



OPEN

## Semi-automated classification of colonial *Microcystis* by FlowCAM imaging flow cytometry in mesocosm experiment reveals high heterogeneity during seasonal bloom

Yersultan Mirasbekov<sup>1,11</sup>, Adina Zhumakhanova<sup>1,11</sup>, Almira Zhantuyakova<sup>1,10,11</sup>, Kuanysb Sarkytbayev<sup>1,2</sup>, Dmitry V. Malashenkov<sup>3</sup>, Assel Baishulakova<sup>1</sup>, Veronika Dashkova<sup>1,4</sup>, Thomas A. Davidson<sup>5</sup>, Ivan A. Vorobjev<sup>1,2</sup>, Erik Jeppesen<sup>5,6,7,8</sup> & Natasha S. Barteneva<sup>1,9</sup>✉

A machine learning approach was employed to detect and quantify *Microcystis* colonial morphospecies using FlowCAM-based imaging flow cytometry. The system was trained and tested using samples from a long-term mesocosm experiment (LMWE, Central Jutland, Denmark). The statistical validation of the classification approaches was performed using Hellinger distances, Bray–Curtis dissimilarity, and Kullback–Leibler divergence. The semi-automatic classification based on well-balanced training sets from *Microcystis* seasonal bloom provided a high level of intergeneric accuracy (96–100%) but relatively low intrageneric accuracy (67–78%). Our results provide a proof-of-concept of how machine learning approaches can be applied to analyze the colonial microalgae. This approach allowed to evaluate *Microcystis* seasonal bloom in individual mesocosms with high level of temporal and spatial resolution. The observation that some *Microcystis* morphotypes completely disappeared and re-appeared along the mesocosm experiment timeline supports the hypothesis of the main transition pathways of colonial *Microcystis* morphoforms. We demonstrated that significant changes in the training sets with colonial images required for accurate classification of *Microcystis* spp. from time points differed by only two weeks due to *Microcystis* high phenotypic heterogeneity during the bloom. We conclude that automatic methods not only allow a performance level of human taxonomist, and thus be a valuable time-saving tool in the routine-like identification of colonial phytoplankton taxa, but also can be applied to increase temporal and spatial resolution of the study.

Studying plankton organisms is critical to assess the health of ocean and freshwater ecosystems. Over the past decade, a combination of image analysis technologies and machine learning algorithms has been applied to characterize zooplankton<sup>1–5</sup> and phytoplankton organisms<sup>6–13</sup>. Light microscopy is still considered the golden

<sup>1</sup>School of Sciences and Humanities, Nazarbayev University, Nur-Sultan 010000, Kazakhstan. <sup>2</sup>National Laboratory Astana, Nur-Sultan 010000, Kazakhstan. <sup>3</sup>Department of General Ecology and Hydrobiology, Lomonosov Moscow State University, 119991 Moscow, Russian Federation. <sup>4</sup>School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan 010000, Kazakhstan. <sup>5</sup>Department of Bioscience, Aarhus University, 8600 Silkeborg, Denmark. <sup>6</sup>Institute of Marine Sciences, Middle East Technical University, Mersin 33731, Turkey. <sup>7</sup>Limnology Laboratory, Department of Biological Sciences and Centre for Ecosystem Research and Implementation, Middle East Technical University, Ankara 06800, Turkey. <sup>8</sup>Sino–Danish Centre for Education and Research, Beijing 100049, China. <sup>9</sup>The Environmental Research and Efficiency Cluster (EREC), Nazarbayev University, Nur-Sultan 010000, Kazakhstan. <sup>10</sup>Present address: University of British Columbia, Vancouver, Canada. <sup>11</sup>These authors contributed equally: Yersultan Mirasbekov, Adina Zhumakhanova and Almira Zhantuyakova. ✉email: natalie.barteneva@nu.edu.kz

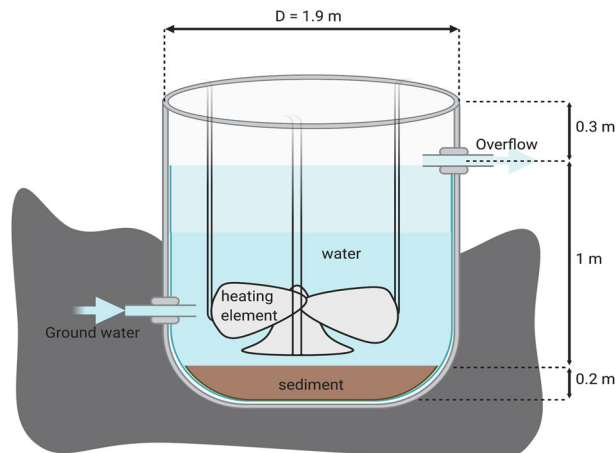
standard technique providing high-resolution plankton images for qualitative and quantitative assessment. However, microscopy is a time-consuming approach that requires a high level of taxonomic skills and can result in human-based misclassification and underestimation of rare species<sup>14–16</sup>. Moreover, microscopy identification of plankton is limited by an increased sample variability and diversity of the spatial orientation of the plankton organisms in the imaging plane, presence of organic matter particles in the water samples, and decay. There is, therefore, a high demand for automating the process of classification to enable high-throughput data processing. In the last decades imaging cytometers such as FlowCytobot<sup>6</sup>, FlowCAM<sup>5,8,12,13</sup>, and Imagestream X Mark II<sup>9</sup> have been used to improve and speed up phytoplankton image acquisition. The FlowCAM instrument has become a valuable tool in marine and freshwater plankton studies because it enables researchers to classify, count, and monitor different plankton organisms<sup>17,18</sup> in the preferred detection size range of 20–300  $\mu\text{m}$ <sup>19</sup>. Image analysis and classification of large image datasets are primarily sensitive to high variations, manual misclassifications, and biased interpretations<sup>20</sup>. However, so far, phytoplankton analysis with imaging cytometers has mainly been limited to the genus level and not included differentiation of colonial morphospecies<sup>5–13</sup>.

*Microcystis* spp. is a dominant cyanobacterial genus appearing in all regions of the world. *Microcystis* can form toxic blooms whose occurrence is expanding; thus, more than 100 countries worldwide have documented such toxic blooms in freshwater lakes and streams<sup>21</sup>. Toxic strains of *Microcystis* produce hepatotoxins and neurotoxins<sup>22,23</sup> that constitute a serious threat to human health by contaminating drinking water resources. The toxins of *Microcystis* spp. have harmful effects on different trophic levels in an aquatic food web, such as phytoplankton, zooplankton, fish, and mollusks<sup>24–28</sup>. Depending on the prevailing environmental conditions, *Microcystis* tend to form colonial structures covered by a thick polysaccharide sheath (mucilage)<sup>29</sup>. Colony formation by *Microcystis* can be induced by low temperatures, low light intensity, high lead ion concentrations, and the presence of other cyanobacterial species. According to Zheng et al.<sup>30</sup>, under laboratory conditions *Microcystis* spp. occur only as single or paired cells, preventing replication and study *Microcystis* spp. bloom formation. Morphospecies (or morphoforms) were identified in the *Microcystis* genus<sup>31</sup>, and their physiology, growth, and toxicity vary greatly<sup>32</sup>. Seasonal dynamics and increasing occurrence of water-bloom forming *Microcystis* is of great concern for the ecosystem due to the potential production of potentially toxic microcystins<sup>33</sup>, and *M. aeruginosa* is considered to be a major toxic morphospecies. *Microcystis* spp. occurs in the freshwater bodies mainly in a colonial form<sup>34</sup>, and their bloom dynamics were monitored by different research groups. Thus, in China, *Microcystis* blooms development and sustainment were studied in many lakes, including large Taihu and Dianchi lakes<sup>25,35,36</sup>. Thus, Otten and Paerl<sup>36</sup> studied by genotyping the single colonies of four different morphoforms of *Microcystis* spp. that comprised seasonal blooms in Lake Taihu, and reported that one morphospecies was genetically unique (*M. wesenbergii*) and three (*Microcystis aeruginosa*, *Microcystis flos-aquae*, and *Microcystis ichthyoblabe*) were genetically indistinguishable (96.4% identity of 16S–23S ITS sequences). Ishikawa et al.<sup>37</sup> examined *M. aeruginosa* and *M. wesenbergii* colonies in the Lake Biwa, Yamamoto, and Nakahara<sup>38</sup> investigated *Microcystis* spp. in Hirosawa-no-ike Pond in Japan. Kurmayer and co-authors, Via-Ordorica and others studied *Microcystis* colonies in European freshwater bodies<sup>39,40</sup>, and Alvarez and co-authors in Uruguay<sup>41</sup>.

