Cytometry Part A

ISAC CYTOMETRY Journal of Quantitative Cell Science PART A

ORIGINAL ARTICLE OPEN ACCESS

Morphological Profiling of Imaging Flow Cytometry Data Uncovers Heterogeneity in Infected *Gephyrocapsa huxleyi* Cultures

Maxim Lippeveld[1,2] | Daniel Peralta[3,4] | Assaf Vardi[5] | Flora Vincent[5,6] | Yvan Saeys[1,2]

[1]Data Mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium | [2]Department of Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium | [3]Department of Computer Science and Artificial Intelligence, University of Granada, Spain | [4]DaSCI Andalusian Institute in Data Science and Computational Intelligence, Granada, Spain | [5]Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot, Israel | [6]Developmental Biology Unit, European Molecular Biological Laboratory, Heidelberg, Germany

Correspondence: Daniel Peralta (dperalta@ugr.es)

Received: 30 October 2024 | **Revised:** 28 May 2025 | **Accepted:** 3 June 2025

Funding: This work was supported by Fonds Wetenschappelijk Onderzoek, 1SB9421N. European Regional Development Fund, C-ING-250-UGR23. Flanders AI Research (FAIR) Program, 174B09119.

ABSTRACT

Phytoplankton, such as the coccolitophore *Gephyrocapsa huxleyi* (*G. huxleyi*), has a major ecological impact through photosynthesis—the production of oxygen and organic material. A significant threat to *G. huxleyi* populations is viral infection with the specific *Gephyrocapsa huxleyi* virus (GhV). Previous research has provided important insight into the infection cycle of *G. huxleyi*. However, research including quantitative morphological information on infected cells is lacking, potentially masking heterogeneity in the infection cycle. In this study, we propose a machine learning (ML) pipeline to incorporate morphological profiling into the analysis of spatially resolved single-molecule mRNA fluorescence in situ hybridization (smFISH)—imaging flow cytometry (IFC) data acquired on infected *G. huxleyi* populations. First, we propose to simplify infection monitoring by using a classification model that does not rely on mRNA staining. Second, we propose an exploratory data analysis pipeline to disentangle two modes of cell death in infected cultures and a subpopulation of healthy cells that potentially will not die from infection, but from programmed cell death (PCD). Overall, we show that morphological profiling of smFISH–IFC data is highly suited for studying microbial interactions in phytoplankton populations.

1 | Introduction

Photosynthetic microeukaryotes and cyanobacteria have a major ecological impact through the conversion of atmospheric and aqueous carbon dioxide to organic material and oxygen [1–3]. These organisms, such as the coccolitophore *Gephyrocapsa huxleyi* (*G. huxleyi*), form enormous blooms in the ocean [4]. A significant driver of bloom collapse is *G. huxleyi*'s susceptibility to infection by the large double-stranded DNA (dsDNA) *Gephyrocapsa huxleyi* virus (GhV) [5, 6]. This lytic viral infection can lead to the viral shunt, the release of algal biomass to the ocean's dissolved organic matter (DOM) pool, which makes GhV infection an important regulator of nutrient flux in biogeochemical cycles, microbial communities, and carbon export to the deep ocean [7–9]. Therefore, optimizing the tracking of viral infection in algal blooms is of vital importance to monitoring and understanding the marine ecosystem.

In recent years, image-based cell profiling [10, 11] has led to numerous biological discoveries [10, 12]. For example, it has been applied to drug discovery [13], stain-free classification of leukocytes in human blood [14], and to perform cell sorting based on spatial phenotypes, such as nuclear translocation [15]. Image-based profiling summarizes high-throughput, spatially resolved

imaging information into a vector of quantitative measurements representing cellular morphological characteristics, such as shape and texture, as well as fluorescence signal intensity and distribution. It uses machine learning (ML) techniques, such as clustering and classification, to analyze these profiles and extract meaningful patterns.

Imaging flow cytometry (IFC) is an ideal candidate for image-based profiling experiments because it combines the fluorescence imaging capabilities of microscopy with the high throughput of conventional flow cytometry. This allows it to capture an abundance of single-cell, spatially resolved data [16, 17]. IFC has been used extensively in phytoplankton research, for instance, for characterizing functional traits [18, 19], environmental monitoring [20–22], and species classification [23, 24]. IFC is also used to quantify gene expression at high throughput in morphologically intact cells by combining it with single-molecule mRNA fluorescence in situ hybridization (smFISH) [25]. Specifically, smFISH–IFC has been used to quantify active GhV infection in *G. huxleyi* blooms, leading to numerous insights on GhV infection [5].

However, while fluorescence staining and manual gating can effectively classify infection cell states, it is labor-intensive, expensive, and a possible point of failure in an experiment. Furthermore, the downstream manual gating analysis to identify infected cells is rather slow, prone to subjectivity, and prone to error for the untrained eye. This raises the need for an automated approach that relies exclusively on brightfield (BF), darkfield (DF), and potentially 4′,6-diamidino-2-phenylindole (DAPI) information.

Along with [5], previous research has improved understanding of GhV's infection strategies and transcriptome regulation of the host [26], and how it affects viral-induced DOM [27]. However, research including ML analysis of quantitative, morphological information on infected cells is lacking. In related research, ML has achieved outstanding performance in both supervised and unsupervised analyses, which highlights its great potential to optimize GhV's infection monitoring and to understand the dynamics of the infected populations.

In this study, we propose an ML pipeline to augment the analysis of the spatially resolved smFISH–IFC data collected in [5] with image-based, morphological profiling. First, we trained and validated a model to classify cells into different infection states to substantially simplify the time-consuming fluorescent staining protocols used for monitoring viral infection in *G. huxleyi* populations. Second, we used a SHapley Additive exPlanations (SHAP)-based analysis [28] to reveal which feature types have the most impact on cell state classification, bringing insight into what channels and features drive classification performance. Finally, we perform an unsupervised clustering and dimensionality reduction analysis to uncover novel biological heterogeneity in the GhV infection cycle driven by morphology. We find evidence for two modes of cell death in infected cultures: lysis after infection and programmed cell death (PCD) without infection. These findings point to a potential seed of resistant cells that could regrow the culture.

Overall, our study demonstrates that image-based profiling is a valuable and highly suited approach for an in-depth analysis of smFISH–IFC data of microbial interactions in phytoplankton.

2 | Results and Discussion

In this section, we present classification and exploratory analysis results obtained on a time course smFISH–IFC dataset monitoring GhV infection in *G. huxleyi* phytoplankton populations [5]. Infection monitoring was done from 0 to 72 h post infection (hpi) by tracking the expression of the viral major capsid protein (*mcp*) gene and the host photosystem II protein D1 (*psbA*) gene. *mcp* encodes the major capsid protein, a structural protein in the viral capsid, and indicates active viral infection. *psbA* is a chloroplast-encoded gene that encodes a core protein of the host photosynthetic machinery, and indicates host metabolic activity. Figure 1 shows an overview of the dataset. See Section 4.1 for details on the data acquisition protocol.

In the following sections, cells are categorized based on their culture origin and manually gated infection state. Two cultures are used in this analysis: *infected* and *control*, with and without virus added, respectively. The infection state was determined based on *mcp* and *psbA* expression and is divided into four categories: healthy (*mcp*−/*psbA*+), dead (*mcp*−/*psbA*−), early infection (*mcp*+/*psbA*+) or late infection (*mcp*+/*psbA*−) (Figure 1a). Because the control culture is not infected with GhV, it contains cells in only two possible infection states: dead (*mcp*−/*psbA*−) and healthy (*mcp*−/*psbA*+). Due to the heterogeneity of the infection process, the infected cultures contain cells in all four infection states. See Section 4.1 for details on the manual gating procedure.

