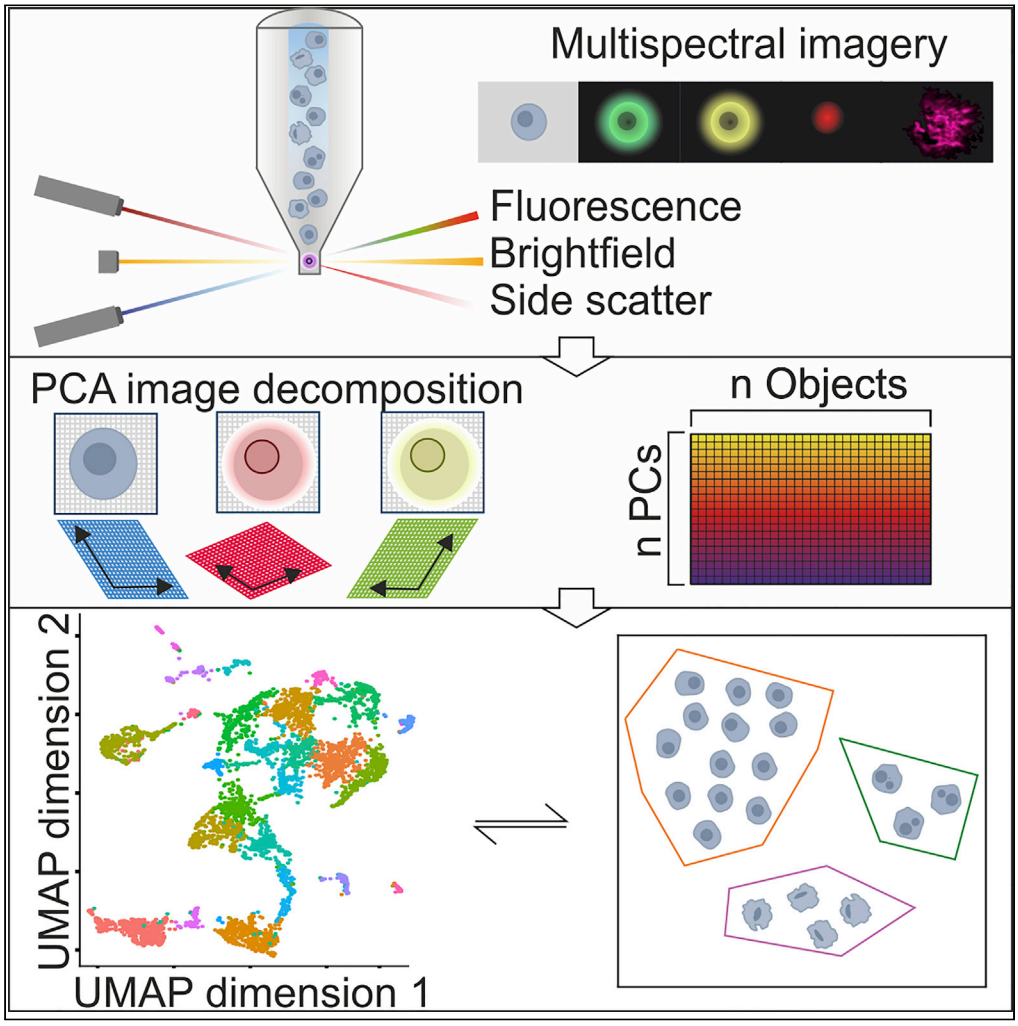
iScience

CellPress
OPEN ACCESS

Article

Dimensionality reduction by UMAP for visualizing and aiding in classification of imaging flow cytometry data



Ireneusz Stolarek, Anna Samelak-Czajka, Marek Figlerowicz, Paulina Jackowiak

paulinaj@ibch.poznan.pl

Highlights

UMAP dimensionality reduction provides fast and accurate method of IFC data analysis

UMAP yields improved object clustering and tagging of the multispectral IFC data

PCA decomposition allows multispectral signals merging for direct image embedding

Stolarek et al., iScience *25*, 105142
October 21, 2022
https://doi.org/10.1016/j.isci.2022.105142

iScience

CellPress
OPEN ACCESS

Article

Dimensionality reduction by UMAP for visualizing and aiding in classification of imaging flow cytometry data

Ireneusz Stolarek,[1] Anna Samelak-Czajka,[1] Marek Figlerowicz,[1] and Paulina Jackowiak[1,2,*]

SUMMARY

Recent advances in imaging flow cytometry (IFC) have revolutionized high-throughput multiparameter analyses at single-cell resolution. Although enabling the discovery of population heterogeneities and the detection of rare events, IFC generates hyperdimensional datasets that demand innovative analytical approaches. Current methods work in a supervised manner, utilize only limited information content, or require large annotated reference datasets. Dimensionality reduction algorithms, including uniform manifold approximation and projection (UMAP), have been successfully applied to analyze the large number of parameters generated in various high-throughput techniques.

Here, we apply a workflow incorporating UMAP to analyze different IFC datasets. We demonstrate that it out-competes other popular dimensionality reduction methods in speed and accuracy. Moreover, it enables fast visualization, clustering, and tagging of unannotated objects in large-scale experiments. We anticipate that our workflow will be a robust method to address complex IFC datasets, either alone or as an upstream addition to the deep learning approaches.

INTRODUCTION

Imaging flow cytometry (IFC) is a powerful technique that combines high-throughput and multi-parameter capabilities of conventional flow cytometry with morphological and spatial information from imaging at single-cell resolution. With the potential to reveal sample complexity and detect rare events, IFC has gained an increasing number of applications in multiple areas of biology and biomedicine (Han et al., 2016; Voronin et al., 2020). However, the capability of IFC to acquire multiple images for thousands of objects poses significant difficulties for the analysis of such large and hyperdimensional datasets. These challenges are particularly prominent when the goal of the experiment is to study a heterogeneous sample of unknown complexity, for example, in environmental or cancer research.

The standard analytical process for IFC data utilizes only a subset of the collected parameters in the form of predefined features of the imaged objects, for example, measures of the size, shape, texture, and localization (Blasi et al., 2016; Hennig et al., 2017). They are manually selected by the user and further applied to discriminate between the populations of interest. Such an approach can be employed for a wide variety of biological problems, but it can be subjective, time-consuming, and limited to simple applications (Blasi et al., 2016; Hennig et al., 2017). In the case of addressing more advanced research problems, a discriminant feature can be a complex combination of several simpler features. Consequently, a significant user experience is required to engineer such a feature. To some extent, the analysis can be automated, yet it will still rely on the same set of predefined or engineered features, therefore utilizing limited information content present in the acquired dataset and usually requiring prior knowledge about the sample composition. Recently, both commercial and open-source deep learning architectures have successfully been applied for analyzing complex samples (Mochalova et al., 2021); however, their use is hindered by the need for large annotated reference datasets and long network training times. Therefore, most current approaches work in supervised mode, rendering them ineffective when applied to complex unannotated samples.

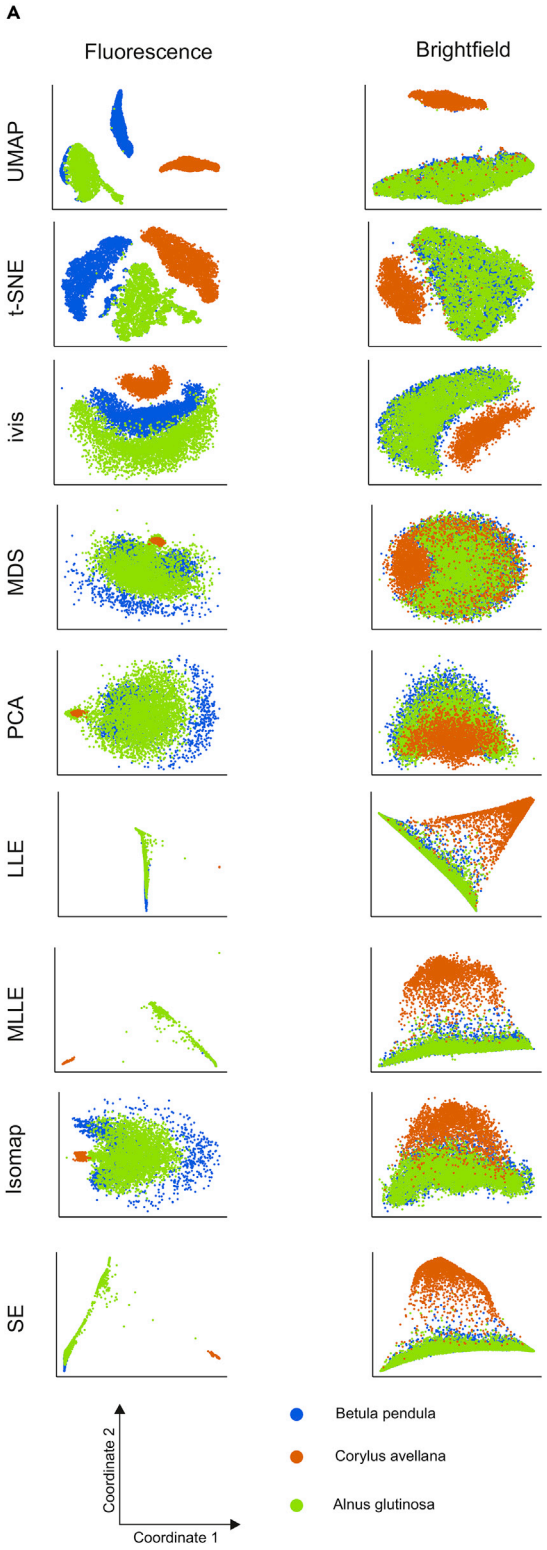
An improvement to the presented workflows can be found in unsupervised dimensionality reduction, which is pivotal for visualizing and clustering unannotated high-dimensional data (Akhbardeh and Jacobs, 2012;

[1]Institute of Bioorganic Chemistry Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland

[2]Lead contact

*Correspondence: paulinaj@ibch.poznan.pl

https://doi.org/10.1016/j.isci.2022.105142


A
Fluorescence
Brightfield
UMAP
t-SNE
ivis
MDS
PCA
LLE
MLLE
Isomap
SE
Coordinate 2
Coordinate 1
Betula pendula
Corylus avellana
Alnus glutinosa