We used a unique LMWE mesocosm experiment (Aarhus University, Denmark)<sup>42</sup> that, in contrast to the limited laboratory conditions, provides a dynamic system for the study of colonial phytoplankton. Previously, in 2018 we attempted classification of morphospecies in a study of the seasonal dynamics of different *Microcystis* spp.<sup>43</sup>. Five different *Microcystis* morphospecies (*M. aeruginosa*, *M. novacekii*, *M. smithii*, *M. wesenbergii*, and *M. ichthyoblabe*) were also detected and identified during the 2019 season. This study aimed to develop and validate a semi-quantitative machine learning algorithm for differentiation of *Microcystis* colonies intragenerically and from other phytoplankton colonial phytoplankton taxa (*Micractinium* genus) as well as from unicellular phytoplankton (*Cryptomonas* spp.). We distinguished five morphological colonial forms of *Microcystis* and found that the proposed intergeneric classification showed higher performance using minimized filter sets, whereas intrageneric differentiation had lower accuracy using high complexity filter sets. These semi-automated imaging cytometry-based classification results are comparable with the traditional human-based level of classification. We attribute the variations in intrageneric colonial analysis accuracy to the high heterogeneity of *Microcystis* spp. in the seasonal *Microcystis* bloom. Moreover, our observation that some *Microcystis* morphotypes completely disappeared and re-appeared along the mesocosm experiment timeline supports the hypothesis of the main transition pathways of colonial *Microcystis* morphoforms.

## Materials and methods

**Mesocosm experimental setup.** We collected phytoplankton samples from the AQUACOSM Lake Mesocosm Warming Experiment (AQUACOSM LMWE experiment) in the experimental facility of Aarhus University in Central Jutland, Denmark (56°140 N, 9°310 E) to study different phytoplankton genera with a FlowCAM imaging cytometer. The mesocosm facility consists of 24 artificially mixed flow-through mesocosms that were established in August 2003. The factorial experimental set-up combines three temperature scenarios and two nutrient levels, all in four replicates (detailed description of experimental design and set-up can be found in Liboriussen et al.<sup>42</sup>). Overall, the LMWE experiment includes six different types of tanks with low and high nutrient levels that are each divided into three sub-types (depending on the temperature of the water: unheated, heated according to IPCC climate scenario A2, and eight heated according to A2 + 50%)—all with three replicates. In the present study, we collected seasonal (from May 23, 2019, to September 17, 2019) samples on 13 dates in the 12 high nutrient tanks, in a total of 168 samples. The tanks are named A1–3, D1–3, F1–3, and G1–3. A schematic representation of a mesocosm is shown in Fig. 1. The samples were preserved with glutaraldehyde solution (Sigma-Aldrich, USA) at a final concentration of 1% and analyzed using a FlowCAM imaging cytometer (Yokogawa Fluid Imaging Inc., USA).



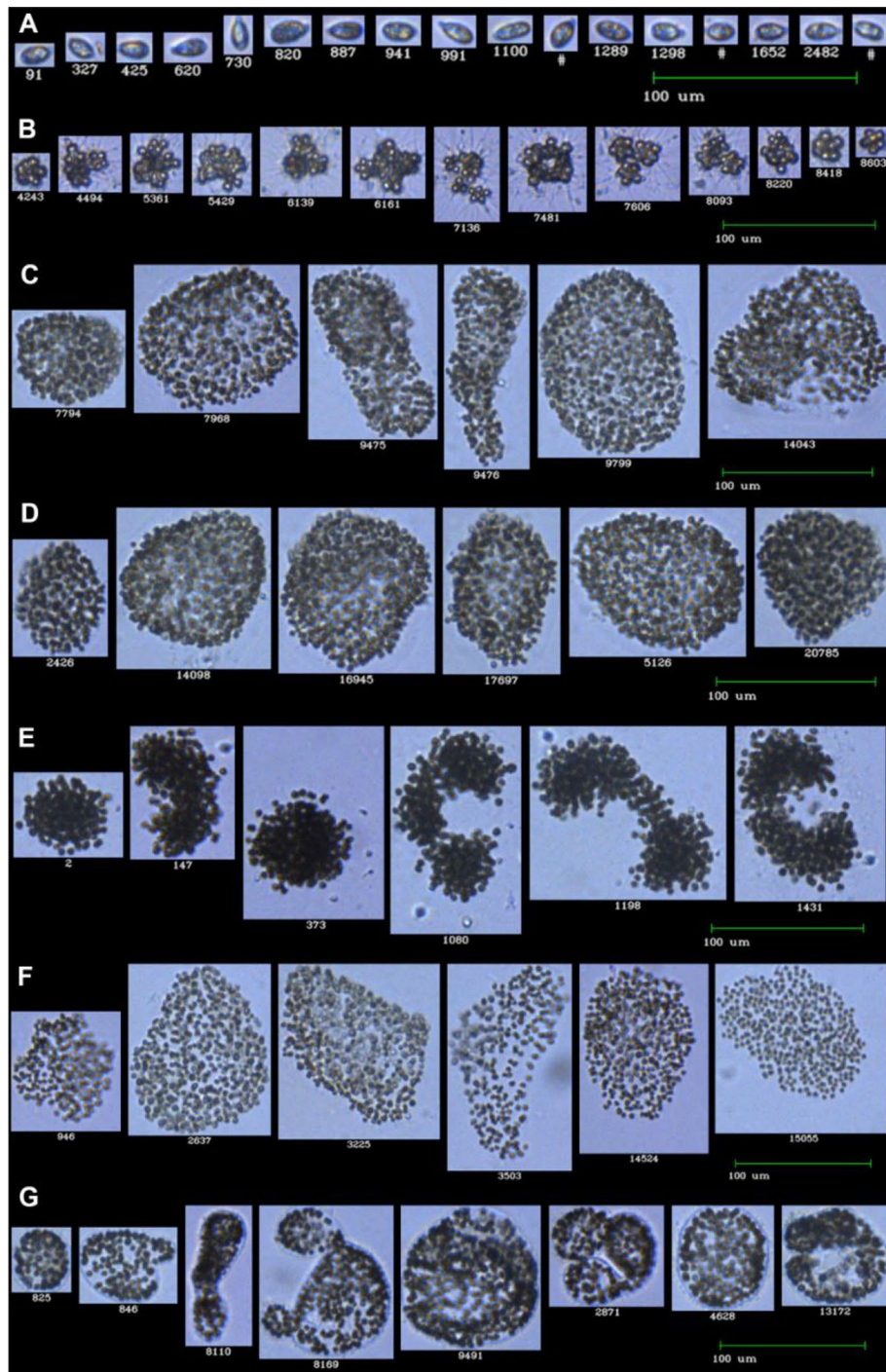
**Figure 1.** An image of one of the 24 flow-through tanks in the LMWE experiment run at the experimental facility belonging to Aarhus University, Denmark (modified from<sup>42</sup>). The collection tank was placed at right side (not included). This image was created with BioRender (<https://biorender.com/>).

**Instrumentation.** We used a benchtop FlowCAM imaging cytometer equipped with VisualSpreadsheet software (Yokagawa Fluid Imaging, USA). Samples were recorded in autoimage mode using combinations of 10× objective (NA = 0.3; resolution 1 pixel equals to 0.554 μm)/100 μm flow cell and/or 20× objective/50 μm flow cell for identification, classification, and quantification. Identification and quantification of phytoplankton cells by light microscopy were performed under Leica DM500 (Leica Microsystems, Germany) equipped with phase contrast and series of objectives.

**Phytoplankton morphological classification.** As stated before, we elucidated semi-automated classification for morphological analysis between a genus-level and colonial morphospecies dataset. Therefore, phytoplankton classification in this study was focused on on *Cryptomonas* sp., *Micractinium*, and *Microcystis* morphospecies. *Microcystis* spp. was divided into five morphospecies, namely *M. aeruginosa*, *M. ichthyoblabe*, *M. novacekii*, *M. smithii*, and *M. wesenbergii*; and used this morphological difference between colonial morphospecies for intrageneric (within genus) classification. Training sets were developed with an expert taxonomist's participation (with > 10 years of experience). On the images, *Cryptomonas* sp. (hereafter *Cryptomonas*) were defined as brown-green cells with two flagella<sup>44</sup>. Examples from FlowCAM imaging are given in Fig. 2A. For *Micractinium*, spines/bristles were used for the identification, as illustrated in Fig. 2B. Five major morphospecies of *Microcystis* spp. were separated into classes of *M. novacekii*, *M. ichthyoblabe*, *M. smithii*, *M. aeruginosa*, and *M. wesenbergii* (see Fig. 2C–G). During a seasonal *Microcystis* bloom, some of the images show colony remnants with few or no cells. These images were assigned to the class “Membrane”. Small and dispersed non-colonial forms of *Microcystis* spp. classified as “Undefined”. The two latter classes were not used in the training sets in order to ensure clear separation between the five colonial morphospecies.