2.1 | Gradient Boosting Models Enable the Use of a Simplified and Automated Infection Monitoring Protocol

To simplify and automate infection monitoring in *G. huxleyi* blooms, we explore the use of image-based, ML models, reducing the need for fluorescent staining to rely only on BF, DF, and optionally, DAPI information. To explore this, we set up a classification pipeline that predicts one of four ground-truth infection states, as defined above. The classification is based on each cell's extensive morphological profile derived from BF, DF, and DAPI images. These images capture forward scattered light, side scattered light, and light emitted from stained nucleotides, respectively. We test two settings with profiles derived from different subsets of images: (i) the stain-free BD profile (BF and DF images), and (ii) the BD + D profile (BF, DF, and DAPI images). In all settings, features are weighted equally when training the classification models.

We specifically test the impact of DAPI as it may significantly improve classification performance without adding too much complexity: it is a routinely added, low-cost stain, and could potentially be applied to live cells [29]. Finally, we assess the impact of training the models using cells from the infected culture only and using cells from both the infected and control cultures, as the heterogeneity between both cultures might affect classification performance.

An eXtreme gradient boosting (XGB) [30] model was trained to predict the four states achieving a maximum cross-validated balanced accuracy of 0.81 (±0.002) using the BD + D profiles from infected culture cells. With the BD profile we achieved a balanced accuracy of 0.75 (±0.003) training on

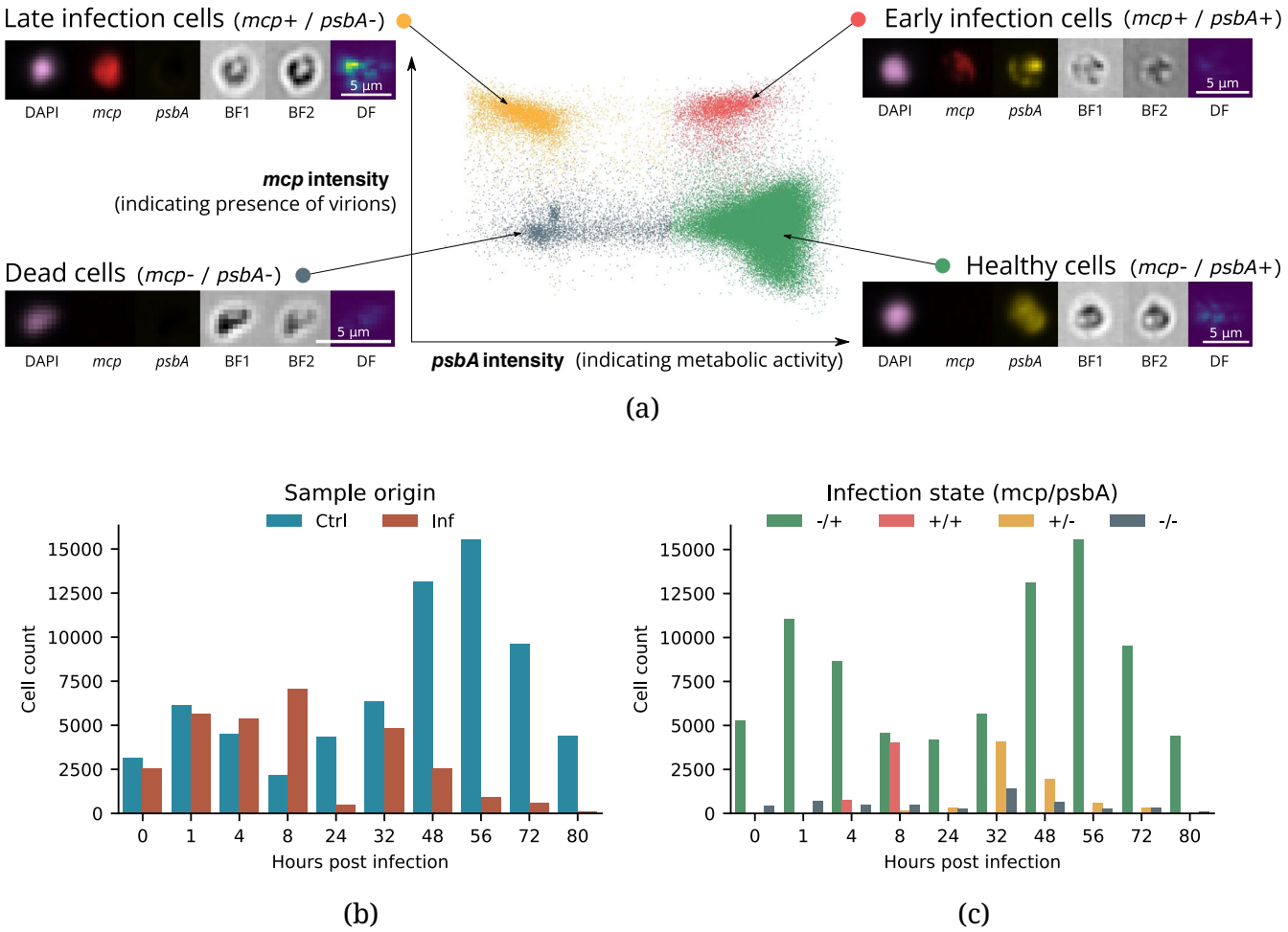


FIGURE 1 | Overview of the single-molecule mRNA fluorescence in situ hybridization imaging flow cytometry dataset of *Gephyrocapsa huxleyi* phytoplankton infected with the *G. huxleyi* virus. (a) Scatter plot colored according to manually gated infection states showing logicle-transformed expression of metabolic (*psbA*) and viral (*mcp*) activity probes. Example images for each infection state show spatially resolved expression of probes and cell morphology. (b) Cell counts for control and infected cultures per time point post infection. (c) Cell counts per time point post infection grouped by infection state. [Color figure can be viewed at wileyonlinelibrary.com]

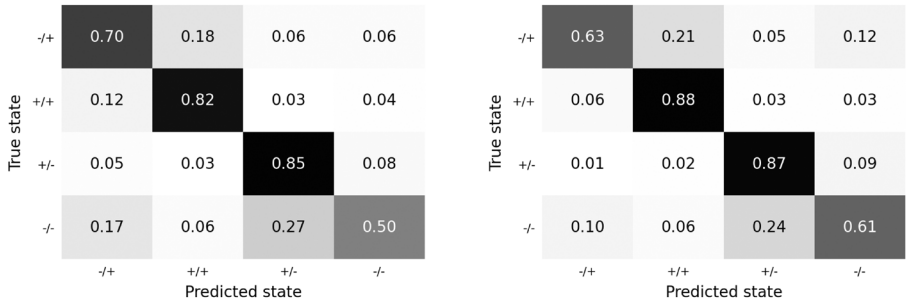
infected and control cultures. Figure 2 shows obtained accuracies for all settings, along with confusion matrices. For both feature profiles, we found that including control culture cells for training decreased healthy (*mcp*−/*psbA*+) state prediction performance. For the BD feature set, this performance drop was offset by a performance gain in the remaining states, leading to a better overall performance. The latter was not the case for the BD + D profile.

The confusion matrices indicate that dead (*mcp*−/*psbA*−) cells are the most difficult to predict. This is likely due to the smaller number of dead cells in the dataset and large morphological variability within this class. The recall of dead and healthy cells drops most between BD + D and BD settings, indicating that the model relies strongly on DAPI information for these states. This is expected as DAPI correlates with infection due to increased *de novo* nucleotide synthesis for virion production during infection [5, 31]. For reference, a small analysis of the incorrectly predicted cells is shown in Supplementary Figure 4.

In conclusion, we find it is possible to simplify and automate infection monitoring using ML models by using only DAPI staining, or no staining at all. However, accuracy can still be improved, mainly in dead cell classification.

