and the phytoplankton groups, even with missing values. By leveraging the topology conservation property of SOM, we ensure that the estimated relationships are consistent with the overall structure of the data.

Specific comments:

The title is a bit catchy and inaccurate. It is rather about pigments, which are typical for color groups, but which can be very different in type of phytoplankton and corresponding genes.

Indeed, in this work, pigments were used, but not for training the method. We appreciate the referee's concerns regarding clarity, and we would like to address them.

The method we introduce in this manuscript is based on a dataset of phytoplankton groups quantified using the psbO molecular method and expressed in terms of Chla fraction, in combination with satellite variables. As described in the text, psbO is a single-copy gene that is present across all phytoplankton groups. This gene encodes proteins that structure a compartment of the Chloroplast photosystems.

It is important to note that pigments were not used for training the SOM method. Instead, they were employed solely for comparative purposes. The outputs of SOM-psbO (previous version of the algorithm) were compared to in-situ phytoplankton groups estimated using diagnostic pigment analysis (DPA), as described in studies such as Soppa et al. (2014). DPA methods have been widely used in remote sensing studies to estimate phytoplankton functional types (PFT) or size classes, and current operational methods such as Xi et al. (2020) and PHYSAT are based on them. However, it is crucial to acknowledge the high uncertainties associated with DPA methods, as highlighted by Chase et al. (2020). These uncertainties can lead to misleading interpretations of real PFT and phytoplankton size class relative abundance.

Therefore, it is important to clarify that diagnostic pigment data were not utilized in the development of the SOM method in both versions of the algorithm (old and revised). Their inclusion was solely for comparison purposes, to highlight the differences and uncertainties associated with the pigment-based approach.

We hope this clarification addresses the concerns regarding the use of pigments in our study.

The figures should all be revised, e.g. Fig. 5. Axis labels with units are often missing. Partly chlorophyll concentrations are given in log10, this is better in Fig. 2.

All figures were revised according to the referee's suggestion.

Line 87: Only as a comment that size fractioning often damages the cells, and such data should therefore be treated with caution.

We are aware of the drawbacks of size fractionation. The filters may retain cells smaller than the nominal pore because of net clogging, or because they were trapped in fecal pellets. On the contrary, long needle-like species and broken cells and colonies can pass through small mesh sizes. The patterns that we described in the current work based on size-fractionated samples can be complemented in the future by exploring non-fractionated samples. However, there is still no equivalent standardized sampling covering the main ocean regions as the size-fractionated samples from Tara Oceans.

For the discrimination of absorption features, rather the central visible region is necessary (e.g. Xi et al. 2015). In this respect, the use of the GlobColour data set with Rrs only up to 555 nm is unfavorable, as the correlation plots show. The OC-CCI dataset has more (MERIS) bands here and corresponding differences could be underlined. References to GlobColour and matchup procedure are missing.

As the referee correctly pointed out, the SOM-psbO method was trained using a dataset that included 17 variables, including satellite reflectance at 412, 443, 490, and 555nm. We apologize for not clearly indicating in the initial version of the manuscript that this method is specifically developed for open ocean applications.

In the clear open ocean, beyond 555nm, the information contained in the remote sensing reflectance (Rrs) bands is limited due to the strong absorption by water, as also mentioned in Xi et al. (2015). Our choice of the range and number of satellite reflectance bands was inspired by the work conducted in the PHYSAT method (Alvain 2005, Ben Mustapha et al., 2013), which is a classification method that utilizes reflectance anomalies in the four selected bands to identify dominant phytoplankton functional types.

To further support our argument, we rebuilt and cross-validated our methodology using different combinations of 15 bands ranging from 412 to 709nm. However, we found that increasing the number of bands did not lead to a significant improvement in performance. It should be noted that the Rrs bands selected, including the additional 670 nm band, are commonly measured by all sensors used to build the Rrs product of Globcolour. This overlapping of different sensors enhances data availability and coverage, thus increasing the importance of these Rrs bands within the initial dataset. The inclusion of the Rrs at 670 nm did not significantly impact the performance of either SOMRCA or SOMChlF, primarily due to the open ocean nature of the dataset.