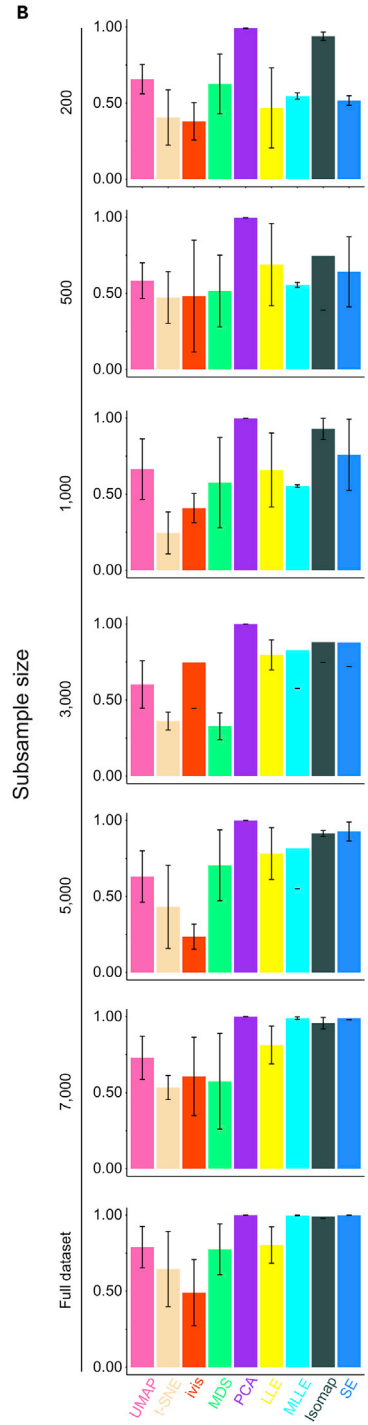

B
Subsample size
200
500
1,000
3,000
5,000
7,000
Full dataset
1.00
0.50
0.00
UMAP
t-SNE
ivis
MDS
PCA
LLE
MLLE
Isomap
SE

Figure 1. Comparison of the performances of unsupervised dimensionality-reduction methods on IFC image data
(A) Two-dimensional embeddings of IFC images of pollen from three plant species (*Betula pendula*, *Corylus avellana* and *Alnus glutinosa*) from the fluorescence channel and brightfield channel.
(B) Reproducibility of large-scale structure embeddings.
Error bars are represented as +/− 1SD See also Figure S2.

Saeys et al., 2016). Dimensionality reduction is aimed at transforming the data into relevant and reduced dimensional space by discovering intrinsic data structure and limiting redundancy. It allows extracting useful variables while discarding noise and correlated features (Huang et al., 2019; Sumithra and Subu, 2015). Multiple dimensionality-reduction techniques have been used to analyze biological data, including PCA, t-SNE (van der Maaten and Hinton, 2008), isomap (Tenenbaum et al., 2000), LLE, MLLE, MDS, SE, ivis, and UMAP (McInnes et al., 2018). UMAP is a manifold learning technique based on Riemannian geometry and algebraic topology. It allows UMAP to preserve short run times as well as the local and global data structures. Accordingly, it has been employed for analyzing single-cell sequencing, mass cytometry, and spectral flow cytometry data and is the state-of-the-art method (Becht et al., 2018; Ferrer-Font et al., 2020). The application of UMAP to the analysis of image data, including multispectral imagery, has not yet been investigated.

In this work, we present the application of UMAP to the analysis of various IFC datasets (including pollen, red blood cells, and monocyte cell line (THP1), demonstrating its ability to facilitate visualization, classification, and tagging of objects from large-scale experiments on complex populations, while working directly on pixel-based image data. In contrast with the existing IFC data analysis methods, our approach does not require training, nor does it use any numerical features. The only feature engineering step incorporated is the preprocessing of the images with PCA, and vectors of their pixel values are used as input for UMAP. The proposed method can either be used alone (for less complex data) or followed by further steps, covering numerical features or applying deep learning (for more demanding datasets). In addition, we provide a means for integrating UMAP with popular IFC analysis software IDEAS into a single pipeline.

RESULTS

Performance of dimensionality-reduction algorithms on various IFC image types

To determine how the dimensionality-reduction approaches handle different types of IFC data, we tested the ability of UMAP and several other common algorithms, t-SNE, Isomap, LLE, MLLE, MDS, SE, PCA, and ivis, to discriminate between 3 plant species, *Betula pendula*, *Alnus glutinosa,* and *Corylus avellana,* based on brightfield and fluorescence images of their pollen. As input, we used 8-bit raw TIFF images exported from the IDEAS software (see STAR Methods). The pixel value distribution of IFC images is largely different from classical images, with the majority of pixels having values similar to the background signal (Figure S1).

UMAP and t-SNE outperformed the rest of the algorithms in the quality of low-dimensional data representation (Figure 1A). Both methods achieved the best results on the fluorescence images, producing clearly separated clusters of objects. None of the methods properly recognized 3 distinct groups of objects in the brightfield images (Figure 1A).

Next, to quantify the reproducibility of the embeddings, we measured the correlation of the object coordinates in embeddings on random dataset subsamples of varying sizes with those from embeddings of the full dataset, which included 9,000 images (Figures 1B and S2). With growing subsample sizes, the resulting embeddings were closer to the representations of the full dataset.

Furthermore, we measured the correctness of the low dimensional representation of the IFC images by clustering the points with hdbscan (McInnes and Healy, 2017) (an unsupervised density-based clustering algorithm) and calculating cluster purity and normalized mutual information (NMI) metrics (Figures 2A and S3) (see STAR Methods). Cluster purity evaluates whether a cluster contains objects representing the same class. NMI quantifies the similarity between the two groupings, i.e., by comparing true versus hdbscan predicted object labels NMI gives a measure of how much of the original grouping information is captured. Among the tested algorithms, only UMAP and t-SNE obtained high values for both purity and NMI in fluorescent and brightfield channel images. The embedding from the brightfield channel showed higher cluster purity, whereas higher NMI was obtained when reducing the dimensions of the fluorescence channel image dataset.

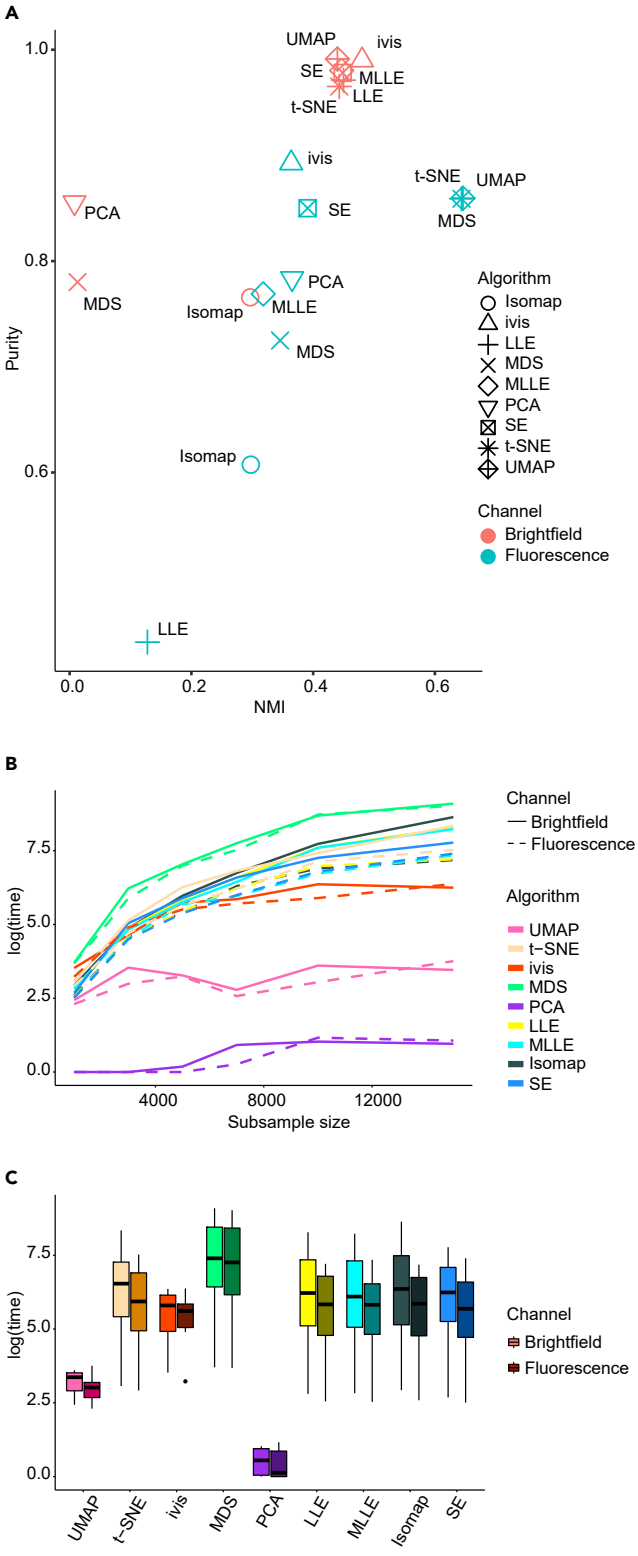

A
1.0
0.8
0.6
Purity
0.0
0.2
0.4
0.6
NMI
UMAP
ivis
SE
MLLE
LLE
t-SNE
ivis
t-SNE
UMAP
PCA
SE
MDS
MDS
PCA
Isomap
MLLE
MDS
Isomap
LLE
Algorithm
Isomap
ivis
LLE
MDS
MLLE
PCA
SE
t-SNE
UMAP
Channel
Brightfield
Fluorescence
B
7.5
5.0
2.5
0.0
log(time)
4000
8000
12000
Subsample size
Channel
Brightfield
Fluorescence
Algorithm
UMAP
t−SNE
ivis
MDS
PCA
LLE
MLLE
Isomap
SE
C
7.5
5.0
2.5
0.0
log(time)
UMAP
t−SNE
ivis
MDS
PCA
LLE
MLLE
Isomap
SE
Channel
Brightfield
Fluorescence

Figure 2. Dimensionality-reduction method runtimes on IFC image data

(A) Cluster purity and NMI values. Brightfield (red), fluorescence (blue). See also Figure S3.