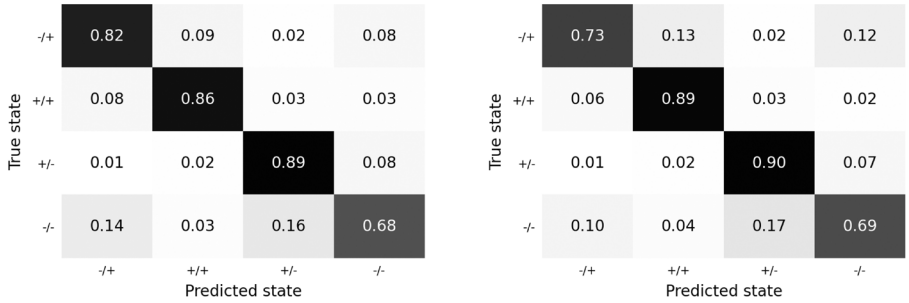
2.2 | SHAP Analysis Provides Insight Into Important Features for Classification

To gain more insight into the importance of the features used by the XGB classification models, we performed a SHAP analysis [28]. SHAP is an approach rooted in game theory that fairly distributes a model's output value over the input features according to the impact each feature has on the output. The SHAP values allow us to assess which features were important for the classification of a particular input and to evaluate the overall impact of the feature by aggregating values across the dataset. For easier interpretation, we group the features into three categories: intensity (e.g., mean, standard deviation, skewness), shape (e.g., eccentricity, major axis length, and area) and texture (e.g., gray level co-occurrence matrix (GLCM) dissimilarity, Sobel mean, and Sobel standard deviation). Figure 3 shows two summarizing views of the SHAP values for the 20 features with the highest mean absolute



(a) BD, trained on infected only

(b) BD, trained on infected and undersampled control



(c) BD+D, trained on infected only

(d) BD+D, trained on infected and undersampled control

Feature set	Train set	Balanced accuracy (mean ± s.e.m.) Test	Train
BD	Infected only	0.72 (0.002)	0.75 (0.001)
	Infected + undersampled control	0.75 (0.003)	0.85 (0.007)
BD+D	Infected only	0.81 (0.002)	0.89 (0.018)
	Infected + undersampled control	0.80 (0.002)	0.90 (0.005)

(e) Cross-validated balanced accuracy

FIGURE 2 | Confusion matrices and balanced accuracy show performance of infection state classification obtained with an XGB classifier trained without features derived from mcp and psbA images. Models were compared when trained with features derived from two sets of images: (i) brightfield and darkfield, and (ii) brightfield, darkfield, and DAPI. Both feature sets were also compared when trained with cells from only infected cultures and from both control and infected cultures. Test metrics are obtained on infected culture cells only. Classification performance is best when DAPI is included and the classifier is trained on infected cultures only.

SHAP value. Figures S5 and S6 show the distribution of SHAP values for these 20 features in more detail. Figure S7 depicts the distribution of two features identified by SHAP as having high impact on classification.

The composition of the 20 features with the highest contribution for both the BD and BD + D models shows that for each infection state, the intensity features contribute most, followed by the texture and shape features. In the BD model, all channels contribute equally, with a bias towards DF features for dead (*mcp*−/*psbA*−) cell classification. In the BD + D model, this is also the case, but the contribution of the DF channel is partially shifted to the DAPI channel. This is likely due to the aforementioned correlation of DAPI with infection through *de novo* nucleotide synthesis. DAPI also has a high contribution to healthy (*mcp*−/*psbA*+) cell classification through the texture features. In both models, all shape features in the top 20 are derived from both BF channels. This is expected, as the BF-derived masks capture the morphology of the cell most clearly. Morphology captured by the DAPI and DF masks proved to be less interesting. This could be due to the small size of *G. huxleyi* combined with IFC's limited resolution.

In addition to this analysis, it also appeared in the data that the skewness of the BF pixel distribution was bimodal, with one negative peak and one positive peak (Figure S14a). This bimodality does not correspond to batch effects nor to the masking process, and more experiments would be necessary to identify its cause.

Following the SHAP analysis contributing most importance to the intensity features, we trained classifiers with only those features and found that this causes a limited, yet non-negligible drop in performance. For the BD + D model, average balanced accuracy drops from 0.81 to 0.79, and for the BD profile, it drops from 0.75 to 0.72. This result confirms that an optimal classification pipeline should include texture and shape features.

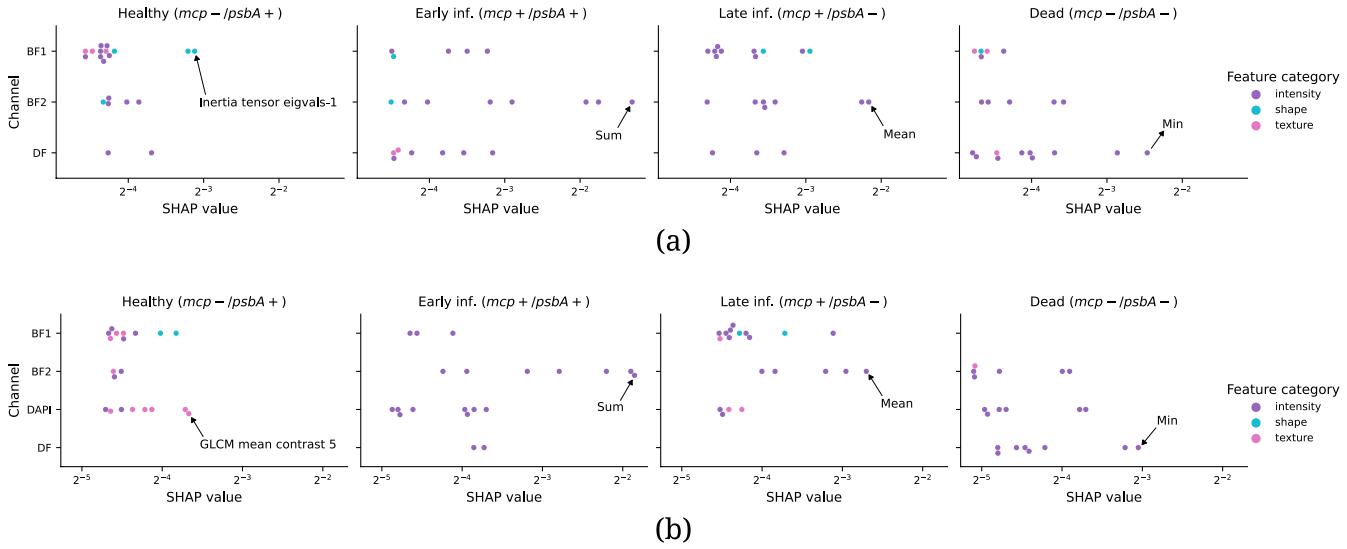


FIGURE 3 | SHAP was used to compute the contribution of each feature to the classification of each infection state for the (a) BD and (b) BD + D model. To ease interpretation, the features are subdivided into three categories: Intensity, shape, and texture. We show the 20 features with the highest absolute SHAP values averaged across the dataset. The arrow annotations show the most impactful feature for each state. Figure S7 shows example images of the BF inertia feature identified in (a) and the DAPI contrast feature identified in (b). [Color figure can be viewed at wileyonlinelibrary.com]

Furthermore, it highlights the potential of ML to analyze multidimensional data to unravel properties that are not identified by manual analyses.