It is important to note that one of the advantages of using machine learning methods such as SOM is to reduce the complexity of the problem while capturing non-linear relationships that are present in the environment. The correlations with the Rrs bands, which the referee mentioned as unfavorable in Figure 6, are indeed essential and statistically significant. It is crucial to consider that the problem we are addressing is multivariate. Preserving the inter-variable relationships, even those with lower correlations, is a major advantage of utilizing such a machine learning method.

It is a Case-1 approach for a medium range of chlorophyll concentrations, which should be communicated in a better way. Maybe flagging and an uncertainty product would be useful. Indeed, as clarified in the previous comment, the method developed in this paper is specifically designed for open ocean (case 1) applications. This statement has been further clarified in the revised version of the manuscript.

The approach proposed in this paper to estimate phytoplankton groups from satellite data is based on an unsupervised neural classification technique, specifically the Self-Organizing Map (SOM). The SOM summarizes the non-linear relationship between the satellite data and phytoplankton groups, effectively reducing noise and mitigating the influence of uncertainties within the dataset.

The function that links the predictors (satellite data) to the predicted variables (phytoplankton groups) is represented by an allocation function based on a weighted Euclidean distance. In other

words, this function searches for and associates the closest neuron in the SOM to a new or unfamiliar observation.

The main source of uncertainties in the estimation process lies in the allocation function. Among hundreds of neurons in the SOM, one neuron is chosen as the assignment based on the minimum distance between the neurons and the pixel, regardless of whether the distance is strong or weak. Since one of the properties of SOM is the preservation of topology (where neighboring neurons are similar), a pixel can be assigned to several adjacent neurons, with a distance order, representing a neighborhood of close neurons.

Now, how do uncertainties in the satellite variables influence the allocation function and, consequently, the results?

If the distance between a pixel and a neuron is small, the influence of uncertainties is minimal and will not significantly affect the assignment of the pixel. However, if a large distance is observed between the observation and the assigned neuron, uncertainties in the variables can have a greater impact on the choice but remain within the bounds of the chosen neuron's neighborhood.

To consider all the uncertainties associated with the allocation function, we have chosen to associate each pixel with a weighted standard deviation based on the first 10 closest neurons. The weights correspond to the distances between the first 10 matching neurons and the pixel. This allows us to incorporate uncertainties into the assignment process and provide a measure of confidence for each pixel's assignment.

By considering the weighted standard deviation, we account for the influence of uncertainties in the satellite variables and provide a more comprehensive understanding of the allocation process within the SOM.

