Responses to reviewers

We thank the reviewer for their additional comments on the revised manuscript. Below we respond to each point individually, outlining the changes made to the manuscript (line numbers refer to lines in the manuscript with tracked changes).

Thank you for considering the suggestions. The manuscript has improved and most of my suggestions were noted or reasonably justified. There were still a few points that, in my opinion, were insufficiently addressed or answered, perhaps due to misunderstandings. So, still a few comments for the authors to consider.

My comment on acknowledging also the smaller efforts of publishing open, manually annotated datasets was addressed as follows "Earlier plankton image datasets were modest in size, typically containing a dozen or a few dozen of classes (Benfield et al., 2007), but were crucial for establishing the first classification methods." Based on this, I must assume that the authors have misunderstood my point. I did not mean earlier, first works, but all different efforts there are to expand the publicly available image libraries for reproducible and open science, as well as for the use of other users of the instruments. I know it is an issue that, within the wide field of publications on plankton recognition, it is hard to compare results between them, especially when the datasets have not been published. It is also hard even if the datasets are published, but if all the different methods are tested on different datasets. Therefore, benchmark datasets are valuable. However, to promote open science and also advance the development of models with a diverse training set, it is also important to promote smaller efforts in publishing open datasets for the purpose of model development. For future avenues of developing automated recognition of plankton, the more diverse training datasets we have available from multiple sources, the better. I believe this has also been improving a lot in recent years. Therefore, it would be nice to mention that, besides the "three major plankton image datasets" used in this study, there is an expanding effort to publish manually annotated datasets.
To make it easier and get some idea of how much effort there has been on this, I compiled a list (surely not comprehensive) for you to have a look at.
-European IFCB users have gathered links to open datasets
https://nordicmicroalgae.org/annotated-images/
-Table 2 in Eerola et al. 2024: listed published datasets of plankton that were related to publications.
-Table 1 in Kareinen et al. 2025 (Self-Supervised Pretraining for Fine-Grained Plankton Recognition https://arxiv.org/html/2503.11341v2) contains some more recently published datasets.
-A new version of the ZooLake dataset https://doi.org/10.25678/000C6M
-A FlowCam dataset https://doi.org/10.5281/zenodo.16679297
-A Flowcam dataset https://zenodo.org/records/16840846

We thank the reviewer for raising this issue, the we have inserted the following paragraph at line 154 to highlight the importance of publishing smaller open datasets:
Beyond publishing large reference datasets, as we strive to do in this work, another avenue for progress is the collection of many diverse, albeit smaller, datasets. This is typically the first step for the creation of "universal" foundation-type models. The push towards more open and reproducible science has helped in this respect and several local datasets have been published: e.g. Table 1 in Kareinen 2025, Table 2 in Eerola et al. 2024.

My previous comment: "Figure 2: Why did you choose to show accuracy and not F1 score in the first panel (the same comment also goes for the subsequent figures)? What is the Random classifier? It was mentioned in a paragraph starting from line 350, but it would require a better explanation." -I would still argue on behalf of adding F1 scores. I don't think that it would be redundant, as it is very difficult to go and look at the F1 scores from the tables. I think that if you want to highlight the fact that the naive baseline can be misleading, you could add the F1 on top of the accuracy bars as dots. The F1 score is a very common metric to report, and it would make sense to have it in the figures as well to make comparisons easier. It would also be interesting to show how different picture F1 gives versus balanced accuracy, thus not being a redundant addition to the first panel in figures 2, 3, and 4.

We appreciate the suggestion to display F1 scores alongside accuracy-type metrics. To evaluate their added value, we generated supplementary plots that show micro-F1 versus accuracy and macro-F1 versus balanced accuracy for every model (Figure S2, presented below).