2.3 | Clustering Analysis Suggests That a Group of Resistant Cells in Infected Cultures Likely Die From NVCD

To uncover heterogeneity beyond the previously identified states of the GhV infection cycle, we performed an unsupervised clustering and dimensionality reduction analysis of the dataset. The obtained clustering is shown on the 2D uniform manifold approximation (UMAP) reduction in Figure 4. We hypothesize, based on this analysis, that there are two mechanisms of cell death in infected cultures: one involving direct lysis from GhV infection and one involving PCD prior to any viral infection. We find evidence that could support this hypothesis in the clusters of healthy (*mcp*−/*psbA*+) and dead (*mcp*−/*psbA*−) cells.

First, we find a group of clusters predominantly containing healthy (*mcp*−/*psbA*+), non-infected cells, which are likely in later stages of the cell cycle. These clusters contain 75% of all healthy control culture cells, which are mainly expected to go through the cell cycle. Figure S10 shows the distribution of culture origin in these clusters. Interestingly, in these clusters, we find a small fraction of cells from the infected culture (0.3% of all infected culture cells) that are not infected (*mcp*−) and age beyond 24 hpi. Additionally, a SHAP analysis of the clustering indicates that these cells are on average larger and more metabolically active compared to other healthy cells (Figure S11). Cell-to-cell heterogeneity within a monoclonal population of *G. huxleyi* can lead to varying degrees of resistance [32], and size may convey higher fitness to the host. Larger cells may also be a consequence of resistance: during their life cycle, cells double their volume before dividing, and resistance could be accompanied by cell division arrest. These uninfected cells could be a seed for regrowing the culture after infection.

Additionally, the clusters containing the dead (*mcp*−/*psbA*−) cell population provide more evidence for the presence of cells in the infected culture not dying from infection. An analysis of this observation is shown in Figure 5. The dead cells are divided over 5 clusters, which vary in culture origin, as shown in Figure 5a. Two of these clusters [18 and 25] are composed of cells originating equally in control samples from all time points and infected culture samples from early time points (mainly before 24 hpi). Given that control culture cells can only die from PCD without infection, we hypothesize that the dead cells of the infected culture clustered with them also die from PCD without infection. To verify that *mcp*−/*psbA*− cells indeed represent dead cells, it was previously shown with Sytox staining – a marker for cell death – that the proportion of Sytox-positive cells correlated well with *mcp*−/*psbA*− cells [5].

In addition to a majority of dead cells, clusters 18 and 25 also contain some healthy (*mcp*−/*psbA*+), uninfected cells, likely close to PCD.

Besides this, we also find that cluster 28 is composed entirely of cells originating from the infected culture that likely died after viral infection. The SHAP analysis shows that this cluster contains small cells with a high DAPI intensity (Figure 5b–d and Figure S9). Indeed, large dsDNA viruses, such as GhV, require large amounts of nucleotides to meet demand for virion production. To this end, they hijack the cell to promote *de novo* nucleotide synthesis, increasing DAPI signal intensity [5, 31]. Furthermore, Figure 5e,f shows that cells in cluster 28 are smaller and less structurally intact compared to those of cluster 18 and 25, supporting the fact that cluster 28 contains cells that died after lysis. We refer the reader to the supplementary

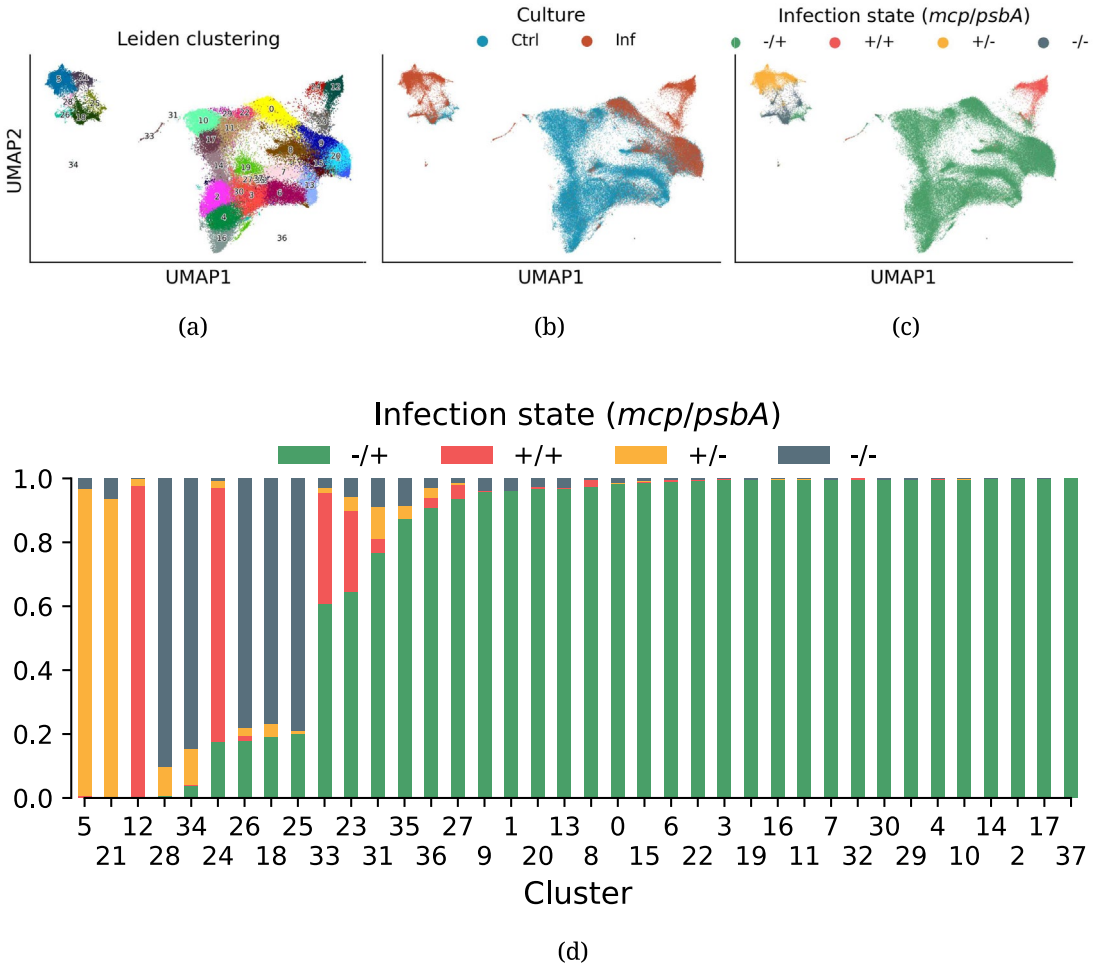


FIGURE 4 | We performed unsupervised Leiden clustering and UMAP dimensionality reduction of the full dataset with features derived from all available images (brightfield, darkfield, DAPI, *mcp* and *psbA*). The obtained 2D reduction is shown colored according to the (a) Leiden clustering, (b) culture origin (control or infected), and (c) manually gated labels for healthy (*mcp*−/*psbA*+), early infection (*mcp*+/*psbA*+), late infection (*mcp*+/*psbA*−) or dead (*mcp*−/*psbA*−) cells. (d) Fraction of manually gated infection states in the clusters obtained with Leiden clustering. [Color figure can be viewed at wileyonlinelibrary.com]

material (Section 3, Figures S12–S15) for a complementary analysis of some of the clusters and features.

Our clustering-based detection of subpopulations likely dying from PCD in infected cultures aligns with previous studies documenting resistance mechanisms in *G. huxleyi*. Earlier work demonstrated that *G. huxleyi* can escape viral infection through a strategy where the calcified diploid phase transitions to a resistant haploid phase [33, 34]. Cell-to-cell heterogeneity within monoclonal populations was also shown to drive stable host-virus coexistence [32]. These findings support our hypothesis of discovering resistant subpopulations within infected cultures. To validate these mechanisms, future work should combine the smFISH–IFC approach with genomic and transcriptomic analyses to find a molecular basis of resistance to GhV infection, and potentially link them to morphological differences we observed.