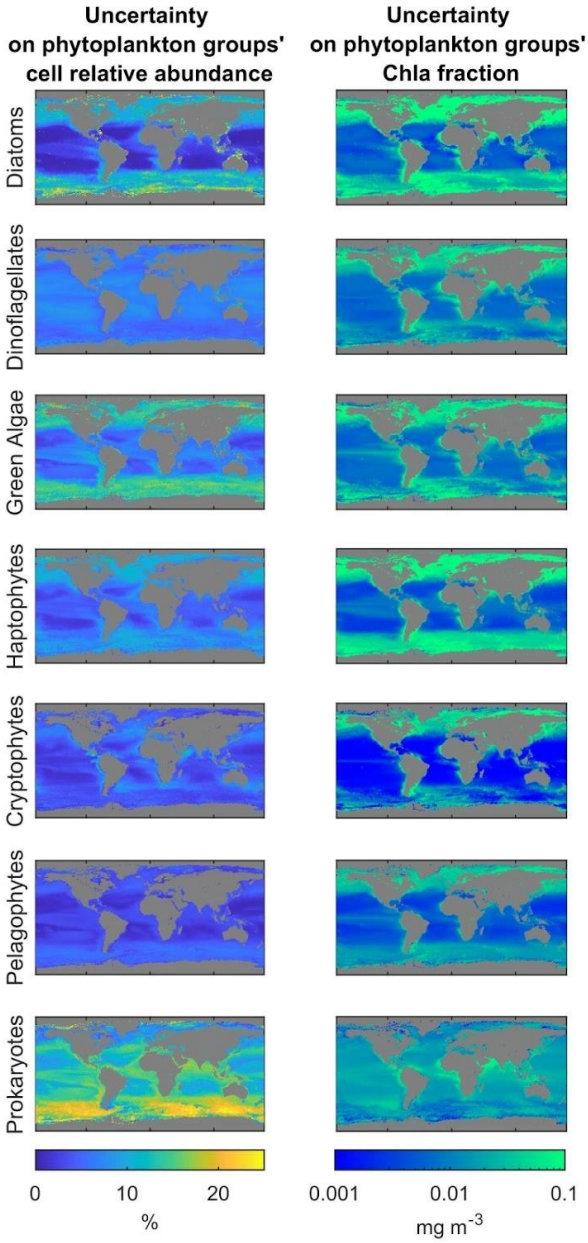


Figure 2: Global uncertainties regarding phytoplankton groups' cell relative abundance, Chla contribution (SOMRCA), and Chla fraction (SOMChlF). In this context, the following uncertainties on the outputs represent the interval (defined with a standard deviation calculated on the neighboring associated neurons per satellite pixel) of SOMRCA and SOMChlF to estimate the different phytoplankton groups.

However, in such open ocean conditions, HPLC methods are often at the limit (if low volumes of water are filtered) – extreme uncertainties may exist in the fundamental training data.

The very deep sequencing of the *Tara* Oceans metagenomes (between ~10^8 and ~10^9 total reads per sample) allows high detection power (e.g., for rare species). In addition, filter volumes were high: 100 L for 0.22-3, 0.8-5 and 5-20, 1-20 m3 for 20-180, and 10-100 m3 for 180-2000.

Besides SST is salinity actually a strong indicator for some PFTs.

SSS is a strong indicator of some PFTs due to intervariable correlations, and their patterns are related to physical conditions, like the ones of SST. However, SSS satellite products are not as accurate as SST products and at a lower resolution (best resolution at 25kms vs. 4kms). The addition of Satellite SSS products might corrupt the output of the operational phase.

Method part is unclear, especially lines 163-212. A part of the problem could be that less common naming conventions are used, e.g. do you refer to neural network architecture if you optimize the size map? How does the final map or architecture look like?

We acknowledge the referee for highlighting these communication issues. We introduced a clearer definition of the SOM size; We refer to the number of neurons represented by n=p x q, where p and q are the dimensions of the SOM 2D neuron grid.

Line 269: The more parameters we utilize, the more we must trust the data quality. Nevertheless, seen over the global ocean, there are many uncertainties in all mentioned parameters and regions. Especially Rrs in blue bands and the retrieved chlorophyll concentration must be considered as critical, even more because reflectances are derived from multi-mission merged data with sensor-specific atmospheric correction.

The question raised highlights the importance of considering data quality when utilizing parameters. In the context of the global ocean, there exist numerous uncertainties associated with the mentioned parameters and regions. As mentioned in the previous comment regarding uncertainties, the SOM process attunes uncertainties and enables the possibility to estimate uncertainties in the outputs. This has been implemented in the second version of our algorithm.

The marine model of ocean color algorithms is for atmospheric correction and chlorophyll retrieval is mostly based on a diatom-like chlorophyll-specific absorption and scattering behavior (e.g. Bricaud et al., 1995). Thus, good that there is relatively high correlation of diatoms and chlorophyll concentration. But what is actually with features that are not captured, e.g. specific optical properties of Coccolithophores (e.g. Balch, 2018)? There is a high abundance, e.g. in The Great Calcite Belt, where Fig. 7 indicates high reliability of the model with a C2 distribution in Fig. 10, that seems to be different. I see some question marks and would ask for more careful discussion about the model uncertainty.