(B) Average run times for embedding brightfield and fluorescence images with respect to the subsample size. Brightfield is depicted as a solid line, fluorescence is depicted as a dotted line.

(C) Comparison of embedding runtimes for brightfield (brighter shade) and fluorescence images (darker shade). Whiskers represent 95th percentile.

UMAP, together with PCA, was significantly faster than the other tested approaches. Importantly, its speed did not decrease significantly with the growing size of the data (Figure 2B). All methods ran faster on the fluorescence image dataset than on the brightfield dataset (Figure 2C). Differing run times and clustering results of brightfield and fluorescence images suggested that the two types of image data had significantly different structures. To better characterize these structures, we decomposed both datasets with PCA (Figure S4). We noted that the number of principal components (PCs) needed to retain 95% of the variance present in the dataset was much higher for the brightfield than for the fluorescence images. Additionally, the relation between the number of PCs and the percent of the explained variance differed between images acquired from the two considered detection channels. Whereas many PCs were needed to capture even small amounts of variance in the brightfield images, even up to 50% of the variance in the fluorescence images was captured by the first PC. This demonstrated that brightfield pixel-based data contained a high level of noise coming from the background pixels, therefore posing a significant challenge for any dimensionality-reduction technique.

Analysis software usually offers several image export options. To determine how the choice of export parameters impacts the downstream dimensionality-reduction procedure, we exported from the IDEAS software the fluorescence images of *C. avellana* pollen with various settings: (1) raw 8-bit, (2) padded 8-bit, (3) raw 16-bit, and (4) padded 8-bit with manually enhanced contrast data. The manual modulation of contrast was applied, as this is a common augmentation of the image that users perform during the exploratory data analysis (Figure S5). Next, the generated datasets were analyzed with UMAP. The two-dimensional embedding distinguished between various image export options (Figures 3A–3C). A change in the image depth from 8 to 16 bits (Figure 3A) or the application of padding on the edges of the image (Figure 3B) changed the local representation of points in the cluster. The manual change in contrast (Figure 3C) resulted in a separation of the two 8-bit datasets; however, their local structure remained unchanged, which indicated that the embedding properly recognized the proportional change in the signal strength across the whole image. Any difference introduced to the images through export options or image augmentation resulted in UMAP recognizing such a population of objects as a separate cluster (Figure S6). Altogether, these results showed the high sensitivity of UMAP to file parameters that cannot be visually inspected by the user and demonstrated the importance of consistent data manipulation protocols to avoid false clustering outcomes.

Influence of the IFC experimental technical variability on UMAP dimensionality reduction

To test how much noise in IFC data associated with the equipment and the laboratory procedures impacts the dimensionality-reduction results, we applied UMAP to a dataset consisting of 4 technical replicates from a single sample of *C. avellana* pollen, analyzed by IFC over a month immediately after material acquisition, after one week, two weeks and four weeks (Figure 4A). Here, we used brightfield images because, for this channel, the LED light intensity changes dynamically during the experiment, which is more likely to generate technical variability, as opposed to a user-defined setup of laser intensities in the fluorescent channels. According to the expectations, UMAP recognized a single cluster of objects (Figure 4A). The distribution of points within the cluster in the absence of biological variation was driven mostly by the differences in the background signal and position of the object in the image (Figures 4B and 4C). This result showed that technical noise associated with the equipment and laboratory procedures did not impact the low-dimensional representation of the input data by UMAP.

Multispectral analysis of pixel-based data with UMAP

IFC acquires the image data from several channels, capturing light at different wavelengths. Thus, to take full advantage of this technology, the data analysis approach should utilize the merged content of multispectral data. To demonstrate how to use UMAP for analyzing multispectral imagery, we used brightfield and fluorescence images of *B. pendula*, *A. glutinosa*, and *C. avellana* pollen.

We explored two methods to perform the analysis. In the first method, we blended the brightfield and fluorescence images for each object and directly reduced the dimensions of composite images with UMAP

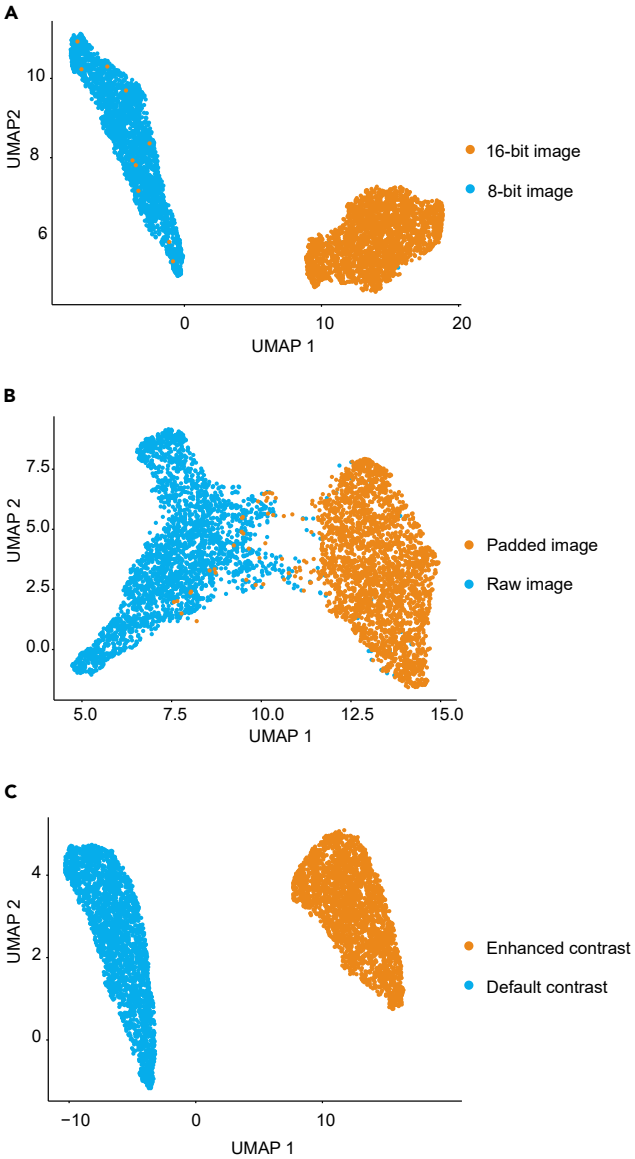


Figure 3. Effects of image augmentation on UMAP embeddings

(A–C) 2D embedding of the same dataset with (A) changed image depth from 8 to 16 bits, (B) addition of padding on the image edges, and (C) manually enhanced contrast of the images. Unmodified raw control images are depicted in blue and augmented images are in orange.

(Figures 5A and S7). In the second method, we utilized PCA to decompose the brightfield and fluorescence images into PCs capturing 95% of the variability in the dataset (Figures 5B, 5D, and 5E). Next, the PC matrices from both image types were passed to UMAP for 2-dimensional embedding.

The first method using a composite image could not separate clusters belonging to *B. pendula* and *A. glutinosa* (Figure 5A), similar to brightfield images (Figure 1A). In contrast, the second approach employing PCA-preprocessing of imagery led to a clear separation of all of the true object groups, thus allowing the efficient use of the multispectral dataset (Figure 5B).

To validate the results generated based solely on images, in the next step, we also applied UMAP to the numerical features (i.e., object area, length, width, average pixel intensity) automatically calculated by the IDEAS software for the same pollen dataset (STAR Methods) (Figure 5C). We found that using PCA-transformed pixel data

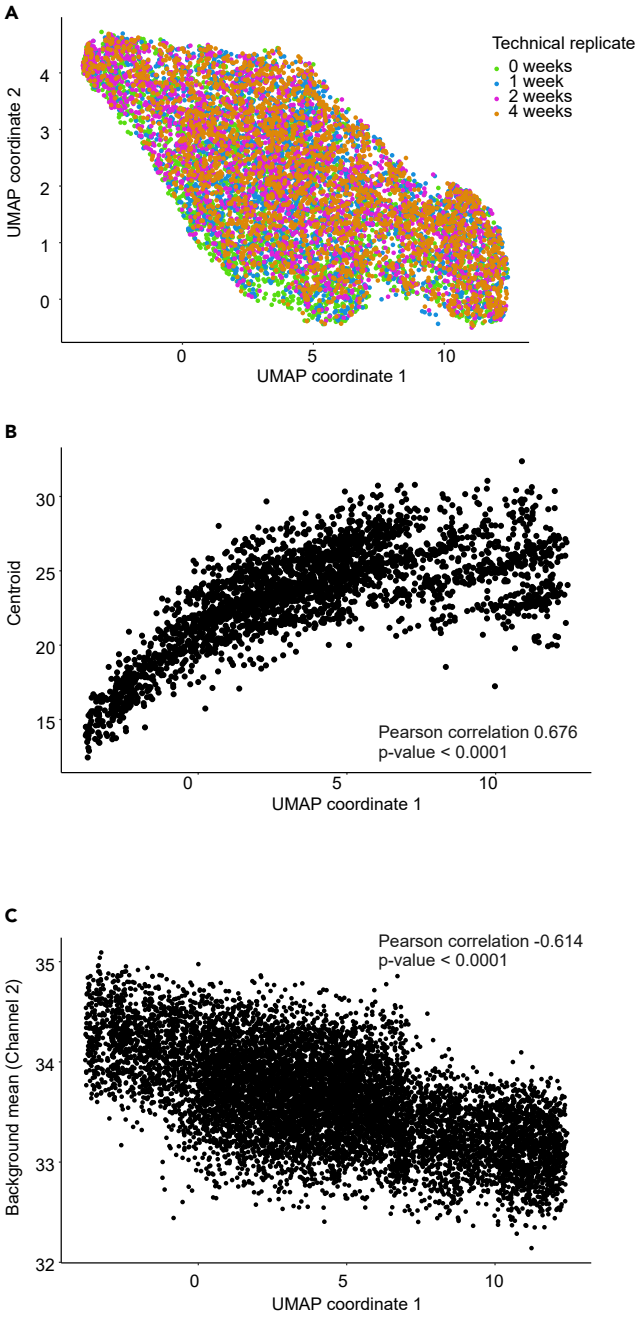


Figure 4. Assessment of technical variability effects on UMAP embedding

(A) 2D embedding of a single biological replicate imaged over a period of a month: directly after material collection, one week, two weeks, and four weeks.

