Dear Editors and Referees,

We would like to express our sincere appreciation for the insightful comments and suggestions that have significantly contributed to improving our manuscript. We have carefully considered the referees' feedback, and in this revised version, we have addressed the concerns and provided comprehensive clarifications as suggested.

Specifically, we have restructured the referencing in line with the referee's recommendations, aiming to provide a more accessible and illustrative discussion rather than relying solely on specific algorithms. Responding to the referee's insightful suggestion, we have introduced metrics to evaluate our SOM methodology, which has shed light on the relative errors inherent in both psbO-based algorithms. This key insight underscores the complexities associated with the errors of SOMRCA in estimating phytoplankton relative abundances when compared to the estimation of Chla fractions per phytoplankton group using SOMChlF. Last, we have discussed the uncertainty associated with SOMRCA and SOMChlF, emphasizing its implications in contrast to previous studies that utilized the DPA approach.

We have taken great care to ensure that the manuscript and the responses provided in this document are aligned and effectively address the concerns raised by the referees.

# Response to referee #1

Hourany et al. have been developing a machine learning based algorithm trained on several remotely sensed products (RRS, bbp, Kd490, SST, CHL) combined with omics-based biomarker developed from the RV Tara Ocean data set to obtain cell abundance and fraction to total Chla of seven major marine phytoplankton groups. They have evaluated their algorithm with cross-comparison, independent validation and intercomparison to similar satellite products. While I think overall the method development seems to be robust and documented, the manuscript lacks especially:

a) correctly referencing other work done in the field of phytoplankton measurements, analysis and especially PFT algorithm development,

b) several details in the two chapters "Materials" and "Methods", and

c) discussion on their algorithm performance regarding pixel uncertainty, cross-validation, independent validation and intercomparison results.

Below I detail further these shortcomings.

Because of this I think the manuscripts require in these aspects substantial revision before it can become accepted, while most of the other parts can mostly remain.

Detailed comments:

1. It would be good also to have a list of abbreviations in the supplement. There are so many abbreviations used and parameters listed in the manuscript, it becomes confusing.

We added a list of acronyms in the end of the main document (Table 4)

2. Introduction: at several sentences the references provided are not clear or correct or do not merit former work executed in the field:

a) Line 34-35: that is a very sloppy statement "… a range of ecological and biogeochemical problems" What is meant by problems?

We meant by the use of the word "problems" to address various scientific questions

This has been changed in the text to make it clearer:

*This interest has facilitated the integration of the concept of phytoplankton functional types (PFT) and taxonomic groups (PG) into studies exploring various ecological and biogeochemical aspects (Le Quéré et al., 2005; Hood et al., 2006).*

b) Line 39 ff. it is not clear if the methods developed to detect "… abundance of PFT and SC are also meant to be based optical characteristics – since this is clearly stated for the "specific taxa" this should also be clarified here and the references provided then should match the specific method principle. I recommend then to cite here overview papers (see IOCCG 2014, Mouw et al. 2017, Bracher et al. 2017) or at least to put "e.g." since the citations provided are far from complete. In addition, Alvain et al. 2005 and Ben Mustapha et al. 2013 retrieve dominant groups and no abundances, and Chase et al. 2020 method does retrieve PSC from satellite ocean color data, it assessed the diagnostic pigment method based on in-situ data for phytoplankton size classes.

The paragraph has been modified according to the referee's suggestions.

c) Line 48 ff. : should also merit Brewin et al. 2010. A three-component model of phytoplankton size class for the Atlantic Ocean. Ecological Modelling, 221(11), pp.1472-1483. – I would put "e.g." since this list is far from complete!

The statement has been modified accordingly.

d) Line 52 should reference to Brewin at al. 2015 not 2014!

The reference has been modified accordingly.

e) Line 62 (also Methods chapter 2.3.1): You say you downloaded the Xi et al. product from the Copernicus website – if it was after July 2021, it most probably is the product based on Xi et al. 2021 which includes the SST as variable to constrain the algorithm.

We apologize for the confusion, we indeed used the newest version of Xi et al (Xi et al., 2021).

Therefore, we rectified the description of this product.

3. Material & Method sections:

a) a flow chart (Fig. 4) is provided for the SOM DRCA & DChlF data sets – however, everything else connected to methods applied in study is lacking. Since you did many different other parts (DPA three coefficients averaging for HPLC data global and Tara, uncertainty assessment, satellite product intercomparison, cross validation, etc.) – it would be good to have an overview.

To enhance the transparency and comprehensibility of the methodology, we have updated the flowcharts according to the referee's suggestion, providing a detailed overview of each step in the algorithm.

To manage the complexity of the process, we have introduced three sub-flowcharts: one outlining the general training procedure, another focusing on the parametrization of the Self-Organizing Map (SOM) and the selection of variables within the training procedure, and a third delineating the operational phase. These figures have been included in the supplementary materials, and in the main text, a statement has been added referring to these flowcharts.

These flowcharts are intended to provide an accessible and comprehensive understanding of the algorithm, ensuring that readers can navigate and comprehend the various stages of our approach.

b) Chapter 2.1.1- line 75 ff.: It is not clear why stations are discarded when not all 5 size fractions were contained in a station sample – for me it does not make sense from an ecological standpoint. In addition, you do not mention how many stations were then excluded.

