scientific data

OPEN

DATA DESCRIPTOR

Annotated IFCB plankton images from the Mediterranean Sea

Melpomeni Sofia Mente[1,4], Emilie Houliez[2,4], Eleonora Scalco[1], Leonilde Roselli[2,3] ✉ & Diana Sarno[1,3] ✉

The Imaging FlowCytobot (IFCB), supported by machine learning-based classifications, has revolutionized plankton research by automating plankton monitoring and considerably increasing sampling resolution. However, building a training set of labeled IFCB images to train the machine learning algorithms remains time-consuming and challenging. Consequently, there is a growing demand within the IFCB user community for shared datasets of taxonomically annotated IFCB images. Currently, such datasets are scarce and lack Mediterranean coverage. This data descriptor introduces MedPlanktonSet, a dataset comprising 77,271 taxonomically annotated IFCB images provided with their associated features. Data were collected from November 2022 to February 2025 at six stations in the Gulf of Naples (Western Mediterranean Sea) and the IFCB images were classified into 139 categories. MedPlanktonSet will support the development of various machine learning classifiers, 3D plankton reconstructions, training of plankton taxonomists and trait-based ecological studies. Ultimately, by facilitating the use of IFCBs with their associated classifiers, MedPlanktonSet will contribute to advancing research on plankton biodiversity and ecology.

Background & Summary

Plankton (phytoplankton, mixoplankton and zooplankton) play critical roles in marine ecosystems. They form the base of most marine food webs, support fisheries and aquaculture and drive essential processes such as carbon dioxide fixation, oxygen production and nutrient cycling. Monitoring plankton diversity and dynamics in relation with biotic and abiotic factors, is thus essential to understand marine ecosystems functioning and resilience. As such, plankton monitoring is an integral component of various initiatives aimed at safeguarding water quality and ocean health (e.g. EU Water Framework Directive 2000/60/EC, EU Marine Strategy Framework Directive 2008/56/EC, the OSPAR Convention and the Harmful Algal Bloom and Hypoxia Research and Control Amendments Act in the USA). Plankton has also recently been recognized as a powerful actor in supporting efforts to address the so-called "planetary triple crisis: loss of biodiversity, climate change and pollution"[1].

Traditionally, plankton monitoring relies on microscopy-based methods. This approach is generally considered the reference for quantifying plankton because it has provided the foundation of plankton ecology and offers detailed taxonomic identifications and quantitative abundance estimations[2,3]. But it is time-consuming, requires trained taxonomists, and can miss rare or delicate species due to sample handling and fixation[4–6]. In the last decades, the use of molecular methods has expanded. While these techniques have improved the detection of rare species, revealed cryptic taxa and contributed to the discovery of many new species, they primarily provide presence-absence data or relative abundances estimated based on genes copy numbers[2,5,7]. In addition to being more expensive than microscopy, quantification with these techniques remains challenging and identification heavily relies on existing sequence databases in which plankton organisms currently lack adequate representation. Analyzing the large datasets generated by molecular methods also requires powerful computers and skills in bioinformatics[2,5,7]. Both microscopy and molecular based monitoring are labor-intensive. Consequently, with these techniques, samples are often collected with a resolution not always adequate for detecting short-term changes in plankton communities such as the rapid onset and development of blooms[8].

These limitations led the plankton ecologists to increasingly use plankton imaging systems[9]. One of these instruments is the Imaging FlowCytobot (IFCB)[10]. The IFCB is an imaging flow-cytometer which combines

[1]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Napoli, Italy. [2]Stazione Zoologica Anton Dohrn, Brindisi Marine Center, Via Duomo 20, 72100, Brindisi, Italy. [3]NBFC, National Biodiversity Future Center, Piazza Marina 61, 90133, Palermo, Italy. [4]These authors contributed equally: Melpomeni Sofia Mente, Emilie Houliez. ✉e-mail: leonilde.roselli@szn.it; diana.sarno@szn.it

flow-cytometric and video technologies to capture high resolution images of particles in the size range of < 10 to 150 μm. Designed for *in situ* deployments, the IFCB autonomously samples and analyzes water samples, around the clock, during unmanned deployments lasting up to six months without maintenance. The IFCB has revolutionized plankton research by automating the *in situ* high-frequency monitoring of plankton organisms and has provided significant new insights into their ecology. For example, it has contributed to the detection and early warning of Harmful Algal Blooms[11–14]. It successfully offered the possibility to study *in situ* live cells of delicate planktonic organisms that are frequently damaged using the traditional net- and microscope-based methodology[15]. It has also highlighted the importance of understudied processes in plankton bloom dynamics, such as transitions in life cycle stages[16], parasitic infections[17–19] or predator-prey interactions[20,21].