(B) Correlation between the position of the object in the image and UMAP coordinate 1.

(C) Correlation between the background signal and UMAP coordinate 1.

allowed for even better clustering than that obtained with numerical data. There was less species overlap, and the objects were properly separated based on their fluorescence intensity. Quantitatively the PCA preprocessed dataset showed high values of cluster purity, comparable to composite images, and had the highest NMI (Figure 5F).

This observation showed that the proposed approach, including PCA-preprocessing of imagery followed by dimensionality reduction with UMAP, can be successfully applied for unsupervised IFC data analysis. As

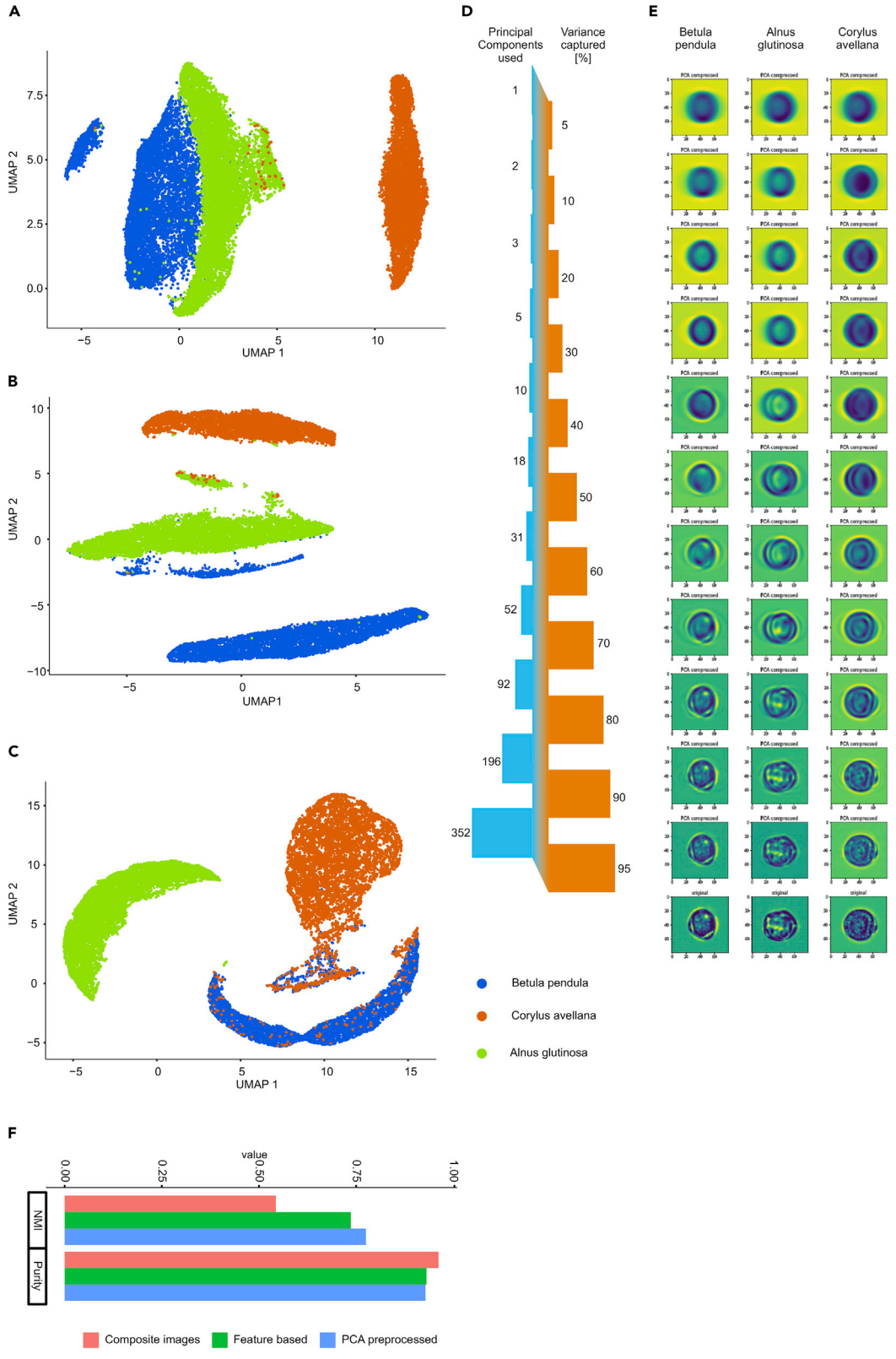

A
UMAP 2
UMAP 1
B
UMAP 2
UMAP1
C
UMAP 2
UMAP 1
D
Principal Components used
Variance captured [%]
1
2
3
5
10
18
31
52
92
196
352
5
10
20
30
40
50
60
70
80
90
95
Betula pendula
Corylus avellana
Alnus glutinosa
E
Betula pendula
Alnus glutinosa
Corylus avellana
F
value
NMI
Purity
Composite images
Feature based
PCA preprocessed

Figure 5. Dimensionality reduction in multispectral IFC images

(A–C) UMAP 2D embeddings of multispectral images for (A) composite images from overlayed channel data (See also Figure S5), (B) PCA merged channel data, using PCA components explaining 95% of the variance in images from each channel, and (C) IDEAS numerical features calculated for all channels used.

(D) Relation of the number of PCA components used and their association with the amount of explained variance in the images.

(E) The gallery represents examples of reconstructed images from the given number of PCA components.

(F) NMI and purity values calculated for tested data preprocessing workflows.

a proof of concept, in the next steps, we demonstrated the performance of UMAP in different use cases that employed multispectral datasets.

UMAP use case – Visualization of complex multispectral IFC dataset

IFC is often used to gather data on multiple object types within one experiment. An example of such an experiment is a high-throughput screening of pollen samples, which is important for ecological studies and allergology. Such samples pose a significant difficulty to standard image analysis, as they are typically composed of an unknown number of pollen species. To demonstrate the ability of UMAP to visualize and cluster complex populations, we applied UMAP to embed one of the largest and most complex reference sets of IFC images consisting of over 35,000 objects from 35 different plant species imaged in three detection channels: brightfield, fluorescence and darkfield (Dunker et al., 2021). Again, PCA-transformed pixel data were used for dimensionality reduction by UMAP.

UMAP produced a low-dimensional representation of objects in a short time, with multiple clearly visible separate clusters (Figures 6A and S8). Although the total number of clusters did not match the number of species, objects from a single species most often occupied a specific region in the embedding. Only for a small fraction of species were the pollen images allocated by UMAP into separate clusters. In addition, some clusters were heterogeneous and included objects representing different species. Notably, in such cases, the objects from each species still occupied distinct areas of a cluster. We found that the species with very low object counts could not be properly assigned to the clusters. This indicated that in some complex datasets, a certain minimum number of objects in each class is required for correct clustering.

Because UMAP is a graph-based dimensionality-reduction method, the relations between points and clusters can be studied in detail by generating a connectivity plot. The one obtained for the examined dataset clearly showed the embedding areas where clusters were strongly locally connected (Figure 6B). We tested if restricting the clustering to these areas would improve the clustering results. Using a full dataset, hdbscan reached purity value of 0.73 and NMI of 0.54. When we restricted the clustering to points with strong local connections, the purity and NMI values increased to 0.94 and 0.61, respectively. Therefore, these sites are good starting points for sampling objects in an unsupervised manner to guide downstream analysis and classification in open-source or commercial tools such as IDEAS software.

UMAP use case – Unsupervised screening of red blood cell morphology

IFC has multiple applications, which implies that the analyzed objects can have significantly different characteristics, dependent on the sample type. For example, the images of cells have lower contrast than those of pollen. Consequently, distinguishing blood cells from the background is much more challenging than in the case of pollen. There is a growing interest in the objective analysis methods of large-scale datasets for biomedical applications. A prominent example is the assessment of the quality of stored red blood cells (RBCs) for transfusions, which is done by counting the six RBCs subclasses: Smooth disc, crenated disc, crenated discoid, crenated spheroid, crenated sphere, and smooth sphere. A change in discs toward spheres is associated with RBC degradation. To determine whether UMAP can provide meaningful representations of such objects, we used a dataset of human RBC imagery classified both by experts and an automated deep learning approach (Doan et al., 2020). UMAP properly distinguished between the main morphological classes: disc/discoid versus spheroid/sphere objects (Figure 7). In addition, UMAP placed cells imaged in a side view at equal distances between the two major clusters. A further distinction between the RBC classes was, however, impossible. A subgroup of objects annotated as crenated and smooth discs were put in a separate cluster (group of objects in the bottom left of the embedding). Through observation of the images from this cluster one can conclude that the pictures represent a view from the top of the RBCs, whereas the rest of the crenated and smooth discs were imaged from the bottom. Hdbscan clusters corresponded well with main RBC classes. The few objects which lowered the purity of the calculated clusters were located in the embedding near objects of different RBC class (as defined by UMAP), most likely