We apologize for this error, we indeed verified and there were no stations among the 145 stations that have been discarded. All stations have been utilized and size fractions were aggregated into an average value of relative psbO read per group. We have described the source data in a supplementary figure S1.

***Added:*** *Among the 210 Tara Oceans stations, 145 stations sampled psbO reads in different ocean regimes from oligotrophic to eutrophic waters (Chla from 0.01 to 10 mg.m-3, median at 0.3 mg.m-3, from 2009 to 2013. Seawater samples were filtered in order to differentiate five planktonic size fractions (0.22-3um, 0.8-5um, 5-20um, 20-180 um, 180-2000 um). For the purpose of this study, we pooled the five size fractions into a single aggregated sample.*

Also add the information what exact values for the weights were taken for each size fraction to obtain their chl-a fraction.

First, as mentioned in the text, all phytoplankton groups in each size class were weighted equally by the mid-value of the size range, i.e., x0.9 for the first size class [0.6-1.2], x2.9 for the [0.8-5] size class, x12.5 for the [5-20] size class, and last x100 for the [20-180] size class. Applying equation 1 pools all size fractions per group while considering the *psbO* read values and the size factors mentioned above.

Why do these weight values make sense for the conversion?

The *psbO* measurements are proxies of relative cell abundance since this protein-encoding gene is generally present as a single-copy and is found in all phytoplankton groups. For example, if we take a huge diatom compared to a tiny *Synechococcus*, both have 1 *psbO* gene and therefore are counted as 1 within the *psbO* quantification. However, we know that a diatom's Chla content is way greater than that of *Synechococcus* (Agustí, 1991; Fujiki and Taguchi, 2002; Dairiki et al., 2020; Bock et al., 2022). This is where the conversion via size-dependent weights is essential in the case of Chla content estimation.

In Line 82 it is not clear what 5% here means – relative to the total abundance in each size class or for each size class?

5% of the total cell relative abundance among all size classes.

**In response to the questions of the referee, we decided that it is essential to add further clarification on this aspect to the manuscript (see section 2.1.1)**

c) Chapter 2.1.2, line 205: Add more information by providing exactly the 11 bands used from 412 to 670 nm from the RRS data set.

The 11 Rrs bands were: 412, 443, 469, 490, 510, 531, 547, 555, 620, 645, and 670 nm. Added accordingly in the text.

d) Chapter 2.2: Overall, I wonder why not much more HPLC data have been used for your algorithm validation. E.g., you cite Xi et al. 2020 – then you should be aware of the much bigger pigment data set used in this work (taking advantage of the compilation in Losa et al. 2017). Further check also identification on the error in LTER Palmer HPLC data in Xi et al. (2021) – it may also affect already your compiled data set.

Thank you for pointing out the existence of a more extensive dataset in Losa et al., 2017. It is important to clarify that the HPLC dataset was not solely employed for validation purposes, but rather for comparing the estimations of phytoplankton groups using two different methodologies: the DPA and the *psbO*-derived satellite algorithm. These methods are based on distinct assumptions and resolutions of phytoplankton groups. Using the DPA as a direct validation for the *psbO* data presents challenges. The estimation of phytoplankton groups using pigments is inherently imperfect and relies on assumptions that introduce considerable variability and bias in determining the contribution of specific pigments to the assessment of phytoplankton groups.

During the first phase of the review process, referee #2 raised concerns regarding labeling this comparison as an independent validation and suggested a thorough review of the validation scheme of our algorithm. Consequently, in the revised version of the paper we have re-evaluated the validation process of our algorithm as recommended by referee #2 while introducing a test set validation as explained in the paper.

Therefore, the HPLC dataset was primarily utilized to compare different levels of information and demonstrate the agreement between HPLC and *psbO* data. The database we employed is deemed sufficient to address the questions posed.

NB: Both Losa et al., 2017 and our compiled HPLC database share many common sources, particularly the compiled database of MAREDAT, which constitutes a major part of both datasets.

Finally, before your paper becomes accepted, the compiled HPLC data set with the diagnostic pigments, total chl, and retrieved PFT chl-a conc. should be made available to the readers (e.g., by storage in a public repository).

All *psbO*, HPLC, and satellite matchups datasets will be made available to the community in a public repository, alongside the SOMChlF and SOMRCA algorithms with their operational functions. A statement will be added in the acknowledgment.

e) Chapter 2.2: Why did you choose to apply for the dpa method using the 3 sets of coefficients proposed by Uitz, Brewin, Soppa and that then taking from these calculations the average fraction. You should at least somewhere discuss why you followed this method, instead of just using the coefficient proposed by one of author (I would rather recommend then the newest citation – actually newer ones have been published since then).

The selection of the three sets of coefficients was based on their estimation using global HPLC datasets. While there are newer data sets available, such as those derived by Brewin et al. (2017) and Chase et al. (2020), it is important to note that these are primarily developed using HPLC data at regional or basin scales, as demonstrated in the case of the northern Atlantic Ocean in the examples mentioned.