**Preparation of training set and dataset.** Different mesocosm samples were mixed to achieve an optimal number of representatives from all three to examine intergeneric classification between colonial forms of *Microcystis*, *Micractinium*, and single cells of *Cryptomonas* genera. It is important to note that *Microcystis novacekii* was used in the training set as the only representative of the genus *Microcystis*. None of the tanks were found to contain all three representatives in the high image counts. Using the preliminary abundance analysis, four samples (D1\_17/09/2019, D1\_22/08/2019, G3\_11/06/2019, G3\_17/06/2019) were mixed in equal volumes, with a final volume of 10 mL. Then, bright-field images were recorded applying FlowCAM imaging flow cytometry in autoimage mode. Each sample was passed through a 100 μm filter and then recorded with a 10× magnification objective (we observed only a few large colonies with a light microscope in unfiltered samples and for safety reasons (to prevent clogging), the samples were filtered for use on FlowCAM). Intergeneric classification was performed manually, and the distribution of classes is given in Suppl. Table 1. In total, the dataset included 972 images.

The same procedure was conducted to acquire intrageneric data of *Microcystis* spp. Images were recorded from 168 Mesocosm samples, out of which 69 samples were positive for the presence of *Microcystis* spp. The overall dataset included 119,135 images of *Microcystis* spp. Excluding *Sheaths* and *Non-colonial clusters* images, there were 70,305 images of colonial *Microcystis* separated into five classes based on the previous section's classification. The D1\_17/09/2019 sample (collected from Mesocosm tank D1 in 17/09/2019) was used for training and test datasets as it contained a high proportion of all five *Microcystis* morphospecies in the amount of 5068 images; the detailed data distribution can be found in Suppl. Table 1. The image recordings from D2\_09/03/2019 sample (collected from Mesocosm tank D2 on 09/03/2019) was used to assess intrageneric classifiers in a different dataset with 2,552 images of colonial *Microcystis*.



**Figure 2.** FlowCAM imaging flow cytometry of phytoplankton from the LMWE 2019 experiment. (A) *Cryptomonas* sp.; (B) *Micractinium*; (C) *Microcystis aeruginosa*; (D) *M. ichthyoblabe*; (E) *M. novacekii*; (F) *M. smithii*; (G) *M. wesenbergii*.

**Feature extraction and evaluation of filter sets.** The imaging and cytometric data from LMWE samples were acquired using the FlowCAM instrument. To achieve an even distribution of representatives, 150 images were randomly moved from the Classification Window to Open view in the VisualSpreadsheet software. So, we used 150 representative images for each of the abundant phytoplankton species (*M. aeruginosa*, *M. ichthyoblabe*, *M. novacekii*, *M. wesenbergii*, *M. smithii*, *Cryptomonas*, and *Micractinium*) to train VisualSpreadsheet software to differentiate listed phytoplankton taxa. Then, 25 or 50 images were randomly selected as training data (further referred to as “25” and “50”), and the VisualSpreadsheet software generated an initial set of classification parameters specific for each species based on selected images. After auto-filtering, there were 48 image

features left, which are based on five different categories: size, shape, texture, gray-scale signal, and color signal measurements.

The Filter Dialog box contained image features and their ranges between the minimum and maximum values. The filter sets were reduced by systematic selection to leave the minimal number of features until the filter set's accuracy started deviating significantly. In other words, the procedure was performed till the overall Accuracy value reached 0.75 and below, according to Eq. (1). The changes in filter sets resulted in an increase in the true positive rate with a drawback of decrease in the true negative rate. The changes in the true positive rate were recorded for each filter set, and they are compared in Suppl. Fig. 1. Shortlisted particle properties were saved in a filter format file. The value ranges were separately recorded for the selection of "25" and "50" images for training. These value ranges provide the basis for selection of particles/images in the dataset. It was decided to include the intersecting ranges between "25" and "50" to create the third type of filter set, named as "Intersection". As this method is based on selected images, the produced classifier is equivalent to selecting more images ( $50 \leq X \leq 75$ ). It was done to remove false positive results and increase the overall accuracy of the classification. So, the highest min. value and lowest max. value were taken manually to decrease the range for each parameter.

$$Accuracy = \frac{True\ Negative\ results + True\ Positive\ results}{Total\ number\ of\ results} \quad (1)$$

Equation (1): Equation for Accuracy that was used to leave the most important particle properties.

The above-mentioned filter sets were saved in filter format file and used for classification and further evaluation of the test dataset that excluded images from training dataset. The results of the test classification were recorded and used to construct a confusion matrix. Finally, the performance of the classification was evaluated based on reliability (precision) and accuracy of each classifier according to the procedure adapted from Aldenhoff et al.<sup>45</sup>.

Hellinger distance (HD) was used to identify any dataset shifts between the training dataset ( $T$ ) and the test dataset ( $X$ ). Equation (2) was adapted from the work of Cieslak and Chawla<sup>46</sup>. The minimum value for HD is 0, which is mainly observed when datasets are identical. The " $1/\sqrt{2}$ " was added to change the maximum value of the HD from  $\sqrt{2}$  (approximately 1.41) to 1.0.

$$Hellinger\ distance(X, T) = \frac{1}{s} \sum_{f=1}^s HD_f(X, T) = \frac{1}{s} \sum_{f=1}^s \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^c \left( \sqrt{\frac{X_{f,k}}{X}} - \sqrt{\frac{T_{f,k}}{T}} \right)^2} \quad (2)$$

Equation (2): Equation for Hellinger distance between datasets  $X$  and  $T$ , where  $s$  represents the number of different filter sets used,  $HD_f(X, T)$  is the Hellinger distance for a given filter set ( $f$ ),  $c$  is the number of classes (bins) used in classification method,  $X$  is the total number of representative images in the dataset, and  $X_{f,k}$  is the number of classification matches with the feature set  $f$  that belongs to the class (or bin)  $k$  (the same definitions are applied in the training dataset  $T$ ).

Since we used test dataset with an uneven distribution of samples (Suppl. Table 1), the difference-based method was applied. The error for each class was identified using the equation for symmetric mean absolute percentage error (SMAPE), according to Eq. (3). This can be used as a clear indication of classification performance depending on the class<sup>47</sup>. The SMAPE analysis was performed to identify the error for each classification bin.

$$SMAPE(c_i) = \frac{1}{s} \sum_{f=1}^s \frac{|X_{f,c_i} - X'_{f,c_i}|}{X_{f,c_i} + X'_{f,c_i}} \quad (3)$$

Equation (3): Equation for symmetric mean absolute percentage error (SMAPE) for each class, where  $X_{f,c_i}$  is the actual value of the number of representatives for class  $c_i$  using the filter set  $k$ , and  $X'_{f,c_i}$  represents the forecast value.

Bray–Curtis dissimilarity (Eq. 4) and Kullback–Leibler Divergence (Eq. 5) were used as metrics of the overall performance. In both approaches, the zero value indicates that the generated forecast distribution is fully identical with the actual data. The Bray–Curtis dissimilarity determines dissimilarity between the actual data and the data predicted by filter sets using their relative abundance data<sup>47</sup>. The Kullback–Leibler Divergence uses probability distributions to perform natural measures of relative entropy<sup>46</sup>. These values were calculated to assess the performance of the intergeneric and intrageneric (*Microcystis* morphospecies) approaches.

$$Bray - Curtis\ dissimilarity = \frac{\sum_{k=1}^c |X_k - X'_k|}{\sum_{k=1}^c X_k + X'_k} \quad (4)$$

Equation (4): Equation for Bray–Curtis dissimilarity for the overall performance metric, where  $X_k$  is the actual value in the bin  $k$ , and  $X'_k$  represents the predicted value using the filter set  $k$ .

$$Kullback - Leibler\ Divergence(X \parallel X') = \sum_{k=1}^c \left[ \frac{X_k}{\sum_{k=1}^c X_k} \times \left( \log \frac{X_k}{\sum_{k=1}^c X_k} - \log \frac{X'_k}{\sum_{k=1}^c X'_k} \right) \right] \quad (5)$$

Equation (5): Equation for Kullback–Leibler Divergence, where  $X_k$  is the actual value in the bin  $k$ , and  $X'_k$  represents the predicted value using the filter set  $k$ .

Particle properties	Type	Descriptions	Value range for X
Average Blue	Color	Average pixel value for blue color plane	$X \in [0, 255]$ ;
Diameter (ABD)	Size	Circle-based diameter that equal to ABD Area	$X > 0$ ;
Edge gradient	Texture	Average pixels intensity of outside border of a particle after an application of Sobel Edge Detect convolution filter	$X \in [0, 255]$ ;
Intensity	Grayscale	Average grayscale value of pixels of a particle (grayscale sum / number of particle pixels)	$X \in [0, 255]$ ;
Length	Size	Maximum value of 36 feret measurements	$X > 0$ ;
Perimeter	Size	Total length of edges including edges of any hole	$X > 0$ ;
Ratio red/blue	Color	Ratio between Average Red and Average Blue	$X \geq 0$ ;
Ratio red/green	Color	Ratio between Average Red and Average Green	$X \geq 0$ ;
Roughness	Shape	Unevenness/irregularity of a particle's surface, defined as the ratio between perimeter and convex perimeter. Larger values have a non-convex perimeter and/or interior holes	$X \geq 1$ ; $X = 1$ for a filled shape with convex perimeter;
Sigma intensity	Grayscale	Standard deviation of particle's grayscale values	$X \geq 0$

**Table 1.** List of particle properties with corresponding type, description and possible value range. (modified from FlowCAM user manual).