3 | Conclusion

Previous research has studied infection of *G. huxleyi* blooms with its specific coccolithovirus GhV using various technologies: single-cell RNA sequencing of infected cells revealed how infection changes the cellular transcriptome [26], untargeted exometabolomics of infected algal blooms showed how infection contributes to the DOM [27], and smFISH–IFC was used to show the diversity of cell fates after viral infection [5]. In this study, we expanded on this research by employing ML techniques to analyze morphological profiles derived from smFISH–IFC data.

First, we expanded on research in stain-free analysis of phytoplankton [24, 35–37] by setting up a ML pipeline for classifying cells into four infection states based on BF, DF, and, optionally, DAPI information. The achieved performance demonstrates that this pipeline can simplify infection monitoring by mitigating the need for smFISH and automating the quantification of viral infection after acquisition. This result therefore allows for more efficient and less labor-intensive data acquisition, potentially leading to a better understanding of viral infection in algal blooms.

Second, we used clustering techniques to unravel additional biological heterogeneity in the infection cycle. We were able to disentangle modes of cell death within infected cultures,

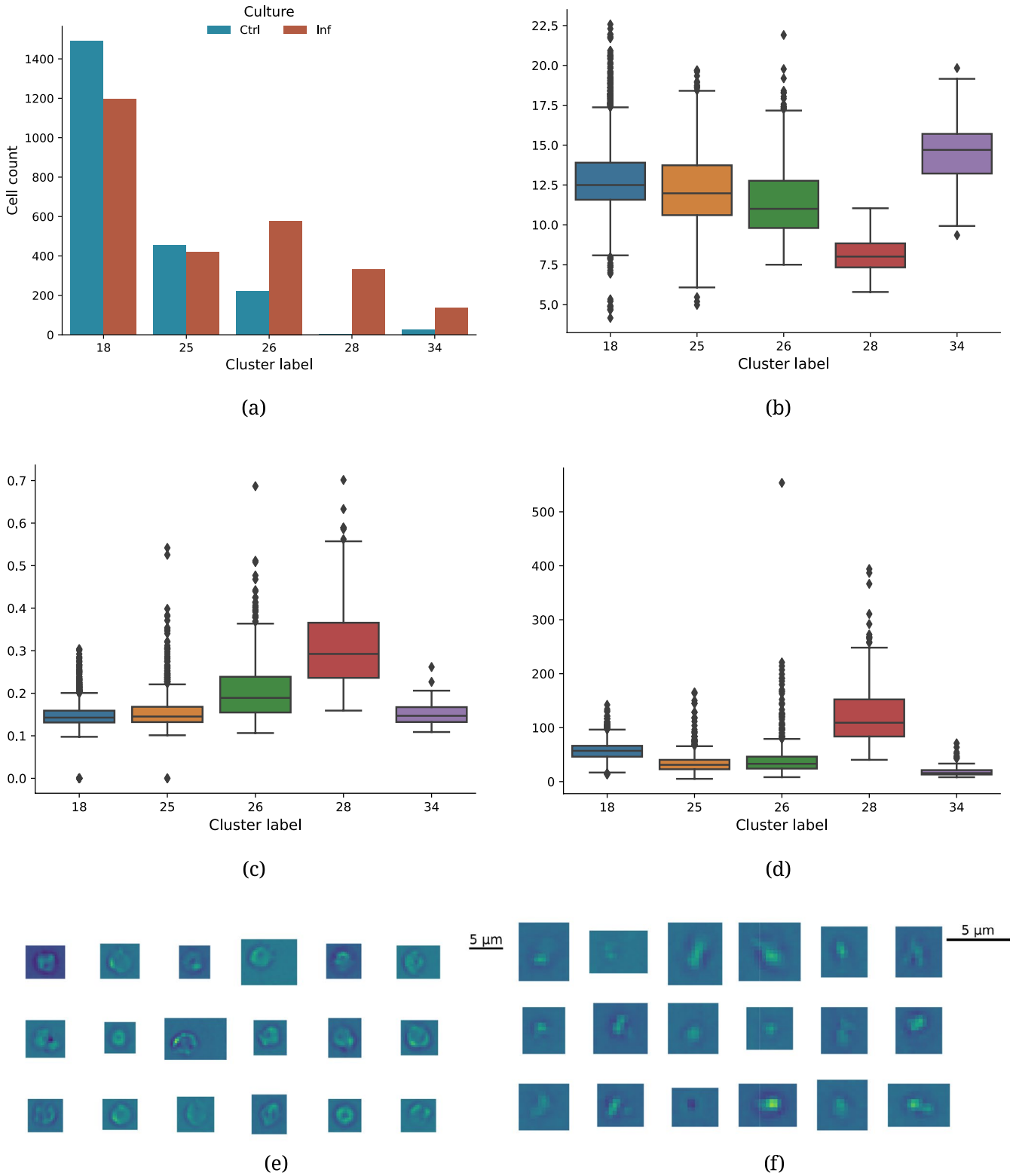


FIGURE 5 | (a) Clustering divides manually gated dead (*mcp*−/*psbA*−) cells over five clusters. Clusters 18 and 25 contain cells originating equally from infected and control cultures. Clusters 26, 28, and 34 contain predominantly cells from infected samples. A SHAP analysis comparing dead cell cluster 28 against other dead cell clusters shows that (b) the minor axis length, (c) the GLCM energy feature derived from the BF, and (d) the mean value of the Sobel map derived from the DAPI image have a high impact on clustering. Examples of BF images from clusters 18 and 25 (e) show structurally intact dead cells that likely died from PCD prior to infection, and lysed cells from cluster 28 (f) that died from infection. [Color figure can be viewed at wileyonlinelibrary.com]

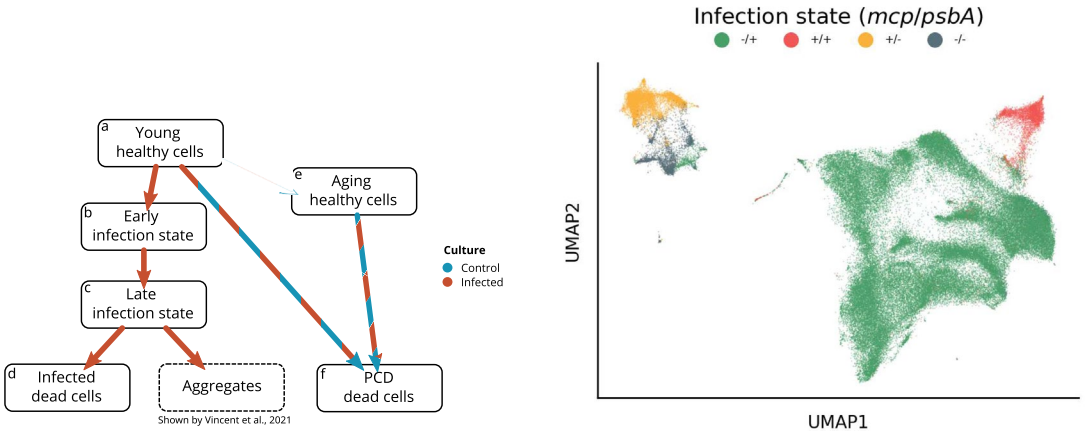


FIGURE 6 | (a) Schematic representation of cell states and transitions leading from healthy (*mcp*−/*psbA*+) to dead (*mcp*−/*psbA*−) cells for which we find evidence in the dataset using the Leiden clustering approach discussed in Section 2.3. Colors of the arrows indicate in which culture each transition occurs. (b) States and transitions from (a) plotted onto the UMAP embedding obtained in Section 2.3. The letters in (b) correspond to the letters on the schematic representation in (a). [Color figure can be viewed at wileyonlinelibrary.com]

differentiating between cells that move directly to PCD versus those that go through viral infection. We also found evidence for two groups of healthy (*mcp*−/*psbA*+) cells in infected cultures: those that follow their development and those that do not, pointing to two potential subpopulations of resistant cells that could regenerate the culture. Figure 6 shows a schematic representation of the subpopulations and transitions between them, for which we found evidence in the dataset.