We are grateful to the referee for bringing the study of Losa et al., 2017 to our attention. In this revised version, we have utilized the coefficients tuned on a global HPLC dataset by Losa et al. (2017).

A thorough examination of the values assigned to the coefficients by these four studies reveals disparities that do not consistently align across all pigments. Notably, while the coefficients for diatoms exhibit similarity across the four sets, differences arise, for instance, in the case of prokaryotes, only Brewin et al. (2015) and Uitz et al. (2006) show close coefficients associated with Zea, while in the case of haptophytes, where only Brewin et al. (2015) and Soppa et al. (2014) estimates similar coefficients attributed to 19HF. The discrepancies can be attributed to variations in the datasets utilized for coefficient estimation and differences in the methodologies employed.

To ensure the robustness of the results and to account for the diverse outputs stemming from the utilization of these coefficients, we opted to compute the average of the outputs from the three sets of coefficients in the previous version, now from four sets of coefficients while adding Losa et al., 2017.

***Added:*** *An examination of the values assigned to the coefficients by these four studies reveals disparities that do not consistently align across all pigments. Notably, while the coefficients for diatoms exhibit similarity across the four sets, differences arise, for instance, in the case of dinoflagellates, only Brewin et al. (2015) and Uitz et al. (2006) show close coefficients associated to Perid, while in the case of haptophytes, where Brewin et al. (2015), Soppa et al. (2014) and Losa et al., (2017) estimates close coefficients attributed to 19HF. The discrepancies can be attributed to variations in the*

*datasets utilized for coefficient estimation and differences in the methodologies employed. We chose to do an average of the output of the four sets of coefficients to increase the robustness of the results while considering the different outputs of the utilization of these coefficients.*

f) Chapter 2.3.1: mind to check if the basis of the CMEMS global PFT product is really Xi et al. 2020 (see comment 2e)– add also the version number of the product in the description. In any case the product is not provided from 1997, but only from 2002 onward. In any case you description that this algorithm uses 15 bands is not correct at all. Please carefully check and provide a correct description.

We apologize for the confusion, we indeed used the newest version of Xi et al.

Therefore, we rectified the description of this product.

**Added:** *This Globcolour product contains the concentration of each phytoplankton functional type (expressed in terms of Chla concentration fraction) based on the Xi et al., 2021 algorithm, processed from 2002 to the present. This algorithm estimates the Chla concentration of diatoms, dinoflagellates, haptophytes, green algae, and prokaryotes. The algorithm was implemented using HPLC-based phytoplankton groups using the DPA approach (Losa et al., 2017, Soppa et al., 2014) merged to OC Rrs products (412, 443, 490, 510, 531, 547, 555, 670, and 678 nm) and accounting for the influence of SST on the derived PFT quantities (product number: OCEANCOLOUR_GLO_BGC_L3_MY_009_103).*

g) Chapter 2.3.2: it is unclear if also the PFT-chla derived from SOM predicted pigments using Hourany et al. 2019a have been produced by using the average value from applying in the DPA the 3 sets of coefficients proposed by Uitz, Brewin, Soppa. Please clarify.

Indeed, the SOM-Pigment outputs from El Hourany et al., 2019 were derived using these 3 sets of coefficients proposed by Uitz et al., 2006, Soppa et al., 2014, and Brewin et al., 2015.

We have added this in the manuscript in section 2.3.2.

h) Chapter 3.1 – line 162: it seems except for matching the data based on 3x3 pixel box +/-1 day no further criteria to select “valid” matchups has been used. Protocols recommend that at least 50% of the pixels are valid (unflagged) and the coefficient of variation is within 20% (e.g., see EUMETSAT protocol: https://www.eumetsat.int/media/44087 ). Can you provide more details or comment why no further quality control had been applied.

Indeed, to extract the match-up for a given observation, a 3x3 pixel box was employed, centered around the observation's coordinates on the same day. The average of the non-outlier pixels was computed. If this approach was unproductive due to a low number of pixels within the 3x3 box or the absence of any pixel, a 3x3 pixel extraction was performed for the adjacent days (+1 and -1). However, we did not enforce any additional strict protocols as per the EUMETSAT protocol, as only a small number of valid matchups were anticipated.

To our knowledge, the *psbO* gene database is a valuable source that provides complete information about the relative phytoplankton cell abundance across 7 taxonomic groups. Thus, the intrinsic value of this database is significant. While recognizing the importance of the EUMETSAT protocol in ensuring data quality and homogeneity in match-up exercises, it is important to highlight that our methodology, based on Self-Organizing Maps (SOM), has proven to be effective in reducing noise through vector quantization. Added to that, given the operational nature of the method and the coherence of results from cross-validation and tests, we believe that the evidence showing that this protocol is convincing.

It is imperative to emphasize that any future generation of *psbO* datasets should adhere to the EUMETSAT protocol or other masking protocols adopted by the OC community in the future.

Following these match-up exercises, we performed a baseline comparison between in-situ Chlorophyll-a (Chla) and satellite-derived Chla. This comparison is deemed satisfactory, with an error rate of 33%.