## Results

**Classification parameters for each morphological class.** Classification parameters were extracted using the feature finder tool in VisualSpreadsheet software. The description and possible range for each acquired particle property are summarized in Table 1. The selected most important particle properties are listed in Table 2. The upper table (Table 2A) summarizes particle properties with corresponding value ranges for intergeneric classification between three different genera, *Cryptomonas*, *Micractinium*, and *Microcystis*. The lower table (Table 2B) includes the same information for intrageneric classification of the five *Microcystis* morphospecies. For the “25” or “50” images selected as training data, the ranges are shown as a minimum and maximum values. The “intersection” filter with narrowed value ranges is included to assess intercepting images between “25” or “50”.

**Evaluation of filter sets in high-throughput data set.** Generated sets of particle properties were examined on the test dataset that exhibited an uneven distribution (more detailed in Suppl. Table 1). The image collection covered only described genera-level and morphospecies-level classification. In other words, images of other species were excluded from the comparison of selected types of data.

The “Intersection” filter setting with narrowed ranges was used as the representative data. Results of semi-automated classification, “predicted” by classifiers, were compared with the manual classification data, that is considered as “True label”. The confusion matrices were constructed based on true positive results and misclassifications by calculating precision and false discovery rate, respectively (Table 3A,B). Overall performance for intergeneric and intrageneric classification in percentage is provided in Tables 3C,D, and more detailed information on the other two filter settings, when “25” and “50” images were used for training, can be found in Suppl. Tables 2 and 3.

The classification of intergeneric *Cryptomonas*, *Micractinium*, and *Microcystis* images was performed using only two-particle properties. Setting ranges for diameter (ABD) and intensity was enough to discriminate between one and the two other classes with an overall 96–100% performance for these classifiers. *Cryptomonas* has the smallest cell size (8–13  $\mu\text{m}$ ) of the three genera, and size was consequently used as a feature. On selected 25 images, the diameter ranges of the *Micractinium* and *Microcystis* filter sets did not intersect. However, when the training set was increased to 50, the intensity feature's inclusion became necessary to ensure accurate classification, implying combined use of the size and signal strength feature categories for the larger filter sets. The confusion matrices showed a low misclassification rate with an overall 96–100% performance for the dataset of 971 images. The differences between the three species (genera) were sufficient for the VisualSpreadsheet software to perform useful classification.

The performance of the intrageneric *Microcystis* spp. classification was considerably lower than for the intergeneric classification. Firstly, the filtering pipeline for classification included a wider variety of particle properties. Since the colony size of *Microcystis* spp. colonies varied, the basic diameter (ABD) parameter was not applied. However, the size-based parameters of perimeter and length were used for differentiation of *M. aeruginosa* and *M. ichthyoblabe*, respectively. In the training dataset, the false discovery rate was as low as 0.47. However, it was increased to 0.71 when *M. aeruginosa* classifier was applied to the test dataset. Additionally, Shape features were used in classifiers for *M. aeruginosa*, *M. ichthyoblabe*, and *M. smithii*. The initial Shape classifier had Edge gradient parameter due to a semi-transparent halo's appearance around *M. aeruginosa* colonies. The second shape feature was Roughness, which has increased values when bigger interior holes are present. In our prediction system, the classifier for *M. smithii* had a higher value range between 1.58 and 10.70 compared to *Microcystis ichthyoblabe* (1.38–4.16). The Roughness particle property differentiated the other two colonial morphospecies.

All other particle properties were based on Signal strength, namely Average Blue, Intensity, Ratio Red/Blue, Ratio Red/Green, and Sigma Intensity. The classifier for *M. wesenbergii* had a combination of Signal strength features, including Intensity, Average Blue (average value for blue color pixels), and Sigma Intensity (standard deviation of particle's grayscale values). These features could not adequately filter out *M. novacekii* images,

A						
Particle property	25 selected		50 selected		Intersection	
	Min	Max	Min	Max	Min	Max
<i>Cryptomonas</i>						
Diameter (ABD) (µm)	8.23	11.53	8.22	12.72	8.23	11.53
<i>Micractinium</i>						
Diameter (ABD) (µm)	16.19	40.52	15.85	40.52	16.19	40.52
Intensity	79.20	105.85	73.64	105.85	79.20	105.85
<i>Microcystis novacekii</i>						
Diameter (ABD) (µm)	45.35	85.69	33.48	83.89	45.35	83.89
Intensity	49.12	92.47	40.55	97.59	49.12	92.47
B						
Particle property	25 selected		50 selected		Intersection	
	Min	Max	Min	Max	Min	Max
<i>M. aeruginosa</i>						
Average blue	76.28	101.01	76.08	97.85	76.28	97.85
Edge gradient	99.93	159.87	96.11	159.87	99.93	159.87
Perimeter (µm)	391.00	1620.14	391.00	1460.73	391.00	1460.73
Ratio red/blue	1.15	1.28	1.15	1.30	1.15	1.28
Sigma intensity	23.72	34.09	23.68	35.70	23.72	34.09
<i>M. ichthyoblabe</i>						
Average blue	72.02	96.30	77.64	96.30	77.64	96.30
Intensity	78.76	110.43	83.81	110.43	83.81	110.43
Length (µm)	56.85	138.37	60.17	128.39	60.17	128.39
Ratio red/blue	1.17	1.32	1.15	1.32	1.17	1.32
Roughness	1.42	4.16	1.38	4.16	1.42	4.16
Sigma intensity	22.14	28.37	21.58	30.97	22.14	28.37
<i>M. novacekii</i>						
Intensity	48.36	80.51	46.45	74.81	48.36	74.81
<i>M. smithii</i>						
Intensity	94.55	117.32	93.98	120.00	94.55	117.32
Ratio red/green	1.17	1.35	1.16	1.39	1.17	1.35
Roughness	2.19	10.70	1.58	10.70	2.19	10.70
<i>M. wesenbergii</i>						
Average blue	65.24	85.13	61.31	85.13	65.24	85.13
Intensity	73.59	98.04	67.25	98.04	73.59	98.04
Ratio red/blue	1.19	1.36	1.19	1.39	1.19	1.36
Ratio red/green	1.17	1.23	1.15	1.27	1.17	1.23
Sigma intensity	26.89	43.70	26.10	47.12	26.89	43.70

**Table 2.** Set of particle properties within each filter set for intergeneric classification (A) and intrageneric classification (B).

resulting in a lower accuracy and reliability values of 73.6% and 43.3%, respectively. However, these classifier features were efficient in elimination of *M. aeruginosa*, *M. ichthyoblabe*, and *M. smithii* images from *M. wesenbergii* classifier bin, which resulted in considerably low false discovery rate of 0.02–0.07. Finally, *M. novacekii* was identified through the single gate of Intensity feature because the small value range did not intersect with other morphospecies, and provide a relatively high precision of 98.1%.

Since the test dataset was imbalanced with an uneven distribution of representatives, balanced accuracy was calculated for each filter set application by averaging the true positive rate and true negative rate. Different validation methods were used to examine both the training dataset and the test dataset for each filter set and the summary for the calculations is given in Table 4.

The results show that intergeneric classifications between three different genera have lower Hellinger distance, which indicates that small data shifts can influence the performance of filter sets. The data shift value for morphospecies (intrageneric) classification was considerably higher than for the intergeneric classification, which affected their classification accuracy. However, there was no strong linear correlation between Hellinger distance and balanced accuracy, especially in the intergeneric classification.

For both the intergeneric and intrageneric approaches, the accuracy values for *M. novacekii* were around 91%. Other genera (*Cryptomonas* and *Micractinium*) had balanced accuracy values of 87–90%, while the other

A	True label			
	Genus:	<i>Cryptomonas</i>	<i>Micractinium</i>	<i>Microcystis</i>
Predicted label [int.]	<i>Cryptomonas</i>	66	1	0
	<i>Micractinium</i>	0	236	0
	<i>Microcystis</i>	0	0	100

B	True label					
	Morphospecies:	<i>M. aeruginosa</i>	<i>M. ichthyoblabe</i>	<i>M. novacekii</i>	<i>M. smithii</i>	<i>M. wesenbergii</i>
Predicted label [int.]	<i>M. aeruginosa</i>	110	363	5	27	6
	<i>M. ichthyoblabe</i>	79	377	2	15	9
	<i>M. novacekii</i>	0	23	2438	0	24
	<i>M. smithii</i>	48	79	16	84	14
	<i>M. wesenbergii</i>	21	25	142	6	148

C	True label				Reliability
	Genus:	<i>Cryptomonas</i>	<i>Micractinium</i>	<i>Microcystis</i>	
Predicted label [int.]	<i>Cryptomonas</i>	16.4	0.2	0.0	98.5
	<i>Micractinium</i>	0.0	58.6	0.0	100.0
	<i>Microcystis</i>	0.0	0.0	24.8	100.0
	<b>Accuracy</b>	<b>100.0</b>	<b>99.6</b>	<b>100.0</b>	<b>99.8</b>

