The Biofilm Image Library: Enabling Machine Learning for 3D Biofilm Analysis

Albert Parker[1, 2], Heidi Smith[1,3], Bettina Buttaro[4] and Joseph Picone[5]

1. The Center for Biofilm Engineering, Montana State University, Bozeman, Montana, USA
2. Department of Mathematical Sciences, Montana State University, Bozeman, Montana, USA
3. Department of Microbiology and Cell Biology, Montana State University, Bozeman, Montana, USA
4. Sol Sherry Thrombosis Research Center, Lewis Katz School of Medicine, Temple University, Phil., PA, USA
5. The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA

{albert.parker, heidi.smith}@montana.edu, {bettina.buttaro, picone}@temple.edu

Abstract

Microbial biofilms are complex spatially heterogeneous communities that play critical roles in medical, industrial, and environmental systems. Advances in three-dimensional (3D) imaging technologies, particularly confocal microscopy, have enabled the collection of high-resolution 3D biofilm image datasets. However, quantitative analysis of these datasets remains constrained by the lack of standardized repositories, scalable annotation tools, and robust machine-learning methods for 3D segmentation and characterization. Existing biofilm analysis tools rely heavily on threshold-based approaches and often fail when confronted with multi-species or multi-domain biofilms with complex architectures.

This chapter introduces the Biofilm Image Library (BIL), presenting the framework for a critical and transformative open-access, community-driven infrastructure designed to enable the development of next-generation computational tools for biofilm analysis. The overarching goal of the BIL is to provide curated 3D microscopy datasets, standardized metadata, expert annotations, and integrated machine-learning pipelines to facilitate reproducible research in computational microbiology. In this chapter, we identify key challenges and research gaps in biofilm image analysis, including cell segmentation, spatial co-localization across imaging channels, fungal morphology quantitation, and multiscale modeling of microbial communities. We further describe how emerging machine-learning methods, including deep learning-based segmentation and annotation frameworks, can enable quantitative characterization of biofilm structure, organization, and function.

By combining standardized imaging resources with accessible computational tools, the BIL accelerates advances in biofilm science, promotes cross-disciplinary collaboration among microbiologists, engineers, and computer scientists, and supports transformative science in applications including antimicrobial resistance, disease diagnostics, environmental monitoring, spaceflight life support systems and astrobiology.

Keywords: biofilm imaging, confocal microscopy, multiscale modeling, computational microbiology

Corresponding Author: Al Parker, 304 Barnard Hall, Center for Biofilm Engineering, Montana State University, P.O. Box 173980, Bozeman, Montana, USA 59717-3980. Tel: 406-994-5145; Email: albert.parker@montana.edu.

Table of Contents

1. Introduction 1
1.1. A Brief Historical Perspective 1
1.2. Limitations of State of the Art in Machine Learning 2
1.3. Limitations of Analytic Methods 3
1.4. Microscope Acquisition Parameters and Their Impact on Image Data 3
2. Limitations of Current Image Analysis Tools 5
2.1. Bias in Biomass Estimation 5
2.2. Absence of Uncertainty Quantification and Ad hoc Design of Imaging Studies 6
2.3. Poor Understanding of Spatial Correlation of Multiple Species 7
3. Machine-learning Tools Easily Accessible from the BIL 7
3.1. Identification of Individual Microbial Cells 8
3.2. Species Identification and Spatial Correlation of Multiple Species 8
3.3. Predictive Models of Biofilm Behavior 9
3.3.1. Data Collection: Why 3D? 10
3.3.2. Observation and Pattern Discovery: Towards Mechanistic Understanding 10
3.3.3. Predictive Biofilm Science: Open Research Questions 12
4. BIL Resources 12
4.1. Types of Resources 13
4.1.1. Image Data 13
4.1.2. Software Tools 15
4.1.3. Knowledge Discovery 15
4.2. How to Subscribe to BIL Resources**?** 16
4.3. How to Submit Data to the BIL**?** 16
5. BIL Outcomes and Deliverables 16
5.1. Standardization 16
5.2. Transformative Science 17
5.3. Education 17
6. Conclusions 18
Acknowledgments 19
References 20
List of Figures 27

1. Introduction

Microbial biofilms are structured communities of diverse microorganisms that attach to surfaces and to each other, are embedded within extracellular polymeric substances [1,2,3] and are widespread across natural, human, and engineered environments [1,4,5]. Biofilms represent a strategic growth phase that not only allows cells to survive in hostile environments, but also to colonize new niches. Collectively, biofilms are far more resilient than planktonic cells [6-9]. For example, biofilms have increased resistance to antimicrobials and heavy metals, higher desiccation tolerance, and different substrate-utilization kinetics than their planktonic counterparts. Furthermore, biofilms can retain water and nutrients, and can tolerate rapid changes in salinity, osmolarity, pH, nutrient availability, and redox conditions. Together, these adaptive properties drive the persistence and ecological success of biofilms. Mitigating biofilms costs the global economy $4 trillion annually across many market sectors (Table 1 in [10]): medical and human health ($387 billion), consumer products ($211 billion), built environment ($49 billion), food and agriculture ($324 billion), water and wastewater treatment ($117 billion), energy and waste ($5 billion), marine systems ($34 billion), and finally, infrastructure (including oil and gas), which is the largest sector with costs arising from microbially influenced corrosion ($2.7 trillion). Conversely, biofilms also present opportunities for beneficial applications including bioremediation of contaminated environments [11], microbiome restoration, wastewater treatment [12], and bioenergy production [13,14].

1.1. A Brief Historical Perspective

For over 30 years, biofilm research has largely focused on single-species bacterial systems. While single organism studies have provided fundamental insights into biofilm formation, structure, and physiology, as shown in Figure 1, most biofilms in natural and engineered environments exist as complex, multi-species, multi-domain consortia (e.g., composed of bacteria, archaea, fungi and viruses). Increasingly, the community of biofilm researchers is shifting towards the study of these more relevant systems that exhibit complex microbial associations that influence metabolite exchange, interspecies signaling, and cooperative interactions that may contribute to biofilm survival in extreme environmental conditions [15,16]. The complexity of multi-species and multi-domain biofilm systems has exposed the limitations of single-species models, highlighting critical knowledge gaps in understanding how spatial organization, interspecies interactions, and multiscale processes collectively govern biofilm behavior.

To understand bacterial biofilm behavior (e.g., growth, maturation, spatial structuring, substrate and nutrient transport, and responses to chemical gradients) mathematical modeling has played an important role. Most existing models assume that biofilms can be described as a homogeneous medium in which case dimension-reduction approaches can be used to describe phenomena in one or two spatial dimension(s) [7,17]. However, these simplified models do not incorporate the heterogeneity and complexity of the 3D structure that are likely to have a significant impact on the biofilm's function.

Over the last two decades, imaging has become an essential instrument in biofilm research because of the ability to provide three-dimensional (3D) information of living hydrated *in situ* biofilms non-invasively and in real time [18,19,20]. Given that biofilms have high phenotypic and genotypic heterogeneity, there are few analytical techniques besides imaging that can capture these variations in 3D space and time. Imaging by confocal microscopy has allowed gene expression localization [21,22], structural analysis of extracellular material [23-26], community identification [27], fluid flow and diffusivity pattern recognition [18], and visualization of structural heterogeneity in response to antibiotics and hydrogen peroxide production among others [28,29].

At present, analyzing even a single 3D image can be a daunting computational task, with no standardized image data format and very few easy-to-use accessible tools. Several 3D biofilm image analysis tools have been developed, including the venerable COMSTAT [30,31], Center for Microbial Ecology Image

Analysis System (CMEIAS) [32], Daime [33], Beer's Analysis [34] and BiofilmQ [35]. Fundamentally, these existing approaches seek to answer two key questions: (1) *Where is the biomass?* and (2) *Where are the microbial cells in the biofilm?* Answering both questions is essential for understanding community spatial organization and for informing mathematical models to infer function. But answers remain surprisingly elusive. The existing image analysis approaches above attempt to address these questions primarily through intensity-based thresholding and pixel-based volumetric analysis, which fail spectacularly when the biofilm is very sparse (due to a low signal-to-noise ratio); very dense (due to signal heterogeneity); or very thick (due to laser attenuation). Other imaging artifacts may also be present, such as cellular movement, substratum autofluorescence, photobleaching, and point spread function-based distortion, that can result in inaccurate biomass quantification and cell identification [36].

To overcome these challenges, there has been a flurry of research and development of *cell segmentation* tools using conventional and deep learning 3D image analysis methods: BCM3D 2.0 [37], StarDist OPP [38], DeepSeeded [39] and Cellpose–SAM [40]. Training these systems using manually annotated data remains a human resource bottleneck. These methods and their training sets are still being benchmarked [41], with cell segmentation continuing to be an active area of research. Nonetheless, even these state-of-the-art approaches are severely limited and currently constrained to biofilms composed of single color channel image data from single species biofilms.

1.2. Limitations of State of the Art in Machine Learning

Applying machine-learning (ML) approaches to perform data intensive tasks such as cell segmentation to a large database of 3D images requires that new computationally efficient tools be developed. Central to understanding biofilm function in 3D is the ability to analyze spatial and organizational information derived from noisy biological 3D images. This is a task that can only be done by developing ML pipelines that enable analysis of these complex structures across different species and different complex environmental systems. Current limitations in manual and 2D-based analyses highlight the need for scalable ML pipelines capable of analyzing 3D structure-function relationships, ultimately linking biofilm spatial organization to metabolic activity, interspecies interactions, and functional outcomes. A natural starting point is the development of such tools for model systems (e.g. composed of single bacterial species). Examples of biofilm architectures are shown in Figure 1.

Biofilm image analysis pipelines need to manage 3D data from multiple color channels. That is, each color in a 3D confocal image is captured as a separate channel by a laser with a specific wavelength. Currently, most image analysis approaches analyze each color channel separately. Though many tools exist for cell segmentation of 2D images, far fewer tools exist for 3D cell segmentation of multi-color images. At present, the available image analytical tools that provide state-of-the-art performance are either commercial and/or computationally expensive [38,42,43,44], and in our experience do not deliver accurate cell segmentations on biofilm images.

Highly accurate cell segmentation of these types of images cannot be done by simply decomposing the signal into monochromatic channels and applying models trained on specific colors. To separate a cluster of cells with overlapping or adjacent boundaries, it is desirable to fuse information from multiple channels. We have used existing tools, such as those in ImageJ 3DSuite [45,46], to manually segment and annotate images, and have found performance, defined by the accuracy of boundaries identified by the ML algorithms compared to reference annotations, to be very dependent on the image and channel. An example is shown in Figure 2, where an image is composed of red and green channels and then segmented using a combination of traditional signal processing and morphological operations. Building ML algorithms that work across a wide range of 3D images (e.g., different organisms, microscopy, magnification, dimensionality, channel complexity), potentially with many colored channels, will be a challenge and requires large data sets for training.

1.3. Limitations of Analytic Methods

A critical gap remains in methods for analyzing multi-species and multi-domain biofilms. In addition to answering *Where is the biomass* and *Where are the cells?* in multi-species and multi-domain systems, biofilm researchers also want to know *Which species?* Answering is non-trivial, as accurate species identification first depends on cell segmentation, and often depends on manual tuning of algorithm parameters that vary wildly across environmental conditions and imaging scenarios. While advanced ML algorithms offer the potential to overcome this limitation and to approach human performance on cell segmentation, large amounts of annotated data are needed to train such algorithms. To annotate enough complex structures to train the tools to derive rules of spatial and structural organization for cooperation, competition, and protection in response to nutrients, flow and stressors, one needs a large collection of images across diverse model systems.

The lack of image analysis tools has also impacted the progress of more challenging research questions such as *How does biofilm morphology and subcommunity arrangement affect pathogenesis?* To successfully answer such questions will require new robust image analysis methods that are easily accessible, centralized, computationally efficient, capable of micro-segmentation of high resolution images and reproducible across multiple laboratories. While ML and artificial intelligence research is pervasive today, the success of advanced ML algorithms tends to correlate with the amount and quality of training available (and the economic interest in the application area). Many important problems in the life and health sciences simply do not have sufficient data available to support the application of advanced ML algorithms involving long-distance spatial context necessary for macroscale modeling of biofilm communities across entire surfaces such as catheters [42,43]. Advanced techniques using foundation models, pre-training and fine-tuning are still in their infancy for biological imaging.