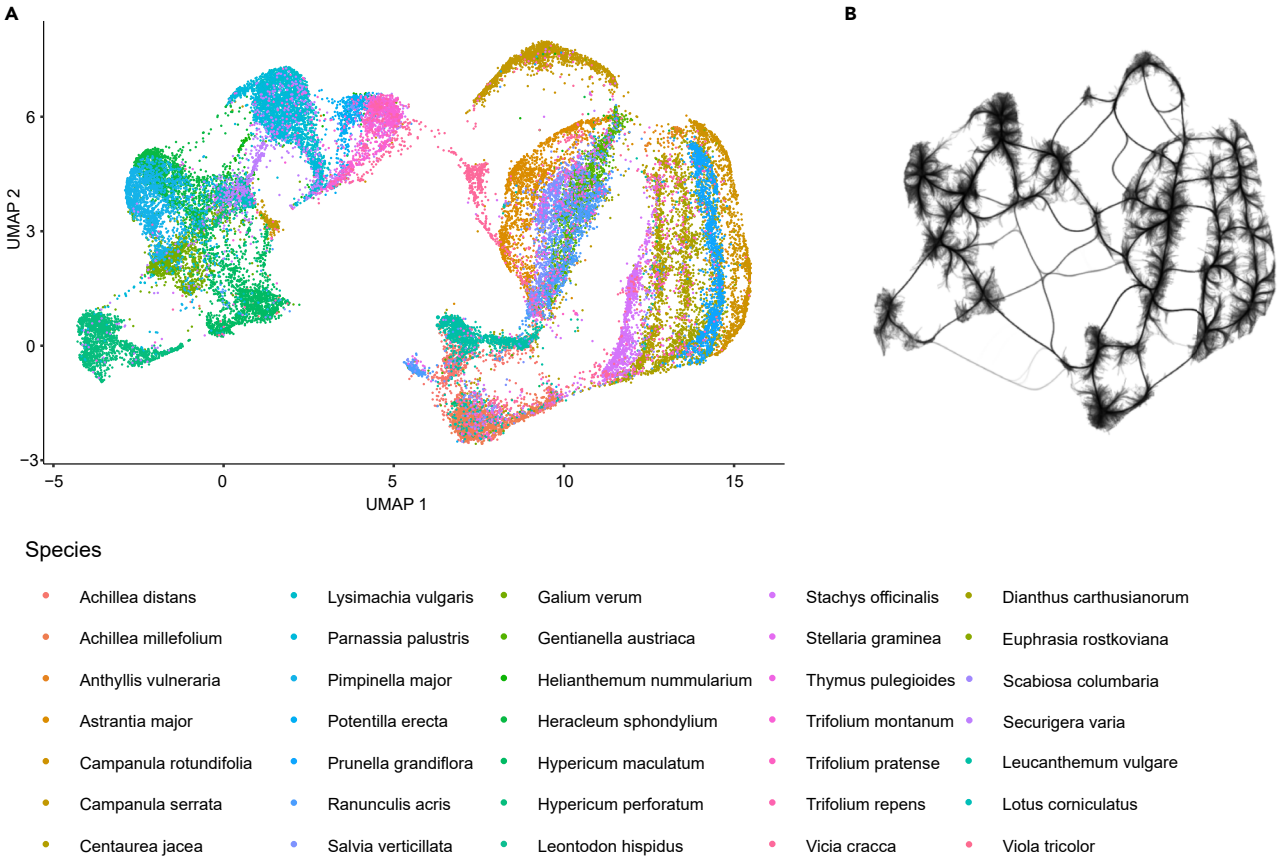


Figure 6. Visualization of the complex multispectral dataset

(A and B) (A) 2D UMAP embedding of images from 35 pollen species (See also Figure S8), (B) connectivity plot of the UMAP embedding.

because of the varied object illumination by the LED in brightfield channel, therefore rendering the distribution of background pixel values more similar to those from objects belonging to a different class.

UMAP use case – Multidimensional outlier detection

IFC, a high-throughput technique, creates a need for filtering the input data of outlying objects, i.e., those that deviate significantly from the other objects and, thus, can hinder result interpretation. However, because the image data are hyperdimensional, identifying outliers should also consider that an object can be an outlier with respect to non-human readable features. To demonstrate how UMAP can be employed to identify outliers in IFC data, we applied it to the THP1 cell images captured in 5 detection channels (Figures 8A–8J). UMAP users can decide whether, on graph building, the combination of local point sets is performed through the union of graph edges (default behavior, see STAR Methods) or their intersection. A change toward intersection decreases the graph connectivity. To identify outliers, we gradually forced UMAP to perform intersection (Figures 8A–8F). As a result, the loosely connected objects were pushed from the clusters. Next, we analyzed the identified outliers in the IDEAS software. The inspection of the image gallery revealed that the outlying objects included highly irregular cells and doublets (Figure 8G). As expected, an object outlying in value for one feature appeared normal when analyzed with respect to another feature (Figures 8H and 8I). Having already found the outliers, we were able to guide the IDEAS Feature Finder wizard to identify key image features and distinguish them from normal objects (Figure 8J).

UMAP use case – Visualization of complex multispectral IFC dataset

To demonstrate the ability of UMAP to represent complex populations, we used a dataset consisting of brightfield, fluorescence and darkfield images of bee pollen (Figure 9A). We preprocessed the multispectral data with PCA as described above, used UMAP for dimensionality reduction and clustered the points

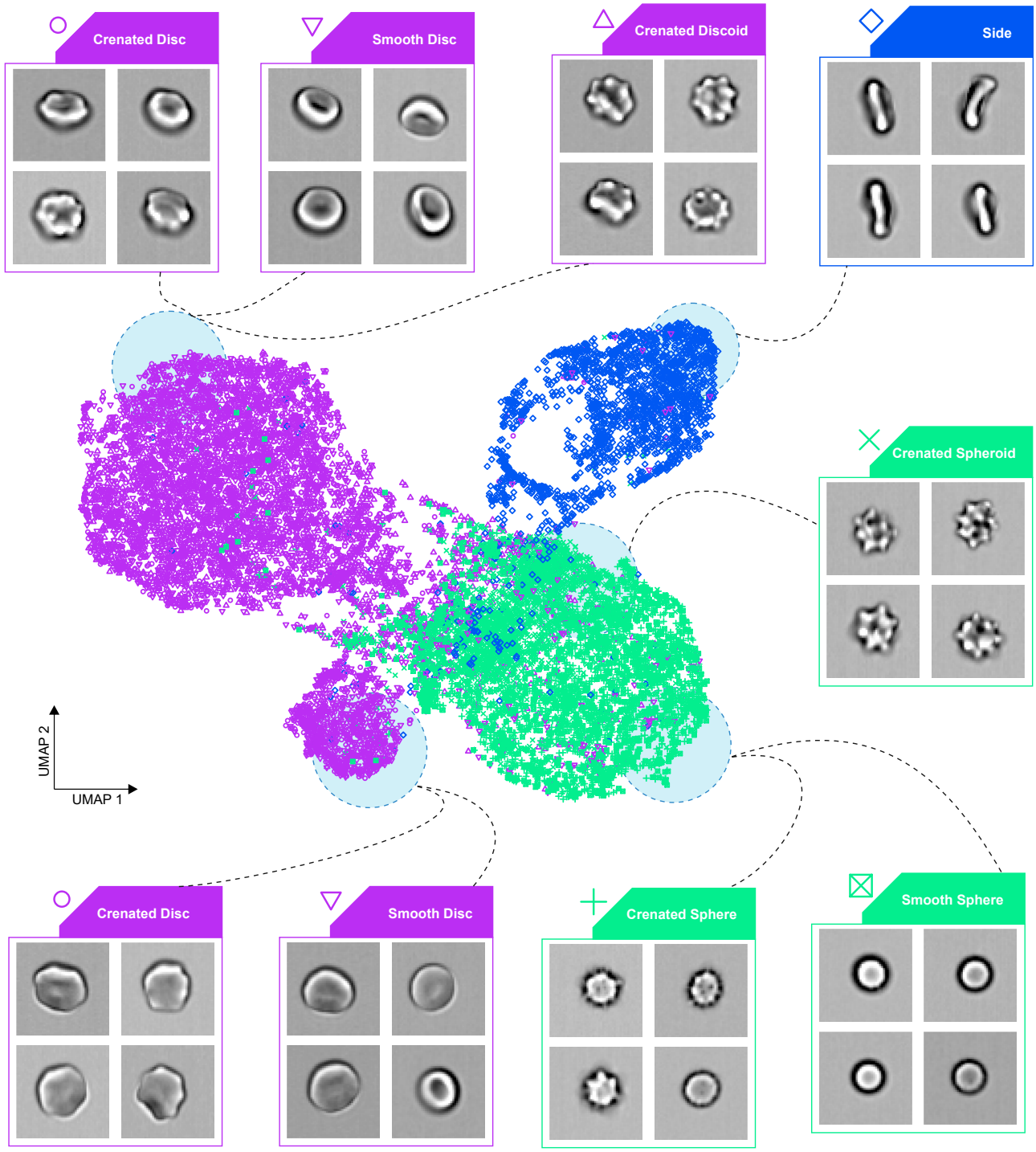


Figure 7. UMAP embedding of low-contrast objects

2D embedding of the RBCs. Example images of the six main types of RBSs and cells captured in the side view are provided. Light blue circles depict locations from which sample images were selected.

with the hdbscan. Furthermore, to validate the correctness of the UMAP embedding, we performed the analysis of this dataset in IDEAS software and overlayed the results with UMAP clusters (Figures 9B–9E, STAR Methods section). Hdbscan distinguished 4 clusters of points. The manual analysis of the dataset in the IDEAS software identified 3 subpopulations based on the area of the objects and the proportion