However, high frequency sampling has also brought new challenges. When deployed *in situ*, depending on the concentration of particles, a single IFCB can generate up to 10,000 high resolution images per sample and analyzes approximately 60 samples per day[22].This volume of data rapidly precludes the manual inspection of each image for taxonomic identification. Consequently, to process such large datasets, scientists rely on machine learning algorithms to develop classifiers[23,24]. These algorithms use supervised classifications meaning that, before being operational, the algorithm must be trained with a training set of images manually labeled by experts[25]. To reach good classification performances, it is advised to train the algorithm using at least 1000 images per category and to collect the images over a full year of sampling to account for seasonal variations in plankton communities[26,27]. But identifying and manually annotating IFCB images is time-consuming as this requires a careful examination of the images by a group of experts trained in plankton identification[27,28]. With plankton taxonomy being a discipline in decline[29], the requirement of taxonomic experts to train the IFCB's classifiers can be a limitation. In fact, some research institutes lack the required knowledge to properly identify their IFCB images[30]. In addition, reaching 1000 images per category can be challenging for rare plankton taxa and often results in imbalanced datasets with potential detrimental effects on classification performances[31].

The increasing popularity of the IFCB, coupled with the challenges encountered when creating a training set of labeled IFCB plankton images, has led to a growing demand within the IFCB user community for shared datasets of taxonomically annotated IFCB images. However, currently, such datasets are scarce and have limited geographic coverage. The most complete dataset to date is WHOI-plankton[32]. WHOI-plankton comprises over 3.4 million IFCB images collected 3 km south of Martha's Vineyard (Massachusetts, USA) over nine years and labeled by experts into 103 categories. Two other datasets collected in different locations of the Baltic Sea are also available: SYKE-plankton[33] and SMHI IFCB plankton image reference library[34]. SYKE-plankton contains approximately 63,000 IFCB images classified into 50 categories. The SMHI IFCB plankton image reference library is separated into 3 subsets containing a total of 76,032 images. In the three subsets, the IFCB images were classified into a different number of categories (61, 83 or 39 categories). Consequently, no dataset of labeled IFCB images was available for the Mediterranean Sea.

This data descriptor introduces MedPlanktonSet, a new dataset of 77,271 IFCB images taxonomically annotated (to the lowest possible taxonomic level), along with their associated features, collected in the Gulf of Naples (Mediterranean Sea). MedPlanktonSet can have multiple applications. The first type of application is in the development of various classifiers to categorize IFCB plankton images using machine learning algorithms. MedPlanktonSet can be used to develop a local classifier specifically optimized to recognize IFCB images from the Mediterranean Sea. MedPlanktonSet can also be used to augment other datasets of labeled IFCB images, enabling the development of a classifier for a broader geographic area that includes the Mediterranean Sea. Subsets of MedPlanktonSet can be used (alone or in combination with data from other IFCB datasets) to create classifiers that focus on specific taxa, such as targeted genera or species of harmful algae. MedPlanktonSet is also useful as a "negative dataset" to control the specificity of classifiers developed to categorize IFCB images from other localities than the Mediterranean Sea (i.e. to evaluate the ability of these other classifiers to not classify the Mediterranean taxa into the categories they were trained to recognize) or to develop an open-set model for plankton recognition[35]. Similar to how ImageNet[36] is used for transfer learning to train machine learning algorithms with different goals than classifying the images present in this large dataset, MedPlanktonSet can complement other datasets and have the same type of wide applications. The second possible type of application is the reconstruction of the 3D structure of plankton taxa[37,38]. MedPlanktonSet provides a rich library of plankton taxa imaged in different orientations. Another possible application is in the training of plankton taxonomic analysts. A study made during the COVID pandemic showed that using IFCB images can shorten by up to half the time required to train new plankton taxonomic analysts[30]. Given that each image is provided with its associated features, MedPlanktonSet has also the potential to be integrated into large databases of plankton size measurements such as the Pelagic Size Structure database[39]. Finally, the features provided in MedPlanktonSet can be leveraged in trait-based ecological studies[27,40,41].

Methods

Surface seawater samples were collected from November 2022 to February 2025 at six stations in the Gulf of Naples (Western Mediterranean Sea): the Long-Term Ecological Research Station MareChiara (LTER-MC), mouth of the Sarno River, Torre del Greco, Capri canyon, Ammontatura and Napoli Porto (Fig. 1). Images were acquired with the IFCB configured to analyze 5 mL water samples by triggering only on fluorescence (PMT B = 0.65, Trig B = 0.13).