The BIL (*www.biofilmimagelibrary.org*) overcomes these hurdles and lays the groundwork for an open-access, community-driven repository of standardized, high-quality biofilm images, metadata, and associated computational models, and accelerates the development of open source easy-to-use computationally efficient tools aided by ML. As shown in Figure 3, the BIL is focused on providing three types of resources: (1) *Data:* manually annotated 3D image data that can be used to train sophisticated ML models, (2) *Services:* including image search, similar-image retrieval, and ML tools that automatically annotate images using powerful foundation models; and (3) *Resources:* including software tools for statistical and geometric analysis of annotated images, along with community annotation and evaluation standards. Contributions from academia, hospitals, industry, government labs, and research centers feed the library, and analysis of its large, diverse image inventory enables knowledge discovery, predictive models, and scientific impact. In this way, the BIL fosters computer science innovations that significantly enhance the scientific community's ability to study biofilms across a wide range of environments, from industrial and environmental to medical.

1.4. Microscope Acquisition Parameters and Their Impact on Image Data

Confocal microscope (CM) settings can be tuned and the selected settings impact how the resulting 3D image data are collected. Confocal microscopy encompasses several imaging modalities, including scanning laser, spinning disk, and slit-scanning systems, each differing in scanning strategy, speed, and resolution to suit specific applications. The experimental setup of a confocal scanning laser microscope, one of the most common modalities, consists of a laser excitation source, optical components for selective detection of emitted fluorescence, and a scanning system that acquires sequential optical slices (called *z*-slices) to reconstruct 3D images [47].

Independent of the specific CM, the microscopist collecting the images chooses several aspects of the imaging experiment: the stain or fluorophore used when preprocessing the microbiological specimen, the

laser emission bandwidth and intensity, detector gain, the objective lens to use (resulting in a specific optical magnification), where to image on a surface or substratum (i.e., the field of view (FOV)), how deep into the specimen to image, and the pixelation of the image. These user-defined collection parameters are critical because they determine the sensitivity, resolution, and how representative the acquired data is, which influences the accuracy and variability of downstream quantitative image analysis.

For example, when the microscopist chooses to use a 25X objective lens on the CM, the corresponding FOV being imaged is approximately $620\mu m \times 620\mu m$. There is a very high probability of detecting an individual bacterium (commonly $1\mu m - 2\mu m$ in diameter) in this sized FOV. When the microscopist uses a 63X objective, the FOV is about $250\mu m \times 250\mu m$. Such higher magnification objective lenses will typically be capable of visualizing individual bacterial cells in the FOV, although now the FOV is much smaller, requiring more FOVs and more time on the microscope to maintain a low limit of detection across the entire substratum (e.g., of a medical device or a food contact surface) [48] . For the image data visualized in Figure 1B, the horizontal (xy) FOV was captured by a 25X objective lens resulting in approximately a $620\mu m \times 620\mu m$ FOV with individual cells being clearly visible. Objective lenses with lower magnification than 25X have a much lower chance of distinguishing individual bacteria, making the resulting images problematic for detecting low levels of biofilm, such as residual contamination after cleaning a medical device.

The CM images in Figure 1A-C were collected over different lattice sizes $n_x \times n_y \times n_z$ chosen by the microscopist. A lattice is a regular array of points describing the spatial arrangement of pixels that compose a 3D CM image. The z-dimension is associated with the vertical dimension in line with the CM's objective lens. The x-and y-dimensions of the lattice are orthogonal to the z-dimension. When biofilms are grown on a flat substratum, images are usually collected (as in Figure 1) so that x and y are parallel to the flat substratum where the biofilm is attached and z is orthogonal to the substratum [47]. A CM constructs a 3D image by collecting multiple z-slices, each an $n_x \times n_y$ planar image at focal planes and depths chosen by the microscopist, at n_z discrete vertical locations (z) throughout the biofilm. Out-of-focus light is filtered out using a pinhole. The n_z z-slices are stacked together to form a z-stack. The number of x and y pixels is the same in each z-slice, $n_x = n_y$.

At each location (x, y, z) in the lattice, the CM records the intensity of the biofilm's fluorescence after excitation, I_{xyz}, as an 8-bit integer (i.e., an integer value between 0 and 255). It is convenient to index the lattice locations (x, y, z) by integers (i, j, k), in which case the intensities I_{xyz} are conveniently denoted by I_{ijk}. Each location in the lattice corresponds to a 3D pixel in the image, also called a voxel, with physical size $\Delta x \Delta y \Delta z$. Because microscopists often desire high spatial resolutions to achieve appealing visualizations, they typically choose fine pixelations, $n_x = n_y = 1024$ (e.g., Figure 1A), which sets the spatial resolution in the z-slices ($\Delta x = \Delta y$). When generating a video of images over time, however, the microscopist often lowers the spatial resolution, e.g. $n_x = n_y = 512$, to increase the temporal resolution (Δt) of the time lapse image frames in the video. In Figure 1, the physical depth of the vertical dimension (z) ranges from just a few microns to $200\ \mu m$. The microscopist sets the number (n_z) of z-slices over this range depending on the thickness of the biofilm which sets the spatial resolution in the vertical dimension (Δz). Finding specific temporal resolutions (Δt), pixelations (i.e., sizes of the lattice $n_x \times n_y \times n_z$), numbers of fields of view and numbers of biological replicates that give sufficient information about the biofilm while minimizing uncertainty have been studied by the authors [49,50].

The rest of this chapter is organized into five sections. Section 2 discusses the limitations of current image analysis tools. Section 3 describes the types of software tools that will be available from the BIL and the analyses they support. Section 4 details the types of data hosted by the BIL and describes how other institutions can share data in an interoperable way. Section 5 outlines the potential scientific impact of the

BIL as a community-wide resource. In Section 6, we conclude the chapter with a brief discussion of future directions for the BIL.

2. Limitations of Current Image Analysis Tools

Biofilm image analysis tools have lagged the advancement of technologically advanced image collection instrumentation. Existing tools, such as COMSTAT, CMEIAS, Daime, Beer's Analysis and BiofilmQ, are fundamentally limited in their ability to resolve and compare 3D biofilm structures and instead focus primarily on quantifying biomass using pixel-based (voxel-based) representations, without explicitly identifying individual cells. Consequently, these approaches fail to capture the spatial organization and structural relationships essential to biofilm function. Currently available analyses also lack the capacity for comparative analysis across datasets, limiting the discovery of shared architectural features and generalizable rules governing biofilm organization. Addressing these limitations requires a shift toward cell-resolved, 3D analysis frameworks capable of supporting cross-system comparisons. The following subsections provide more details of the current limitations of existing image analysis tools.

2.1. Bias in Biomass Estimation

The simplest question one may ask from 3D image data is *How much biomass or biovolume is there?* The answer provided by most existing image analysis approaches is to simply count bright pixels. Perhaps not surprisingly, counting bright pixels can be problematic.

<here> "Bright pixels" are those pixels with intensities at or above some threshold θ. A fundamental assumption when calculating biomass or biovolume with respect to a threshold is that pixels with intensities at or above θ are spatial locations associated with biomass, while other pixels are associated with background or "noise". Threshold selection often lacks robustness with a strong dependence on how specimens were prepared and digitized. Hence, to accurately estimate biomass or biovolume, it is crucial to determine the threshold in a biologically meaningful way. Very commonly (e.g., when using COMSTAT or BiofilmQ), Otsu's method [51] is applied to determine a threshold θ that minimizes the intra-class variability of two classes: biomass or background. As discussed below, Otsu's thresholding method appears to accurately identify biomass for thin biofilms but fails when the biomass is much thicker than $60\ \mu m$.

In addition to biovolume, conventional biofilm quantitation from CM images involves other global metrics calculated from pixel intensities. Metrics and their corresponding formulae are:

- average intensity, $\bar{I} = \frac{1}{n_x n_y n_z}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\sum_{k=1}^{n_z} I_{ijk}$ (see e.g. [18]); (1)
- biovolume = volume of bright pixels (pixels with intensities above a threshold θ),

  $V = \Delta x \Delta y \Delta z \sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\sum_{k=1}^{n_z} \delta_{ijk}(I_{ijk} \geq \theta)$; (2)

  where δ(·) is an indicator function;
- average thickness over the FOV is biovolume/FOV, $\bar{T} = \frac{V}{n_x n_y \Delta x \Delta y}$; (3)
- surface coverage (%), $C = \frac{1}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} \delta_{ij}\left(I_{ijk} \geq \theta \text{ for any } k\right) \times 100\%$. (4)

The quantities in Eqs. (1)-(4) are provided by publicly available image analysis software like COMSTAT, BiofilmQ and Beer's Analysis; or when using commercial software like Imaris [52].

For thick biofilms, it is known that the global metrics in Eqs.(1)-(3) can be biased due to limited laser penetration or attenuation deep into the biofilm due to Beer's Law [20, 53]. That is, in thick biofilms, the

intensities associated with biomass deep in the biofilm, typically at more than $60\ \mu m$ below the biofilm's surface, are lower than expected. This bias due to attenuation yields an artificially lower average intensity ($\bar{I}$), lower biovolume (V) and lower thickness estimates ($\bar{T}$). One approach for mitigating this bias for thick biofilms is to first estimate the top of the biofilm and then calculate biovolume and thickness by integrating under the top, which is performed by both COMSTAT [30,31] and Beer's Analysis [53].

This bias in biomass estimation was investigated [50] when assessing changes in biovolume of a very thick biofilm undergoing expansion and contraction when intermittently subjected to a saltwater treatment. A video of 2D representations of the 3D image data of the biofilm undergoing these events is available in the Supplementary Materials at [36]. Two different approaches were used to calculate biovolume from the video of 3D biofilm image data. Figure 4 shows biovolumes estimated from 40 frames of the video that capture expansion of the biofilm as a saltwater treatment is removed and then reapplied. The saltwater treatment was applied sometime between frames 31 and 32 that is associated with a large drop in volume between frames 31 and 33. The expansion and contraction of the biofilm was captured by first estimating the top surface of the biofilm and then integrating underneath. However, simply counting bright pixels severely underestimated the biovolume and failed to even detect the expansion or contraction events.

Figure 4 suggests that biofilm surface fitting then integrating the biovolume could be a panacea for calculating biovolumes for thick biofilms. However, this approach has at least two limitations: (1) every pixel under the biofilm's top surface contributes to the biovolume – i.e., any holes in the biofilm will bias the biovolume estimate upwards; and (2) the biofilm is assumed to be attached to a flat substratum, with the first z-slice captured just above the substratum. The second limitation may be mitigated by providing an estimate of the potentially non-linear surface of the substratum where the biofilm is attached. This may be provided when a color channel is used to image the substratum itself.

In summary, even for the simplest pixel-based global metrics (see Eqs. (1)-(4) above), the underlying assumption is that pixel intensities are relatively homogenous throughout the biofilm and can be aggregated together regardless of spatial location. This assumption breaks down in thick biofilms, resulting in biased estimates of the most ubiquitous biofilm characteristic ever: biovolume. Additional sources of bias when calculating global metrics are photobleaching (i.e., photon-induced chemical damage to fluorophores) and phototoxicity (light induced cellular damage). Even if a global metric like biovolume is estimated well, this still does not allow for the enumeration of individual cells in 3D, or resolve cell density, clustering, or spatial organization within the biofilm. We will investigate in Sections 3 and 5 how the BIL can address these limitations.

2.2. Absence of Uncertainty Quantification and Ad hoc Design of Imaging Studies

Quantitation of the global metrics in Section 2.1 from a single image is almost always reported as a single point estimate with no error bars or estimate of variability – i.e., with no uncertainty quantification. For example, the software COMSTAT, BiofilmQ and Imaris provide no measures of uncertainty when calculating biovolume, thickness or surface area of a single 3D CM image. A main driver for this omission is that it is not straightforward how to account for spatial correlation when calculating error bars. Put another way, intensities for pixels close together tend to be similar but vary independently for pixels that are far apart. Such spatial correlation tends to inflate the error bars associated with biovolume, thickness or surface area. There are few instances in the literature where uncertainty quantification is provided for image analysis outcomes. The error bars for biovolumes in Figure 4 were calculated by fitting a Laplacian model of 2D spatial correlation to the top surface of the biofilm separately at each time point [50]. Beer's analysis calculates biovolume error bars by fitting a regression model to all intensities in the 3D image data as a function of depth [53] and then fitting an AR(1) model of 2D spatial correlation to the top surface of the biofilm [19].