D	True label						Reliability
	Morphospecies:	<i>M. aeruginosa</i>	<i>M. ichthyoblabe</i>	<i>M. novacekii</i>	<i>M. smithii</i>	<i>M. wesenbergii</i>	
Predicted label [int.]	<i>M. aeruginosa</i>	2.7	8.9	0.1	0.7	0.1	21.5
	<i>M. ichthyoblabe</i>	1.9	9.3	0.0	0.4	0.2	78.2
	<i>M. novacekii</i>	0.0	0.6	60.0	0.0	0.6	98.1
	<i>M. smithii</i>	1.2	1.9	0.4	2.1	0.3	34.9
	<i>M. wesenbergii</i>	0.5	0.6	3.5	0.1	3.6	43.3
	<b>Accuracy</b>	<b>42.6</b>	<b>43.5</b>	<b>93.7</b>	<b>63.6</b>	<b>73.6</b>	<b>77.7</b>

**Table 3.** The results of classification using intersecting ranges for the intergeneric classification of three classes (upper row) and the intrageneric classification of five *Microcystis* morphospecies (lower row). (A, B) Numbers represent results when corresponding classifier was applied. Confusion matrices for used filter sets with precision values for correct predictions (highlighted in green) and false discovery rate for misclassifications (highlighted in orange). The coloring was assigned depending on relative frequency in each column. (C, D) Percentage values for accuracy and reliability (precision) of used methods. Overall performance for intergeneric and intrageneric classification in percentage is highlighted in blue.

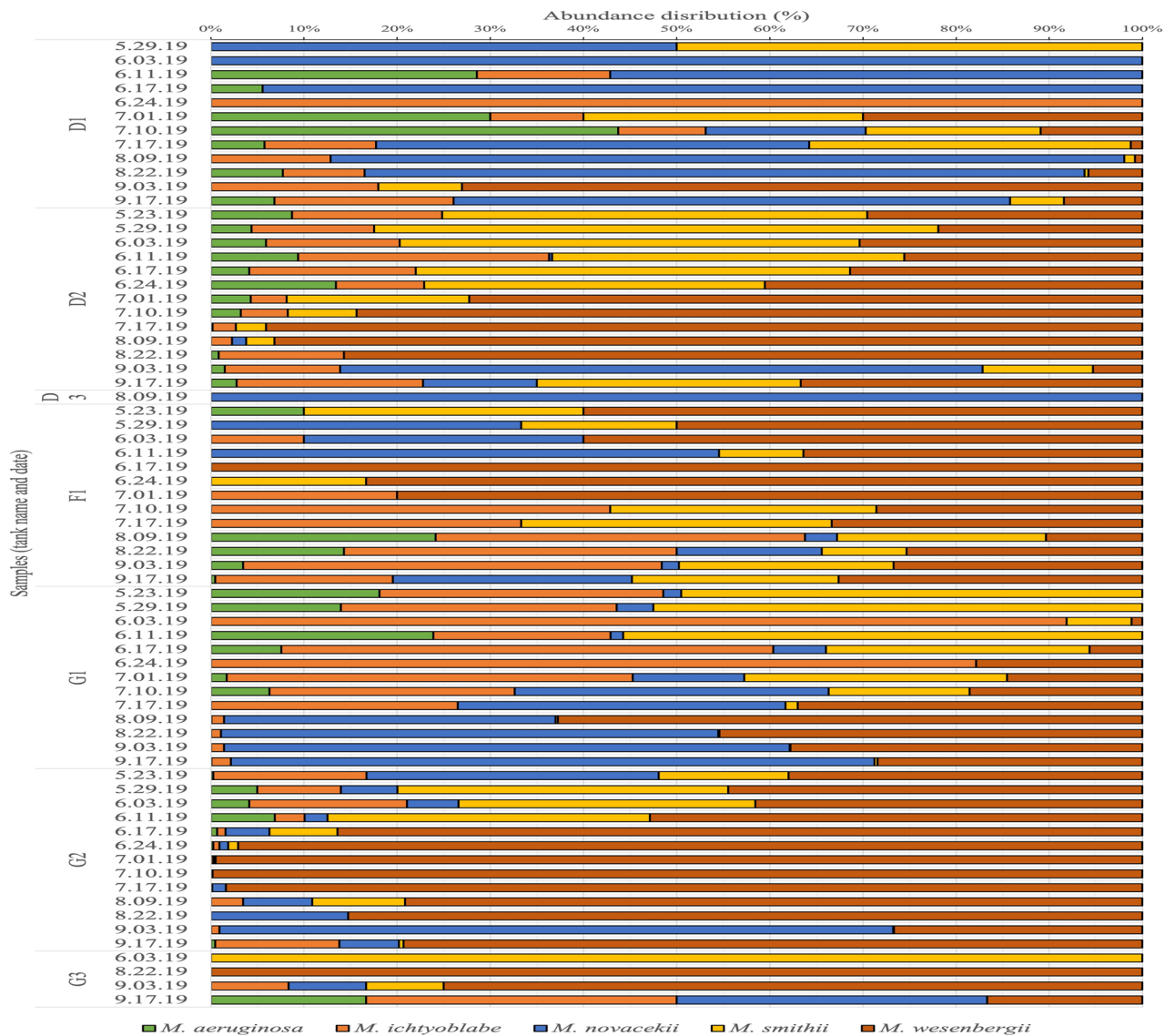
Filter set used for classification	Hellinger distance $x \in [0,1]$	Balanced accuracy (%)	SMAPE (%)	Bray–Curtis dissimilarity $x \in [0,1]$	Kullback–Leibler Divergence $x \geq 0$
<b>Intergeneric classification</b>					
<i>Cryptomonas</i>	0.13	90.13	3.60	0.128	0.0006
<i>Micractinium</i>	0.10	87.26	5.04		
<i>M. novacekii</i>	0.07	91.32	3.17		
<b>Intrageneric classification</b>					
<i>M. aeruginosa</i>	0.19	73.2	17.24	0.145	0.0281
<i>M. ichthyoblabe</i>	0.15	71.3	16.44		
<i>M. novacekii</i>	0.24	90.7	2.75		
<i>M. smithii</i>	0.17	77.7	11.05		
<i>M. wesenbergii</i>	0.16	74.3	9.68		

**Table 4.** Summary of the evaluation of filter sets using balanced accuracy, Hellinger distance and SMAPE, Bray–Curtis dissimilarity, and Kullback–Leibler Divergence.

four *Microcystis* morphospecies has lower values of 73–78%. The SMAPE analysis exhibited an error percentage of  $\leq 5.0\%$  for the intergeneric classification, rising to 17.2% for the *M. aeruginosa* filter set in the morphospecies classification.

The results of the two classification approaches were evaluated by Bray–Curtis dissimilarity, and the intrageneric classification had a higher dissimilarity (0.145) than the classification between genera (0.128). The results were checked by calculating Kullback–Leibler Divergence (KLD) to compare the probability distributions of predicted data and actual data. The calculations also indicated a lower KLD value of 0.0006 for the intergeneric





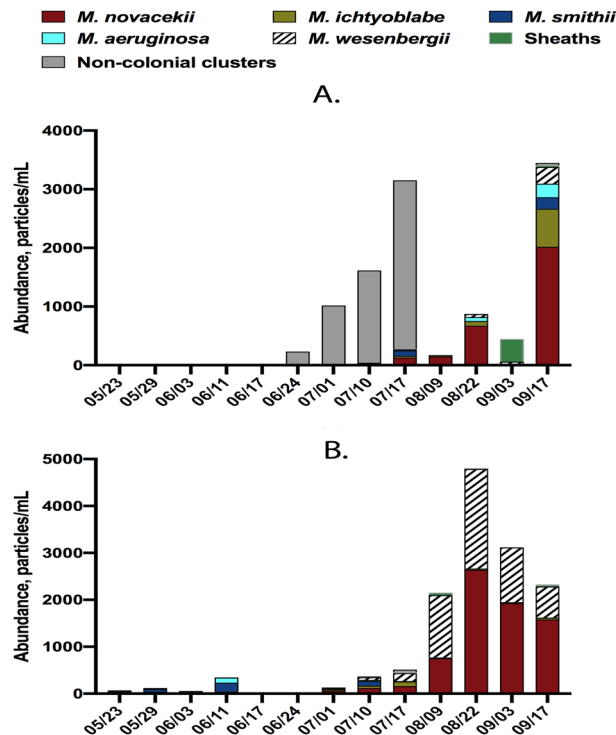
**Figure 3.** The distribution of *Microcystis* morphospecies in samples from the LMWE experiment in 2019. Grouping is based on type of tank and the date when the samples were taken.

classification compared to the intrageneric classification value of 0.0281. In other words, the classification information lost using the method of genera classification is lower than when using the classification between morphospecies.