Images were then manually annotated and classified into 139 categories using the publicly available IFCB MATLAB-based annotation tool: "startMC" (https://github.com/hsosik/ifcb-analysis/wiki/Instructions-for-manual-annotation-of-images). Plankton organisms were identified to the species level, genus level, or higher taxonomic groups, based on the visibility of distinctive morphological features. For *Pseudo-nitzschia spp*., a potentially toxic diatom that forms chains, a further division into seven categories was conducted to facilitate precise counting of the number of cells per chain. Following classification, the images

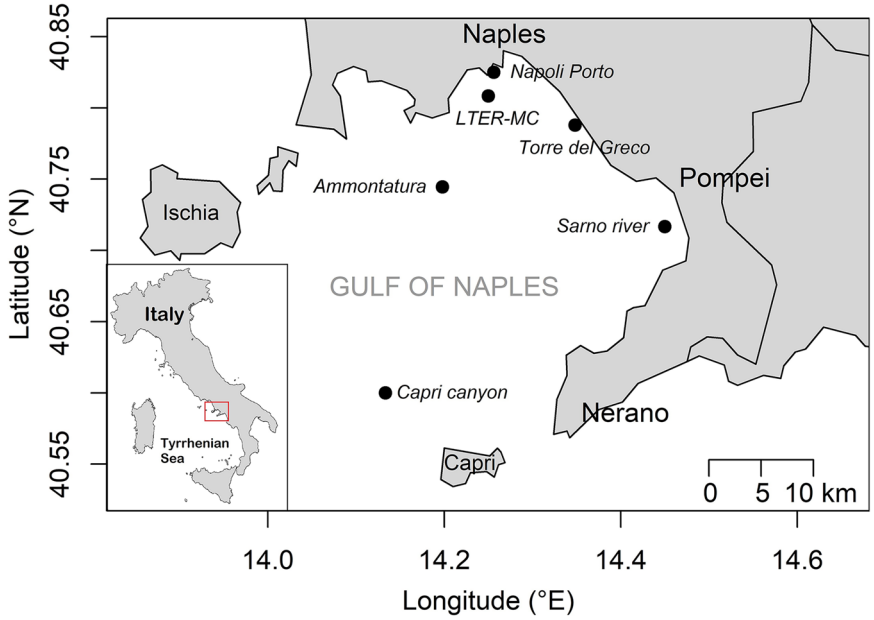

Fig. 1 Map of the Gulf of Naples with the six sampling stations where the IFCB data were collected. LTER-MC = Long-Term Ecological Research Station MareChiara, Capri canyon = NEREA-Capri (Dohrn Canyon), Sarno river = NEREA-Sarno.

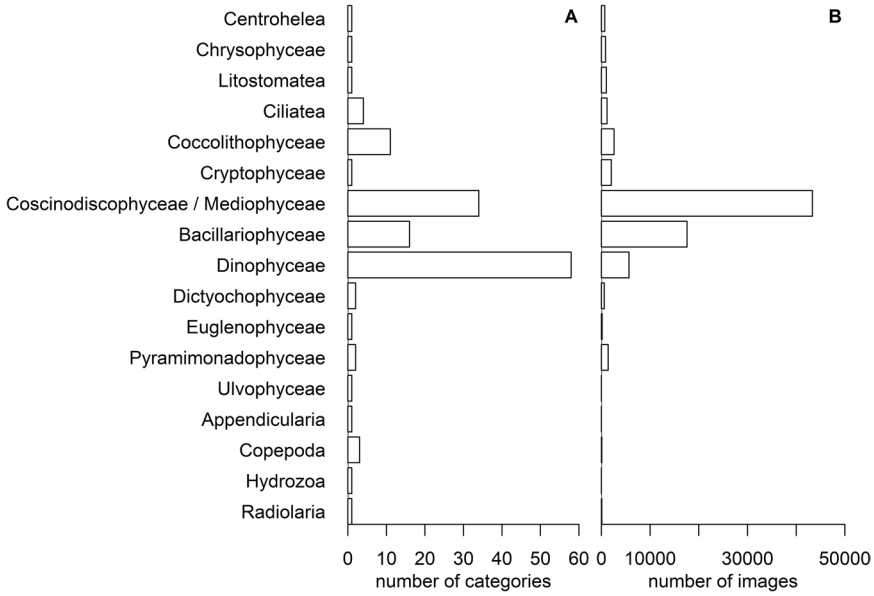

Fig. 2 Number of categories (**A**) and images (**B**) per taxonomic class of MedPlanktonSet.

were extracted in the png format into separate folders corresponding to each category using a modified version of the IFCB MATLAB code: "export_png_manual_fromROI" publicly available on the IFCB-analysis github (https://github.com/hsosik/ifcb-analysis).