In addition to the factors that contribute to the variability in a single image (heterogeneity of biomass and the resulting pixel intensities; and the pixelation chosen (n_x, n_y, n_z)), there are other sources of variability that dominate even more. To mitigate, a microscopist must carefully choose the appropriate number of fields of view and the number of biological replicates when designing an imaging study [19,49]. In our experience, whether studying colony-forming units (CFU) or imaging outcomes, the amount of uncertainty also depends on the microbial species being studied and the presence of environmental stressors such as the antimicrobial efficacy of any applied treatments [49,53]. Unfortunately, unlike standardized tests that generate CFUs, the variability of imaging outcomes, even for simple pixel-based metrics as described in Section 2.1, are not well known.

Well-designed imaging studies (with adequate temporal resolution, pixelation, numbers of z-slices, FOVs and biological replicates) are essential for understanding the reproducibility of biofilm formation, organization and tolerance. These designs require a detailed knowledge of the sources of variability for different species and environments. The BIL will bridge this gap by harboring rich imaging datasets for different microbial species and environments, allowing researchers to estimate sources of variability and then design statistically powerful imaging studies to efficiently test hypotheses desperately needed by clinicians, antimicrobial formulators and medical device manufacturers. As additional datasets continue to be generated, they will be incorporated into the BIL.

2.3. Poor Understanding of Spatial Correlation of Multiple Species

Not uncommonly, microscopists collect image data from multiple species from the same biofilm sample. Ideally, the spatial location of each species is captured in a separate color channel. For example, in Figure 2, cyanobacteria are colored red and heterotrophic bacteria are colored green. In this example, both a red intensity and a green intensity are collected from each pixel location in the $n_x \times n_y \times n_z$ lattice. Given multiple color channel image data, a common question is *How do different species correlate in 3D space?*

This is not an easily answered question using any of the existing image analysis tools because they focus on a pixel-based analysis of each channel separately (Section 2.1). Pixel-based metrics are available that answer a different question: *How do different color channels correlate in 3D space?*. To calculate this correlation, each channel in an image is first thresholded and analyzed separately, then a nearest neighbor approach is applied [54] to estimate how often pixels in each colored channel are associated with similarly or differently colored pixels as a function of distance. Unfortunately, the spatial correlation of the pixels from multiple channels may not accurately describe how cells from different species correlate in space. For example, in a multiple-species model where the cells of one species are physically much larger than the other, there will tend to be more clustered pixels associated with the larger celled species than for the smaller celled species even if the clustering/aggregation patterns of the two species are the same. This highlights one goal of the BIL, which is to first identify where individual cells are using cell segmentation. After individual cells are identified, then it is possible to directly estimate how cells of different species correlate in space.

3. Machine-learning Tools Easily Accessible from the BIL

There is a need to move beyond the simple pixel-based biomass measures and provide quantitation that predicts biofilm adhesion, architecture, and functional behavior. A summary of the technical capabilities we envision for the BIL is shown in Figure 3.

3.1. Identification of Individual Microbial Cells

Answering the question *Where are the individual cells?* is referred to as cell segmentation. Cell segmentation from 3D biofilm images is a very challenging ML problem because it requires a two-step process: (1) low-level frame or pixel level classification, and (2) aggregations of these local decisions into complex shapes. To do this accurately, the algorithms must learn to emulate human judgements, which are based on sizes, shapes, density, micro-level and macro-level information. Unfortunately, adequate data to train and develop such models across a wide range of biofilms does not exist. This requires a curated data set with manual annotations or ML-assisted annotation, although it is argued by some that training by synthetic data is a viable alternative [55]. Further, rapidly evolving research in self-supervision, pre-training and vibe research are reshaping how we approach the process of annotating data [56].

Bacterial cell segmentation remains an active area of research, with existing approaches including BCM3D 2.0, StarDist OPP, DeepSeeded, and Cellpose–SAM. Figure 5 shows an example of raw image data and predictions made with an ML algorithm, StarDist OPP. This example was generated by adapting a standard StarDist OPP model using a small number of images ($N = 3$) that were partially annotated by human annotators. While the results are promising, the modified DICE score [43] for this small data set was in the 60% range, which indicates there is still significant room for improvement before these segmentations are ready for downstream quantification analysis.

Another active area of research is the identification and quantification of the heterogeneous morphologies of filamentous fungi, as depicted in Figure 6. Currently, few tools are available for fungal quantitation. One example, the Fungal Feature Tracker (FFT) [57], can comprehensively characterize mycelia, including hyphal segments, hyphal tips, total mycelial length, and total area covered by the mycelium. However, FFT requires high-contrast, low-noise images for accurate analysis.

Quantifying fungi within biofilms presents additional challenges because fungal cells often exist in multiple growth states within the same image. For example, images of *Candida albicans* grown in the presence of *Enterococcus faecalis* may contain yeast cells, chaining yeast, pseudohyphae, and hyphae. As growth progresses, hyphae can become large and entangled, forming dense mycelial mats.

To address these challenges, image analysis techniques are being developed to accurately classify yeast cells, yeast chains, pseudohyphae, hyphae, and hyphal mats while providing both counts and volumes. However, several obstacles remain before the process can be fully automated. Chains and pseudohyphae occasionally require manual reassignment between classification categories. When yeast cells and hyphae are so densely packed together that they are completely touching, the existing algorithms cannot reliably distinguish between them. ML approaches are currently being trained to recognize and remove yeast cells from these images to improve hyphal analysis. Similarly, when hyphae become highly entangled, total hyphal volume can be measured, but individual hyphae cannot be resolved. This limitation may ultimately be best addressed through ML-based segmentation and classification approaches.

3.2. Species Identification and Spatial Correlation of Multiple Species

Understanding multi-species and multi-domain biofilms has profound implications across every scientific domain from space travel to the human microbiota including endodontic, peritoneal, cystic fibrosis and chronic wound infections. In vitro biofilm studies typically utilize engineered reactor systems with well-characterized (often genetically tractable) single microbial species (e.g., *Escherichia coli, Candida albicans*, *Pseudomonas* spp., *Staphylococcus* spp., *Legionella).* The authors have developed and standardized such systems [58-61].

While much has been learned from single species studies, biofilms exist naturally as multi-species, multi-domain consortia, as shown in Figure 7. Hence, truly understanding mixed species and multi-domain biofilms (MDB) requires advanced analytical and analysis approaches. Simple natural systems such as subaerial biofilms and in vitro studies of mixed species and mixed domain biofilms are already being used to begin to understand the principles underlying these types of biofilms and their complex mixture of smaller interdependent subcommunities. The number and types of microbial interactions in multi-species biofilms grow exponentially with the addition of each new organism [62], making unraveling system interactions as large as the human microbiome challenging.

The structural and spatial organization of the communities may indicate their functions. For example, metabolic cooperativity may be recognized by closely associated interdigitated communities. Protection may be suggested by outside communities surrounding inside communities. For example, multi-domain biofilm's (e.g., bacteria with fungi) minimization of extracellular space promoting cellular associations that impact metabolism and cooperation may contribute to the survival of such films in (harsh) environmental conditions [15,16]. These patterns are observed in natural communities from subaerial biofilms on marble monuments (Figure 2) to dental and gastrointestinal tract biofilms.

The biggest challenges with determining the spatial co-location of multiple species in the same biofilm are not just accurate identification of the different species (Figure 2), but also accurate cell segmentation of each species. Traditional nonlinear techniques such as color balancing, automatic gain control, erosion, dilation, channel differencing, etc. do not result in robust image-independent algorithms. Robust methods are needed for 3D segmentation and comparison of community structural organizations between images as well as metabolic cooperativity as suggested by 3D Raman spectroscopy. For MDB involving bacteria and fungi, the quantitation of fungal hyphae (filaments of various lengths running through biofilms) presents an interesting problem in morphological analysis (Section 3.1). This remains a consistent computational challenge for all investigators examining fungi in any setting. By allowing cell segmentation training and fungal morphological training on large amounts of data, the BIL facilitates assessing how the cells of different species co-locate in 3D space.

3.3. Predictive Models of Biofilm Behavior

The BIL ultimately allows images to be transformed into mechanism-based predictive models by providing large, standardized datasets that enable ML and mathematical modeling to identify general principles governing biofilm formation, organization, function, and resilience. Additionally, if trained on reliable empirical data, ML can identify growth trajectories and infer the environmental variables that drive developmental transitions, enabling prediction of future community structure under changing conditions. The BIL is a collection of *observations* in the form of 3D images with associated metadata. ML is a powerful tool for *pattern discovery* in the organization of communities. These organizations can be used to gain a *mechanistic understanding* of specific community function and allow a comparison across communities to develop overarching biophysical principles that drive community function which can be mathematically modeled. Mathematical modeling can be used to explain and validate experimental and observational data, generate novel hypotheses for potential gaps in knowledge, and ultimately be used for *prediction* of community behavior. Ultimately, these models will be used to address multiple questions about biofilms: *How do they get there? How do they grow? How are microbiomes organized? How do factors such as physical surfaces, nutrients, temperature, atmosphere, and toxic exposures affect these structures?* And finally, *how do these microscopic observations scale to ecological niches and available surfaces?*

3.3.1. Data Collection: Why 3D?

Biofilms are inherently 3D structures, and understanding biofilm behavior spanning environmental, medical and industrial settings requires analytical techniques that can preserve the spatial relationships that govern microbial interactions and function. Traditional two-dimensional (2D) imaging captures a small fraction of biofilm architecture. Relying on 2D image analysis can obscure and miss important information about cell positioning, vertical stratification, pore structure, connectivity, and the spatial organization of subpopulations. For example, it has been demonstrated that cells that appear adjacent in a 2D projection may not actually be neighbors in 3D space [63]. As a result, critical processes that occur within the depth of the biofilm cannot be accurately measured or modeled from 2D observations alone.

Therefore, the gold-standard in biofilm science is 3D imaging, which preserves the spatial relationships that determine microbial function. Depending on biofilm thickness and cellular density, as well as the spatial resolution, penetration depth, and optical sectioning characteristics of the imaging modality employed (e.g., multiphoton versus point-scanning confocal microscopy), 3D imaging can provide quantitative, spatially resolved characterization of biofilm architecture. Such measurements can capture both the organization and spatial heterogeneity of cellular communities and the diffusion-reaction length scales that govern local microenvironmental conditions, including nutrient availability, oxygen transport, metabolite exchange, pH gradients, quorum-sensing signaling, antimicrobial penetration, reactive oxygen species dynamics, and waste-product accumulation. Collectively, these measurements provide a framework for relating biofilm structure and cellular organization to the physicochemical gradients and transport processes that regulate community function and survival.

Importantly biofilm models must also incorporate emergent dynamics and passive physical processes such as cell clustering, jamming, mechanical buckling, and the forces governing transitions from motile populations to mature 3D biofilms [64]. Together, these physical, chemical, and biological interactions shape 3D biofilm architecture, function, and resilience across multiple spatial and temporal scales.

3.3.2. Observation and Pattern Discovery: Towards Mechanistic Understanding

The first step towards discovering biofilm behavior is observation in the form of 3D images coupled with metadata about important growth conditions for the biofilms. For example, for laboratory grown biofilms, metadata is required on medium, temperature and flow. For environmental biofilms, metadata is required on sampling location, time of year, and measurements of environmental conditions such as temperature or humidity. ML inherently depends on data collection, high-quality annotation and informative metadata.

Once properly trained, ML is a powerful tool for recognizing patterns. Particularly important for accurate mathematical modeling and prediction of biofilm function is the spatial organization of the biofilm community in 3D. Once the individual cells are identified, the shape, size, spacing and organization of the community members can be discovered. The conserved patterns within a system can provide clues to their biological function and conserved patterns across biological systems suggest underlying mechanistic principles.

The next step is to understand mechanistically how 3D structural organization shapes biofilm function. Machine learning (ML) models, including large language models (LLMs) and AI-assisted coding tools such as Claude.ai [65], can help integrate data on the spatial and cellular organization of biofilms with mathematical models to generate testable hypotheses. In this approach, ML acts as a reasoning layer that links quantitative simulations with experimental observations.

For example, our group has generated imaging datasets capturing the spatial organization of biofilms, together with mathematical and agent-based simulation models [28, 29, 66]. Based on these data and

model predictions, we used ML to identify spatial and phenotypic patterns, relate these patterns to predicted mechanisms, and propose explanations for the emergence and spatial organization of antibiotic resistance.

This workflow becomes particularly powerful when many different cellular arrangements are available for analysis. Each biofilm can be described using quantitative features such as nearest-neighbor distances, cell density, cluster size, species adjacency, local nutrient concentrations, channel geometry, EPS distribution, and spatial gene expression. ML can identify relationships among these variables, while mechanistic models can test whether those relationships could plausibly arise from processes such as diffusion, growth, mechanics, signaling, or competition. ML can then integrate these results with existing biological knowledge to suggest mechanisms that can be tested experimentally.