***Added:*** *To extract the match-up for a given observation, a 3x3 pixel box was employed, centered around the observation's coordinates on the same day. The average of the non-outlier pixels was computed. If this approach was unproductive due to a low number of pixels within the 3x3 box or the absence of any pixel, a 3x3 pixel extraction was performed for the adjacent days (+1 and -1). Following these match-up exercises, we performed a baseline comparison between in-situ Chlorophyll-a (Chla) and satellite-derived Chla. This comparison is deemed satisfactory, with an error rate of 33%.*

i) Chapter 3.2.2 – line 227 ff: Since you noticed that using 670nm in the algorithm did not improve it, why did you keep it? Further, in Line 230 the reference of Xi et al. (2015) is not suited since the paper is focusing on simulated data sets across many (all) water types – probably much better to cite here Torecilla et al. (2011) or Taylor et al. (2011)

where the HCA method (or Alvain et al. 2005 with Physat) has been applied to RRS data from the open ocean in order to derive information on phytoplankton community structure.

As previously mentioned in the manuscript, the 670 nm band was excluded from the algorithm. However, during the initial round of the review process, one of the referees emphasized the importance of utilizing the remote sensing reflectance spectrum, extending up to the near-infrared range. In response to this suggestion, we referenced the work of Xi et al. (2015), as recommended by this referee.

We simplified the explanation regarding the final selected bands in our algorithm to address potential queries that readers might have on this matter while omitting the discussion about the RRS at 670 nm that was not included in the algorithm.

***Added:*** *The choice of Rrs bands aligns with previous work conducted on the PHYSAT method by Alvain et al. (2005) and Ben Mustapha et al. (2013). The PHYSAT method utilizes reflectance anomalies in the same four selected bands to identify dominant phytoplankton functional types. In the clear open ocean, the information contained in the remote sensing reflectance (Rrs) bands beyond 555 nm is limited due to the strong absorption by water (Torrecilla et al., 2011; Taylor et al., 2011). It should be noted that the Rrs bands selected 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.*

j) Chapter 3.2.4: I missed a discussion about the input data uncertainty influencing the uncertainty of the retrieved PFT products (should be put in chapter 4).

Currently, no comprehensive uncertainties encompass all the associated steps in the quantification of *psbO*, including filtration, extraction, and the accuracy of *psbO*-based analyses. To address these uncertainties, a statement has been included in section 3.2.4 to provide further elaboration on the complexities and potential variations in the quantification process of *psbO*.

**Added:** *However, we should acknowledge the importance of addressing the uncertainties in the psbO measurements and their potential impacts on the algorithm's outputs, that are not taken into account in this study. This exclusion is primarily due to the absence of a comprehensive framework that accounts for all the associated steps in the quantification of psbO, including aspects such as filtration, extraction, and the accuracy of psbO analysis. Pierella Karlusich et al. (2022) conducted a thorough comparative study, evaluating psbO quantities against data obtained from confocal and optical microscopy, as well as cytometry, revealing an agreement of 70% (Spearman's Rho =0.64–0.71, p-value <.001).*

*However, it is essential to recognize that, like psbO, every quantification method is subject to uncertainties stemming from the various steps of the quantification process, emphasizing the necessity of comprehensive assessments within every in-situ measurement protocol.*

k) Chapter 3.4: The cross-validation results should also provide information of the mean or median relative deviation (MRD) in order to be comparable to other approaches (e.g., Xi et al. 2020, 2021, Lange et al. 2020) – it would be good to have here more statistical measured.

MRD has been incorporated across the study.

All the metrics, old and newly added, were further discussed in text section 4.1. Section 4.1 was modified according to the newly added information.

Direct comparisons with other approaches based on error values remain challenging. It is essential to recognize that this algorithm is rooted in a genomic dataset, delineating taxonomic groups differently from the HPLC DPA method, as exemplified in studies such as Xi et al. (2020, 2021) and Lange et al. (2020). A notable bias between HPLC DPA-derived PFT and *psbO*-derived PFT groups arises from the contrasting definitions of these PFT groups. As well as the differences in PFT group definition, the quantified errors also show the sensitivity specific to each algorithm and methodology followed and can be associated to the coherence of the dataset used in the study.

The comparability of these methods lies within the patterns observed at a global scale and the seasonal variations, enabling to highlight the convergence and divergence between the DPA and *psbO* methods.

***Added:*** *The cross-validation and test exercises demonstrated an average R2 of 0.68 for SOMRCA and 0.74 for SOMChlF across all phytoplankton groups (Fig. 5, table 3). Aggregating all Chla fractions showcased a satisfactory agreement between estimated total Chla and in-situ values (R2= 0.83), indicating the preservation of the initial phytoplankton quantity expressed in total Chla. For SOMRCA, the RMSE ranged between 2% and 23% in the test set and between 2% and 19% in cross-validation. The highest errors were observed for Prokaryotes, reaching 24% due to their high relative cell abundance in the initial dataset. In the case of SOMChlF, the RMSE ranged between 0.02 and 0.24 mg m-3 in cross-validation and 0.02 and 0.31 in the test set, with the highest error associated with the estimation of Chla, stemming from the cumulative Chla fractions of phytoplankton groups. Notably, the largest RMSE among phytoplankton groups was observed for the Diatom Chla fraction, attributed to their substantial Chla content and its exponential relationship with total Chla. The MRD highlighted a distinct contrast between SOMRCA and SOMChlF performance. Notably, SOMRCA exhibited a significantly higher*