Beyond providing images of each particle present in a water sample, the IFCB offers the possibility to extract the features associated with each image. These features are an ensemble of numerical characteristics or properties of the image that reflects various aspects of the particle's appearance and morphology including its size, shape, symmetry, texture and orientation[22]. Features are required to develop random forest classifiers but can also have other applications, such as in datasets of plankton size measurements, 3D reconstructions of plankton taxa or trait-based approaches. The features of MedPlanktonSet were extracted by following the IFCB blobs and features extraction procedure and using MATLAB codes publicly available on the IFCB-analysis github (https://github.com/hsosik/ifcb-analysis/wiki/Blob-extraction,-feature-extraction,-and-classifier-application). Two versions (V2 and V4) of the algorithms are available for extracting blobs and features. Version V4 refines the segmentation algorithm of the V2, aiming to improve edge detection and is recommended for applications

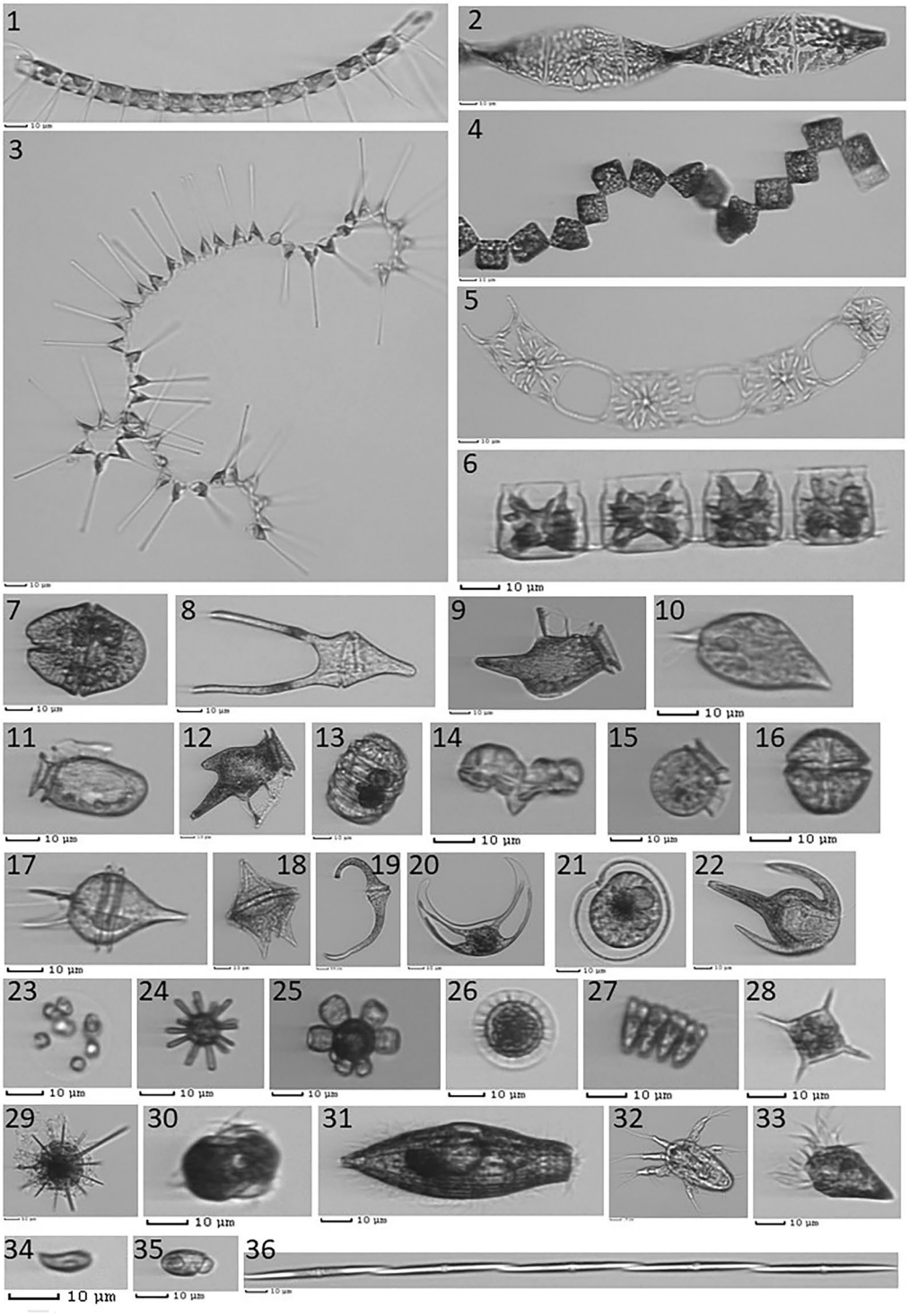

Fig. 3 Examples of images for 36 categories of MedPlanktonSet. (1) *Chaetoceros curvisetus / Chaetoceros pseudocurvisetus* (2) *Helicotheca tamesis* (3) *Asterionellopsis* cf. *glacialis* (4) *Neobrightwellia alternans* (5) *Eucampia cornuta* (6) *Lithodesmium variabile* (7) *Akashiwo sanguinea* (8) *Ceratoperidinium margalefii* (9) *Dinophysis caudata* (10) *Prorocentrum* cf. *gracile* (11) *Dinophysis sacculus* (12) *Dinophysis tripos* (13) *Polykrikos kofoidii* (14) *Karenia papilionacea* (15) *Dinophysis parva* (16) *Alexandrium* spp. (17) *Protoperidinium diabolus* (18) *Protoperidinium* cf. *divergens* (19) *Pseliodinium fusus* (20) *Pyrocystis* cf. *lunula* (21) *Pyrophacus horologium* (22) *Tripos azoricus* (23) *Phaeocystis jahnii* (24) *Rhabdosphaera clavigera* (25) *Scyphosphaera apsteinii* (26) *Pterosperma* spp. (27) *Oltmannsiellopsis viridis* (28) *Dictyocha fibula* (29) Radiolaria (30) *Mesodinium* spp. (31) *Tiarina* spp. (32) Copepod nauplii (33) Ciliates (34) Cryptophyceae (35) *Kapelodinium vestifici* (36) *Pseudo-nitzschia* spp. 5 cells.

prioritizing quantitative particle sizing. However, the V4 version provides a smaller set of features, lacking some of the non-geometric cell properties available in V2 that are important for input in random forest training and application. Therefore, users intending to work with random forests classifiers are advised to continue to use the features from V2. For these reasons, MedPlanktonSet includes the features obtained with both algorithms (V2 and V4).