Some open research questions that this workflow will address:

- *How do biofilms initially colonize surfaces?* The workflow will allow for identification of common attachment patterns informed by features such as surface charges and physical features both on synthetic surfaces as well as natural environments such as crevices in rocks and polygons in permafrost. After initial adherence patterns, early common structural changes associated with successful colonization can be investigated leading to understanding of the general and specific principles driving biofilm colonization.
- *What physicochemical factors influence biofilm growth and maturation?* Determination of how factors such as nutrient availability, fluid flow, surface properties, temperature, and humidity shape biofilm structure. For example, how can nutrient access be modeled based on diffusion reaction coupled with cell packing and EPS composition in 3D? In microbiomes, how does 3D structure influence organization of species specialized in use of primary nutrient sources vs. secreted volatile and non-volatile metabolic end-products from primary carbon and energy source users? How does flow create microenvironments at leading edges, around, and behind structures?
- *How are biofilms and subpopulations spatially organized?* Integration of 3D microscopy with complementary datasets such as Raman spectroscopy and spatial transcriptomics enables ML to identify recurring patterns of biofilm organization across structural, chemical, and genetic environments. By linking spatial architecture with local metabolic activity and gene expression, ML can uncover how distinct subpopulations form, interact, and specialize within biofilms. Identifying important, conserved properties such as structurally organized reaction diffusion rates coupled with metabolic rates allows for prediction of subcommunity and whole community organization and metabolic activity.
- *How do biofilms tolerate or resist stressors and toxic exposures?* Biofilms face multiple environmental challenges including antimicrobial exposure from other members of the microbiome or human disease treatment as well as survival in harsh environments such as Antarctic ice, desert sands, surface of marble monuments (Figure 8). Spatial organizations that harbor persister cells and cysts within 3D structures is a mechanism for survival of toxic assaults or temporary harsh environmental conditions [8]. Mathematical models in 3D can help establish other mechanisms by accounting for influences of heterogeneous cellular packing, reaction and absorption barriers, and even protection by resistant surrounding subcommunities. The formation of protected microenvironments could be particularly important under flow where protected microenvironments could form behind resistant structures.
- *How do microscale cellular interactions scale to ecosystem-level behaviors?* By integrating microscopy, mesoscale information (e.g., clinical imaging), and macroscale environmental conditions (e.g., photography or mathematical modeling), researchers can determine how local micro-interactions between cells and subcommunities give rise to larger-scale structures and emergent behaviors. For example, Figure 8 shows how 3D microscopic images of biofilm communities can be combined with mesoscale photographs of biofilms on national monument surfaces to enable ML to identify the effect of biofilms across scales. Another example is understanding how microscopic 3D images of biofilms on intravenous catheter surfaces [67] can be used to predict catheter related bloodstream infections, the deadliest hospital acquired infection in the US [68]. Mathematical models can be developed that bridge the scales allowing for prediction of activity beyond the microscale. These multiscale analyses will bridge the gap between microscopic observations and biofilm function at the scales relevant to human health, industry, and natural ecosystems.

This workflow integrates 3D biofilm microscopy, metadata, and ML to identify individual cells and their spatial organization, revealing conserved structural patterns linked to biological function. Coupled with physical principles (e.g., diffusion, reaction kinetics) and iterative experimentation, this ML-driven mathematical modeling approach generates testable hypotheses addressing key questions: how biofilms colonize surfaces, what physicochemical factors drive growth and maturation, how subpopulations are spatially organized (via integration with spectroscopy and transcriptomics), how biofilms resist stressors, and how microscale interactions scale to ecosystem- and health-relevant macroscale behaviors.

3.3.3. Predictive Biofilm Science: Open Research Questions

A major goal of the BIL is to enable the community to develop accurate predictive models. Continuous iterations of the mechanistic discovery process will produce the data needed for accurate prediction. When trained on large, high-quality datasets, coupled with accurate mathematical models, our understanding of biofilms has the potential to move beyond description toward prediction, enabling researchers to anticipate how biofilms will organize, function, and respond to change.

This pipeline has applications to complex microbiomes in the environment and human host as well as to single species and multi-domain biofilms, pathogenicity and biofouling. The BIL allows biological imaging, mathematics and ML to be used to accelerate discovery in these multiple complex arenas. Key open research questions include:

- How does the gastrointestinal microbiome produce enough metabolic byproducts to influence human health and cause disease when the microbiome becomes dysbiotic?
- What is the impact of biofilm communities on environmental cycles?
- How do microbiome communities react to their environments such as different seasons or human nutrient intake?
- How do biofilms look in human tissue?
- How to predict the distribution of biofilms in water systems and across entire catheters?
- How to predict the pathogenic capabilities of clinical isolates based on how their virulence factors influence 3D plate colony morphology?
- How to predict what combinations of treatments are effective on different types of biofilms causing industrial biofouling or human infections?

The foundation of ML is training on abundant, robust images that can only be provided by a community of scientists through a repository such as the BIL. Taken together, this framework establishes an iterative, multiscale approach for transforming 3D observations of biofilms into mechanistic models of their structure and function. Systematic acquisition of 3D imaging together with comprehensive environmental and experimental metadata, coupled with high-quality cell-level annotation, provides the foundation for ML methods to identify conserved patterns of cellular organization. These structural insights can then be integrated with physicochemical constraints, including diffusion, reaction kinetics, transport, and flow, to develop predictive mathematical models of biofilm growth, maturation, stress response, and community function. Ultimately, this approach provides a general framework for identifying both conserved and system-specific principles governing biofilm colonization, development, resilience, and ecological function across diverse biological and environmental contexts.

4. BIL Resources

The authors have been in the business of distributing data and ML software in various capacities since the early 1980s [70,71,72]. For data to be useful in research and technology development, it must be augmented with meaningful metadata and curated so that it remains relevant to current research and

algorithm development. This often means updating annotations as our understanding of the basic science evolves [73], and equally important, updating file storage standards to make the data easily interoperate with existing tools. A good example of this is our use of simple spreadsheet formats based on the open source csv standard [74] rather than complex database structures. Use of these kinds of simple standards makes it much easier for small research groups and people new to the field to get started. In this section, we briefly describe the types of resources available in the BIL.

4.1. Types of Resources

The BIL's conceptual framework, summarized in Figure 3, is designed to foster collaboration and synergy across the biofilm research community through three primary outcomes:

- *Image Data*: manually annotated curated 3D datasets with expert annotations that can be used to train sophisticated ML models for biofilm image analysis. Maintaining consistency across datasets is one of many important issues that must be addressed to facilitate studies across multiple datasets.
- *Software Tools*: tools to support statistical and geometric analysis of annotated images, as well as the ability to automatically annotate new images using powerful foundational models. Reproducibility is a huge challenge for modern AI-based tools, so our software will be designed and distributed in a manner that emphasizes validation and reproducibility.
- *Knowledge Discovery*: scientific insights generated through the analysis of the BIL's large and diverse inventory of biofilm images. Building a community-wide repository for knowledge relevant to our resources promotes new science and technology development. A generation of AI tools is emerging that can exploit such open source knowledge.

An overview of these outcomes is described briefly below.

4.1.1. Image Data

Although data itself is an important resource, for ML applications annotated data is the most important resource because it enables training and the development of powerful ML technology. We have successfully distributed such data for many decades. Most of our most popular corpora (i.e., a large, structured collection of data, including metadata and annotations) have undergone major revisions four or five times before the annotation standards have stabilized. Therefore, we cannot overestimate the importance of the curation aspects of the BIL.

A very common file format for microscopy image data is the Tagged Image File (TIF) format. 3D images are saved as multi-page (or multi-plane) TIFs, with each z-slice in the 3D z-stack saved as a page (see Section 1.4). The preferred format for the BIL is Open Microscopy Environment (OME) TIF files because certain image metadata are stored directly in the header of the TIF, most importantly being the physical dimensions of the field of view (i.e., the planar dimensions x and y), the image depth (in the vertical z dimension) and the number of channels. In our experience, all microscope manufacturers (e.g., Leica, Zeiss, Nikon, Olympus) by default store 3D images in a proprietary format but also allow saving image data as a TIF. Therefore, it has become a standard interchange format.

These image data are stored in an uncompressed format. A single file contains a 3D image typically consisting of 1024×1024 pixel slices with as many as 200 z-slices per image and at least two color channels (see Section 1.4). Though typical file sizes vary depending on the type of imaging done, a typical file size is about 250 Megabytes. To fit into the multi-page TIF format, the z-slices of the different color channels are either stacked as a block, or interwoven.

In addition to the raw image data, we provide relevant metadata via an Excel spreadsheet. Though the details of these spreadsheets vary with the corpus and imaging system, the general types of fields we attempt to capture include:

- Investigator
- Original data provider
- Geographic location
- Date of submission
- Species being imaged
- Stain or tag used
- Type of microscope used
- Field of view size (μm× μm) in planar dimensions *x* and *y*
- Image depth, *z* (μm)
- Pixelation of each z-slice
- Voxel (Pixel) Width, Δx=Δy (um)
- Voxel Height (Depth), Δz (um)

One example of a popular microscope used by the imaging community is manufactured by Leica Microsystems which saves image data in a proprietary Leica Image File (LIF) format [75]. Other microscope manufacturers use similar proprietary image file formats. Several open source toolkits, including ImageJ [46], support most of these common proprietary formats, as does the BIL. The file structure begins with a magic number followed by a UTF-16 XML header that describes the images and metadata, with raw pixel data stored in contiguous data blocks. Metadata can be extracted and viewed using several open source tools such as lifreader [76]. The metadata for the LIF files is extensive and contains information about the microscope used to collect the image:

Spatial & Dimensional Metadata

- X, Y, Z pixel dimensions (logical size)
- Physical size of each dimension (e.g., μm per axis)
- Physical origin coordinates
- Pixel/voxel size (the scale conversion factor, px/μm for X, Y, Z)
- Bit depth per channel (e.g., 8-bit, 12-bit, 16-bit)
- Number of channels and timepoints

Acquisition Settings

Confocal settings stored in the XML include scan mode (e.g., xyz), scan direction, scan speed, zoom level, pinhole size (in μm and Airy units), magnification, objective name, immersion medium, numerical aperture, and refractive index.

Stage & Instrument

Stage position metadata includes X, Y, and Z coordinates, the microscope model and software version.

Channel Information

Each channel stores LUT (lookup table/color) assignment, intensity range (min/max), emission wavelength, and detector settings. Laser intensity and excitation wavelength are also present, though there are known issues in some readers where per-channel laser metadata can be overwritten if the same key names are reused across channels.

Timestamps

Timestamps are recorded per frame (i.e., multiple frames or images can be stored within a LIF), including both relative time (in seconds) and absolute time with date.

After importing image data from an LIF, we split the data by color channels using Fiji in ImageJ and save each channel as a separate multi-page TIF using lossless compression.

Whether the data is imported as a multi-page TIF or as a LIF, the data may then be manually annotated, and the annotations are stored as bit masks in a TIF file. Once annotation is complete, the original image files and TIFs with annotations are converted to OME-TIFs for distribution.

Manual annotation is perhaps the most time-consuming part of the process. For annotation of 3D images, we have focused on using the open source tool napari [77]. Instead of annotating every cell, our strategy is to annotate roughly 100 representative cells from each image. Our goal is to provide the ML algorithms enough variety that they can learn high performance segmentation rules. Preliminary experiments indicated that there was minimal loss in performance by switching from annotating every cell to

annotating only a selected number. This results in a significant reduction in labor costs and allows us to annotate significantly more images. Similarly, we do not annotate every *z*-slice but instead annotate approximately every fifth *z*-slice equispaced throughout the 3D *z*-stack, capturing the most significant aspects of the geometry, and then we interpolate across these *z*-slices in the *z*-stack to construct a 3D annotation. A comparison of ML predictions after training on fully annotated images versus partially annotated images is shown in Figure 9.

4.1.2. Software Tools

In addition to 3D image data and metadata described in the previous section, the BIL provides models trained using various deep learning packages that can segment data at the cellular level. Currently, we are using StarDist OPP [38] for segmentation. The primary programming environment for our software is Python. Some legacy Matlab and Java software will be provided since these are still in use throughout the community. But we expect all our new software development to be done in Python. We currently rely on Python 3.11 or greater and use standard implementations of ML code developed in CUDA so that the software can transparently run on CPUs and GPUs. We distribute a requirements.txt file with each tool to make it easy to create a compatible Python environment.