We admit that within the first version of the algorithm, since we didn't take into consideration the effect of size per group and per sample, the Chla fraction concentration per group was biased. The pos-training classification (Figure 12 in the revised manuscript, section 4.3) into dominant phytoplankton communities was revised accordingly after incorporating the phytoplankton size information as described in Sommeria-Klein et al 2021 Science:

$$Chla\ fraction_{PFT} = \mathrm{Chla}_{in-situ} * \frac{\sum_{s=1}^{4}\left(\frac{psbO_{PFT} * size_s}{\sum_{PFT=1}^{7}(psbO_{PFT} * size_s)}\right)}{\sum_{s=1}^{4}\sum_{PFT=1}^{7}(psbO_{PFT} * size_s)}$$

Therefore, when converting psbO reads to relative abundance, considering the size of the phytoplankton cell for each group, we highlight the contribution of each group's size to the total chlorophyll-a (Chla) concentration.

Compared to the previous version, and due to the data conversion, five clusters turned out to be sufficient to describe the dominant patterns. In the Southern Ocean, the C3 group emerges and dominates, while there is also a higher relative abundance of Haptophytes and Diatoms. In the Arctic Ocean, the C4 group dominates. Although the phytoplankton communities of both C3 and C4 clusters were relatively similar, the optical signal was significantly different, allowing us to distinguish between the two clusters.

It is unclear how the new method behaves compared to the mentioned operational model by Xi et al. (2020). What are the advantages of the presented method?

Xi et al., (and the SOM-Pigments method) is based on the DPA pigment approach to identify phytoplankton groups (4 functional types).

The method described in this paper is developed with a harmonized database on the phytoplankton taxonomic community structure based on the psbO gene quantification. Molecular methods like this have a deep taxonomic resolution (including for cryptic species) as well as high detection power (e.g., for rare species). In addition, this particular gene is present in all phytoplankton groups, eukaryotes, and prokaryotes alike, with a single copy per cell.

Quantifying it using satellite data provides an unbiased picture of phytoplankton cell relative abundances.

Comments Referee #2

The authors develop a machine learning approach to link ocean colour data and in situ omics to improve detection of phytoplankton functional types and groups from space. The topic they are dealing with is innovative. However, the methodology and algorithm development steps are hard to follow and need to be revised to make the workflow clearer to the reader. In this scope, a flowchart is essential.

I am not fully convinced by the validation approach of the method. The training is done using the whole omics database and cross-validation statistics show the good prediction capabilities of the model. Then, the validation is made with an external database built on HPLC-based information. From my point of view, this cannot be considered a proper validation because one quantity is based on HPLC data, the estimated one on omics data. Such a comparison thus implies that the two approaches bring the same level of information on phytoplankton taxonomy. In this case, there would be no need to develop a new approach based on omics. However, as discussed at the end of the paper, HPLC- and omics-based phytoplankton information have some degree of correlation, which is good because this means that OMICS information can be found in optical properties to some extent and OMICS based approaches are welcome because they will bring new and complementary information on phytoplankton from space.

I realize that the OMICS database used to develop the new ML approach is small, but probably the authors might think to train the model over 70% of the database and validate it with the remaining 30%.

Results need to be discussed more and the text about retrieved global distribution of phytoplankton and biomes needs to be profoundly checked and revised.

The work thus needs to be deeply revised to improve the methodology and make the validation stronger as well as the text more readable.

We would like to express our gratitude for the valuable review. We acknowledge the referee's observations regarding technical issues in the initial version, and we have prepared this revised version while applying the referee’s suggestions.

We would like to admit that the reasoning behind validating with a pigment-based approach was misleading. For that, we chose to follow the referee’s major comment and evaluate the algorithm using a two-step procedure:

We split the Tara Oceans psbO dataset into 80% to train the SOM, and 20% as a test set.

1- During the SOM training based on 80% of the dataset, a different combination of satellite variables was used to determine the best set of variables to estimate the 7 phytoplankton groups in terms of relative cell abundance and Chla fraction.

- Per a combination of variables, we increase the number of neurons to determine the optimal size of the SOM from 10 neurons to 1000 neurons.

- For each number of neurons used, the quantization and topographic errors related to the SOM are calculated and a one leave-out cross-validation procedure is performed to assign performance metrics (R2 and RMSE) to help choose the best SOM size and satellite variables combination.

The best SOM configuration and variable combination are based on the lowest errors and highest R2 values.

2- The chosen SOM is tested using the 20% test set, providing an independent set of performance metrics.