Data Records

MedPlanktonSet is publicly available on Zenodo[42]. It comprises three zip files and a csv file.

The zip file "IFCB_images.zip" contains taxonomically annotated IFCB images organized into 139 folders, with each folder containing the png images for a specific category. Each png image follows the same naming format. Let's use the image "D20230719T072550_IFCB181_00148.png" as an example to explain this format: "D20230719" indicates the date the image was recorded by the IFCB (July, 19 2023); "T072550" represents the recording time (07:25:50); "IFCB181" is the identifier of the IFCB instrument; and "00148" is the Region Of Interest (ROI) number.

The zip file "FeaturesV2.zip" contains a MATLAB file and a csv file. Both files hold the 237 features extracted with the V2 algorithm for each image provided in "IFCB_images.zip". Detailed information about the features, including their meaning and units, is available in Table S1.

The MATLAB file "FeaturesV2" includes six arrays:

- *class2use*: a vector listing the categories used for classifying the images (ROIs).
- *class_vector*: a list of indices indicating the category to which each ROI has been classified.
- *featitles*: the column names for the variables presented in the *train* array.
- *nclass*: the total number of ROIs for each category listed in *class2use*.
- *targets*: the names of the ROIs found in *train*.
- *train*: an array where each row corresponds to a single ROI, and columns contain the corresponding features values.

The csv file "FeaturesV2" has 240 columns:

- *ROIname*: contains the names of the images (ROIs).
- *IFCB_category* indicates the name of the category to which the ROI has been classified.
- *Class_vector* a list of indices indicating the category to which each ROI has been classified.
- The remaining columns are the features values.

Similarly, the zip file "FeaturesV4.zip" contains a MATLAB file and a csv file structured identically to those in "FeaturesV2.zip" but they provide the 28 features extracted with the V4 algorithm.

The csv file named "List_of_images.csv" is organized in three columns. The first column contains the names of the images (ROIs). The second column is the category to which the image has been classified. The third column is the sampling location.

Note that the size features provided in "FeaturesV2.zip" and "FeaturesV4.zip" are given in pixels. They can be converted into microns by applying a conversion factor which accounts for the magnification and camera resolution of the IFCB. At the time of this writing, the recommended conversion factor is 2.77 pixels per micron.