**Dynamics of a seasonal *Microcystis* bloom succession in LMWE-2019 mesocosm.** The results of dominant species analysis by light microscopy in the tanks containing *Microcystis* spp. are listed in Suppl. Table 4; and often included *Micractinium* spp. (detailed in Suppl. Table 4) or *Cryptomonas* spp. (D3, F3, G3 tanks, at different dates). The semi-automatic classification based on well-balanced training sets from *Microcystis* seasonal bloom provided a high level of intergeneric accuracy (96–100%) (in comparison with *Micractinium* spp. and *Cryptomonas* spp.) but relatively low intrageneric accuracy (67–78%). The percentage distribution of colonial *Microcystis* morphospecies in samples from the LMWE experiment in 2019 is presented in Fig. 3 below. Importantly, there was a sequential appearance of the *Microcystis* spp. morphotypes, and also some *Microcystis* morphotypes (*M. aeruginosa*, *M. novacekii*, *M. smithii*) completely disappeared during certain periods of time and re-appeared later Fig. 4A and B.

## Discussion

This study demonstrated the proof-of-concept of using a machine learning approach in the analysis of colonial morphospecies of *Microcystis*. The results described here are based on previous researchers' work demonstrating the application of machine learning in the identification and counting of different taxa of plankton organisms<sup>1,2,5,7,15</sup>. The current gold standard for phytoplankton taxonomy is light microscopy of algal samples; however, there is a huge interest to apply semi-automated and automated approaches for *Microcystis* colonial



**Figure 4.** Seasonal changes of abundances of *Microcystis* spp. colonial morphoforms and water temperature from May to September 2019 in mesocosm tanks D1 (A) and G1 (B).

forms classification. Light microscopy's biggest disadvantages are the extensive training and time period required for a taxonomist to become a proficient expert, the high cost of training, and the large component of manual work involved. Although sequencing and the following molecular biological identification have become more popular in recent years, microscopy and visual morphological analysis remain the most important and widely available tools. In the context of saving time during taxonomic analysis, imaging cytometers constitute a faster and efficient way to receive the morphological information required for taxonomic identification<sup>5</sup>. The imaging cytometer in our study was a FlowCAM instrument used by many research groups worldwide<sup>5,8,12,19</sup>.

In recent years, automatic classification of plankton has attracted increasing attention, with the development of methods including both handcrafted features<sup>48–50</sup> and deep learning architectures<sup>51–54</sup>. The former was used for semi-automatic classification by Gorsky et al.<sup>3</sup>, who applied the ZooProcess and Plankton Identifier software for feature extraction and zooplankton taxonomic characterization. The latter, being based on convolutional neural networks, used input images to extract features for several classifiers, but this was a task that required a considerably higher number of annotated images as training datasets for each class<sup>55</sup>. However, the authors found it difficult to create a well-balanced training dataset for deep learning from natural samples with both high diversity and a high abundance of plankton taxa. The images extracted from field samples often showed a natural class imbalance of phytoplankton taxa. For example, Lee et al.<sup>51</sup> used the WHOI-Plankton database with 3 million plankton images, where >90% of all images were annotated for only 5 different classes. In the recent study by Kerr and co-authors<sup>12</sup>, the class imbalance issue was addressed by constructing deep learning algorithms in a collaborative model to achieve the classification of under represented classes found in FlowCAM images. However, this prediction model showed poor performance in certain minority classes. If the non-target training instances heavily outnumber the target classes' training instances, the deep learning algorithms can be ineffective in determining class boundaries. Several studies demonstrate that balanced image distributions yield the best performances<sup>56–58</sup>. We had the advantage of observing seasonal blooms in the mesocosm samples, which helped create well-balanced training sets of *Microcystis* morphospecies for use in a semi-automated classification approach. In the 2019 LMWE experiment, we followed a *Microcystis* seasonal bloom represented by a changing ratio of colonial morphospecies at different dates (Fig. 3). This allowed us to create class-balanced training sets by choosing time points with sufficient amounts of all five *Microcystis* morphospecies. It is a first attempt to apply a semi-automatic algorithm for intrageneric analysis of colonial *Microcystis*, the majority of previous studies being focused on the analysis of colonial phytoplankton taxa at genus level<sup>10,12,53</sup> or used for analysis training sets build with single-celled *Microcystis* laboratory cultures<sup>9,11</sup>.

Here, we presented an identification logic and statistical evaluation of the accuracy and reliability of the approach used for the classification of five colonial morphospecies of *Microcystis* available from a seasonal mesocosm experiment. To verify a machine learning approach for intergeneric classification, we also used plankton from different genera, namely, unicellular *Cryptomonas* and colonial *Micractinium*, available from the mesocosm experiment plankton samples taken during the 2019 season. *Cryptomonas* was represented by brown-green

colored asymmetric cells with a transparent membrane on the outside and an average size of about 40  $\mu\text{m}$ . It is non-toxic freshwater algae with two flagella and is usually consumed by zooplankton<sup>44</sup>. The representative from the second algal genus was the colonial green algae *Micractinium*, which has proteinaceous spines to prevent grazing by planktonic rotifers<sup>59</sup>. Based on microscopy analysis, both algae were dominant or co-dominant with *Microcystis* in LMWE-2019 tanks at many dates (Suppl. Table 4 for *Micractinium* and F3 tanks for *Cryptomonas* spp.). We developed a machine learning approach based on the simple brightfield-related morphological descriptors that demonstrated high performance at the intergeneric level of phytoplankton taxa with a training set of image samples derived from different time points of the 2019 LMWE season. Overall, the accuracy of intergeneric classification of *Microcystis* spp. in the mesocosm samples compared to other colonial and/or unicellular algae showed high performance of 96–100%, stressing the value of using minimized filter sets including 1–2 features. However, this semi-automated classification demonstrated 65–75% accuracy for intrageneric morphospecies within colonial *Microcystis* spp. This type of classification required significantly more filter descriptors (up to 5 particle properties). Nevertheless, the obtained results are comparable with those of analysis by human taxonomists, which, according to Culverhouse and co-authors<sup>14</sup>, is between 67 and 83%. It means that it is possible to evaluate the automatically significant percentage of acquired during seasonal bloom *Microcystis* images and save 70–80% of researcher time.

By contrast, the suggested machine learning approach using well-balanced training sets covering the whole seasonal bloom demonstrated a higher level of accuracy of up to 93% for intrageneric differentiation of *Microcystis* morphospecies, if a training set was created and applied to the images of the five algal forms taken as they occurred at a one-time point in the samples during the bloom. However, a set of classification parameters tends to be less optimal to a particular tank and sampling date. It has less accuracy when applied to other sample sets (detailed description is provided in Suppl. Table 5). We hypothesize that the decrease in the accuracy can be explained by a significant level of colonial phenotypic variability, i.e., high heterogeneity of toxic and non-toxic *Microcystis* morphospecies<sup>60–62</sup> during the seasonal bloom. *Microcystis* heterogeneity shows up, evidenced by differences in image features patterns encountered when data sampling dates are separated by a few weeks.

The described machine learning approach was applied to produce a long-term dataset aimed to understand the colonial *Microcystis* development in relation to environmental factors (manuscript in preparation). The obtained data revealed a sequential seasonal disappearance/reappearance of the certain colonial *Microcystis* morphoforms (Figs. 3 and 4A,B). Morphological variability of *Microcystis* colonies induced by laboratory conditions have been described recently by different groups<sup>63,64</sup>. Similar observations of sequential changes and disappearance of certain colonial *Microcystis* morphologies were reported in Lake Taihu study<sup>65</sup>. Together these observations and our results obtained with machine learning analysis of colonial *Microcystis* are supporting the hypothesis of main transition pathways of colonial *Microcystis* morphoforms<sup>61</sup>. The classification of cyanobacteria strains that was done exclusively by morphological characteristics is not always sufficient<sup>61,66,67</sup>, and our observations emphasize the early formulated suggestions that previously distinct morphospecies may belong to single species<sup>68</sup>.

Colony formation of *Microcystis* is thought to contribute to the global success of this genus in freshwater ecosystems<sup>69,70</sup>. With an increase of environmental problems related to climate change and water scarcity<sup>71</sup>, we need to understand better the factors and mechanisms affecting *Microcystis* colonial forms evolution and dominance. This study provides a useful approach for quantitative analysis of *Microcystis* diversity.

## Conclusions

The estimation of speed/type for phenotypic changes in colonial *Microcystis* requires a high spatial and temporal resolution, and mesocosm studies of seasonal *Microcystis* spp. succession together with semi-automated machine learning algorithm of colonial forms analysis may provide much more detailed and less prone to user bias analysis. Morphological analysis of phytoplankton along time and at recording seasonal changes of single species represents an important tool to study dynamics of aquatic ecosystems<sup>72,73</sup>. Our results suggest that by combining intrageneric classification with the relatively simple set of descriptors in imaging flow cytometry, we can provide an opportunity to examine the colonial morphoforms of *Microcystis* at a higher resolution and temporal level during seasonal bloom. Although previous studies have developed machine learning and deep learning approach to classify plankton<sup>1–12</sup>, our study is the first to differentiate colonial morphoforms of freshwater *Microcystis* at the intrageneric level. The accuracy of the approach is raising to experienced human taxonomists' performance level, thereby reducing the time of analysis and subjectivity. As one of the significant outcomes of this work, such results further highlights a high level of *Microcystis* spp. heterogeneity during a seasonal bloom and support the hypothesis of main transition pathways of colonial *Microcystis* morphoforms. The classification algorithm's accuracy depends on the increased diversity of images features, which can be enriched in the future by including a variety of fluorescence-correlated morphological parameters in the filter sets. We expect that automated methods will be increasingly used in the future, allowing early detection of toxic morphospecies of colonial *Microcystis* and other harmful algae.