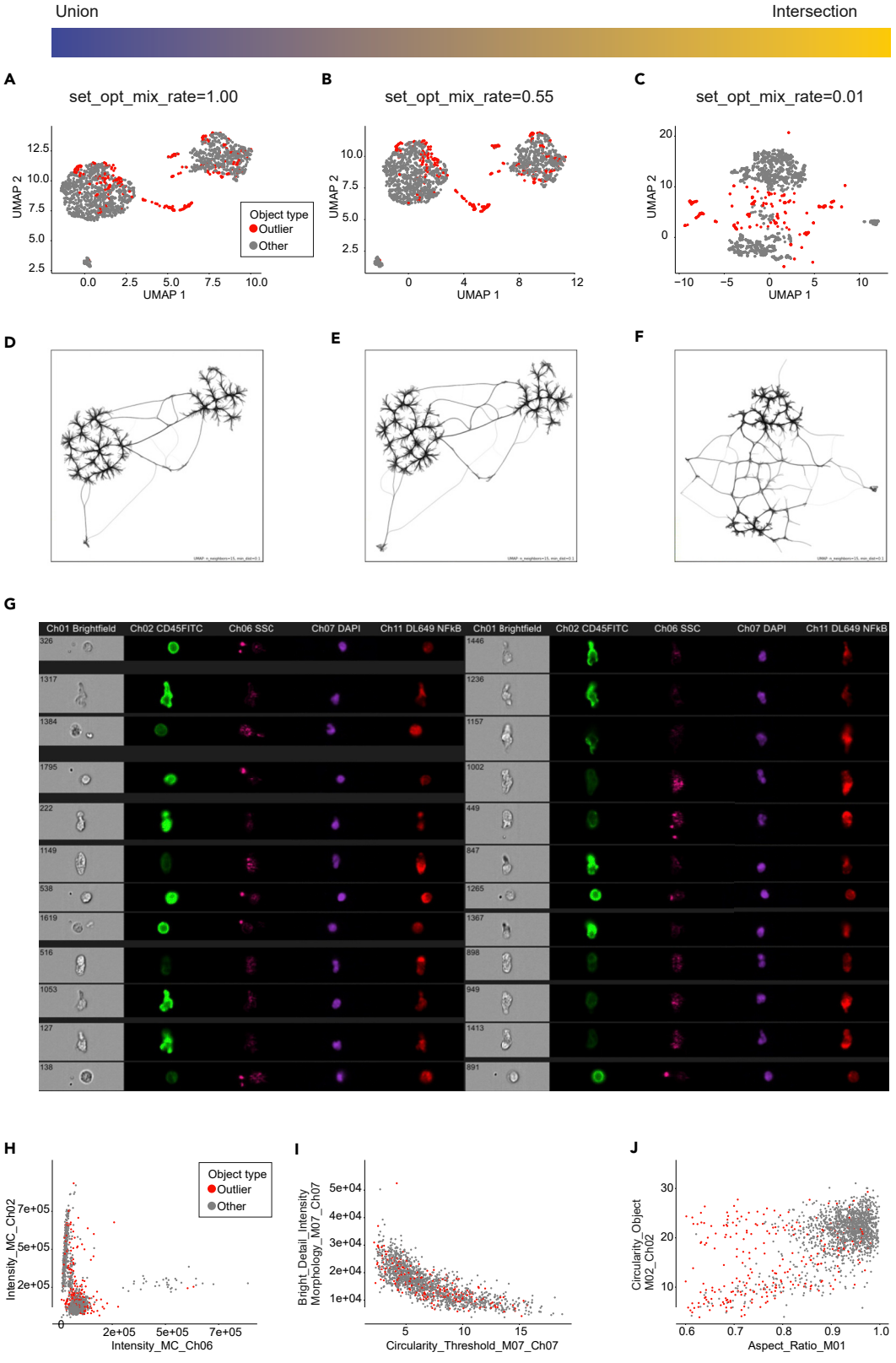

Union
Intersection
A
set_opt_mix_rate=1.00
B
set_opt_mix_rate=0.55
C
set_opt_mix_rate=0.01
UMAP 1
UMAP 2
Object type
Outlier
Other
D
E
F
G
Ch01 Brightfield
Ch02 CD45FITC
Ch06 SSC
Ch07 DAPI
Ch11 DL649 NFkB
H
Intensity_MC_Ch02
Intensity_MC_Ch06
I
Bright_Detail_Intensity_Morphology_M07_Ch07
Circularity_Threshold_M07_Ch07
J
Circularity_Object_M02_Ch02
Aspect_Ratio_M01

Figure 8. Application of UMAP for outlier detection

(A–F) UMAP embeddings (A–C) and corresponding connectivity plots (D–F) of the images of THP1 cells show changes in the data representation as a response to UMAP working in union or intersection mode.

(G) Gallery of identified outlying objects.

(H and I) Scatterplots of the top features discriminating between round and irregular cells, with UMAP-identified outliers marked in red.

(J) Scatterplot of the two top features discriminating between UMAP-identified outliers and the rest of the objects.

of their length and width (aspect ratio). The first subpopulation was characterized by high area and high aspect ratio, the second one consisted of objects with low area and low aspect ratio, whereas the third included pollen with low area and high aspect ratio. Furthermore, we delineated that the first population (high area and high aspect ratio) was composed of two types of pollen: spiky edge and smooth edge ones. Cluster 1 corresponded to the pollen characterized in IDEAS by low area and high aspect ratio, clusters 3 and 4 consisted of pollen with high area and high aspect ratio, whereas cluster 2 contained objects with low area and low aspect ratio, with a fraction of pollen from the other subpopulations. Notably, UMAP correctly placed spiky edge and smooth edge pollen types into separate clusters (3 and 4 respectively). Quantitatively, in each cluster >90% of objects were from single pollen type (as classified by the manual expert analysis done in IDEAS).

DISCUSSION

With the increasing challenges posed by the rapidly growing number of studies of complex populations at single-cell level, there is a demand for further advances in IFC data analysis methods. To address the trade-off between automation and analytical power, several data analysis approaches are available. The basic pipelines utilize low information content and are often time consuming. The advanced deep learning methods take advantage of the high dimensionality of data, but require user-defined data labels. Here, we propose an alternative solution that has the advantages of both approaches, yet lacks the challenges associated with them. Our work demonstrates that dimensionality reduction by UMAP is a robust approach to analyzing complex IFC datasets in an unsupervised manner, bringing immediate informative representations, even in the case of the complex datasets. The applications of dimensionality reduction techniques in the analyses of multispectral imagery have been extensively examined in the context of satellite images, geoscience, and agriculture (Ding et al., 2021; Liu et al., 2021; Xi et al., 2021). These methods have also been used for biological data, including single-cell imaging, however in most cases they involve the extraction of a vector of features from each image (Bandyopadhyay et al., 2014; Carpenter et al., 2006; Mitra et al., 2022; Peralta and Saeys, 2020; Velliangiria et al., 2019). In contrast, we used vectors of the pixel values as the input for UMAP. Of course, feature-based algorithms provide insight into biologically relevant characteristics, thus allowing us to directly address a scientific question of interest. We tested the UMAP performances on two input types: numerical features generated by the IDEAS software and pixel-based image data. Whereas UMAP produced meaningful embeddings in both cases, it achieved superior results when run on images, although in such events extracting information content was hindered to some extent by the prevalent background pixels. Despite this, as outlined here, pixel-based approach is an attractive solution in many use cases, especially when the biological background is well understood and the analysis is aimed at determining the composition of the sample. There are no reports comparing different dimensionality reduction algorithms for the purpose of IFC imagery processing step. We benchmarked several widely used methods. UMAP was the fastest among them and provided the most accurate data representations, which outcompeted the very popular t-SNE. Notably, the growing size of the embedded datasets did not increase the computational time of UMAP even when working with >40,000 hyperparameter objects. In addition to the ability to resolve major subpopulations, UMAP also preserved the continuity of the morphology changes, making the plots easy to interpret.

We found that the UMAP performance can be further improved if the input is preprocessed by PCA. By combining UMAP with an upstream PCA image decomposition, we created a workflow in which IFC data from multiple fluorescence channels could be analyzed jointly. PCA preprocessing of the images has been successfully applied previously, showing that it is a general trend that majority of the variation present in the images can be omitted yielding high-speed gains without sacrificing the accuracy (Benito and Peña, 2005; Khaing et al., 2020; Ng, 2017; Zhao et al., 2022). Our approach resolved major subpopulations with effectiveness comparable to the classification based on deep learning or expert judgment, without reference to prior knowledge. This ability makes UMAP a promising alternative to lengthy procedures and can be particularly useful in high-throughput screenings and biomedical applications.

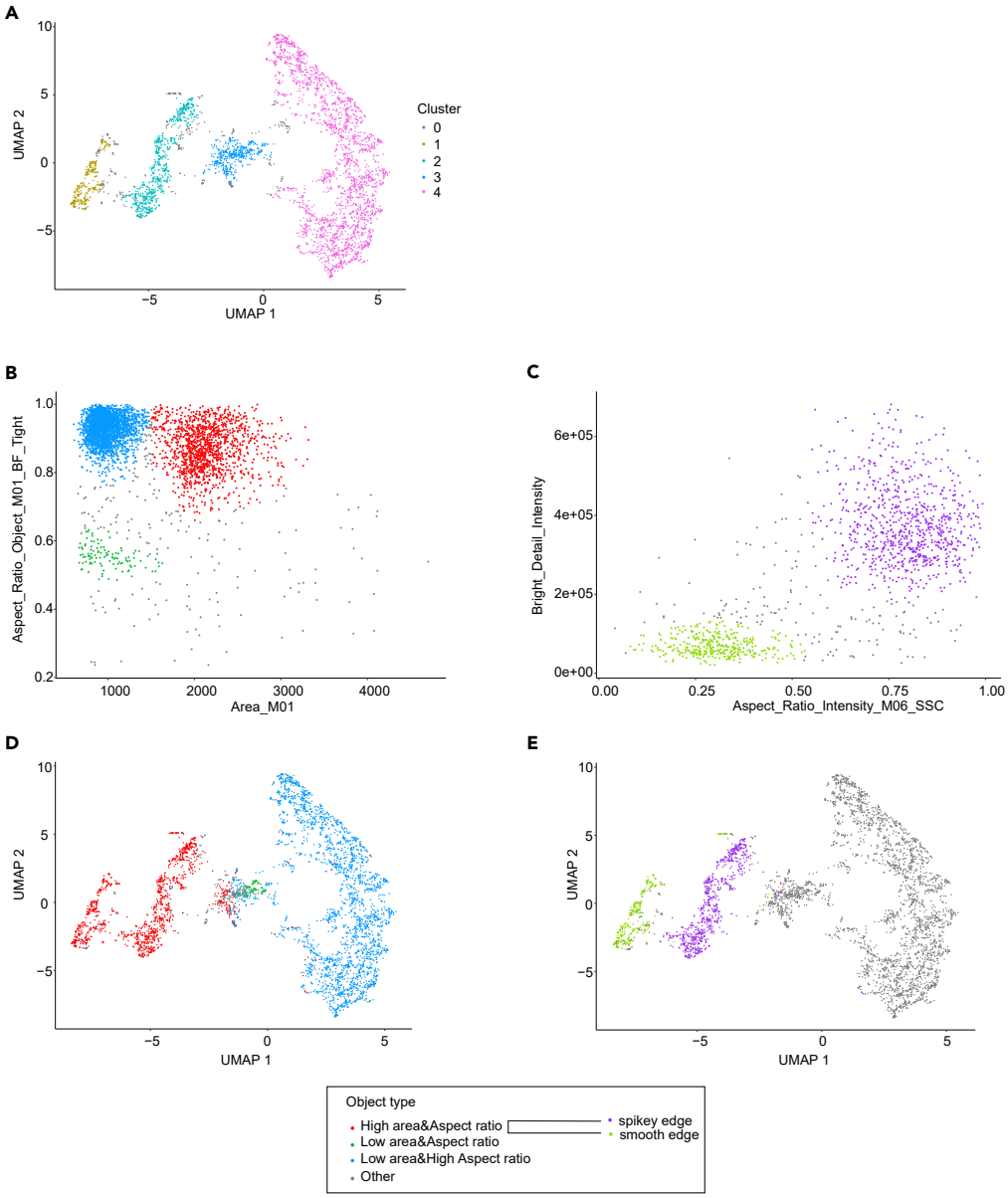


Figure 9. Visualization of complex multispectral IFC dataset

(A) UMAP embedding (A) of the bee pollen dataset. Colors correspond to the hdbscan detected clusters.