Data Overview

MedPlanktonSet comprises a total of 77,271 taxonomically annotated IFCB images along with their associated features (V2 and V4 version). The IFCB images were classified into 139 categories spanning 18 taxonomic classes (Fig. 2). The complete list of categories with the number of images per category can be found in Table S2. Figure 3 shows examples of images for 36 of these categories. Example images for the entirety of the categories can be viewed on the MedPlanktonSet website (https://szn-ifcb.weebly.com/medplanktonset.html).

Technical Validation

All the images in the dataset were visually inspected and annotated by a single person using the IFCB MATLAB-based annotation tool "startMC". All the image classifications were then verified and corrected by at least two other taxonomic experts. The accuracy of taxonomic classification and names was verified using AlgaeBase[43].

Data availability

The dataset entitled "MedPlanktonSet - A dataset of labeled IFCB images from the Mediterranean Sea" is publicly available on Zenodo (https://doi.org/10.5281/zenodo.15471023). It comprises three zip files and a csv file.

Code availability

The images and features provided in MedPlanktonSet were analyzed and extracted using codes that were already publicly available on the IFCB-analysis github (https://github.com/hsosik/ifcb-analysis/wiki).

Received: 11 June 2025; Accepted: 11 September 2025;
Published online: 24 October 2025

References

1. The Plankton Manifesto. A call for Plankton-Based Solutions to address the triple planetary crisis (biodiversity, climate & pollution). Ocean Steward Coalition. United Nations. UN Global Compact. 20 (2024).
2. Campbell, L., Gaonkar, C. C. & Henrichs, D. W. Chapter 5 - Integrating imaging and molecular approaches to assess phytoplankton diversity in *Advances in Phytoplankton Ecology* (eds Lesley A. Clementson, Ruth S. Eriksen, & Anusuya Willis) 159–190 (Elsevier, 2022).
3. Suthers, I., Rissik, D. & Richardson, A. *Plankton: A guide to their ecology and monitoring for water quality*. (CSIRO publishing, 2019).
4. Choi, J. W. & Stoecker, D. K. Effects of Fixation on Cell Volume of Marine Planktonic Protozoa. *Applied and Environmental Microbiology* **55**, 1761–1765, https://doi.org/10.1128/aem.55.7.1761-1765.1989 (1989).
5. Holland, M. M. *et al*. Mind the gap-The need to integrate novel plankton methods alongside ongoing long-term monitoring. *Ocean & Coastal Management* **262**, 107542, https://doi.org/10.1016/j.ocecoaman.2025.107542 (2025).
6. Stoecker, D. K., Gifford, D. J. & Putt, M. Preservation of marine planktonic ciliates: losses and cell shrinkage during fixation. *Marine Ecology Progress Series* **110**, 293–299, https://doi.org/10.3354/meps110293 (1994).
7. Medlin, L. K. & Kooistra, W. H. Methods to estimate the diversity in the marine photosynthetic protist community with illustrations from case studies: a review. *Diversity* **2**, 973–1014, https://doi.org/10.3390/d2070973 (2010).
8. Sonnet, V., Mouw, C. B., Ciochetto, A. B. & Carney-Almeida, J. Hit or miss? Impact of time series resolution on resolving phytoplankton dynamics at hourly, weekly, and satellite remote sensing frequencies. *Limnology and Oceanography: Methods* **22**, 254–267, https://doi.org/10.1002/lom3.10604 (2024).
9. Lombard, F. *et al*. Globally consistent quantitative observations of planktonic ecosystems. *Frontiers in Marine Science* **6**, 196, https://doi.org/10.3389/fmars.2019.00196 (2019).
10. Olson, R. J. & Sosik, H. M. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods* **5**, 195–203, https://doi.org/10.4319/lom.2007.5.195 (2007).
11. Campbell, L., Henrichs, D. W., Olson, R. J. & Sosik, H. M. Continuous automated imaging-in-flow cytometry for detection and early warning of Karenia brevis blooms in the Gulf of Mexico. *Environmental Science and Pollution Research* **20**, 6896–6902, https://doi.org/10.1007/s11356-012-1437-4 (2013).
12. Campbell, L. *et al*. First harmful Dinophysis (Dinophyceae, Dinophysiales) bloom in the US is revealed by automated imaging flow cytometry. *Journal of Phycology* **46**, 66–75, https://doi.org/10.1111/j.1529-8817.2009.00791.x (2010).
13. Fachon, E. *et al*. Tracking a large-scale and highly toxic Arctic algal bloom: Rapid detection and risk communication. *Limnology and Oceanography Letters* **10**, 62–72, https://doi.org/10.1002/lol2.10421 (2025).
14. Fischer, A. D. *et al*. Nutrient limitation dampens the response of a harmful algae to a marine heatwave in an upwelling system. *Limnology and Oceanography* **9999**, 1–17, https://doi.org/10.1002/lno.12604 (2024).
15. Brownlee, E. F., Olson, R. J. & Sosik, H. M. Microzooplankton community structure investigated with imaging flow cytometry and automated live-cell staining. *Marine Ecology Progress Series* **550**, 65–81, https://doi.org/10.3354/meps11687 (2016).
16. Brosnahan, M. L. *et al*. Rapid growth and concerted sexual transitions by a bloom of the harmful dinoflagellate Alexandrium fundyense (Dinophyceae). *Limnology and Oceanography* **60**, 2059–2078, https://doi.org/10.1002/lno.10155 (2015).