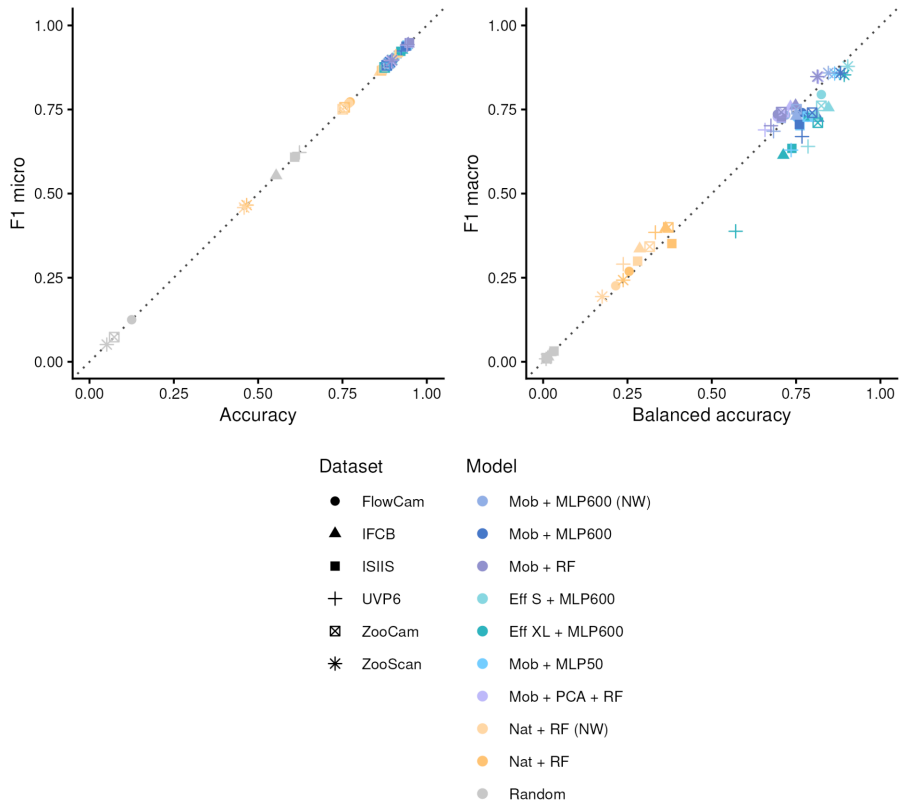

Figure S2: Relationship between F1 and accuracy-type metrics. (a) Micro-averaged F1 against accuracy and (b) macro-averaged F1 against balanced accuracy; for every model across all datasets. The dotted diagonal represents the 1:1 line, highlighting where the two metrics convey identical information.

The plots reveal two key observations that support our decision to retain accuracy and balanced accuracy as the primary figures:

1. Micro-F1 vs. accuracy: all models fall on the 1:1 line, indicating that micro-F1 conveys exactly the same information as accuracy.
2. Macro-F1 vs. balanced accuracy: again, every model lies close to the 1:1 line, confirming that macro-F1 closely resembles balanced accuracy for the class-imbalanced scenario we consider.

Because both F1 variants collapse onto the same trends as the accuracy-based measures for all models, we conclude that adding F1 bars would not improve interpretability but only hemper readability of Figures 2-4. We therefore keep accuracy and balanced accuracy in said figures, we add Figure S2 to the supplementary materials and we provide the full set of F1 values (alongside with other metrics) in the supplementary Table S8 for readers who wish to inspect them. We think that this decision preserves clarity in the main figures while still offering complete metrics for full transparency. The following edits were made to the manuscript.

Paragraph updated at line 336:
"Usual metrics were computed: accuracy score (percentage of objects correctly classified), balanced accuracy, macro-averaged F1-score, micro-averaged F1-score, class-wise precision (percentage correct in the predicted class) and recall (percentage correct within the true class)."

Paragraph updated at line 352:
"To focus on these classes, we also computed the average of precision and recall per class, weighted by the number of objects in the class, but using only plankton classes, i.e. the target classes (Owen et al., 2025)."

Inserted in legends of figures 2, 3 and 4:
"All values, including F1-scores, are reported in Table S8."

Inserted at line 391:
"The same applies for F1-scores: macro-F1 captures the failure of the random classifiers, while micro-F1 mirrors accuracy (Fig. S2)."

My previous comment: "375-385: Wouldn't it be important to find a harmonic mean between precision and recall rather than emphasize the importance of precision and detection of rare classes over recall?" -Yes, but my point was that you chose to present results only of the weighted models from here onwards, which had lower precision but higher recall (it seems I wrote them the other way round in my comment). The question was why you didn't choose

the model based on the balance between these two, but chose the strategy highlighting recall. I know that if you want to get better recall for the plankton classes, you chose the weighted model, but I would not want my classes to be disturbed by many false positives either. This is why I asked why not choose a strategy highlighting the harmonic mean between these two instead of choosing a strategy that highlights getting correct hits of the rarer classes, but with the consequence of getting false positives? This is also a topic: with a closed set classification, one needs to choose which strategy to follow; with open set classification methods, the target is to accurately identify images belonging to the existing classes, but also to identify or filter out the ones that don't. I think this is a topic that should be raised in the section on costs and benefits of using CNNs, a limitation of closed-set classification systems, and also an alternative approach of filtering out images of too low prediction scores (thresholding), which is, of course, also nonideal.

We thank the reviewer for this valuable comment. We agree that favouring recall is not the only possible strategy; it is the one we chose in this study because it aligns with common goals when training such models. We have inserted the following paragraph in the discussion at line 601 to discuss this trade-off:
"Weighting improves the recall of rare classes but reduces their precision, reflecting the classic precision–recall trade-off. When downstream analysis involves manual verification, higher recall is advantageous because a few false positives in rare classes can easily be corrected while missed detections would likely be lost among the most numerous classes and not easily recovered. Conversely, in high-throughput monitoring through imaging, where human review of all samples is infeasible, emphasizing precision reduces spurious detections at the cost of under-estimating true abundances. In such settings, post-hoc confidence thresholding (e.g. Faillettaz et al., 2016; Luo et al., 2018) offers a pragmatic compromise, albeit an imperfect one. In all situations, using various intensities of class weighting is a flexible solution to adapt the classifier to the study's objective."