Our initial software offering includes several Python tools to display image data in various ways, and to calculate analytics on these data. For example, we provide a tool that reads annotated images and computes several popular geometric measures that quantify the abundance and shape of cells. These programs are command line-driven and designed following standard principles to make it very easy to process large numbers of files. They can be downloaded from the BIL web site [78].

An important part of our software portfolio is tools to automatically segment images. This is a complex problem that is often not addressed well by modern deep learning technology [72]. We are currently leveraging tools provided in StarDist OPP to segment cells. Due to the fact that our images are partially annotated and that StarDist OPP is somewhat limited in the way it manages annotation labels, we preprocess manually segmented images by condensing all annotations for a given image into a single 2D file that contains a concatenation of each annotated region with a nominal amount of background image surrounding the annotated region. These concatenated images are then used as training data for StarDist's standard segmentation algorithm. We are in the early stages of hyperparameter tuning and optimization of this process, but thus far it is producing acceptable results as shown in Figure 9.

4.1.3. Knowledge Discovery

Even just a year ago, knowledge discovery was hampered by substantial resource requirements to develop algorithms for analysis of complex data sets. It was the case that there was a score of characteristics that a biofilm scientist wanted to estimate from a large data set collected from a complex system but only had access to algorithms to estimate a handful of those characteristics. In today's world where artificial intelligence platforms like Claude can easily generate algorithms and the associated software in negligible development time, biofilm researchers now have easy access to the required algorithms to estimate all desired characteristics. Algorithm development is now hindered by researchers' query refinement. The crucial limiting factor in this new world is access to relevant large, curated data sets generated by basic science. For the analysis of 3D image data of biofilms, this limiting factor is alleviated by the BIL. Data remains the engine to knowledge discovery.

Foundation models, pre-trained models, and self-supervised learning are among the AI technologies rapidly transforming how data are prepared, represented, and analyzed. In parallel, we are developing AI-based tools that provide natural-language, chatbot-style interfaces for interacting with and querying complex datasets. Retrieval-Augmented Generation (RAG) is an AI architecture that augments Large

Language Models (LLMs) with an external information-retrieval mechanism [79], enabling models to incorporate relevant information from domain-specific or continuously updated knowledge sources at inference time rather than relying solely on knowledge encoded in their parameters. Although these capabilities are still in the early stages of development, our goal is to provide a web-based interface through which users can upload an image and obtain an automated analysis of its contents. We anticipate making this service available through the BIL website in 2027.

4.2. How to Subscribe to BIL Resources?

Users must register to receive access to BIL resources. Registration is easy. The registration form, which is emailed after completion to *help@biofilmimagelibrary.org*, can be found at [80]. Users need to provide a minimal amount of contact information and sign a document describing the open source terms of use. Users receive an email response requesting they submit a public key. This key is added to our authorized keys file. Once this is completed, users can download data using an anonymous rsync command [81].

This method of distribution was adopted for two reasons. First, rsync is an incredibly useful tool because it only downloads updates to the data, which are identified in a variety of ways. It makes it very easy to keep users updated with the latest releases without requiring them to download the entire corpus. Similarly, if the download job fails at some point, which happens often with international users, the job can be easily restarted from where it left off.

Second, prior to our adoption of SSH keys as our primary distribution mechanism, it was difficult to prevent unauthorized redistribution of the data on sites such as Kaggle. The current approach, which places more accountability on users, prevents unauthorized sharing of passwords, etc. Such measures are required in today's complex world of open source data where provenance and data integrity are critical issues.

4.3. How to Submit Data to the BIL?

Users can initiate a request to share their data with the BIL and can contact *help@biofilmimagelibrary.org* to begin discussions about a dissemination process. Data is transferred via a similar rsync process as described in Section 4.2, or in the event the data is extremely large, by hard drive. Instructions can be found on the BIL resources page [80]. We generally prefer data that can be released using our standard data use agreements. Proprietary data or data requiring unique subscription models do not fit well with the goal of the BIL to be an open source repository.

5. BIL Outcomes and Deliverables

A key challenge in advancing biofilm research is the lack of standardized, accessible resources for managing and analyzing the increasingly diverse imaging data being generated. The BIL has started to address this gap by establishing common standards, curated datasets, and computational tools that enable reproducible analysis and facilitate the integration of biofilm imaging across research domains. These capabilities provide the foundation for both standardized data-driven research and transformative scientific discovery.

5.1. Standardization

Unlike omics data that have a set of defined data formats and public well-maintained repositories where they can be deposited and made available to the scientific community, there are no specific standards, formats, or public repositories that exist for biofilm images. While there are some popular open source

standards [82], there are no standards across corpora or tools for common file formats such as annotations, metadata, derived data, etc.

Ultimately the BIL will address this gap by providing:

- A curated database of biofilm images;
- A collection of annotated images for ML;
- Easy to access machine-learning analytical tools for microbiology;
- Newly developed biofilm image analysis tools in easy-to-use formats;
- Standard tools for annotation, storage and validation of data.

This will include prepacked toolboxes such as a technique for the analysis of biomaterial properties of biofilms and prepackaged programs for biofilm quantitation [34,83].

Another frequently underappreciated issue is the consistency of annotations across corpora. The BIL has developed standard conventions for how image data is annotated [43] and is applying these annotations to several corpora collected from our partnering institutions.

5.2. Transformative Science

Biofilms are ubiquitous in natural and industrial systems. The BIL significantly aids in translating fundamental findings about biofilms across medicine, industry, environment and even in space by providing a platform for the exchange of tools, information and images. The understanding gained from quantifying spatial organization and function of complex communities that live as a biofilm via ML and mathematical models will not merely be translational (from one field to another) but transformative (paradigm shifting). Examples include antimicrobial resistance and biofouling, device-associated infections, diagnostic technologies, cultural heritage preservation, and long-duration space exploration systems.

The BIL is uniquely positioned to serve as a repository of the diverse, high-quality datasets needed to develop, train, and benchmark these computational approaches. ML-driven analyses can also support multiscale modeling efforts that connect observations from laboratory biofilm systems to larger-scale applications, including medical devices, gastrointestinal environments, and built or cultural heritage structures. A current application of ML is the segmentation of individual cells within biofilms, a longstanding challenge that limits quantitative analyses. Section 3.1 describes our ongoing development of single-cell segmentation approaches trained on datasets spanning diverse cellular morphologies. These outputs can be integrated with established analysis platforms, such as BiofilmQ [84,85], to generate the quantitative measurements required for multiscale modeling.

As the BIL grows, developed models can be extended to increasingly complex and multi-species communities, ultimately providing robust, community-accessible tools and workflows for biofilm research. Beyond single-cell analyses, these approaches can be leveraged to identify higher-order patterns of community organization, including conserved spatial arrangements associated with surface colonization, biofilm maturation, and stress tolerance. For example, ML-driven analyses may reveal how biofilms establish protective microenvironments that shield sensitive populations from environmental or toxic stressors, enabling continued growth and coordinated protection across colonized surfaces.

5.3. Education

The BIL is also an educational resource. It houses a series of educational videos, providing students and educators with background on biofilms in natural, medical and industrial settings. It provides researchers with access to a database of biofilm images and analytical tools, enhancing learning in microbiology,

bioinformatics, and ML. The BIL's open-access platform promotes cross-disciplinary research, fostering collaboration among biologists, mathematicians, and engineers.

The combination of mathematical modeling and ML allows for rich laboratory experiences for cross-training students in biology, engineering and mathematics. For undergraduate and graduate students, it provides an environment for connecting biological experiments with tool development and applied mathematics. We have trained both undergraduate and graduate students from many different disciplines (including math and statistics) to perform benchtop biological and imaging experiments coupled with image analysis. The students work on various projects on the biology side, the math side, the ML side or all three. Examples include early biofilm colonization, species identification from MDBs, determining the impact of 3D structures on protection from antibiotics [28], and subcommunity organization in multi-species biofilms [65,69]. Additionally, students participate in interdisciplinary research discussions and gain hands-on experience at the interface of biology and data science, including opportunities that may not be available at institutions with limited biological research infrastructure, such as annotating microscopy datasets for ML model development.

By transforming imaging datasets into interactive learning experiences, the BIL can help students understand how 3D biofilm structures shape biological function. For example, in an exercise designed for a 3000 level microbiology laboratory, students work in teams and are provided with image datasets of antibiotic-treated biofilms showing both protective biofilm structures and protected microenvironments where bacterial growth continued despite treatment. Using the images as a guide, students identified the protected regions and filled them in with modeling clay on a corresponding 3D-printed protective biofilm structure. This hands-on activity illustrates how biofilm architecture creates localized protective niches and helps students visualize the relationship between 3D structure and antibiotic resistance. More broadly, the exercise demonstrates how imaging data can be transformed into quantitative and physical models, illustrating a central BIL concept: linking biofilm structure to biological function and, ultimately, to predictive modeling.

6. Conclusions

Advances in microscopy, image analysis, and machine learning are transforming the ability to quantitatively characterize the structure, function, and dynamics of complex microbial communities. As highlighted throughout this chapter, major challenges remain, including the accurate segmentation of cells in dense 3D datasets, characterization of multi-species and multi-domain communities, integration of heterogeneous imaging modalities, and development of reproducible workflows that enable meaningful comparison across studies. Addressing these challenges will require standardized datasets, open-access computational resources. The BIL satisfies these requirements, representing an important step toward advancing quantitative, reproducible, and scalable analysis of complex 3D biofilm imaging data. By combining curated microscopy datasets, standardized metadata, expert annotations, and accessible ML tools, the BIL will address critical limitations in current biofilm image analysis workflows. Accurately segmenting cells, characterizing multi-species and multi-domain communities, and quantifying spatial organization in 3D enables researchers to move beyond simple biomass measurements toward predictive models of biofilm behavior, function, and virulence. Equally important, the BIL establishes a collaborative framework that brings together microbiologists, engineers, mathematicians, and computer scientists to accelerate innovation in computational microbiology.

The long-term impact of the BIL extends well beyond the development of improved image analysis tools. By providing a centralized, community-driven infrastructure for biofilm imaging research, the BIL supports transformative advances in medicine, environmental science, industrial systems, and space exploration. ML models trained on standardized datasets have the potential to reveal conserved structural and functional patterns across diverse biofilm systems, enabling new approaches to understanding

antimicrobial resistance, microbial cooperation, and protective microenvironments. Through its commitment to open-access resources, education, and cross-disciplinary collaboration, the BIL establishes the computational foundations necessary for accurate predictive modeling and the next generation of biofilm discovery.

Acknowledgments

This collaboration is currently sponsored by the National Science Foundation (Award No. 2450549). Imaging completed within the Center for Biofilm Engineering Bioimaging Facility (RRID:SCR_026519) at Montana State University is supported by funding from industry members of the CBE, the National Science Foundation MRI Program (Award No. 2018562), the M. J. Murdock Charitable Trust (Award No. 202016116), the US Department of Defense (Award Nos. 77369LSRIP and W911NF1910288), and the National Institutes of Health (Award No. P20GM162003). Imaging completed within Temple University was supported by the National Science Foundation (Award No. DMS 2325170) and National Institutes of Health (Award No. R35HL155694). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We also wish to thank Dr. Darla Goeres and Lindsey Miller for sharing the image of E. coli in Figure 1B; Dr. Ghazal Vahidi for sharing the multi-domain biofilm image in Figure 7A; and Dr. Elizabeth Sandvik for sharing the multi-domain biofilm images in Figure 8.