17. Catlett, D. *et al*. Temperature dependence of parasitoid infection and abundance of a diatom revealed by automated imaging and classification. *Proceedings of the National Academy of Sciences* **120**, e2303356120, https://doi.org/10.1073/pnas.2303356120 (2023).
18. Catlett, D. *et al*. Concurrent DNA meta-barcoding and plankton imaging reveal novel parasitic infection and competition in a diatom. *Limnology and Oceanography* **9999**, 1–16, https://doi.org/10.1002/lno.12629 (2024).
19. Peacock, E. E., Olson, R. J. & Sosik, H. M. Parasitic infection of the diatom Guinardia delicatula, a recurrent and ecologically important phenomenon on the New England Shelf. *Marine Ecology Progress Series* **503**, 1–10, https://doi.org/10.3354/meps10784 (2014).
20. Houliez, E., Fischer, A. D., Bill, B. D. & Moore, S. K. Does prey availability influence the detection of Dinophysis spp. by the imaging FlowCytobot? *Harmful Algae* **130**, 102544, https://doi.org/10.1016/j.hal.2023.102544 (2023).
21. Ladds, M., Sosik, H. M. & Gobler, C. J. Prey morphotype and abundance controls plastid retention and bloom dynamics for a mixotrophic dinoflagellate. *Limnology and Oceanography* **69**, 2732–2747, https://doi.org/10.1002/lno.12708 (2024).
22. Sosik, H. M. & Olson, R. J. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods* **5**, 204–216, https://doi.org/10.4319/lom.2007.5.204 (2007).
23. Eerola, T. *et al*. Survey of automatic plankton image recognition: challenges, existing solutions and future perspectives. *Artificial Intelligence Review* **57**, 114, https://doi.org/10.1007/s10462-024-10745-y (2024).
24. Irisson, J.-O., Ayata, S.-D., Lindsay, D. J., Karp-Boss, L. & Stemmann, L. Machine Learning for the Study of Plankton and Marine Snow from Images. *Annual Review of Marine Science* **14**, 277–301, https://doi.org/10.1146/annurev-marine-041921-013023 (2022).
25. Belcher, B. T. *et al*. Demystifying image-based machine learning: a practical guide to automated analysis of field imagery using modern machine learning tools. *Frontiers in Marine Science* **10**, 1157370, https://doi.org/10.3389/fmars.2023.1157370 (2023).
26. Kraft, K. *et al*. Towards operational phytoplankton recognition with automated high-throughput imaging, near-real-time data processing, and convolutional neural networks. *Frontiers in Marine Science* **9**, 867695, https://doi.org/10.3389/fmars.2022.867695 (2022).
27. Orenstein, E. C. *et al*. Machine learning techniques to characterize functional traits of plankton from image data. *Limnology and Oceanography* **67**, 1647–1669, https://doi.org/10.1002/lno.12101 (2022).
28. Kenitz, K. M. *et al*. Convening Expert Taxonomists to Build Image Libraries for Training Automated Classifiers. *Limnology and Oceanography Bulletin* **32**, 89–97, https://doi.org/10.1002/lob.10584 (2023).
29. McQuatters-Gollop, A. *et al*. From microscope to management: The critical value of plankton taxonomy to marine policy and biodiversity conservation. *Marine Policy* **83**, 1–10, https://doi.org/10.1016/j.marpol.2017.05.022 (2017).
30. Clayton, S. *et al*. Imaging Technologies Build Capacity and Accessibility in Phytoplankton Species Identification Expertise for Research and Monitoring: Lessons Learned During the COVID-19 Pandemic. *Frontiers in Microbiology* **13**, 823109, https://doi.org/10.3389/fmicb.2022.823109 (2022).
31. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* **106**, 249–259, https://doi.org/10.1016/j.neunet.2018.07.011 (2018).
32. Sosik, H. M., Peacock, E. E. & Brownlee, E. F. WHOI-Plankton: Annotated plankton images - data set for developing and evaluating classification methods [dataset]. https://doi.org/10.1575/1912/7341 (2014).
33. Kraft, K. *et al*. SYKE-plankton [dataset]. https://doi.org/10.23728/b2share.abf913e5a6ad47e6baa273ae0ed6617a (2022).
34. Torstensson, A., Skjevik, A.-T., Mohlin, M., Karlberg, M. & Karlson, B. SMHI IFCB Plankton Image Reference Library. Swedish Meteorological and Hydrological Institute (SMHI). *[Dataset]*. https://doi.org/10.17044/scilifelab.25883455.v4 (2024).
35. Badreldeen Bdawy Mohamed, O., Eerola, T., Kraft, K., Lensu, L. & Kälviäinen, H. Open-Set Plankton Recognition Using Similarity Learning. *Advances in Visual Computing* **13598**, 174–183, https://doi.org/10.1007/978-3-031-20713-6_13 (2022).
36. Deng, J. *et al*. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 https://doi.org/10.1109/CVPR.2009.5206848 (2009).
37. Borrelli, F. *et al*. 3D holographic flow cytometry measurements of microalgae: strategies for angle recovery in complex rotation patterns. *Lab Chip* D5LC00559K. https://doi.org/10.1039/D5LC00559K (2025).
38. Ronen, R., Attias, Y., Schechner, Y. Y., Jaffe, J. S. & Orenstein, E. Plankton reconstruction through robust statistical optical tomography. *Journal of the Optical Society of America A* **38**, 1320–1331, https://doi.org/10.1364/JOSAA.423037 (2021).