*median relative deviation, approximately three times that of SOMChlF's MRD. The MRD for SOMRCA fluctuated between 0.36 and 0.81 for cross-validation and between 0.28 and 0.92 for the test set, with Dinoflagellates exhibiting the highest MRD. In contrast, SOMChlF's MRD per group ranged between 0.13 and 0.24 for phytoplankton Chla fraction and 0.33 for Chla in the test set. This discrepancy emphasizes the complexity of determining the phytoplankton community structure in terms of relative cell abundance, indicating the likelihood of diverse community structures responding to the same satellite-derived environmental context.*

***Added:*** *Uncertainty values reached 30% relative cell abundance for SOMRCA and 0.15 mg m-3 of Chla for SOMChlF, revealing distinct regional patterns in both cases. Notably, the observed uncertainties generally aligned with the concentration gradient in Chla fraction and cell abundance per group. The uncertainty associated with SOMRCA's outputs corresponded to the high relative deviation noted in the test and cross-validation, suggesting the potential acceptance of multiple community structures represented by the neurons of SOMRCA for a single satellite pixel, thus contributing to increased uncertainty levels. Regions at high latitudes exhibited the highest uncertainties for diatoms, green algae, and haptophyte relative cell abundances, while the Southern Ocean displayed heightened uncertainties specifically for prokaryotic cell abundance. The increased uncertainty within the Southern Ocean, particularly for prokaryotes, could be attributed to the limited sampling conducted in this geographical region. This limitation resulted in a notable dissimilarity between satellite data collected in this area and the data sampled in the initial dataset, aligning with the findings of the reliability index. This finding is consistent with the documented very low abundance of cyanobacteria in the Southern Ocean (Flombaum et al., 2013), which may contribute to heightened model uncertainty for this particular region.*

4. Section Results and Discussion

a) Figure 7 caption: provide n (number of observations) for both data sets, the cross-val set and the test set. As stated above also show (and discuss) results for RMSD and MRD since R^2 is not a very robust measure of accuracy of a product. For the PG-Chla comparisons it should be clearly stated in chapter 3 that R^2 results from calculations based on log-transformed data, while MRD and RMSD are based on non-log-transformed data.

MRD values are provided in Table 3. In the related Figure 7 (Figure 5 in this new version) we added in the caption to refer to Table 3 for further metrics. We added the n values in the caption of Figure 5 and Table 3.

It is clearly stated in section 3.2.1, Figure 5 and Table 3 that $R^2$ and MRD result from calculations based on log-transformed data, and RMSE is based on non-log-transformed data.

We did not put the RMSE nor the MRD results in Figure 5 due to the overcrowding of the image.

b) Line 320ff: I think it is difficult to understand what is presented in Figure 8 and discussed here and no values specific for each group and separately for chla-fraction and abundance are provided.
The results in fFgure 8 (now Figure 6 in this version) present a pixel-by-pixel indicator of the applicability of the method. As described in section 3.2.4, this indicator is acquired upon comparing the values for each parameter in a pixel to the initial data set used to train both SOM algorithms. It shows the flaws that are brought by the low coverage of the initial data set in certain regions of the global ocean. It is not in any way an uncertainty estimate, but a potential confidence/validity mask that can be associated to the outputs. This has been explained and discussed in the text in section 4.1.

Your pixel-by-pixel uncertainty assessment in terms of values and what it actually considered should be compared to other PFT/PSC algorithms results (e.g. see Brewin et al. 2017, Xi et al. 2021, Lange et al. 2021) - probably in chapter 4.3.

A comparison of uncertainties has been added in section 4.4. However, one may note that, as mentioned in the previous answers, this algorithm is based on a genomic dataset, with a different definition of the taxonomic groups than seen in the HPLC DPA method and using different algorithms.

***Added:*** *Upon comparing the uncertainty patterns with those observed in Xi et al. (2021), similar trends were identified for the Chla fraction of eukaryotic phytoplankton, displaying consistency in following the Chla concentration gradient as seen in our study. Notably, regions such as the gyres exhibited lower uncertainties, whereas higher uncertainties were evident in high-latitude regions and marginal seas. Conversely, when examining the uncertainty in the retrieval of prokaryote Chla by Xi et al. (2021), lower uncertainties were noted in polar regions, contrasting with higher uncertainties observed in low-latitude regions. Similarly, in Brewin et al. (2017), the uncertainty maps for diatoms and dinoflagellates depicted distribution patterns akin to our uncertainty estimates in the North Atlantic Ocean.*

*The noted coherence in uncertainty patterns between HPLC-based products and our psbO-based product can be attributed to the direct relationship between DPA pigment concentration and total Chla, as well as between psbO-derived Chla fractions and total Chla. Consequently, similar patterns in predictions, as well as in the uncertainties, emerge.*