Next, we highlight some unexpected findings which require future research. First, given the results of the classification model SHAP analyses, we want to investigate how shape and texture-based features could be improved to increase their discriminative power. Second, we find a bimodal distribution of the skewness of the BF pixel distribution feature across the dataset. More investigation is needed to uncover what drove this split.

Finally, in future work, we would also like to generate a dataset with a shorter time interval between samples to allow for automated trajectory inference. This analysis could provide additional evidence for the previously and newly identified cell fates in GhV infection. Furthermore, deep learning methods are popular in phytoplankton research for species and trait classification [38]. In this work, we opted for feature-based ML methods to allow for better interpretability. It would be relevant to investigate whether there are performance improvements when using deep learning-based methods.

In conclusion, our work has provided contributions to the study of infection in *G. huxleyi* blooms by setting up an end-to-end ML pipeline to analyze smFISH–IFC data from a supervised perspective, to make monitoring more efficient, and from an unsupervised perspective, allowing for the discovery of novel biological insights. This study demonstrated that image-based profiling is a valuable approach for an in-depth analysis of smFISH–IFC data, allowing us to identify avenues for further research.

4 | Materials and Methods

4.1 | Monitoring Viral Infection With smFISHMonitoring Viral Infection With

We refer to [5] for all detailed information on the experimental setup and manual gating approach. We provide a summary below.

G. huxleyi CCMP 2090 was used and infected with a 5:1 multiplicity of infection (MOI) ratio of infectious virus per cell, thereby guaranteeing that all cells encountered an infectious particle 30 min post infection. The time courses of infected and noninfected cultures were sampled simultaneously in triplicates.

Cells were stained with two smFISH probes tracking host metabolic activity and active viral infection. Metabolic activity is tracked through the mRNA expression of *psbA*, a chloroplast-encoded gene of the D1 protein. The D1 protein is involved in the early stages of photosynthesis and is essential for optimal viral infection [39, 40]. Viral infection is tracked through the mRNA expression of the *mcp* gene that encodes the major capsid protein, expressed in viral infection [26].

Data were analyzed using IDEAS6.2 (Amnis, Luminex). The compensation matrix was built using the IDEAS wizard and manually checked before being applied to all the acquired files. Based on the area and circularity of DAPI, three populations were identified as single cells (mainly, DAPI area < 60 a.u.), doublets, and aggregates (mainly, DAPI area > 60 a.u.). Single cells were selected in the same focal plane using the BF gradient and contrast. All gates were defined on a single file before being applied to the total data set. Each file was manually inspected to check the accuracy of single-cell and aggregate gating. Classification of *mcp*− and *psbA*− populations was based on negative controls of uninfected and unstained cells respectively.

4.2 | Dataset Acquisition

Images were collected using the Amnis ImageStream MK-II IFC platform at 0, 1, 4, 8, 24, 32, 48, 56, and 72 hpi. Three biological replicates are available for infected cultures, and two for control cultures. Six channels were collected: the DAPI, *mcp*, and *psbA* fluorescence channels; two BF channels capturing transmitted light at wavelengths of 420–480 nm and 57–595 nm; and a DF image capturing scattered light. We refer to Vincent et al. [5] for acquisition details and staining procedure.

Data were analyzed using IDEAS 6.3 (Amnis, Luminex). The compensation matrix was built using the IDEAS wizard and manually checked before being applied to all the acquired files. Based on the area (the number of microns squared in a mask) and circularity (the degree of the mask's deviation from a circle) of DAPI, three populations were identified as single cells (mainly, DAPI area <60 a.u.), doublets, and aggregates (mainly, 4′,6-diamidino-2-phenylindole (DAPI) area >60 a.u.). Single cells were additionally selected in the same focal plane using the BF gradient and contrast (both gradient and contrast measure the sharpness quality of an image by detecting large changes of pixel values in the image).

The dataset can be downloaded from the Bioimage Archive under accession number S-BIAD617.

4.3 | Morphological Profiling With SCIP

We obtained detailed morphological profiles of the 6-channel IFC images with SCIP [41], a software for processing image cytometry data. This involved export from IDEAS, background masking, and feature computation.

To process the images, we exported them from IDEAS 6.3 to 16-bit, nonpadded `TIFF` files using an AutoHotKey script to automate the point-and-click procedure. This was done on a Windows machine. We stored the images in a Zarr array, an efficient on-disk array storage format, which can be loaded by SCIP.

We computed two types of masks for each image and channel. The first mask (Figure S1) is computed by:

1. smoothing the input with a Gaussian filter with a standard deviation of 0.5 (1 for *mcp* and *psbA*),
2. computing a Sobel map,
3. smoothing the map with a Gaussian filter with standard deviation of 1 (2 for *mcp* and *psbA*)
4. computing the Li threshold,
5. and masking all pixels with a value below the threshold.

The second mask (Figure S2) is computed by:

1. subtracting a median filtered version of the input (filter size 5×5) from itself,
2. smoothing the result with a Gaussian filter with a standard deviation of 0.5 (1 for *mcp* and *psbA*),
3. computing a Sobel map,
4. smoothing the map with a median filter with size 5×5,
5. computing the Otsu threshold,
6. and masking all pixels with a value below the threshold.

Finally, for both masks, small holes (with a maximum area of 25% of the input image) were filled and small objects (with a maximum area of 20 pixels) were removed.

We then computed 406 features per channel and per mask. These features describe the cell's phenotype in terms of shape, texture, and image intensity. We refer to Lippeveld et al. [41] for more details on the features. The profiles were exported to a Parquet file for further downstream processing.

Prior to the downstream analysis, 38,287 events were discarded after being identified as multiplets or debris based on the BF and DAPI mask's major over minor axis ratio and eccentricity. We also removed 113 cells from control cultures that were positive for *mcp*, which was likely due to residual smFISH probes remaining in the sample after the washing step. After quality control filtering, 104,164 cells remained for further analysis.

4.4 | Classification With XGBoost

The classification pipeline was trained and evaluated with nested fivefold cross-validation (CV) stratified for the infection states. Models are trained using either only infected culture cells or infected and control culture cells. In both cases, the identical CV folds of infected culture cells are used. When controls are also used, a separate CV is done and training folds of infected and control cultures are concatenated. Performance is recorded separately on infected and control validation folds. To avoid overrepresentation of control cells in the training data, they are undersampled to a maximum of 20,000 cells per timepoint.

The pipeline starts with a filter removing zero-variance features, followed by a random undersampling of healthy (*mcp*−/*psbA*+) cells to the level of the second most abundant type of cells (early infection (*mcp*+/*psbA*+)), and a random oversampling of all other infection states. The resulting training set is then used to train an XGB classifier. We opted for XGB because it has a fast graphical processing unit (GPU)-accelerated implementation for training and inference, and it can be efficiently explained using the SHAP TreeExplainer algorithm [28].

The inner cross-validation was used to find optimal hyperparameters with the successive halving random search method [42] as implemented in the `scikit-learn` Python package. The search was initiated with 500 randomly sampled candidate hyper-parameter settings and 10 XGBoost estimators for training. At each iteration, the top 50% amount of candidate settings was halved based on the balanced accuracy, and the number

of boosting rounds was doubled. This was repeated until 640 boosting rounds were reached, to keep optimization computationally feasible. Figure S3 shows the search evolution of both configurations and the optimal hyperparameters selected for each outer fold.