(B–E) (B) IDEAS software detected groups of objects, (C) IDEAS software analysis of the High area & Aspect ratio objects group, (D and E) UMAP embedding of the bee pollen dataset with colors corresponding to the groups of objects identified in the IDEAS software.

Dimensionality reduction methods are typically utilized for the purpose of using the reduced embeddings as inputs for the deep learning algorithms. Our workflow can be applied as a standalone solution for less complex datasets, without the necessity for additional analyses that rely on more advanced architectures. However, deep learning will be increasingly utilized in IFC, which can be anticipated based on the successful attempts reported thus far (Dunker et al., 2021; Luo et al., 2021; Otesteanu et al., 2021; Rodrigues et al., 2021). As presented here, UMAP provides a highly valuable guidance to initiate the analysis when no reference is available and enables a very fast tagging of specific cells. Creating an unsupervised UMAP embedding can be readily combined with deep learning as an upstream procedure, to facilitate the tedious manual object annotation or selection of reference populations.

Overall, this work establishes the feasibility of UMAP with a prior PCA image decomposition toward the rapid analysis of IFC datasets. It complements the existing IFC data analysis methods and can contribute to further development of this revolutionary technology.

Limitations of the study

This study focused primarily on the images of objects, where strong morphological differences could be readily observed. Therefore, it remains to be explored how will UMAP behave when more subtle changes are present in the analyzed objects, i.e. differences in granularity or in the proportions between the nucleus and the cytoplasm.

We also note, that in case of complex samples, like the explored dataset of 35 pollen species, postprocessing of the UMAP embedding with unsupervised clustering and additional manual curation by the field expert are indispensable.

It also remains to be studied how IFC data from small objects, where resolution of the camera is too small to capture morphology (i.e., imaging of micronuclei formation, mitochondria counting or spot counting of viral particles) or where signal to noise ratio is small (i.e., imaging of extracellular vesicles) behave when processed with UMAP.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Equipment
  - Dimensionality-reduction algorithms
  - Imaging flow cytometry – Pollen data acquisition
  - Imaging flow cytometry – THP1 data processing in IDEAS
  - Image preprocessing
  - Runtime
  - Reproducibility of large-scale structure
  - Unsupervised clustering

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105142.

ACKNOWLEDGMENTS

This work was supported by the statutory funds of the Institute of Bioorganic Chemistry, Polish Academy of Sciences. The imaging data collected in this study were acquired using the infrastructure developed under the project NEBI - National Imaging Centre for biological and biomedical sciences, POIR.04.02.00-00-C004/19, co-financed through the European Regional Development Fund (ERDF) in the frame of Smart Growth Operational Program 2014–2020 (Measure 4.2 Development of modern research infrastructure of the science sector).

We would like to thank Dr. Łukasz Grewling (Laboratory of Aerobiology, Department of Systematic and Environmental Botany, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland) for collecting and sharing pollen samples.

We are grateful to MSc. Elena Motivans and Dr. Demetra Rakosy for collecting pollen, as well as Dr. Susanne Dunker and Dr. Thomas Hornick (all Helmholtz-Centre for Environmental Research, Leipzig, Germany

& German Center for Integrative Biodiversity Research, Halle-Jena-Leipzig, Germany) for generously sharing the IFC pollen data.

We would like to thank Dr. Ann Power, Mrs. Natascha Steinberg, Dr. Richard Jones and Professor John Love (College of Life and Environmental Sciences, University of Exeter, UK) for granting the permission to use the IFC pollen data.

We would also like to show our gratitude to Dr. Owen Hughes, Brian E. Hall and Dr. Michał Konieczny (Luminex, Seattle, WA, USA) for kindly sharing the IFC data and helpful discussions.

AUTHOR CONTRIBUTIONS

Conceptualization, I.S. and P.J.; Methodology, I.S.; Software, I.S.; Formal Analysis, I.S.; Investigation, I.S. and A.S.-C.; Writing – Original Draft, I.S, A.S.-C. and P.J.; Writing – Review & Editing, I.S., A.S.-C., M.F. and P.J.; Visualization, I.S.; Funding Acquisition, M.F. and P.J.; Resources, I.S., M.F. and P.J.; Supervision, P.J.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 11, 2022
Revised: July 29, 2022
Accepted: September 9, 2022
Published: October 21, 2022

REFERENCES

Akhbardeh, A., and Jacobs, M.A. (2012). Comparative analysis of nonlinear dimensionality reduction techniques for breast MRI segmentation. Med. Phys. *39*, 2275–2289. https://doi.org/10.1118/1.3682173.

Bandyopadhyay, S., Bhadra, T., Mitra, P., and Maulik, U. (2014). Integration of dense subgraph finding with feature clustering for unsupervised feature selection. Pattern Recognit. Lett. *40*, 104–112. https://doi.org/10.1016/j.patrec.2013.12.008.

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44. https://doi.org/10.1038/nbt.4314.

Benito, M., and Peña, D. (2005). A fast approach for dimensionality reduction with image data. Pattern Recognit. Lett. *38*, 2400–2408. https://doi.org/10.1016/j.patcog.2005.03.022.

Blasi, T., Hennig, H., Summers, H.D., Theis, F.J., Cerveira, J., Patterson, J.O., Davies, D., Filby, A., Carpenter, A.E., and Rees, P. (2016). Label-free cell cycle analysis for high-throughput imaging flow cytometry. Nat. Commun. *7*, 10256. https://doi.org/10.1038/ncomms10256.

Bucher, E., Kofler, V., Vorwohl, G., and Zieger, E. (2004). Das Pollenbild der Sudtiroler Honige. Landesagentur for Umwelt und Arbeitschutz (Biologisches Labor).

Carpenter, A., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100. https://doi.org/10.1186/gb-2006-7-10-r100.

Ding, S., Keal, C.A., Zhao, L., and Yu, D. (2021). Dimensionality reduction and classification for hyperspectral image based on robust supervised ISOMAP. J. Indus. Prod. Eng. *39*, 19–29. https://doi.org/10.1080/21681015.2021.1952657.

Doan, M., Sebastian, J.A., Caicedo, J.C., Siegert, S., Roch, A., Turner, T.R., Mykhailova, O., Pinto, R.N., McQuin, C., Goodman, A., et al. (2020). Objective assessment of stored blood quality by deep learning. Proc. Natl. Acad. Sci. USA *117*, 21381–21390. https://doi.org/10.1073/pnas.2001227117.

Dunker, S., Motivans, E., Rakosy, D., Boho, D., Mäder, P., Hornick, T., and Knight, T.M. (2021). Pollen analysis using multispectral imaging flow cytometry and deep learning. New Phytol. *229*, 593–606. https://doi.org/10.1111/nph.16882.

Ferrer-Font, L., Mayer, J.U., Old, S., Hermans, I.F., Irish, J., and Price, K.M. (2020). High-dimensional data analysis algorithms yield comparable results for mass cytometry and spectral flow cytometry data. Cytometry A *97*, 824–831. https://doi.org/10.1002/cyto.a.24016.

Han, Y., Gu, Y., Zhang, A.C., and Lo, Y.H. (2016). Review: imaging technologies for flow cytometry. Lab Chip *16*, 4639–4647. https://doi.org/10.1039/c6lc01063f.

Hennig, H., Rees, P., Blasi, T., Kamentsky, L., Hung, J., Dao, D., Carpenter, A.E., and Filby, A. (2017). An open-source solution for advanced imaging flow cytometry data analysis using machine learning. Methods *112*, 201–210. https://doi.org/10.1016/j.ymeth.2016.08.018.

Huang, X., Wu, L., and Ye, Y. (2019). A review on dimensionality reduction techniques. Intern. J. Pattern Recognit. Artif. Intell. *33*, 1950017. https://doi.org/10.1142/S0218001419500174.

Khaing, T., Nyein, P., Khaing, M., and Wai, K. (2020). Dimension reduction of images using principal component analysis algorithm. Iconic Res. Eng. J. *3*, 39–42.

Liu, H., Li, W., Xia, X.G., Zhang, M., and Tao, R. (2021). Superpixelwise collaborative-representation graph embedding for unsupervised dimension reduction in hyperspectral imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. *14*, 4684–4698. https://doi.org/10.1109/JSTARS.2021.3077460.

Luo, S., Shi, Y., Chin, L.K., Hutchinson, P.E., Zhang, Y., Chierchia, G., Talbot, H., Jiang, X., Bourouina, T., and Liu, A.Q. (2021). Machine-learning-assisted intelligent imaging flow cytometry: a review. Adv. Intell. Syst. *3*, 2100073. https://doi.org/10.1002/aisy.202100073.

McInnes, L., and Healy, J. (2017). Accelerated hierarchical density based clustering. In International Conference on Data Mining Workshops (IEEE), pp. 33–42. https://doi.org/10.1109/ICDMW.2017.12.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://arxiv.org/abs/1802.03426.

Mitra, P., Murthy, C., and Pal, S. (2022). Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/34.990133.

Mochalova, E.N., Kotov, I.A., Lifanov, D.A., Chakraborti, S., and Nikitin, M.P. (2021). Imaging flow cytometry data analysis using convolutional neural network for quantitative investigation of phagocytosis. Biotechnol. Bioeng. https://doi.org/10.1002/bit.27986.

Ng, S. (2017). Principal component analysis to reduce dimension on digital image. Procedia Comput. Sci. *111*, 113–119. https://doi.org/10.1016/j.procs.2017.06.017.

Otesteanu, C.F., Ugrinic, M., Holzner, G., Chang, Y.T., Fassnacht, C., Guenova, E., Stavrakis, S., deMello, A., and Claassen, M. (2021). A weakly supervised deep learning approach for label-free imaging flow-cytometry-based blood diagnostics. Cell Rep. Methods *1*, 100094. https://doi.org/10.1016/j.crmeth.2021.100094.