*However, addressing the similarities and differences between the outputs of the above-cited methods referring to the same phytoplankton group is not a straightforward task. These methods are based on distinct assumptions and resolutions of phytoplankton groups; The estimation of phytoplankton groups using pigments is inherently imperfect and relies on assumptions that introduce considerable variability and bias in determining the contribution of specific pigments to the assessment of phytoplankton groups. For instance, several studies showed that the DPA approach tends to overestimate diatoms (Brewin et al., 2014, Chase et al., 2020). This approach may compromise the relevance of satellite images when used. However, the added value of such an approach resides in the availability of the large HPLC dataset, which allows the development of robust algorithms. On the other hand, the method described in this paper and the generated outputs are based for the first time on a complete and harmonized database of phytoplankton taxonomic community structure on a global scale; an approach that provides an unbiased picture of phytoplankton cell abundances. At this time the major limitation of this approach is the low number of observations from which the metric has been derived.*

c) In addition, in chapter 4.1 and 4.2 a discussion of your two gene-SOM algorithms performance in respect to cross-validation (e.g. as done in Brewin et al. 2015, Xi et al. 2020, 2021) and independent validation to other PFT /PSC algorithms presented in literature (see Mouw et al. 2017 and search newer literature on PSC algorithms) should be added.

All performance metrics, old and newly added, have been further discussed in text section 4.1. Section 4.1 was modified according to the newly added information.

d) Figure 9, also add the number of matchups (at least in the figure caption), add also the MRD!

The number of matchups has been added (N=2671) in the caption of Figure 9 and in the text (Figure 7 in this version) and the MRD values have been added to the figure.

e) Fig. 11 color scale for Chl-a should contain more colors, as in Fig.11 abundance presentation and in Fig. 13, so differences in Chl-a are more visible.

Fixed accordingly.

f) Typos: in line 358 and 370 – this should cite the correct subfigures of Fig. 11.

Corrected

# Response to referee #2

Review of revised manuscript

The work of El Hourany and co-authors presents a machine learning approach (specifically, Self- Organizing Maps) to estimate the relative Chla contribution and cell abundances of seven major taxonomic phytoplankton groups. The results of the trained model are applied to global satellite data, and in turn compared to both a previous SOM model developed using pigment rather than omics-based biomarker data, and a separate DPA-based approach. The study is novel in its use of phytoplankton gene information to train an ML model for assessing phytoplankton community structure from space, and the authors have clearly put thought into comparison with other approaches, and to how uncertainties also play a role in the results. After reviewing the manuscript (and in the context of previous reviewer comments and subsequent revisions), I believe the manuscript is publishable following some minor revisions and corrections. Thanks to the authors for their work on this topic.

General comments

I appreciate the background and description of functional types, the DPA and pigment-based groups in the Introduction text. However, the phrase “phytoplankton functional types” is used several times in the document, when in fact what is meant is phytoplankton taxonomic groups. Although “functional types” has been used rather loosely in the literature, the term “functional” indicates biogeochemical function (e.g. calcifiers, silicifiers), whereas phytoplankton of different sizes or even different taxonomic groups may serve the same ecosystem function. Therefore, I strongly encourage the authors to instead use the phrase ‘phytoplankton taxonomic groups’ when that is actually what is meant, or when referring more broadly to the variety of phytoplankton, ‘phytoplankton community composition/structure’. This attention to phrasing will benefit the community of research working on the topic of phytoplankton community composition from space in the context of interactions with any potential stakeholders and end-users.

We fully agree with the referee regarding the importance of defining the phytoplankton groups as taxonomic groups rather than functional types.

In response to the referee's suggestion, we revised the manuscript replacing instances of "phytoplankton functional types" with "phytoplankton groups" referring to taxonomic groups.

For this matter, we found it important to clarify this difference in the introduction:

*"Recently, ocean color data have also been used to gain information about phytoplankton communities, such as their size structure, and their taxonomic or functional composition. This interest has facilitated the integration of the concept of phytoplankton functional types (PFT) and taxonomic groups (PG) into studies exploring various ecological and biogeochemical aspects (Le Quéré et al., 2005; Hood et al., 2006). Functional types refer to distinct categories associated with biogeochemical processes (e.g., silicifiers, calcifiers) and physiological adaptations to environmental factors (e.g., light, nutrients, turbulence), or to more practical categories identified through specific analytical techniques (e.g., pigment types) (IOCCG report N 14). On the other hand, phytoplankton groups correspond to taxonomic classes (e.g., diatoms, haptophytes, cyanobacteria). It is important to note that phytoplankton from different taxonomic groups can perform the same ecosystem function, e.g., both diatoms and silicoflagellates can biosilicify but represent different taxonomic groups. Specialized algorithms applied to ocean color data have consequently been developed to detect specific taxa with distinctive optical characteristics, e.g., (Brown (1995); Iglesias-Rodríguez et al. (2002)), or the dominance of phytoplankton functional types (e.g., Alvain et al. (2005)) or the relative abundance of phytoplankton groups and size classes in term of their contribution to the Chla e.g., Hirata et al., (2011), Xi et al. (2020, 2021), and lately, plankton assemblages and communities e.g., Kaneko et al., 2023, (Sathyendranath et al., 2014; Bracher et al., 2017; Mouw et al., 2017)"*

Could you comment on the fact that the *psbO* is a proxy of individual cells, but the chain-forming phytoplankton types (e.g., *Chaetoceros*) will contain several, or many, individual cells, and will likely be found in the larger size fractions? For example, if the abundance of diatoms is high in the 20-180 fraction, but the *psbO* represents individual cells (vs. chains), the diatom contribution to Chla could be dramatically overestimated.