39. Dugenne, M. *et al*. First release of the Pelagic Size Structure database: global datasets of marine size spectra obtained from plankton imaging devices. *Earth System Science Data* **16**, 2971–2999, https://doi.org/10.5194/essd-16-2971-2024 (2024).
40. Houliez, E. *et al*. Deploying the Imaging FlowCytobot in the Mediterranean Sea and developing a Trait-Based Classifier for High-Frequency Screening of Phytoplankton Traits and Dynamics. *2024 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*, 401–404. https://doi.org/10.1109/MetroSea62823.2024.10765619 (2024).
41. Sonnet, V., Guidi, L., Mouw, C. B., Puggioni, G. & Ayata, S.-D. Length, width, shape regularity, and chain structure: time series analysis of phytoplankton morphology from imagery. *Limnology and Oceanography* **67**, 1850–1864, https://doi.org/10.1002/lno.12171 (2022).
42. Mente, M. S., Houliez, E., Scalco, E., Roselli, L. & Sarno, D. MedPlanktonSet - A dataset of labeled IFCB images from the Mediterranean Sea [dataset]. *Stazione Zoologica Anton Dohrn*. https://doi.org/10.5281/zenodo.15471023 (2025).
43. Guiry, M. D. & Guiry, G. M. AlgaeBase. *World-wide electronic publication*. https://www.algaebase.org (2025).

Acknowledgements

We thank Iole Di Capua for the annotation of zooplankton images and Isabella Percopo, Adriana Zingone and Marina Montresor for supporting the identification of phytoplankton images. We also thank Ferdinando Tramontano and the crew of the R/V Vettoria for sampling. We thank the Editor and the two anonymous reviewers for their constructive comments. EH was funded by the Italian Ministry of University and Research, PRIN – PNRR 2022 project FOOTMARKS (P2022MA95R). MSM has been supported by a PhD fellowship funded by the Stazione Zoologica Anton Dohrn (Open University – Stazione Zoologica Anton Dohrn PhD Program). This work was partially funded by the National Biodiversity Future Centre (NBFC) Program, Italian Ministry of University and Research, PNRR, Missione 4 Componente 2 Investimento 1.4 (Project: CN00000033). Financial support was also provided by Teresa Romeo through the CRIMAC project funded by the Italian Ministry of University and Research FSC 2014–2020 (grant number 20A02494).

Author contributions

Data collection: M.S.M., E.S.; First annotation of images: M.S.M.; Verification and correction of image annotations: E.S., D.S., E.H., L.R.; Final verification and validation of taxonomic classifications: E.S., D.S. Extraction of images and features: E.H.; Creation of the MedPlanktonSet website: E.H.; Files preparation: E.H., M.S.M.; Writing of the first draft of the manuscript: E.H., M.S.M.; Revision and editing of the manuscript: E.S., D.S., L.R.; Supervision and funding acquisition: D.S., L.R. All the authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05973-y.

Correspondence and requests for materials should be addressed to L.R. or D.S.