4.5 | Dimensionality Reduction With Feature Clustering to Reduce Correlation

To reduce correlation between image-derived features, we perform a feature selection step following the procedure outlined in [43]. First, we compute a Spearman rank correlation [44] matrix of all features. We use the nonparametric Spearman rank correlation since we cannot make assumptions on the distribution of the features. Agglomerative hierarchical clustering with average linkage and Euclidean distance is used to cluster the correlation matrix. We flatten the hierarchy to create clusters that have at least a correlation of 0.9. From each of the clusters, the feature with the highest variance is selected.

4.6 | Dimensionality Reduction and Clustering

To visualize heterogeneity in the dataset, we mapped the cell profiles to a two-dimensional (2D) embedding with UMAP. Figure 4 shows the UMAP reduction colored according to the manually gated infection state and culture origin.

In order to uncover subpopulations of cells in the dataset, we used the Leiden algorithm [45] to cluster cells based on their profile similarity (see Figure 4 and Figure S8). As seen in Figure 4d, the obtained clusters have high homogeneity with respect to the manually gated labels. This means that the unsupervised clustering is able to reconstruct the populations found by manual gating, as well as identify subpopulations within them.

We used the ScanPy package [46] to perform dimensionality reduction and clustering. First, all features were independently Z-score normalized. A principal component analysis was performed on the normalized features [47]. We selected the 81 first components, which explained 90% of the variance. Next, a 20-nearest neighbor graph was constructed using UMAP connectivity and Euclidean distance [48]. This graph was then used to perform Leiden clustering with a resolution of 2.5. UMAP was applied to the PCA components to reduce it to two dimensions using the previously computed 20-nearest neighbor graph. PAGA [49] was used to estimate connectivity structures in the graph and provide an initialization for the UMAP low-dimensional embedding [48, 50].

4.7 | Model Explanations With SHAPModel Explanations With

Shapley values are a concept from game theory that allows for fair credit allocation to players in a cooperative game based on their contribution to the outcome. SHAP values are a reformulation of Shapley values in the context of ML where the game is the ML model, the outcome is the model's prediction, and the players are input features.

SHAP values are a unifying framework for the class of additive feature attribution methods. This class of methods provides local explanations, meaning they attribute feature importances per instance x [51]. SHAP values are challenging to compute exactly, so in general, we need to resort to approximation methods. However, TreeExplainer exploits the structure of tree-based models, such as random forests or gradient boosted trees, to compute exact SHAP values in polynomial time [28]. It is implemented in the SHAP Python package.

In this work, we used TreeExplainer in two contexts. First, to explain the infection state classification with an XGB model. This model was cross-validated, so we concatenated explanations on the test set from all folds to obtain explanations for the full dataset. Features were then prioritized based on their mean absolute SHAP value across the dataset.

Second, we used TreeExplainer to prioritize features that could drive the clustering of cells found with Leiden clustering. To achieve this, we trained a proxy XGB model to predict assigned cluster labels based on input features. The proxy models were validated using a 90/10 train-test split. We then explained the proxy model and prioritized features based on their mean absolute SHAP value across the dataset.

Author Contributions

Maxim Lippeveld: investigation; validation; visualization; software; writing – original draft. **Daniel Peralta:** methodology; investigation; validation; supervision; writing – original draft. **Assaf Vardi:** writing – review and editing; conceptualization; methodology. **Flora Vincent:** writing – review and editing; conceptualization; methodology; investigation; supervision; data curation; validation. **Yvan Saeys:** writing – review and editing; conceptualization; methodology; supervision.

Acknowledgments

The compute resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government. M. Lippeveld is a Predoctoral Fellow of the Fund for Scientific Research FWO Flanders (1SB9421N). This work is partially supported by the Flanders AI Research (FAIR) Program under grant no. 174B09119. This research was also partially supported by the I+D+i project granted by C-ING-250-UGR23 co-funded by "Consejería de Universidad, Investigación e Innovación" and the European Union related to FEDER Andalucía Program 2021-27. Funding for open access charge: Universidad de Granada/CBUA.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. T. Fenchel, "Marine Plankton Food Chains," *Annual Review of Ecology and Systematics* 19 (1988): 19–38. JSTOR: 2097146.

2. C. B. Field, M. J. Behrenfeld, J. T. Randerson, and P. Falkowski, "Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components," *Science* 281, no. 5374 (1998): 237–240.

3. Y. M. Bar-on, R. Phillips, and R. Milo, "The Biomass Distribution on Earth," *National Academy of Sciences of the United States of America* 115, no. 25 (2018): 6506–6511.

4. C. W. Brown and J. A. Yoder, "Coccolithophorid Blooms in the Global Ocean," *Journal of Geophysical Research: Oceans* 99, no. C4 (1994): 7467–7482.

5. F. Vincent, U. Sheyn, Z. Porat, D. Schatz, and A. Vardi, "Visualizing Active Viral Infection Reveals Diverse Cell Fates in Synchronized Algal Bloom Demise," *Proceedings of the National Academy of Sciences of the United States of America* 118, no. 11 (2021): e2021586118.

6. W. H. Wilson, G. A. Tarran, D. Schroeder, M. Cox, J. Oke, and G. Malin, "Isolation of Viruses Responsible for the Demise of an *Emiliania Huxleyi* Bloom in the English Channel," *Journal of the Marine Biological Association of the United Kingdom* 82, no. 3 (2002): 369–377.

7. S. W. Wilhelm and C. A. Suttle, "Viruses and Nutrient Cycles in the Sea: Viruses Play Critical Roles in the Structure and Function of Aquatic Food Webs," *BioScience* 49, no. 10 (1999): 781–788.

8. J. S. Weitz and S. W. Wilhelm, "Ocean Viruses and Their Effects on Microbial Communities and Biogeochemical Cycles," *F1000 Biology Reports* 4 (2012): 17. PMID: 22991582.

9. J. R. Brum, J. C. Ignacio-espinoza, S. Roux, et al., "Patterns and Ecological Drivers of Ocean Viral Communities," *Science* 348, no. 6237 (2015): 1261498.

10. J. C. Caicedo, S. Cooper, F. Heigwer, et al., "Data-Analysis Strategies for Image-Based Cell Profiling," *Nature Methods* 14, no. 9 (2017): 849–863. PMID: 28858338.

11. C. Scheeder, F. Heigwer, and M. Boutros, "Machine Learning and Image-Based Profiling in Drug Discovery," *Current Opinion in Systems Biology* 10 (2018): 43–52.

12. S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, and A. E. Carpenter, "Image-Based Profiling for Drug Discovery: Due for a Machine-Learning Upgrade?," *Nature Reviews Drug Discovery* 20, no. 2 (2021): 145–159.

13. M. A. Bray, S. Singh, H. Han, et al., "Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes," *Nature Protocols* 11, no. 9 (2016): 1757–1774.

14. M. Lippeveld, C. Knill, E. Ladlow, et al., "Classification of Human White Blood Cells Using Machine Learning for Stain-Free Imaging Flow Cytometry," *Cytometry Part A* 97, no. 3 (2020): 308–319.

15. D. Schraivogel, T. M. Kuhn, B. Rauscher, et al., "High-Speed Fluorescence Image–Enabled Cell Sorting," *Science* 375, no. 6578 (2022): 315–320.

16. N. S. Barteneva, E. Fasler-kan, and I. A. Vorobjev, "Imaging Flow Cytometry: Coping With Heterogeneity in Biological Systems," *Journal of Histochemistry and Cytochemistry* 60, no. 10 (2012): 723–733.