The inspection of microscopy images from the same size-fractionated samples showed that long chains (such as those from *Chaetoceros*) are frequently found as shorter fragmented chains and individual cells that pass through small mesh sizes (e.g., Pierella Karlusich et al 2021 *Nat Comm* 12: 4160).

In my opinion, the text of section 3.4 needs to be revised for clarity. Is it not fully clear what the inputs and targets of the random forest regression algorithm are. The following

sentence is not clear: “In the internal node, the selected feature (i.e., pigment in this case) was used to make a decision on how to divide the dataset into separate sets with similar responses in terms of a given phytoplankton group.” Suggest revision to help the reader understand the application of the random forest regression method.

We admit that the statement the referee pointed out and the paragraph dealing with the random forest was lacking clarity. We have added further information to the text and reformulated our ideas as follows:

*Each phytoplankton group’s psbO abundance was associated with its corresponding HPLC pigment measurements performed on the same Tara Oceans station. The ability of pigments to predict a specific phytoplankton group was evaluated using a bagged random forest algorithm (number of learners set to 200), following the permutation-based importance method.*

*Using this method, a pigment composition of the seven major phytoplankton pigments cited in Table 1 was tested to predict the abundance of each of the seven psbO-derived phytoplankton groups and estimate their importance relative to each group. The concentration of each pigment was converted in terms of pigment ratios, a ratio relative to the sum of all pigment concentrations, and in parallel, the psbO-derived relative abundance was used.*

*The bagged random forest algorithm is a set of decision trees, each constituted of internal nodes and leaves. Within the internal nodes, the algorithm uses pigment data as the predictor variable to partition the dataset into subsets based on pigment characteristics. These subsets are then utilized to predict the abundance of specific phytoplankton groups, enabling effective analysis of the importance of pigments to describe the variability of a phytoplankton group. Since this algorithm is used in a case of regression, the training is done while minimizing the error between the psbO-derived phytoplankton group abundance and the predicted one. The permutation-based importance method will randomly shuffle each pigment and compute the change in the model’s performance to predict the abundance of a phytoplankton group.*

Line 351: as I’m sure the authors are aware, the CHEMTAX approach was developed decades ago to address just this. Although it does have its own caveats, it can be a useful tool to compare against, and recent work to improve it and make it more broadly applicable is worth looking into (see Hayward, Pinkerton, and Gutierrez-Rodriguez. 2023. “Phytoclass: A Pigment- Based Chemotaxonomic Method to Determine the Biomass of Phytoplankton Classes.” *Limnology and Oceanography: Methods* 21 (4): 220–41. https://doi.org/10.1002/lom3.10541.) Perhaps this additional analysis is not warranted for this study, but it is worth keeping in mind for future work related to comparison of different approaches used to estimate phytoplankton community structure.

Thank you for your insightful comment regarding the CHEMTAX approach. We have noted your point and acknowledge the significance of evaluating various methodologies in this domain. In future work, it is indeed our intention to make further attempts to define a consensus from different phytoplankton group identification methods, but we consider that this goes beyond the scope of the current manuscript.

Specific comments

While I'm not aware of the specific journal requirements, numbering the equations would make it easier for the reader to reference the equations within the text for future analyses, etc.

We added numbering for each equation as suggested.

L79: start the sentence with "The" to avoid leading with the gene name.

Added accordingly.

Line 100: should be GlobColour (with a capital "C"), throughout.

Modified accordingly throughout the manuscript.

Fig. 1 caption: please define 'DRCA' and 'DChlF' for the reader here

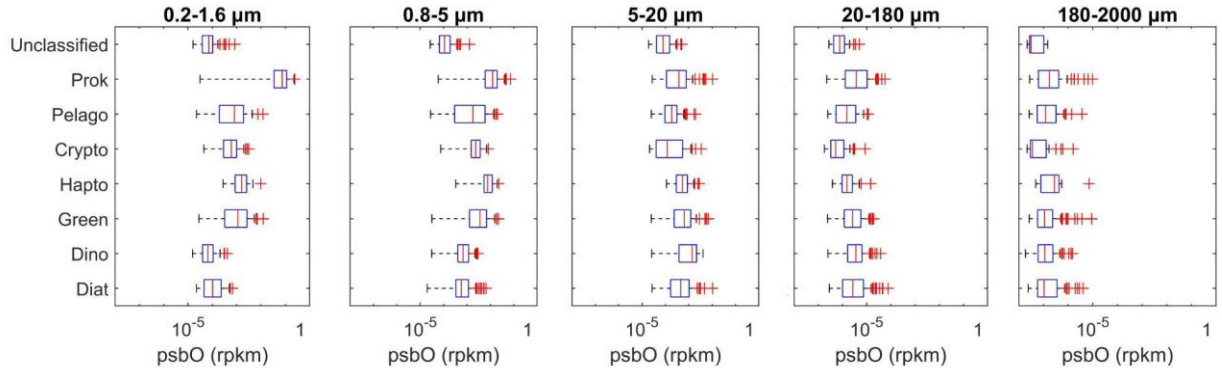
Added in the caption.