References

1. Hall-Stoodley, L., Costerton, J. W., & Stoodley, P. (2004). Bacterial biofilms: From the Natural environment to infectious diseases. *Nature Reviews Microbiology*, 2(2), 95–108. doi: *10.1038/nrmicro821*.
2. Gross, R., Hauer, B., Otto, K., & Schmid, A. (2007). Microbial biofilms: New catalysts for maximizing productivity of long-term biotransformations. *Biotechnology and Bioengineering*, 98(6), 1123–1134. doi: *10.1002/bit.21547*.
3. Limoli, D. H., Jones, C. J., & Wozniak, D. J. (2015). Bacterial Extracellular Polysaccharides in Biofilm Formation and Function. *Microbiology Spectrum*, 3(3). doi: *10.1128/microbiolspec.MB-0011-2014*.
4. Sauer, K., Stoodley, P., Goeres, D. M., Hall-Stoodley, L., Burmølle, M., Stewart, P. S., & Bjarnsholt, T. (2022). The biofilm life cycle: Expanding the conceptual model of biofilm formation. *Nature Reviews Microbiology*, 20(10), 608–620. doi: *10.1038/s41579-022-00767-0*.
5. Flemming, H.-C., & Wuertz, S. (2019). Bacteria and archaea on Earth and their abundance in biofilms. *Nature Reviews Microbiology*, *17*(4), 247–260. doi: *10.1038/s41579-019-0158-9*.
6. Davies, D. (2003). Understanding biofilm resistance to antibacterial agents. *Nature Reviews Drug Discovery*, 2(2), 114–122. doi: *10.1038/nrd1008*.
7. Behnke, S., & Camper, A. K. (2012). Chlorine dioxide disinfection of single and dual species biofilms, detached biofilm and planktonic cells. *Biofouling*, 28(6), 635–647. doi: *10.1080/08927014.2012.700705*.
8. Stewart, P. S., & William Costerton, J. (2001). Antibiotic resistance of bacteria in biofilms. *The Lancet*, 358(9276), 135–138. doi: *10.1016/S0140-6736(01)05321-1*.
9. Costerton, J. W., Stewart, P. S., & Greenberg, E. P. (1999). Bacterial Biofilms: A Common Cause of Persistent Infections. *Science*, 284(5418), 1318–1322. doi: *10.1126/science.284.5418.1318*.
10. Cámara, M., Green, W., MacPhee, C. E., Rakowska, P. D., Raval, R., Richardson, M. C., Slater-Jefferies, J., Steventon, K., & Webb, J. S. (2022). Economic significance of biofilms: A multidisciplinary and cross-sectoral challenge. *Npj Biofilms and Microbiomes*, 8(1), 42. doi: *10.1038/s41522-022-00306-y*.
11. Zelaya, A. J., Parker, A. E., Bailey, K. L., Zhang, P., Van Nostrand, J., Ning, D., Elias, D. A., Zhou, J., Hazen, T. C., Arkin, A. P., & Fields, M. W. (2019). High spatiotemporal variability of bacterial diversity over short time scales with unique hydrochemical associations within a shallow aquifer. *Water Research*, 164, 114917. doi: *10.1016/j.watres.2019.114917*.
12. Allen, C. R., Stein, O. R., Hook, P. B., Burr, M. D., Parker, A. E., & Hafla, E. C. (2013). Temperature, plant species and residence time effects on nitrogen removal in model treatment wetlands. *Water Science and Technology*, 68(11), 2337–2343. doi: *10.2166/wst.2013.482*.
13. Mallette, N., Pankratz, E. M., Parker, A. E., Strobel, G. A., Busse, S. C., Carlson, R. P., & Peyton, B. M. (2014). Evaluation of Cellulose as a Substrate for Hydrocarbon Fuel Production by Ascocoryne sarcoides (NRRL 50072). *Journal of Sustainable Bioenergy Systems*, 4, 33–49. doi: *10.4236/jsbs.2014.41004*.
14. Davis, K. J., Lu, S., Barnhart, E. P., Parker, A. E., Fields, M. W., & Gerlach, R. (2018). Type and amount of organic amendments affect enhanced biogenic methane production from coal and microbial community structure. *Fuel*, 211, 600–608. doi: *10.1016/j.fuel.2017.09.074*.

15. Hooper, S. L., & Burstein, H. J. (2014). Minimization of extracellular space as a driving force in prokaryote association and the origin of eukaryotes. *Biology Direct*, 9(1), 24. doi: *10.1186/1745-6150-9-24*.

16. van Gestel, J., Vlamakis, H., & Kolter, R. (2015). Division of Labor in Biofilms: The Ecology of Cell Differentiation. *Microbiology Spectrum*, 3(2), MB-0002-2014. doi: *10.1128/microbiolspec.MB-0002-2014*.

17. Zhang, T., Parker, A., Carlson, R. P., Stewart, P. S., & Klapper, I. (2020). Multiscale Flux-Based Modeling of Biofilm Communities. *Multiscale Modeling & Simulation*, 18(2), 1025–1052. doi: *10.1137/18M1234096*.

18. Davison, W. M., Pitts, B., & Stewart, P. S. (2010). Spatial and temporal patterns of biocide action against Staphylococcus epidermidis biofilms. *Antimicrobial Agents and Chemotherapy*, 54(7), 2920–2927. doi: *10.1128/AAC.01734-09*.

19. Parker, A. E., Miller, L., Adams, J., Pettigrew, C., Buckingham-Meyer, K., Summers, J., Christen, A., & Goeres, D. (2022). Imaging and plate counting to quantify the effect of an antimicrobial: A case study of a photo-activated chlorine dioxide treatment. *Journal of Applied Microbiology*, 133(6), 3413–3423. doi: *10.1111/jam.15765*.

20. Pettygrove, B. A., Kratofil, R. M., Alhede, M., Jensen, P. Ø., Newton, M., Qvortrup, K., Pallister, K. B., Bjarnsholt, T., Kubes, P., Voyich, J. M., & Stewart, P. S. (2021). Delayed neutrophil recruitment allows nascent Staphylococcus aureus biofilm formation and immune evasion. *Biomaterials*, 275, 120775. doi: *10.1016/j.biomaterials.2021.120775*.

21. McLoon, A. L., Kolodkin-Gal, I., Rubinstein, S. M., Kolter, R., & Losick, R. (2011). Spatial regulation of histidine kinases governing biofilm formation in Bacillus subtilis. *Journal of Bacteriology*, 193(3), 679–685. doi: *10.1128/JB.01186-10*.

22. Hung, C., Zhou, Y., Pinkner, J. S., Dodson, K. W., Crowley, J. R., Heuser, J., Chapman, M. R., Hadjifrangiskou, M., Henderson, J. P., & Hultgren, S. J. (2013). Escherichia coli biofilms have an organized and complex extracellular matrix structure. *mBio*, 4(5), e00645-00613. doi: *10.1128/mBio.00645-13*.

23. Allesen-Holm, M., Barken, K. B., Yang, L., Klausen, M., Webb, J. S., Kjelleberg, S., Molin, S., Givskov, M., & Tolker-Nielsen, T. (2006). A characterization of DNA release in Pseudomonas aeruginosa cultures and biofilms. *Molecular Microbiology*, 59(4), 1114–1128. doi: *10.1111/j.1365-2958.2005.05008.x*.

24. Chen, M.-Y., Lee, D.-J., Tay, J.-H., & Show, K.-Y. (2007). Staining of extracellular polymeric substances and cells in bioaggregates. *Applied Microbiology and Biotechnology*, 75(2), 467–474. doi: *10.1007/s00253-006-0816-5*.

25. Baird, F. J., Wadsworth, M. P., & Hill, J. E. (2012). Evaluation and optimization of multiple fluorophore analysis of a Pseudomonas aeruginosa biofilm. *Journal of Microbiological Methods*, 90(3), 192–196. doi: *10.1016/j.mimet.2012.05.004*.

26. Gloag, E. S., Javed, M. A., Wang, H., Gee, M. L., Wade, S. A., Turnbull, L., & Whitchurch, C. B. (2013). Stigmergy: A key driver of self-organization in bacterial biofilms. *Communicative & Integrative Biology*, 6(6), e27331. doi: *10.4161/cib.27331*.

27. Amann, R. I., Binder, B. J., Olson, R. J., Chisholm, S. W., Devereux, R., & Stahl, D. A. (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and Environmental Microbiology*, 56(6), 1919–1925. doi: *10.1128/aem.56.6.1919-1925.1990*.

28. Shoraka, M., Jean-Baptiste, H., Buttaro, B., & Queisser, G. (2025). Using stochastic cellular automata to model and define sufficient conditions for the survival of Enterococcus faecalis biofilms with the pCF10 plasmid under erythromycin treatment. *PLOS Computational Biology*, 21(9), e1013425. doi: *10.1371/journal.pcbi.1013425*.

29. Jean-Baptiste, H., Henry Megonigal, B., Shoraka, M., Utter, B., Dornblaser, S., Kazi, S., Deep Beeram, V. R., Neishloss, L., Pasyar, A., Kunapuli, S., Voelz, V. A., Queisser, G., & Buttaro, B. A. (2026). pCF10 reprograms Enterococcus faecalis biofilms to promote nonattached aggregate biofilms, oxidative stress adaptation, and spatial antibiotic protection. *Npj Biofilms and Microbiomes*. doi: *10.1038/s41522-026-01121-5*.

30. Heydorn, A., Nielsen, A. T., Hentzer, M., Sternberg, C., Givskov, M., Ersbøll, B. K., & Molin, S. (2000). Quantification of biofilm structures by the novel computer program COMSTAT. *Microbiology* (Reading, England), 146(Pt 10), 2395–2407. doi: *10.1099/00221287-146-10-2395*.

31. Vorregaard, M. (2008). *Comstat2—A modern 3D image analysis environment for biofilms*. Technical University of Denmark. url: *https://www2.imm.dtu.dk/pubdb/edoc/imm5628.pdf*.

32. Liu, J., Dazzo, F. B., Glagoleva, O., Yu, B., & Jain, A. K. (2001). CMEIAS: A Computer-Aided System for the Image Analysis of Bacterial Morphotypes in Microbial Communities. *Microbial Ecology*, 41(3), 173–194. doi: *10.1007/s002480000004*.

33. Daims, H., Lücker, S., & Wagner, M. (2006). Daime, a novel image analysis program for microbial ecology and biofilm research. *Environmental Microbiology*, 8(2), 200–213. doi: *10.1111/j.1462-2920.2005.00880.x*.

34. Sprouffske, K., & Wagner, A. (2016). Growthcurver: An R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics*, 17(1), 172. doi: *10.1186/s12859-016-1016-7*.

35. Hartmann, R., Jeckel, H., Jelli, E., Singh, P. K., Vaidya, S., Bayer, M., Rode, D. K. H., Vidakovic, L., Díaz-Pascual, F., Fong, J. C. N., Dragoš, A., Lamprecht, O., Thöming, J. G., Netter, N., Häussler, S., Nadell, C. D., Sourjik, V., Kovács, Á. T., Yildiz, F. H., & Drescher, K. (2021). Quantitative image analysis of microbial communities with BiofilmQ. *Nature Microbiology*, 6(2), 151–156. doi: *10.1038/s41564-020-00817-4*.

36. Neu, T. R., & Kuhlicke, U. (2026). Biofilm Analysis by Confocal Microscopy—Basics and Practical Aspects. *Microscopy Research and Technique*, 89(2), 183–206. doi: *10.1002/jemt.70064*.

37. Zhang, J., Wang, Y., Donarski, E. D., Toma, T. T., Miles, M. T., Acton, S. T., & Gahlmann, A. (2022). BCM3D 2.0: Accurate segmentation of single bacterial cells in dense biofilms using computationally generated intermediate image representations. *Npj Biofilms and Microbiomes*, 8(1), 99. doi: *10.1038/s41522-022-00362-4*.

38. Jelli, E., Ohmura, T., Netter, N., Abt, M., Jiménez-Siebert, E., Neuhaus, K., Rode, D. K. H., Nadell, C. D., & Drescher, K. (2023). Single-cell segmentation in bacterial biofilms with an optimized deep learning method enables tracking of cell lineages and measurements of growth rates. *Molecular Microbiology*, 119(6), 659–676. doi: *10.1111/mmi.15064*.

39. Toma, T. T., Wang, Y., Gahlmann, A., & Acton, S. T. (2024). DeepSeeded: Volumetric segmentation of dense cell populations with a cascade of deep neural networks in bacterial biofilm applications. *Expert Systems with Applications*, 238, 122094. doi: *10.1016/j.eswa.2023.122094*.

40. Pachitariu, M., Rariden, M., & Stringer, C. (2025). Cellpose-SAM: superhuman generalization for cellular segmentation. *bioRxiv*, 2025.04.28.651001. doi: *10.1101/2025.04.28.651001*.

41. Mazarei, A., Wang, B., Parker, A., Goeres, D., Boas, J., Azevedo, N., & Cardoso, J. (2026). Biofilm3D-Bench: Uncovering trade-offs in 3D biofilm segmentation across classical and deep learning approaches. Bioinformatics. (in review).

42. Picone, J., McKenzie, S., & Desai, M. (2022). Unencumbered Electroencephalogram (EEG) Corpora: Advancing Technology Using Big Data Resources. International Neuroinformatics Coordinating Facility Neuromatics Conference (INCF), 11. url: *https://isip.piconepress.com/publications/presentations_invited/2022/incf/tuh_eeg/*.