As a result, we present in the paper the performance metrics of the best SOM configuration based on the cross-validation procedure and the test set.

The comparison with the HPLC DPA approach will be introduced for comparative purposes only.

Specific comments:

Figure 1 is misleading as the same color palette has been used for both columns though the % axis are different from left to right. A quick reader could interpret the yellow dots of (e.g.) Cryptophytes as abundant as Green Algae or Diatoms.

Indeed, according to the referee's comment, we homogenized the color scale. And since we derive two types of information from the psbO dataset i.e relative cell abundance and Chla fraction per group, a different color palette was used for each new dataset.

Line 91: this statement means that we have phytoplankton also in the 180-2000 um size class, which is possible in case of diatoms chains. Could you provide a distribution of frequency of phytoplankton groups within each size class? This would help the reader to have a wider image of the type of phytoplankton in the database (and especially for those chain-forming species and classes spanning a wide size range).

The distribution of taxonomic groups between size fractions in the psbO dataset is displayed in Fig 2a-b, Fig 7a, Fig 8 and Figure S16 in Pierella Karlusich et al 2023 Mol Ecol Res.

We provide boxplots to illustrate the distribution of the phytoplankton groups per size filter.

The filters may retain cells smaller than the nominal pore because of net clogging, or because they were trapped in fecal pellets. On the contrary, long needle-like species and broken cells and colonies can pass through small mesh sizes. The patterns that we described in the current work based on size-fractionated samples can be complemented in the future by exploring non-fractionated samples.

Line 115: why normalizing omics data on Chl? Because Chl varies according to the physiological status of phytoplankton, a photoacclimation component is re-introduced (which is a major problem in the DPA analysis). Why not using OMICS-based % of the whole population?

Indeed, this type of normalization introduces physiological uncertainties in the data. However, it was judged important to achieve quantity relative to Chla which is often used as a proxy of biomass, and which is a relevant parameter for energy and matter fluxes (e.g., food webs, biogeochemical cycles). Adding to this, this Chla normalization allows to compare this quantity with what we can observe using DPA pigments approach and current satellite operational products.

But, as mentioned in the general answer two types of algorithms are developed to address this issue:

**To address the concerns regarding chlorophyll-a fractionation and enable the emergence of different levels of information as outputs, we trained two algorithms using the same satellite data and SOM methodology. One algorithm provides the relative cell abundance of phytoplankton (including the estimation of direct psbO relative abundance values), while the other algorithm estimates the phytoplankton Chla fraction per group. Importantly, this revised methodology now considers the psbO occurrence per size fraction, which was not taken into account in the initial version of the manuscript. In both algorithms, uncertainties on the outputs were evaluated and therefore are presented with the outputs.

The outputs from both algorithms will allow us to address questions regarding phytoplankton diversity from an ecological perspective (through relative cell abundance) and a biogeochemical perspective (through Chla fraction per group), while considering physiological uncertainties. **

Lines 121-123: it is not clear which data are interpolated. In situ or satellites?

In-situ and Satellite data are both interpolated within the initial data space during the training phase, since missing values can be present in both cased, and since the number of neurons that we are dealing with is greater than the amount of observation. This is monitored through the training process with both the quantization and topographic errors related to SOM.

Table 2 contains mistakes on the coefficients. From Uitz et al. (2006), the coefficient for Chl-b is 1.01 while 0.35 is for 19-BF. Fixed, we apologize for this mistake.

In addition, 19-BF is here only attributed to the pelagophytes while is also a pigment within haptophytes (except coccolithophores). So, from the current coefficients all haptophytes only contain 19-Hex.

Indeed, in the study by Chase et al. (2020), it was demonstrated that the presence of pigments overlaps within size classes and types of phytoplankton. Several ocean color studies, such as those by Hirata et al. (2011) and Xi et al. (2020), have attributed the 19Hf pigment to Haptophytes.

Taking into account the reviewer's comments and the uncertainties associated with pigments, we decided to list the major phytoplankton groups and indicate the most representative pigment for each group.

Line 155: which cross-validation procedure? Do these statistics refer to all pigments or is it a global indicator for the technique?