Received: 19 December 2020; Accepted: 5 April 2021

Published online: 30 April 2021

## References

1. Benfield, M. C. *et al.* RAPID: Research on automated plankton identification. *Oceanography* **20**, 172–187 (2007).
2. Fernandes, J. A., Irigoien, X., Boyra, G., Lozano, J. A. & Inza, I. Optimizing the number of classes in automated zooplankton classification. *J. Plankton Res.* **31**, 19–29 (2009).
3. Gorsky, G. *et al.* Digital zooplankton image analysis using the ZooScan integrated system. *J. Plankton Res.* **32**, 285–303 (2010).

4. Ellen, J., Li, H. & Ohman, M. D. Quantifying California current plankton samples with efficient machine learning techniques. *IEEE* **1**, 1–9 (2015).
5. Detmer, T. M. *et al.* Comparison of microscopy to a semi-automated method (FlowCAM) for characterization of individual-, population-, and community-level measurements of zooplankton. *Hydrobiologia* **838**, 99–110 (2019).
6. Sosik, H. M. & Olson, R. J. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry: Phytoplankton image classification. *Limnol. Oceanogr. Methods* **5**, 204–216 (2007).
7. Buskey, E. J. & Hyatt, C. J. Use of the FlowCAM for semi-automated recognition and enumeration of red tide cells (*Karenia brevis*) in natural plankton samples. *Harmful Algae* **5**, 685–692 (2006).
8. Álvarez, E., Moyano, M., López-Urrutia, A., Nogueira, E. & Scharek, R. Routine determination of plankton community composition and size structure: A comparison between FlowCAM and light microscopy. *J. Plankton Res.* **36**, 170–184 (2014).
9. Dunker, S., Boho, D., Wäldchen, J. & Mäder, P. Combining high-throughput imaging flow cytometry and deep learning for efficient species and life-cycle stage identification of phytoplankton. *BMC Ecol.* **18**, 51 (2018).
10. Göröcs, Z. *et al.* A deep learning-enabled portable imaging flow cytometer for cost-effective, high-throughput, and label-free analysis of natural water samples. *Light Sci. Appl.* **7**, 66 (2018).
11. Thomas, M. K., Fontana, S., Reyes, M. & Pomati, F. Quantifying cell densities and biovolumes of phytoplankton communities and functional groups using scanning flow cytometry, machine learning and unsupervised clustering. *PLoS ONE* **13**, e0196225 (2018).
12. Kerr, T., Clark, J. R., Fileman, E. S., Widdicombe, C. E. & Pugeault, N. Collaborative deep learning models to handle class imbalance in FlowCam plankton imagery. *IEEE Access* **8**, 170013–170032 (2020).
13. Camoying, M. G. & Yñiguez, A. T. FlowCAM optimization: Attaining good quality images for higher taxonomic classification resolution of natural phytoplankton samples. *Limnol. Oceanogr. Methods* **14**, 305–314 (2016).
14. Culverhouse, P. F., Williams, R., Reguera, B., Herry, V. & González-Gil, S. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Mar. Ecol. Prog. Ser.* **247**, 17–25 (2003).
15. Embleton, K. V., Gibson, C. E. & Heaney, S. I. Automated counting of phytoplankton by pattern recognition: A comparison with a manual counting method. *J. Plankton Res.* **25**, 669–681 (2003).
16. Stanislawczyk, K., Johansson, M. L. & MacIsaac, H. J. Microscopy versus automated imaging flow cytometry for detecting and identifying rare zooplankton. *Hydrobiologia* **807**, 53–65 (2018).
17. Reynolds, R. A., Stramski, D., Wright, V. M. & Woźniak, S. B. Measurements and characterization of particle size distributions in coastal waters. *J. Geophys. Res.* **115**, C08024 (2010).
18. Dashkova, V., Malashenkov, D., Poulton, N., Vorobjev, I. & Barteneva, N. S. Imaging flow cytometry for phytoplankton analysis. *Methods* **112**, 188–200 (2017).
19. Poulton, N. J. FlowCam: Quantification and classification of phytoplankton by imaging flow cytometry. *Methods Mol. Biol.* **1389**, 237–247 (2016).
20. Doan, M. *et al.* Diagnostic potential of imaging flow cytometry. *Trends Biotechnol.* **36**, 649–652 (2018).
21. Harke, M. J. *et al.* A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *Microcystis* spp. *Harmful Algae* **54**, 4–20 (2016).
22. Ibelings, B. W. & Chorus, I. Accumulation of cyanobacterial toxins in freshwater ‘seafood’ and its consequences for public health: A review. *Environ. Pollut.* **150**, 177–192 (2007).
23. Fan, H., Qiu, J., Fan, L. & Li, A. Effects of growth conditions on the production of neurotoxin 2, 4-diaminobutyric acid (DAB) in *Microcystis aeruginosa* and its universal presence in diverse cyanobacteria isolated from freshwater in China. *Environ. Sci. Pollut. Res.* **22**, 5943–5951 (2015).
24. Christoffersen, K. Ecological implications of cyanobacterial toxins in aquatic food webs. *Phycologia* **35**, 42–50 (1996).
25. Ma, H. *et al.* Growth inhibitory effect of *Microcystis* on *Aphanizomenon flos-aquae* isolated from cyanobacteria bloom in Lake Dianchi, China. *Harmful Algae* **42**, 43–51 (2015).
26. Song, H. *et al.* Allelopathic interactions of linoleic acid and nitric oxide increase the competitive ability of *Microcystis aeruginosa*. *ISME J.* **11**, 1865–1876 (2017).
27. Princiotta, S. D., Hendricks, S. P. & White, D. S. Production of cyanotoxins by *Microcystis aeruginosa* mediates interactions with the mixotrophic flagellate *Cryptomonas*. *Toxins* **11**, 223 (2019).
28. Rohrlack, T., Henning, M. & Kohl, J.-G. Mechanisms of the inhibitory effect of the cyanobacterium *Microcystis aeruginosa* on *Daphnia galeata*'s ingestion rate. *J. Plankton Res.* **21**, 1489–1500 (1999).
29. Doers, M. P. & Parker, D. L. Properties of *Microcystis aeruginosa* and *M. flos-aquae* (cyanophyta) in culture: taxonomic implications. *J. Phycol.* **24**, 502–508 (1988).
30. Zhang, M. *et al.* Biochemical, morphological, and genetic variations in *Microcystis aeruginosa* due to colony disaggregation. *World J. Microbiol. Biotechnol.* **23**, 663–670 (2007).
31. Komárek, J. A review of water-bloom forming *Microcystis* species, with regard to populations from Japan. *Arch. Hydrobiol. Suppl. Algal. Stud.* **64**, 115–127 (1991).
32. Park, H. D. *et al.* Temporal variabilities of the concentrations of intra- and extracellular microcystin and toxic *Microcystis* species in a hypertrophic lake, Lake Suwa, Japan (1991–1994). *Environ. Toxicol. Water Qual.* **13**, 61–72 (1998).
33. Wu, Y. *et al.* Seasonal dynamics of water bloom-forming *Microcystis* morphospecies and the associated extracellular microcystin concentrations in large, shallow, eutrophic Dianchi Lake. *J. Environ. Sci.* **26**, 1921–1929 (2014).
34. Reynolds, C. S., Jaworski, G. H. M., Cmiech, H. A., & Leedale, G. F. On the annual cycle of the blue-green alga *Microcystis aeruginosa* Kütz. Emend. Elenkin. *Philos. Trans. R. Soc. Lond. B* **293**, 419–477 (1981).
35. Wu, H., Wei, G., Tan, X., Li, L. & Li, M. Species-dependent variation in sensitivity of *Microcystis* species to copper sulfate: Implications in algal toxicity of copper and controls of blooms. *Sci. Rep.* **7**, 40393 (2017).
36. Zhu, L., Wu, Y., Song, L. & Gan, N. Ecological dynamics of toxic *Microcystis* spp. and microcystin-degrading bacteria in Dianchi Lake, China. *Appl. Environ. Microbiol.* **80**, 1874–1881 (2014).
37. Ishikawa, K., Walker, R. F., Tsujimura, S., Nakahara, H. & Kumagai, M. Estimation of *Microcystis* colony size in developing water blooms via image analysis. *J. Jpn. Soc. Water Environ.* **27**, 69–72 (2004).
38. Yamamoto, Y. & Nakahara, H. Seasonal variations in the morphology of bloom-forming cyanobacteria in a eutrophic pond. *Limnology* **10**, 185–193 (2009).
39. Kurmayer, R. & Christiansen, G. The abundance of microcystin-producing genotypes correlates positively with colony size in *Microcystis* sp. and determines its microcystin net production in Lake Wannsee. *Appl. Environ. Microbiol.* **69**, 787–795 (2003).
40. Via-Ordorika, L. *et al.* Distribution of microcystin-producing and non-microcystin-producing *Microcystis* sp. in European freshwater bodies: Detection of microcystins and microcystin genes in individual colonies. *Syst. Appl. Microbiol.* **27**, 592–602 (2004).
41. Álvarez, S. D. *et al.* Morphology captures toxicity in *Microcystis aeruginosa* complex: Evidence from a wide environmental gradient. *Harmful Algae* **97**, 101854 (2020).
42. Liboriussen, L. *et al.* Global warming: Design of a flow-through shallow lake mesocosm climate experiment. *Limnol. Oceanogr. Methods* **3**, 1–9 (2005).
43. Barteneva, N. S. *et al.* Modelling of cyanobacterial blooms dynamics in mesocosm experiment. In *Proceedings of 10th US HAB Symposium* (2019).
44. Choi, B., Son, M., Kim, J. I. & Shin, W. Taxonomy and phylogeny of the genus *Cryptomonas* (Cryptophyceae, Cryptophyta) from Korea. *Algae* **28**, 307–330 (2013).