43. Bagritsevich, M., Hackel, D., Obeid, I., & Picone, J. (2024). Annotation of the Fox Chase Cancer Center Digital Pathology Corpus. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium, 1–4. doi: *10.1109/SPMB62441.2024.10842255*.

44. Jeckel, H., Díaz-Pascual, F., Skinner, D. J., Song, B., Jiménez-Siebert, E., Strenger, K., Jelli, E., Vaidya, S., Dunkel, J., & Drescher, K. (2022). Shared biophysical mechanisms determine early biofilm architecture development across different bacterial species. *PLoS Biology*, 20(10), e3001846. doi: *10.1371/journal.pbio.3001846*.

45. Ollion, J., Cochennec, J., Loll, F., Escudé, C., & Boudier, T. (2013). TANGO: a generic tool for high-throughput 3D image analysis for studying nuclear organization. *Bioinformatics* (Oxford, England), 29(14), 1840–1841. doi: *10.1093/bioinformatics/btt276*.

46. Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7), 671–675. doi: *10.1038/nmeth.2089*.

47. Constans, A. (2004, November 21). The Confocal Microscope. *The Scientist*. url: *https://www.the-scientist.com/the-confocal-microscope-49314*.

48. Sharp, J. L., Parker, A. E., & Hamilton, M. A. (2023). Calculating the limit of detection for a dilution series. *Journal of Microbiological Methods*, *208*, 106723. doi: *10.1016/j.mimet.2023.106723*.

49. Pettygrove, B. A., Smith, H. J., Pallister, K. B., Voyich, J. M., Stewart, P. S., & Parker, A. E. (2021). Experimental Designs to Study the Aggregation and Colonization of Biofilms by Video Microscopy With Statistical Confidence. *Frontiers in Microbiology*, *12*, 785182. doi: *10.3389/fmicb.2021.785182*.

50. Parker, A. E., Pitts, B., Lorenz, L., & Stewart, P. S. (2018). Polynomial accelerated solutions to a LARGE Gaussian model for imaging biofilms: In theory and finite precision. *Journal of the American Statistical Association*, *113*(524), 1431–1442. doi: *10.1080/01621459.2017.1409121*.

51. Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66. doi: *10.1109/TSMC.1979.4310076*.

52. Rahman, M. T., Khan, N. A., Razu, I., Mokbul, M. I., Fatima, S. M., Garcia, C. L., Eckard, P., & Hansen, M. R. (2025). Protocol for supervised automation of cell counting in confocal microscopic mice cochlear imaging datasets using macro in Imaris. *STAR Protocols*, *6*(2), 103815. doi: *10.1016/j.xpro.2025.103815*.

53. Parker, A. E., Christen, J. A., Lorenz, L., & Smith, H. (2020). Optimal surface estimation and thresholding of confocal microscope images of biofilms using Beer's Law. *Journal of Microbiological Methods*, 174, 105943. doi: *10.1016/j.mimet.2020.105943*. Analyzed video in Supplementary Materials at *www.tandfonline.com/doi/suppl/10.1080/01621459.2017.1409121?scroll=top*.

54. Liu, W., Russel, J., Burmølle, M., Sørensen, S. J., & Madsen, J. S. (2018). Micro-scale intermixing: A requisite for stable and synergistic co-establishment in a four-species biofilm. *The ISME Journal*, 12(8), 1940–1951. doi: *https://doi.org/10.1038/s41396-018-0112-2*.

55. Hickl, V., Khan, A., Rossi, R. M., Silva, B. F. B., & Maniura-Weber, K. (2025). Segmentation of dense and multi-species bacterial colonies using models trained on synthetic microscopy images. *PLOS Computational Biology*, 21(4), e1012874. doi: *10.1371/journal.pcbi.1012874*.

56. Picone, J., Buttaro, B., Parker, A., & Smith, H. (2026, July 15). Machine Learning Challenges in Biofilm Engineering. url: *https://isip.piconepress.com/publications/presentations_invited/2026/mbm/bil/*.

57. Vidal-Diez de Ulzurrun, G., Huang, T.-Y., Chang, C.-W., Lin, H.-C., & Hsueh, Y.-P. (2019). Fungal feature tracker (FFT): A tool for quantitatively characterizing the morphology and growth of filamentous fungi. *PLOS Computational Biology*, 15(10), e1007428. doi: *10.1371/journal.pcbi.1007428*.

58. Goeres, D. M., Walker, D. K., Buckingham-Meyer, K., Lorenz, L., Summers, J., Fritz, B., Goveia, D., Dickerman, G., Schultz, J., & Parker, A. E. (2019). Development, standardization, and validation of a biofilm efficacy test: The single tube method. *Journal of Microbiological Methods*, 165, 105694. doi: *10.1016/j.mimet.2019.105694*.

59. Goeres, D. M., Parker, A. E., Walker, D. K., Meier, K., Lorenz, L. A., & Buckingham-Meyer, K. (2020). Drip flow reactor method exhibits excellent reproducibility based on a 10-laboratory collaborative study. *Journal of Microbiological Methods*, 174, 105963. doi: *10.1016/j.mimet.2020.105963*.

60. Parker, A. E., Walker, D. K., Goeres, D. M., Allan, N., Olson, M. E., & Omar, A. (2014). Ruggedness and reproducibility of the MBEC biofilm disinfectant efficacy test. *Journal of Microbiological Methods*, 102, 55–64. doi: *10.1016/j.mimet.2014.04.013*.

61. Bodle, K. B., Vahidi, G., Johnson, E., Walker, D., Parker, A., Jones, C., Goeres, D. M., Peyton, B. M., & Smith, H. J. (2026). Assessing a method for single- and multidomain biofilm growth in a novel biofilm reactor: Shear stress, ruggedness, repeatability, and reproducibility. *Biofilm*, 11, 100372. doi: *10.1016/j.bioflm.2026.100372*.

62. Røder, H. L., Sørensen, S. J., & Burmølle, M. (2016). Studying Bacterial Multispecies Biofilms: Where to Start? *Trends in Microbiology*, 24(6), 503–513. doi: *10.1016/j.tim.2016.02.019*.

63. Røder, H. L., Olsen, N. M. C., Whiteley, M., & Burmølle, M. (2020). Unravelling interspecies interactions across heterogeneities in complex biofilm communities. *Environmental Microbiology*, *22*(1), 5–16. doi: *10.1111/1462-2920.14834*.

64. Wong, L. L., Mugunthan, S., Kundukad, B., Ho, J. C. S., Rice, S. A., Hinks, J., Seviour, T., Parikh, A. N., & Kjelleberg, S. (2023). Microbial biofilms are shaped by the constant dialogue between biological and physical forces in the extracellular matrix. *Environmental Microbiology*, *25*(1), 199–208. doi: *10.1111/1462-2920.16306*.

65. *Claude (Sonnet/Opus version)*. (2026). [Computer software]. url: *https://claude.ai*.

66. Tenore, A., Wu, Y., Jacob, J., Bittermann, D., Villa, F., Buttaro, B., & Klapper, I. (2023). Water activity in subaerial microbial biofilms on stone monuments. *Science of The Total Environment*, 900, 165790. doi: *10.1016/j.scitotenv.2023.165790*.

67. Ryder, M., deLancey-Pulcini, E., Parker, A. E., & James, G. A. (2023). Bacterial transfer and biofilm formation in needleless connectors in a clinically simulated in vitro catheter model. *Infection Control & Hospital Epidemiology*, 44(11), 1760–1768. Cambridge Core. doi: *10.1017/ice.2023.60*.

68. Ryder, M., & Battle, J. (2025). Choice of needleless connector technology as a risk reduction strategy for catheter related bloodstream infection, mortality, and cost: A secondary data analysis. *The Journal of Vascular Access*, 26(4), 1194–1205. doi: *10.1177/11297298241261951*.

69. Tenore, A., Russo, F., Jacob, J., Grattepanche, J. D., Buttaro, B., & Klapper, I. (2024). A Mathematical Model of Diel Activity and Long Time Survival in Phototrophic Mixed-Species Subaerial Biofilms. *Bulletin of Mathematical Biology*, 86(10), 123. doi: *10.1007/s11538-024-01348-3*.

70. Doddington, G. R., & Schalk, T. B. (1981). Computers: Speech recognition: Turning theory to practice: New ICs have brought the requisite computer power to speech technology; an evaluation of equipment shows where it stands today. *IEEE Spectrum*, 18, 26–32. doi: *10.1109/MSPEC.1981.6369809*.

71. Melles, A.-M., Paderewski, M., Oymann, R., Shah, V., Salazar, J., Obeid, I., & Picone, J. (2024). Annotation of Ambulatory EEGs. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 1–4. doi: *10.1109/SPMB62441.2024.10842264*.

72. Hackel, D., Bagritsevich, M., Dumitrescu, C., Al Mamun, Md. A., Purba, S. A., Heathcote, D., Obeid, I., & Picone, J. (2026). Enabling Microsegmentation: Digital Pathology Corpora for Advanced Model Development. In *Signal Processing in Medicine and Biology: Applications of Artificial Intelligence in Medicine and Biology* (Vol. 1, pp. 191-232). Springer. doi: *10.1007/978-3-032-18509-9_6*.

73. Shalamzari, S. S., Bagritsevich, M., Melles, Anne-Mai, Obeid, I., Picone, J., Connolly, D., Wu, C., Brown, B., James, J., Gong, Y., & Wu, H. (2023). Big Data Resources for Digital Pathology. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 1–19. doi: *10.1109/SPMB59478.2023.10372721*.

74. Python Software Foundation. (May 28, 2026). *csv — CSV file reading and writing*. Python 3 documentation. Retrieved May 28, 2026. url: *https://docs.python.org/3/library/csv.html*.

75. Open Microscopy Environment. (2026, June 1). *Leica LAS AF LIF (Leica Image File Format)*. Bio-Formats 6.11.0 documentation. Bio-Formats. url: *https://docs.openmicroscopy.org/bio-formats/6.11.0/formats/leica-lif.html*.

76. Negretti, N., & Arcadia Science. (2026, June 1). *readlif* (Version 0.6.6) [Computer software]. PyPI. url: *https://pypi.org/project/readlif/0.6.6/*.

77. Sofroniew, N., Lambert, T., Bokota, G., Nunez-Iglesias, J., Sobolewski, P., Sweet, A., Gaifas, L., Evans, K., Burt, A., Doncila Pop, D., Yamauchi, K., Weber Mendonça, M., Rodríguez-Guerra, J., Liu, L., Buckley, G., Vierdag, W.-M., Anderson, A., Monko, T., Willing, C., … Yadav, A. S. (2026). napari: A multi-dimensional image viewer for Python (Version v0.7.1a6) [Computer software]. *Zenodo*. doi: *10.5281/zenodo.20080448*.

78. Byrchak, D., Roxas, E., & Picone, J. (2026, June 1). Biofilm Image Analytics. *Institute for Signal and Information Processing*. url: *http://www.isip.piconepress.com/projects/biofilm_image_library/data/software*.

79. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*. doi: 10.5555/3495724.3496517

80. Parker, A, Buttaro, B., and Picone, J. (2026, June 1). Biofilm Image Library Resources. url: *https://www.biofilmimagelibrary.org/html/resources.shtml*.

81. Ylonen, T., & Lonvick, C. (2006). *The Secure Shell (SSH) authentication protocol* (RFC 4252). Internet Engineering Task Force. doi: *10.17487/RFC4252*.

82. Open Microscopy Environment. (2026, June 1). *The Open Microscopy Environment*. url: *https://www.openmicroscopy.org*.

83. Buttaro, B., & Queisser, G. (2021). A Bead Movement Based Computational Framework for 3-Dimensional Analysis of Biofilm Material Heterogeneity. *JoVE*, (172), e62454. doi: *10.3791/62454*.

84. Hartmann, R., Jeckel, H., Jelli, E., Singh, P. K., Vaidya, S., Bayer, M., Rode, D. K. H., Vidakovic, L., Díaz-Pascual, F., Fong, J. C. N., Dragoš, A., Lamprecht, O., Thöming, J. G., Netter, N., Häussler, S., Nadell, C. D., Sourjik, V., Kovács, Á. T., Yildiz, F. H., & Drescher, K. (2021). Quantitative image analysis of microbial communities with BiofilmQ. *Nature Microbiology*, 6(2), 151–156. doi: *10.1038/s41564-020-00817-4*.