We acknowledge that the sentence citing the statistics was unclear. The reported statistics, a regression coefficient of 0.75, and an average RMSE of 0.016 mg.m-3, represent a global indicator for the technique. They reflect the mean error and regression coefficient across the 10 estimated pigments and are given following a cross validation procedure conducted using a one-leave-out random pick from a global HPLC dataset constituted of 12 000 HPLC observations.

Line 163: Please indicate and explain better which are the "several machine learning algorithms" you tested and why a SOM has been chosen. This will be very helpful for scientists approaching the same problem.

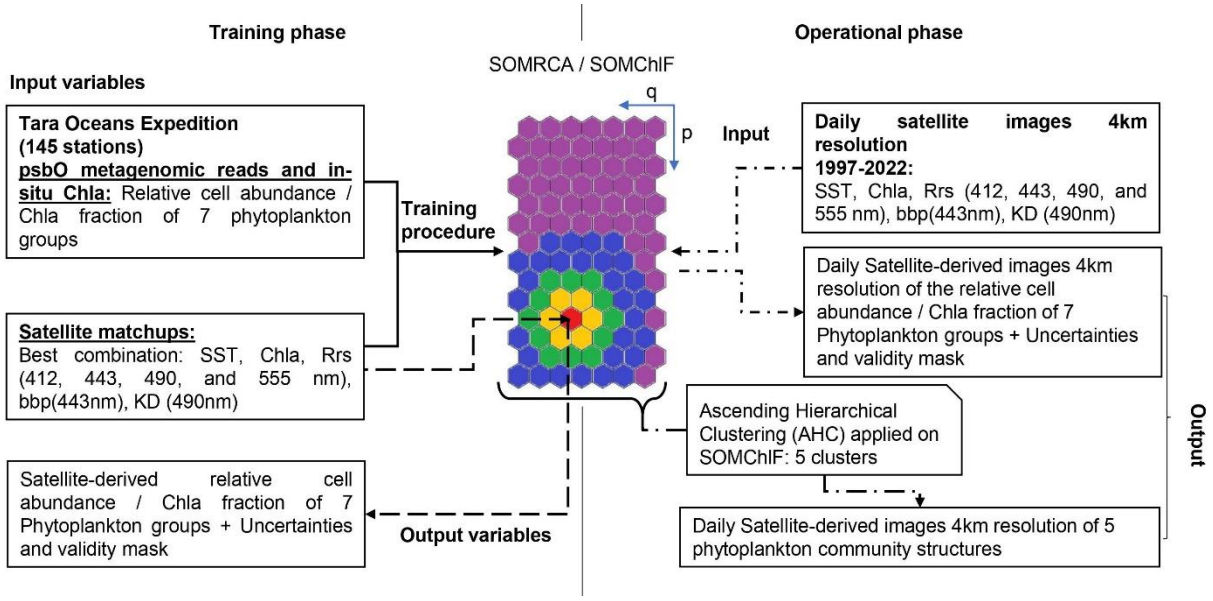
We would like to clarify the sentence introducing the machine learning algorithms used in our study: SOM, hierarchical ascending clustering (HAC), and Random Forest.

Developing an operational algorithm that estimates the abundance of phytoplankton groups from satellite information was achieved using these algorithms. Firstly, the SOM algorithm was utilized to train a model based on the psbO pigment dataset. This allowed us to identify global large-scale patterns and characterize phytoplankton biomes. Last, to explain the potential divergence between the DPA approach and psbO measurements, we employed a Random Forest approach. This analysis highlighted the cumulative importance of pigment composition in estimating the abundance of phytoplankton groups.