Fig. 2 – it would be valuable to also know the absolute *psbO* values as well – for example, it is true that the Prokaryotes are over-represented in the largest size fraction, but are the absolute quantities of *psbO* very low in that size fraction? I guess more generally – what is the range in absolute quantities of the *psbO* gene across the size fractions?

The *psbO* can be used to estimate **absolute** cell abundances with careful normalization and quantitative DNA extraction methods. In the current study, we did not attempt to do it because the metagenomic sampling from *Tara* Oceans was not specifically designed to quantify metagenomic signals per seawater volume due to the lack of "spike-ins" (e.g., DNA internal standards). It's well established that there is an inverse logarithmic relationship between plankton size and abundance (Belgrano et al., 2002; Pesant et al., 2015), so small size fractions represent the numerically dominant organisms in terms of cell abundance (albeit not necessarily in terms of total biovolume or biomass).

We can still express *psbO* abundance in rpkm (reads per kilobase covered per million of mapped reads) to normalize the *psbO* signal by the sequencing depth. There is a decrease in rpkm values towards the larger size fractions, probably explained by the

increase in genome size and complexity in larger size fractions. In addition, prokaryotes are dominant in the smaller size fractions while the larger fractions are characterized by the higher prevalence of eukaryotic phytoplankton.




This figure was added in the supplementary material figure while citing it in the caption of Figure 2.

L116: please define ‘CCI’

Defined as a Climate Change Initiative (CCI)

L130: sentence is awkward as written and ending in “them”; suggest revising to something like “we used two previously published algorithms:”

It was changed as suggested by the referee.

L149: First sentence does not add anything for the reader.

Removed

L 169: is ‘variables’ here referring to the phytoplankton groups, the satellite-derived

parameters, or both?

We chose to normalize every variable, Phytoplankton, and satellite, to reinitialize the weights before using SOM, and to make their values comparable. This has been clarified at the end of the paragraph.

L 174: what is meant by ‘the SOM algorithm that can deal with missing values’? can you give a sentence or two to describe mathematically what is done to account for missing

values? Could you add reference(s) here to back up this widespread use of SOM to complete missing data?

We have formulated and briefly described how SOM can deal with missing values while adding some references. this formulation was added to section 3.2.1

***Added:*** *SOMs have been widely employed to complete missing data, utilizing the truncated distance (Folguera et al., 2015; Charantonis et al., 2015; Saitoh, 2016; Rejeb et al., 2022). The truncated distance is defined as a modification of the standard Euclidean distance between two observations that accounts only for the existing components of the vectors. This modification of the distance measure allows for the comparison of observations with incomplete information by considering only the existing components and effectively handling missing data. The SOM algorithm can then use this truncated distance measure in its learning process to complete missing data and integrate incomplete information, enabling more robust analysis and visualization of the data.*

L 195: increased from 10 to 1000 neurons at what interval?

The interval is 10. We have added this information to the text.

Line 313: Curious how you decided on the threshold of 40%?

We are sorry for this mistake, but due to the change of this figure across the review process, we meant to say 60% instead of 40. The 60% threshold was an arbitrary choice, looking at the shape of the value distributions of this index and their spatial patterns.

In terms of calculation, 60% means that almost 3 out of 8 satellite parameters at a certain pixel are considered as an outlier, and therefore the estimated phytoplankton composition might be biased.

**An explanation was added in section 4.1:** *Threshold arbitrarily chosen while evaluating the frequency histogram of this index's values in Fig.6. A value of 60% roughly translates to the exclusion of 3 out of 8 satellite parameters' values considered outliers at a certain pixel.*

L 357: typo ‘Glocolour’ (missing ‘b’)

Corrected

L 358: typo ‘Fig. 101’

Corrected

L 371: typo ‘Fig. 112’

Corrected

Figure 9. – suggest including a colorbar to show the number of points per pixel based on the color of the dots on the graph

We have added a density color bar.

Figure 12. caption – not clear what is meant by ‘original Rrs spectra’

We meant to refer to denormalized Rrs spectra, as in the original values. We have modified the caption as follows:

*Relative cell abundances per phytoplankton group and normalized and denormalized Rrs spectra were also derived.*

Figure 13. Capitalize the first word of the caption. Could the x-axis of the latitude line graph be revised to label more than just the 10^0 ?

The figure was modified according to the referee’s suggestion.

L 448: suggest revision to ‘launch of NASA’s Plankton, Aerosol, Cloud, ocean Ecosystem (PACE) mission’

Revised according to the referee’s suggestions

L 451: suggest revision to ‘the perspective of the PACE mission,’

Revised according to the referee’s suggestions