Peralta, D., and Saeys, Y. (2020). Robust unsupervised dimensionality reduction based on feature clustering for single-cell imaging data. Appl. Soft Comput. *93*, 106421. https://doi.org/10.1016/j.asoc.2020.106421.

Rodrigues, M., Probst, C., Zayats, A., Davidson, B., Riedel, M., Li, Y., and Venkatachalam, V. (2021). The in vitro micronucleus assay using imaging flow cytometry and deep learning. NPJ Syst. Biol. Appl. *7*, 1–12. https://doi.org/10.1038/s41540-021-00179-5.

Saeys, Y., Van Gassen, S., and Lambrecht, B.N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat. Rev. Immunol. *16*, 449–462. https://doi.org/10.1038/nri.2016.56.

Sumithra, V., and Subu, S. (2015). A review of various linear and non linear dimensionality reduction techniques. Int. J. Comput. Sci. Inf. Technol. *6*, 2354–2360.

Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science *290*, 2319–2323. https://doi.org/10.1126/science.290.5500.2319.

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Velliangiri, S., Alagumuthukrishnan, S., and Thankumar, S. (2019). A review of dimensionality reduction techniques for efficient computation. Procedia Comput. Sci. *165*, 104–111. https://doi.org/10.1016/j.procs.2020.01.079.

Voronin, D.V., Kozlova, A.A., Verkhovskii, R.A., Ermakov, A.V., Makarkin, M.A., Inozemtseva, O.A., and Bratashov, D.N. (2020). Detection of rare objects by flow cytometry: imaging, cell sorting, and deep learning approaches. Int. J. Mol. Sci. *21*, E2323. https://doi.org/10.3390/ijms21072323.

Xi, X., Xia, K., Yang, Y., Du, X., and Feng, H. (2021). Evaluation of dimensionality reduction methods for individual tree crown delineation using instance segmentation network and UAV multispectral imagery in urban forest. Comput. Electron. Agric. *191*, 106506. https://doi.org/10.1016/j.compag.2021.106506.

Zhao, B., Dong, X., Guo, Y., Jia, X., and Huang, Y. (2022). PCA dimensionality reduction method for image classification. Neural Process. Lett. *54*, 347–368. https://doi.org/10.1007/s11063-021-10632-5.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Pollen grains of common hazel (*Corylus avellana*) - part of datasets 1, 2, and 3	Dr. Łukasz Grewling, Laboratory of Aerobiology, Department of Systematic and Environmental Botany, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland	N/A
Pollen grains of silver birch (*Betula pendula*) - part of dataset 1	Dr. Łukasz Grewling, Laboratory of Aerobiology, Department of Systematic and Environmental Botany, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland	N/A
Pollen grains of common alder (*Alnus glutinosa*) - part of dataset 1	Dr. Łukasz Grewling, Laboratory of Aerobiology, Department of Systematic and Environmental Botany, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland	N/A
Deposited data		
Betula pendula, *Corylus avellana*, *Alnus glutinosa* pollen	This study	ifc_umap/example_data at main · istolarek/ifc_umap · GitHub
Pollen grains of 35 plant species - part of dataset 4	Dunker et al., 2021, personal communication	https://doi.org/10.1111/nph.16882
Images of red blood cells - part of dataset 5	Doan et al. (2020)	https://doi.org/10.1073/pnas.2001227117
Images of THP1 cells - part of dataset 6	Luminex - personal communication. Dr. Michał Konieczny (Luminex).	N/A
Images of bee pollen grains - part of dataset 7	Luminex - personal communication. Dr. Owen Hughes (Luminex), with permission granted by Dr. Ann Power, Mrs. Natascha Steinberg, Dr. Richard Jones and Professor John Love (College of Life and Environmental Sciences, University of Exeter, UK).	N/A
Software and algorithms		
Python version 3.7	Python Software Foundation	https://www.python.org
R version 4.1.2	The R Project for Statistical Computing	https://www.r-project.org/
IDEAS version 6.2	Luminex	https://www.luminexcorp.com/imagestreamx-mk-ii/#software
INSPIRE™ version 4.1	Luminex	https://www.luminexcorp.com/imagestreamx-mk-ii/#software

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Paulina Jackowiak (paulinaj@ibch.poznan.pl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This article analyzes existing data, kindly shared by scientists mentioned below. Raw data generated in this study have been deposited in a github repository: https://github.com/istolarek/ifc_umap.

The first dataset consisted of images and image features of the collected pollen from three plant species (Figures 1, 2, and 5). The data were produced as part of this work.

The second dataset contained images from single pollen species, i.e., *C. avellana* imaged in the fluorescence channel and exported as TIFF files from unchanged 8-bit raw images and the same images with augmentations: modified number of bits per image, enhanced contrast, added padding at the image edges (Figure 3). The data were produced as part of this work.

The third dataset contained raw images and their extracted features for *C. avellana* pollen from a single biological replicate imaged in a series of technical replicates (Figure 4). The data were produced as part of this work.

The fourth dataset included TIFF images of annotated pollen from 35 plant species (Figure 6). The data were kindly shared by Dr. Susanne Dunker and Dr. Thomas Hornick (Helmholtz-Centre for Environmental Research, Leipzig, Germany & German Centre for Integrative Biodiversity Research, Halle-Jena-Leipzig, Germany).

The fifth dataset contained annotated TIFF images of RBCs imaged in the brightfield channel (Figure 7). The data were downloaded from the public repository shared in ref. (Doan et al., 2020).

The sixth dataset consisted of multi-channel IFC imagery of THP1 cells (Figure 8). The raw IFC data were kindly shared by Dr. Michał Konieczny (Luminex).

Lastly, the seventh dataset consisted of images and image features calculated in IDEAS software of unannotated bee pollen (Figure 9). The raw IFC data were kindly shared by Dr. Owen Hughes (Luminex), with permission granted by Dr. Ann Power, Mrs. Natascha Steinberg, Dr. Richard Jones and Professor John Love (College of Life and Environmental Sciences, University of Exeter, UK).

The code used to produce the analyses presented in this manuscript with example IFC data is available through a github repository: https://github.com/istolarek/ifc_umap.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact on request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Pollen grains of three species, i.e., common hazel (*Corylus avellana*), silver birch (*Betula pendula*) and common alder (*Alnus glutinosa*), were collected from natural populations growing in Poznań (Western Poland). Pollen grains were collected in 50 mL tubes by gently shaking the catkins (male inflorescences) in the full flowering phase. Until the IFC analysis (datasets 1–3, see chapter below), pollen grains were stored at room temperature. Pollen grains of *B. pendula* are small (~22 μm), triporate, round to oval in shape with thick intines below pores. The general characteristics of hazel pollen grains are similar to birch except that the hazel pollen grains are larger (~28 μm) and have a more triangular shape (in polar view). Pollen grains of *A. glutinosa* (22–25 μm) often have 4–5 pores with distinct thick arches (arci) running between the pores (Bucher et al., 2004).

METHOD DETAILS

Equipment

All presented datasets were produced on ImageStream$^{®X}$ Mark II (Luminex, Seattle, WA, USA) (Bucher et al., 2004; Doan et al., 2020).

Dimensionality-reduction algorithms

We used a total of nine linear and nonlinear algorithms: UMAP - Uniform Manifold Approximation and Projection, t-SNE - t-distributed stochastic neighbor embedding, Isomap – Isometric Mapping, LLE - Locally Linear Embedding, MLLE – modified Locally Linear Embedding, MDS – Multidimensional Scaling, SE – Spectral Embedding, PCA – Principal Component Analysis, ivis – ivis a machine learning library for reducing dimensionality of very large datasets using Siamese Neural Networks.

Imaging flow cytometry – Pollen data acquisition

IFC was performed on a two-camera ImageStream$^{®X}$ Mark II with INSPIRE v4.1 acquisition software (Luminex). Pollen autofluorescence was excited by a 488 nm laser at 7 mW, and emission was captured in the range of 505–560 nm (Ch02). All images were captured with the 40× objective, and a cell classifier (threshold) was applied to the brightfield channel (Ch04) to exclude small particles.

Imaging flow cytometry – THP1 data processing in IDEAS

IFC dataset was kindly shared by Dr. Michał Konieczny (Luminex).

IFC data analysis was performed using IDEAS v6.2 image analysis software (Luminex). The brightfield (Ch01) RMS gradient and area of an object were used to interrogate single-cell events in focus. Single-cell events were gated through Centroid_X and the area of an object to remove clipped images, leaving 4,210 focused, nonclipped single-cell objects. The object area and aspect ratio in brightfield (Ch01) were used to group cells into objects characterized by high area and aspect ratio (1,325 objects), low area and high aspect ratio (2,572 objects), and low area and low aspect ratio (107 objects). The Feature Finder Wizard was applied to engineer features discriminating between sample objects representing images of spiky-edge and smooth-edge pollen types. The Feature Finder calculated aspect ratio intensity (Ch06) and bright detail intensity (Ch01) as the top discriminative features with discriminative RD scores of 2.5 and 2.35, respectively.

Image preprocessing

Except where explicitly noted, all pixel-based images were exported from the IDEAS software as 8-bit raw TIFF images. Each image was processed in Python 3.5 with the opencv2 library to convert it into grayscale, and to standardize each image size, it was reshaped to 64 × 64 pixel dimensions with a "resize" function.

Runtime

For all algorithms, the timing was determined in Python using the "%time" (wall clock) time measurements.

Reproducibility of large-scale structure

Each algorithm was run on a full dataset (9,000 objects). Then, subsamples of sizes 200, 500, 1,000, 3,000, 5,000, and 7,000 were uniformly drawn three times, generating 21 data subsamples of varying sizes. We then ran each algorithm on each data subsample, saving the embeddings. For each subsample and algorithm, we obtained two vectors of embedded coordinates (x,y). For each subsample, we also obtained such a pair of embedded coordinates (x′,y′). From this, we computed $(|cor(x,x')| + |cor(y,y')|)/2$, where $|.|$ denotes the absolute value and cor the computation of the Pearson correlation coefficient. This quantity thus measures the average correlation of coordinates between the full embedding and subsamples from various sizes, maximized across axial symmetries across the x-axis and/or y-axis.

Unsupervised clustering

Hdbscan algorithm (package dbscan, ver. 1.1-10) was used to cluster the points from reduced embeddings. Next, the cluster purity and normalized mutual information (package aricode, ver. 1.0.0) were calculated where feasible.