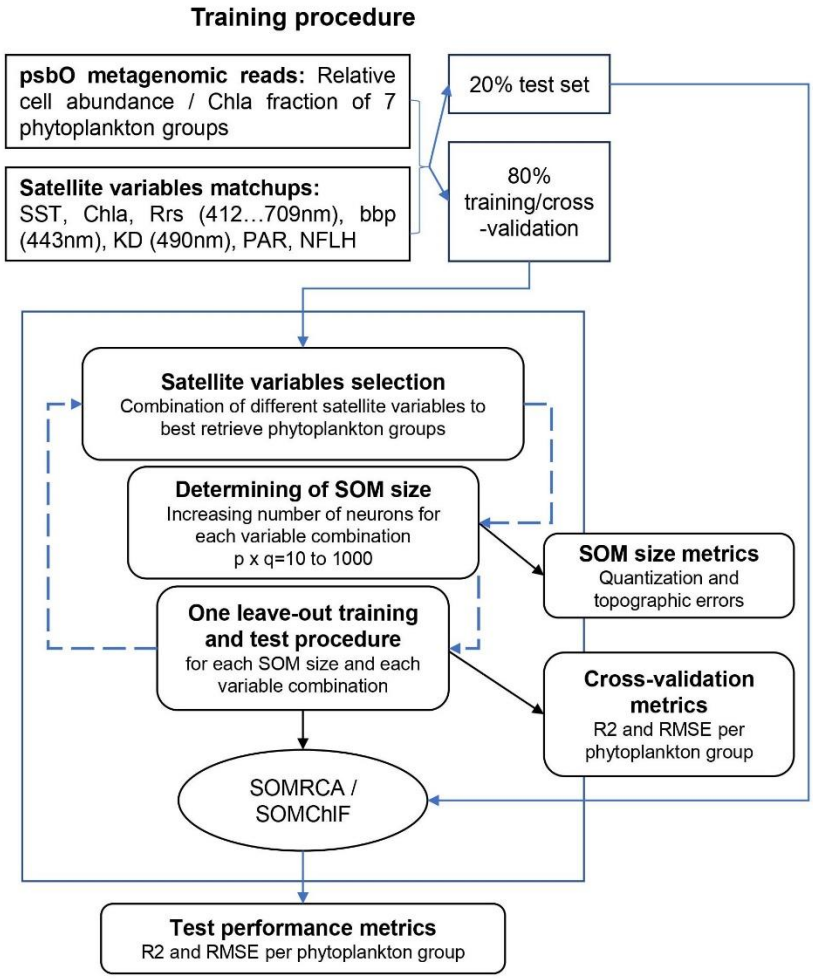
We tested approaches based on Feed-Forward Neural Networks. However, due to the limited number of observations in the dataset, these approaches were not very conclusive. The choice of SOM was based on the previous work by El Hourany et al. (2019), which demonstrated improved performance with an increasing number of neurons, the number of neurons almost twice compared to the observations in the initial dataset, accounting for missing values.

In the following section, each methodology and algorithm are explained in detail. Section 3.1 need to be rewritten and a flowchart added. That's strange to see 3.1.1 and 3.1.2 as two different sections when (if I had well understood) the work is done simultaneously. Figure 5: y- and x- axes should be the same and indicate the name of the solid and dashed lines in the caption.

We apologize for the misleading sectioning. To better clarify the methodology, a flowchart was added and both above-mentioned sections were merged according to the methodology as the reviewer mentioned. Indeed sections 3.1.1 and 3.1.2 are done simultaneously, and iteratively as shown in the new flowchart #2.



Flowchart 1: General scheme of the SOM methodology to estimate phytoplankton groups from satellite data.



Flowchart 2: A focus on the training phase of the SOM which is based on an iterative procedure between different satellite variable combinations and SOM grid size. The choice of the best satellite variables combination and SOM size were based on consensus of low errors and high R2.

Line 191: which several experiments? How many? Please explain better.

The SOM grid size was sampled between 10 to 1000 neurons with a step of 10. Therefore, there were 100 SOM grids that were tested for each variable combination.

Section 3.1.3 needs to be clearer.

Line 269: what is the impact of interpolation on bbp and Kd? (i.e., Interpolation declared in the methods)

Below is a comparison of SOM-psbO and the initial dataset's values for each variable including bbp and Kd. For a SOM grid size of 242 neurons, the SOM was able to catch the values' distribution for both parameters.