17. P. Rees, H. D. Summers, A. Filby, A. E. Carpenter, and M. Doan, "Imaging Flow Cytometry," *Nature Reviews Methods Primers* 2, no. 1 (2022): 1–13.

18. V. Dashkova, D. Malashenkov, N. Poulton, I. Vorobjev, and N. S. Barteneva, "Imaging Flow Cytometry for Phytoplankton Analysis," *Methods* 112 (2017): 188–200.

19. E. C. Orenstein, S. D. Ayata, F. Maps, et al., "Machine Learning Techniques to Characterize Functional Traits of Plankton From Image Data," *Limnology and Oceanography* 67, no. 8 (2022): 1647–1669.

20. S. Dunker, M. Boyd, W. Durka, et al., "The Potential of Multispectral Imaging Flow Cytometry for Environmental Monitoring," *Cytometry Part A* 101, no. 9 (2022): 782–799.

21. A. Lefebvre and E. Poisson-caillault, "High Resolution Overview of Phytoplankton Spectral Groups and Hydrological Conditions in the Eastern English Channel Using Unsupervised Clustering," *Marine Ecology Progress Series* 608 (2019): 73–92.

22. K. Rousseeuw, É. P. Caillault, A. Lefebvre, and D. Hamad, Monitoring System of Phytoplankton Blooms by Using Unsupervised Classifier and Time Modeling, 2013 IEEE International Geoscience and Remote Sensing Symposium—IGARSS, July 2013, pp. 3962–3965.

23. Q. T. K. Lai, K. C. M. Lee, A. H. L. Tang, K. K. Y. Wong, H. K. H. So, and K. K. Tsia, "High-Throughput Time-Stretch Imaging Flow Cytometry for Multi-Class Classification of Phytoplankton," *Optics Express* 24, no. 25 (2016): 28170–28184.

24. H. M. Sosik and R. J. Olson, "Automated Taxonomic Classification of Phytoplankton Sampled With Imaging-In-Flow Cytometry," *Limnology and Oceanography: Methods* 5, no. 6 (2007): 204–216.

25. A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, "Imaging Individual mRNA Molecules Using Multiple Singly Labeled Probes," *Nature Methods* 5, no. 10 (2008): 877–879.

26. C. Ku, U. Sheyn, A. Sebé-pedrós, et al., "A Single-Cell View on Alga-Virus Interactions Reveals Sequential Transcriptional Programs and Infection States," *Science Advances* 6, no. 21 (2020): eaba4137.

27. C. Kuhlisch, G. Schleyer, N. Shahaf, F. Vincent, D. Schatz, and A. Vardi, "Viral Infection of Algal Blooms Leaves a Unique Metabolic Footprint on the Dissolved Organic Matter in the Ocean," *Science Advances* 7, no. 25 (2021): eabf4680.

28. S. M. Lundberg, G. Erion, H. Chen, et al., "From Local Explanations to Global Understanding With Explainable AI for Trees," *Nature Machine Intelligence* 2, no. 1 (2020): 56–67.

29. B. I. Tarnowski, F. G. Spinale, and J. H. Nicholson, "DAPI as a Useful Stain for Nuclear Quantitation," *Biotechnic & Histochemistry* 66, no. 6 (1991): 296–302.

30. T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 16, Association for Computing Machinery, New York, NY, USA, Aug. 2016, pp. 785–794.

31. S. Rosenwasser, C. Ziv, S. G. van Creveld, and A. Vardi, "Virocell Metabolism: Metabolic Innovations During Host–Virus Interactions in the Ocean," *Trends in Microbiology* 24, no. 10 (2016): 821–832.

32. N. Joffe, C. Kuhlisch, G. Schleyer, N. S. Ahlers, A. Shemi, and A. Vardi, "Cell-To-Cell Heterogeneity Drives Host–Virus Coexistence in a Bloom-Forming Alga," *ISME Journal* 18, no. 1 (2024): wrae038.

33. M. J. Frada, S. Rosenwasser, S. Ben-dor, A. Shemi, H. Sabanay, and A. Vardi, "Morphological Switch to a Resistant Subpopulation in Response to Viral Infection in the Bloom-Forming Coccolithophore *Emiliania huxleyi*," *PLoS Pathogens* 13, no. 12 (2017): e1006775.

34. M. Frada, I. Probert, M. J. Allen, W. H. Wilson, and C. de Vargas, "The "Cheshire Cat" Escape Strategy of the Coccolithophore *Emiliania huxleyi* in Response to Viral Infection," *Proceedings of the National Academy of Sciences of the United States of America* 105, no. 41 (2008): 15944–15949.

35. S. D. Grant, K. Richford, H. L. Burdett, D. Mckee, and B. R. Patton, "Low-Cost, Open-Access Quantitative Phase Imaging of Algal Cells Using the Transport of Intensity Equation," *Royal Society Open Science* 7, no. 1 (2020): 191921.

36. Ç. Isıl, K. de Haan, Z. Göröcs, et al., "Phenotypic Analysis of Microalgae Populations Using Label-Free Imaging Flow Cytometry and Deep Learning," *ACS Photonics* 8, no. 4 (2021): 1232–1242.

37. Z. Göröcs, M. Tamamitsu, V. Bianco, et al., "A Deep Learning-Enabled Portable Imaging Flow Cytometer for Cost-Effective, High-Throughput, and Label-Free Analysis of Natural Water Samples," *Light: Science & Applications* 7, no. 1 (2018): 66.

38. S. Dunker, D. Boho, J. Wäldchen, and P. Mäder, "Combining High-Throughput Imaging Flow Cytometry and Deep Learning for Effcient Species and Life-Cycle Stage Identification of Phytoplankton," *BMC Ecology* 18, no. 1 (2018): 51.

39. A. K. Mattoo, J. B. Marder, and M. Edelman, "Dynamics of the Photosystem II Reaction Center," *Cell* 56, no. 2 (1989): 241–246.

40. K. Thamatrakoln, D. Talmy, L. Haramaty, et al., "Light Regulation of Coccolithophore Host–Virus Interactions," *New Phytologist* 221, no. 3 (2019): 1289–1302.

41. M. Lippeveld, D. Peralta, A. Filby, and Y. Saeys, "SCIP: A Scalable, Reproducible and Open-Source Pipeline for Morphological Profiling Image Cytometry and Microscopy Data," *Cytometry Part A* 105, no. 11 (2024): 816–828.

42. K. Jamieson and A. Talwalkar, Non-Stochastic Best Arm Identification and Hyperparameter Optimization, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR, May 2016, pp. 240–248.

43. D. Peralta and Y. Saeys, "Robust Unsupervised Dimensionality Reduction Based on Feature Clustering for Single-Cell Imaging Data," *Applied Soft Computing* 93 (2020): 106421.

44. J. L. Myers and A. Well, *Research Design and Statistical Analysis* (Lawrence Erlbaum Associates, 2003).

45. V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: Guaranteeing Well-Connected Communities," *Scientific Reports* 9, no. 1 (2019): 5233.

46. F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis," *Genome Biology* 19, no. 1 (2018): 15.

47. M. E. Tipping and C. M. Bishop, "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61, no. 3 (1999): 611–622.

48. L. Mcinnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv* (2020). arXiv: 1802.03426 [cs, stat].

49. F. A. Wolf, F. K. Hamey, M. Plass, et al., "PAGA: Graph Abstraction Reconciles Clustering With Trajectory Inference Through a Topology Preserving Map of Single Cells," *Genome Biology* 20, no. 1 (2019): 59.

50. E. Becht, L. Mcinnes, J. Healy, et al., "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP," *Nature Biotechnology* 37, no. 1 (2019): 38–44.

51. S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).

Supporting Information

Additional supporting information can be found online in the Supporting Information section.