45. Aldenhoff, W., Heuzé, C. & Eriksson, L. E. B. Comparison of ice/water classification in Fram Strait from C- and L-band SAR imagery. *Ann. Glaciol.* **59**, 112–123 (2018).
46. Cieslak, D. A. & Chawla, N. V. A framework for monitoring classifiers' performance: When and why failure occurs?. *Knowl. Inf. Syst.* **18**, 83–108 (2009).
47. González, P., Álvarez, E., Díez, J., López-Urrutia, Á. & del Coz, J. J. Validation methods for plankton image classification systems: Validation methods for plankton image classification systems. *Limnol. Oceanogr. Methods* **15**, 221–237 (2017).
48. Duda, R. O. & Hart, P. E. *Pattern Classification* (Wiley, 2006).
49. Li, Z., Zhao, F., Liu, J. & Qiao, Y. Pairwise nonparametric discriminant analysis for binary plankton image recognition. *IEEE J. Oceanic Eng.* **39**, 695–701 (2014).
50. Zheng, H. *et al.* Automatic plankton image classification combining multiple view features via multiple kernel learning. *BMC Bioinform.* **18**, 570 (2017).
51. Lee, H., Park, M. & Kim, J. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. in *2016 IEEE International Conference on Image Processing (ICIP)* 3713–3717 (ieeexplore.ieee.org, 2016).
52. Moniruzzaman, M., Islam, S. M. S., Bennamoun, M. & Lavery, P. Deep learning on underwater marine object detection: A Survey. In: Blanc-Talon, J., Penne, R., Popescu, D. & Schneuders, P. (Eds.) *Advanced Concepts for Intelligent Vision Systems. ACIVS 2017. Lecture Notes in Computer Science*, 10617, 150–160 (2017).
53. Li, Q. *et al.* Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning. *ICES J. Mar. Sci.* **77**, 1427–1439 (2020).
54. Dai, J., Yu, Z., Zheng, H., Zheng, B. & Wang, N. A hybrid convolutional neural network for plankton classification. in *Asian Conference on Computer Vision* 102–114 (2017).
55. Hassaballah, M. & Hosny, K. M. (Eds.) Recent advances in computer vision. Theories and applications. *Studies Comput. Intell.* **804** (Springer, 2019).
56. Vucetic, S. & Obradovic, Z. Classification on data with biased class distribution. In *European Conference on Machine Learning* 527–538 (Springer, 2001).
57. Weiss, G. M. & Provost, F. Learning when training data are costly: The effect of class distribution on tree induction. *J. Artif. Intell. Res.* **19**, 315–354 (2003).
58. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
59. Schlüter, M., Groeneweg, J. & Soeder, C. J. Impact of rotifer grazing on population dynamics of green microalgae in high-rate ponds. *Water Res.* **21**, 1293–1297 (1987).
60. Frangeul, L. *et al.* Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genom.* **9**, 274 (2008).
61. Xiao, M., Li, M. & Reynolds, C. S. Colony formation in the cyanobacterium *Microcystis*. *Biol. Rev.* **93**, 1399–1420 (2018).
62. Le Manach, S. *et al.* Global metabolomic characterizations of *Microcystis* spp. highlights clonal diversity in natural bloom-forming populations and expands metabolite structural diversity. *Front. Microbiol.* **10**, 791 (2019).
63. Otsuka, S. *et al.* Morphological variability of colonies of *Microcystis* morphospecies in culture. *J. Gen. Appl. Microbiol.* **46**, 39–50 (2000).
64. Li, M., Zhu, W. & Sun, Q. Solubilisation of mucilage induces changes in *Microcystis* colonial morphology. *New Zeal. J. Mar. Freshw. Res.* **48**, 38–47 (2014).
65. Zhu, W., Zhou, X., Chen, H. & Li, M. Sequence of *Microcystis* colony formation during recruitment under natural conditions. *Hydrobiologia* **823**, 39–48 (2018).
66. Giovannoni, S. J. *et al.* Evolutionary relationships among cyanobacteria and green chloroplasts. *J. Bacteriol.* **170**, 3584–3592 (1988).
67. Makra, N. *et al.* Molecular taxonomic evaluation of *Anabaena* and *Nostoc* strains from the Mosonmagyaróvár algal culture collection. *South Afr. J. Bot.* **124**, 80–86 (2019).
68. Visser, P. M. *et al.* How rising CO<sub>2</sub> and global warming may stimulate harmful cyanobacterial blooms. *Harmful Algae* **54**, 154–159 (2016).
69. Otsuka, S. *et al.* A proposal for the unification of five species of the cyanobacterial genus *Microcystis* Kützing ex Lemmermann 1907 under the rules of the Bacteriological Code. *Int. J. Syst. Evol. Microbiol.* **51**, 873–879 (2001).
70. Jankowiak, J. G. & Gobler, C. J. The composition and function of microbiomes within *Microcystis* colonies are significantly different than native bacterial assemblages in two North American Lakes. *Front. Microbiol.* **11**, 1016 (2020).
71. Navarro-Ortega, A. *et al.* Managing the effects of multiple stressors on aquatic ecosystems under water scarcity. The GLOBAQUA project. *Sci. Total Environ.* **503–504**, 3–9 (2015).
72. Naselli-Flores, L. Morphological analysis of phytoplankton as a tool to assess ecological state of aquatic ecosystems: the case of Lake Arancio, Sicily, Italy. *Inland Waters* **4**, 15–26 (2014).
73. Naselli-Flores, L., Zohary, T. & Padisak, J. Life in suspension and its impact on phytoplankton morphology: An homage to Colin S. Reynolds. *Hydrobiologia* **848**, 7–30 (2020).

## Acknowledgements

Funding for this work came from Nazarbayev University grant, ORAU #110119FD4513 to N.S.B, and Ministry of Sciences, Kazakhstan grants MES #4350/GF4 and MES #AP05134153/GF4, to I.A.V. and N.S.B. E.J. and N.S.B. were supported by AQUACOSM (Network of Leading European AQUATIC MesoCOSM Facilities Connecting Mountains to Oceans from the Arctic to the Mediterranean) #IFCPHYTO and #SCPCRTNY. D.V.M. is working under #CITIS AAAA-A16-116021660054-4 theme. E.J. was funded from Centre for Water Technology at Aarhus University (WATEC), Sino-Danish Centre for Education and Research, and TŮBĪTAK BIDEB 2232—Outstanding Researchers Programme. We are very thankful to the Core Facilities of Nazarbayev University for access to instrumentation, and to technical members of LMWE Project at Aarhus University (Denmark) for access to mesocosm facilities and technical help.

## Author contributions

Conceptualization, Y.M., A. Zhu., A. Zha., D.V.M., N.S.B., I.A.V., E.J.; Methodology, Y.M., A. Zhu., A. Zha., D.V.M., K.S., A.B., V.D., N.S.B., I.A.V., E.J.; Data curation, Y.M., A. Zhu., D.V.M., V.D., A.B., N.S.B.; Formal Analysis, Y.M., A. Zhu., A. Zha., K.S., A.B.; Resources, I.A.V., T.A.D., E.J., N.S.B.; Writing and Editing, Y.M., A. Zhu., A. Zha., N.S.B., E.J., I.A.V., T.A.D.; Supervision, T.A.D., E.J., I.A.V., N.S.B.; Funding Acquisition, E.J., T.A.D., I.A.V., N.S.B. All authors read and approved 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/s41598-021-88661-2>.

**Correspondence** and requests for materials should be addressed to N.S.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021