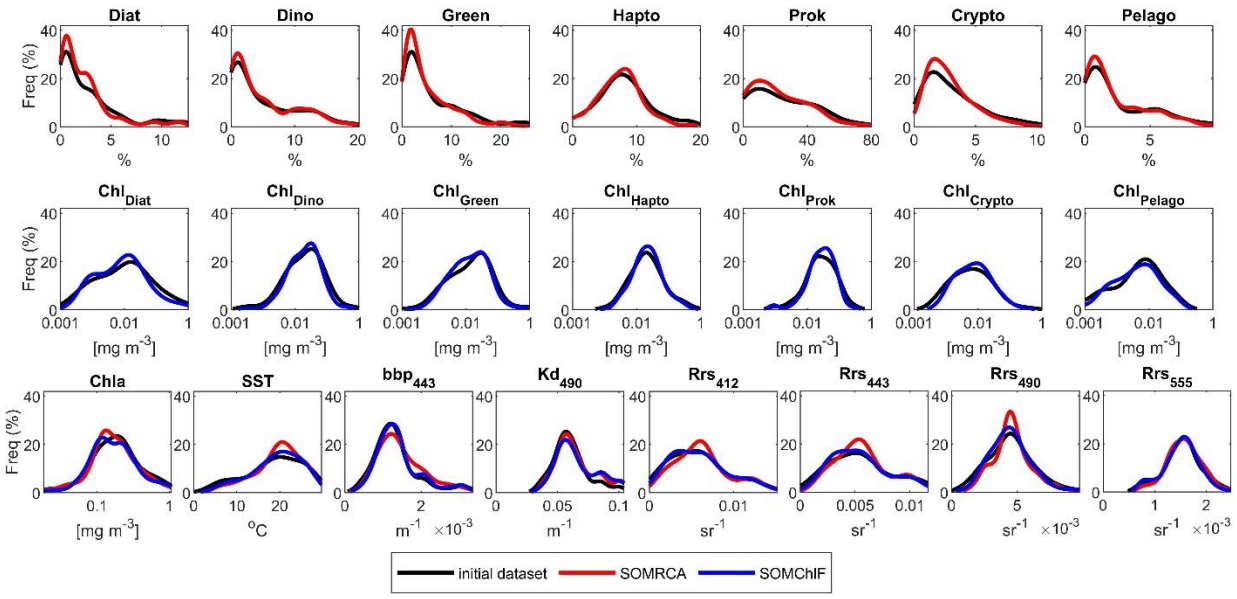


Figure 3: Distribution of values for each variable in the initial dataset and SOMRCA and SOMChlF neurons.

Line 275: from Table 3, pelagophytes instead of cryptophytes

Indeed, we apologize for this error.

Line 314: generally speaking, are you referring to the surface-to-volume ratio?

We have corrected the term: 'biovolume-to-size' was replaced by 'surface-to-size.'

Line 329 and Line 331: please check and discuss: C4, C5 and C6 are dominated by Prokaryotes, but these areas are generally known to be dominated by large phytoplankton. Same for C1, dominated by diatoms but in the subtropics. In addition, it would be nice to see these clusters plotted on map in Figure 10.

We admit that within the first version of the algorithm, since we didn't take into consideration the effect of size per group and per sample, the Chla fraction concentration per group was biased.

The pos-training classification into dominant phytoplankton communities was revised accordingly after incorporating the phytoplankton size information as described in Sommeria-Klein et al 2021 Science:

$$Chla\ fraction_{PFT} = \mathrm{Chla}_{in-situ} * \frac{\sum_{s=1}^{4}\left(\frac{psbO_{PFT} * size_s}{\sum_{PFT=1}^{7}(psbO_{PFT} * size_s)}\right)}{\sum_{s=1}^{4}\sum_{PFT=1}^{7}(psbO_{PFT} * size_s)}$$

Therefore, upon converting psbO reads to relative abundance accounting for the size of the phytoplankton cell per group, we highlight the size contribution of each group to the total Chla.

Compared to the previous version, and due to the data conversion, five clusters turned out to be sufficient to describe the dominant patterns (Figure 4).

Figure 10: How the spectra have been normalized? By the minimum? The spectral shape should be discussed.

Each wavelength was normalized by its values distribution variance within the dataset. We are providing a description and a discussion of both phytoplankton distribution and for the spectral signal in the section 4.3. of the revised manuscript.