85. Hartmann, R., Jeckel, H., Jelli, E., Singh, P. K., Vaidya, S., Bayer, M., Rode, D. K. H., Vidakovic, L., Díaz-Pascual, F., Fong, J. C. N., Dragoš, A., Lamprecht, O., Thöming, J. G., Netter, N., Häussler, S., Nadell, C. D., Sourjik, V., Kovács, Á. T., Yildiz, F. H., & Drescher, K. (2021). Publisher Correction: Quantitative image analysis of microbial communities with BiofilmQ. *Nature Microbiology*, 6(2), 270–270. doi: *10.1038/s41564-021-00863-6*.

List of Figures

Figure 1. 3D confocal scanning laser microscope images demonstrating the heterogeneous structures of biofilms. Even for single species biofilms, including *Pseudomonas aeruginosa* (A,C) or *E. coli* (B), environmental conditions determine structure: (A) in high shear environments, biofilms grow in high density clusters; (B) antimicrobial chemical treatments fail to entirely remove or kill biofilm; (C) in low shear environments, biofilms are less dense. 28

Figure 2. A demonstration of 3D segmentation using conventional tools: (A) CM image of multi-species biofilm composed of (B) Cyanobacteria (red) and (C) the different subcommunities of Cyanobacteria, (D) heterotrophs green small cells, (E) with surrounding green eDNA (extracellular DNA) with holes where the cells were subtracted from the image. 29

Figure 3. A conceptual framework for the Biofilm Imaging Library (BIL), illustrating the integration of community-contributed data, analytical services, and computational resources to enable knowledge discovery, predictive modeling, and scientific impact across diverse application domains. 30

Figure 4. Biofilm volume estimates for 40 frames (about 10 min) of a video of 3D image data are indicated by the solid line. The estimates were found by first finding the top surface of the biofilm and then integrating underneath. Error bars indicate 99% probability intervals. Volume is severely underestimated by an estimator that only counts bright pixels as indicated by the dash-dotted line. 31

Figure 5. Example of 3D image data from Cyanobacteria biofilms as shown in Figure 2 with predictions of individual cyanobacteria cell locations by an ML algorithm. 32

Figure 6. Quantifying fungi within biofilms presents additional challenges: (A) an example of multi-domain biofilm consisting of *Aspergillus aculeatus* and *Pseudomonas aeruginosa* (green = all bacteria, red = membrane compromised bacterial cells, and blue = fungal hyphae); (B) an example of a fungal quantitation tool under development. 33

Figure 7. Images of two MDB from NASA-funded research at the CBE: (A) shows the persistence of an MDB following 252 days of dormancy following a water flush (yellow = all bacteria, magenta = membrane compromised bacterial cells, and cyan = fungal cells); (B) shows the complex 3D structure of fungal hyphae and bacterial associations (green = all bacteria, red = membrane compromised bacterial cells, and blue = fungal hyphae). 34

Figure 8. An example of a multiscale modeling project being undertaken by the authors to estimate the size of Cyanobacteria and heterotroph communities on the roof of Federal Hall in New York City: (A) high magnification image of the community on a particle; (B) lower magnification image of the communities on particles; (C) blackened regions associated with biofilm presence; (D) size of the scraped area for imaged particles; and (E) the roof of Federal Hall with darkened areas containing biofilms. 35

Figure 9. An example of the annotation process. The original image appears on the upper left. A manually annotated image appears on the upper right. An automatically annotated image appears on the lower left that uses a model trained on fully annotated data. A similar automatically annotated image appears on the lower right which is the result of a model trained on partially annotated images. 36

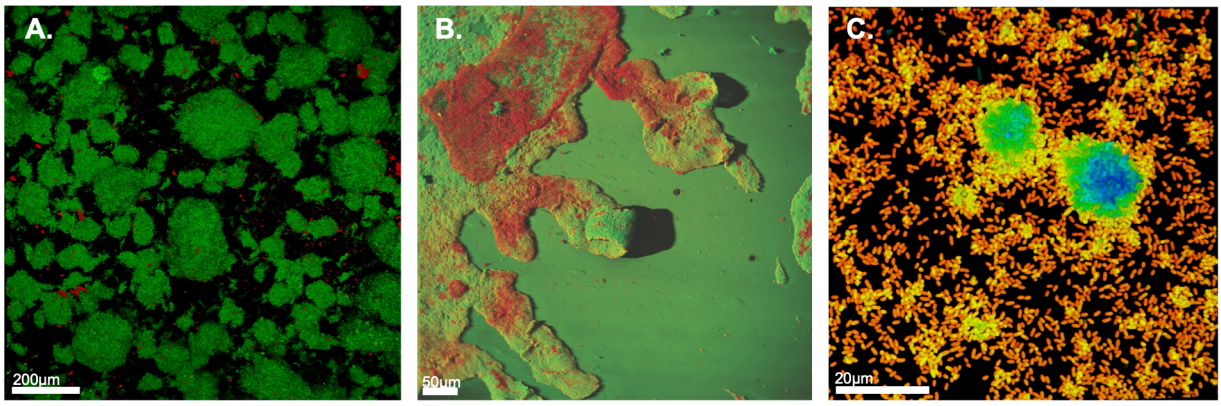
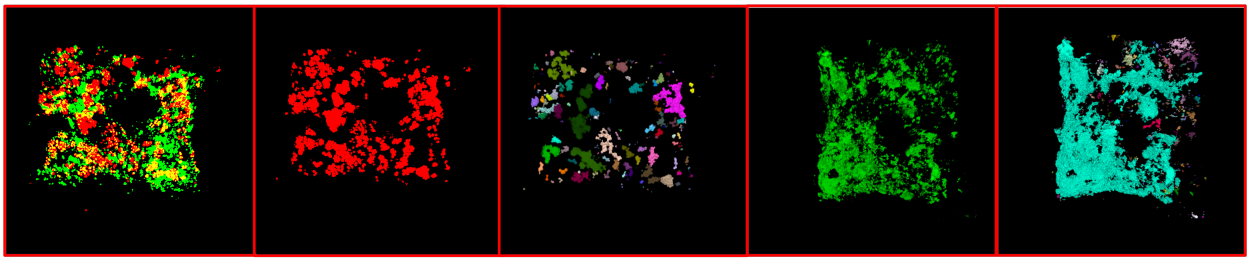


Figure 1. 3D confocal scanning laser microscope images demonstrating the heterogeneous structures of biofilms. Even for single species biofilms, including *Pseudomonas aeruginosa* (A,C) or *E. coli* (B), environmental conditions determine structure: (A) in high shear environments, biofilms grow in high density clusters; (B) antimicrobial chemical treatments fail to entirely remove or kill biofilm; (C) in low shear environments, biofilms are less dense.



(A) (B) (C) (D) (E)

Figure 2. A demonstration of 3D segmentation using conventional tools: (A) CM image of multi-species biofilm composed of (B) Cyanobacteria (red) and (C) the different subcommunities of Cyanobacteria, (D) heterotrophs green small cells, (E) with surrounding green eDNA (extracellular DNA) with holes where the cells were subtracted from the image.

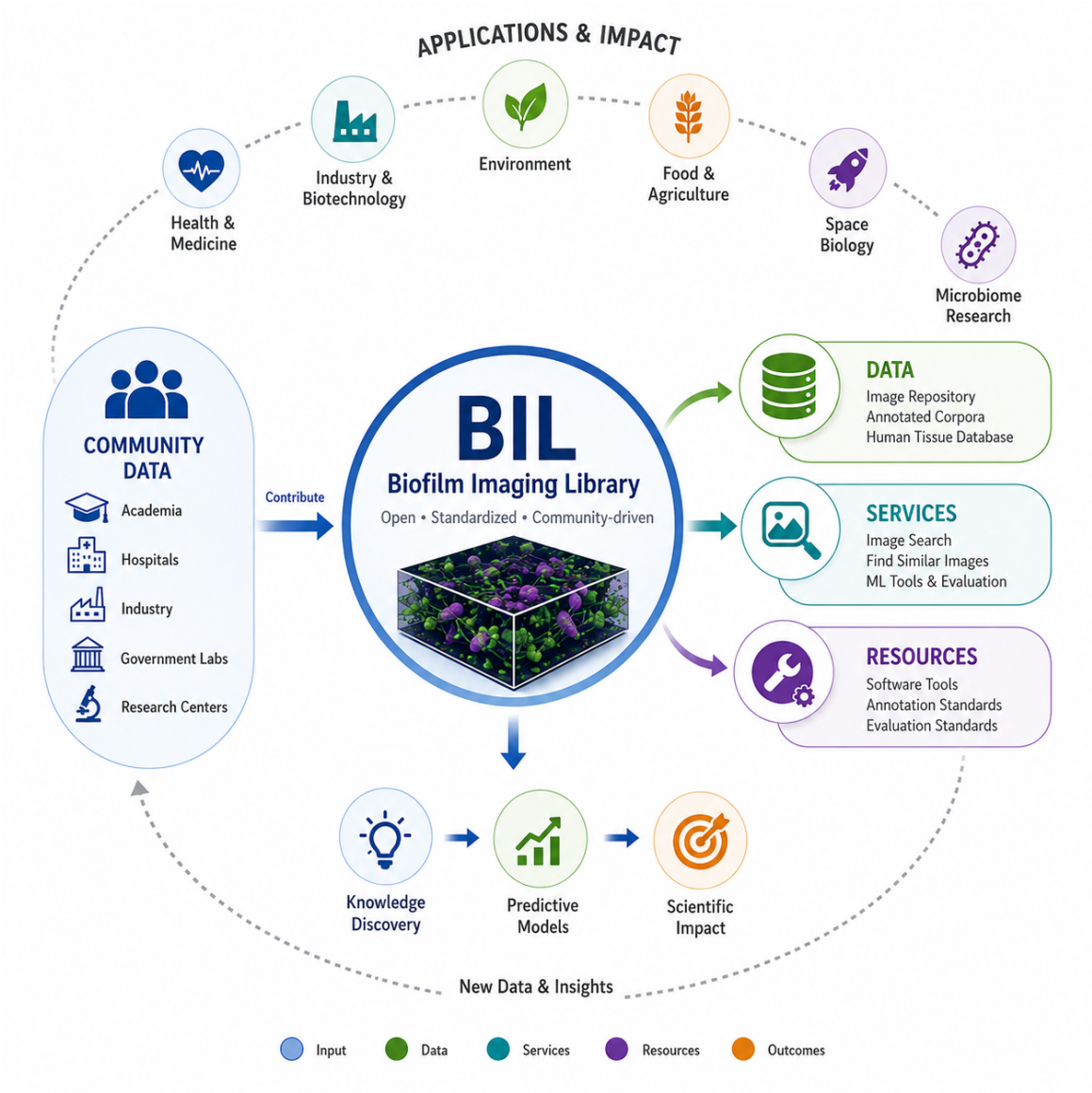


Figure 3. A conceptual framework for the Biofilm Imaging Library (BIL), illustrating the integration of community-contributed data, analytical services, and computational resources to enable knowledge discovery, predictive modeling, and scientific impact across diverse application domains.

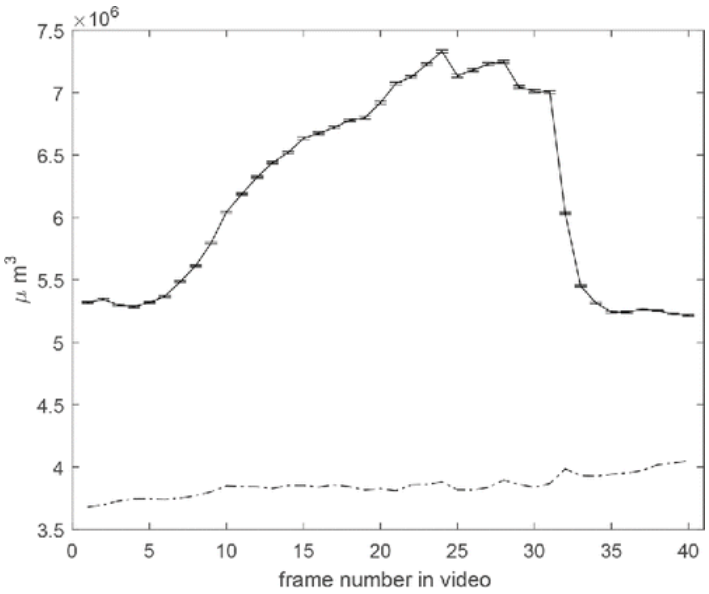


Figure 4. Biofilm volume estimates for 40 frames (about 10 min) of a video of 3D image data are indicated by the solid line. The estimates were found by first finding the top surface of the biofilm and then integrating underneath. Error bars indicate 99% probability intervals. Volume is severely underestimated by an estimator that only counts bright pixels as indicated by the dash-dotted line.

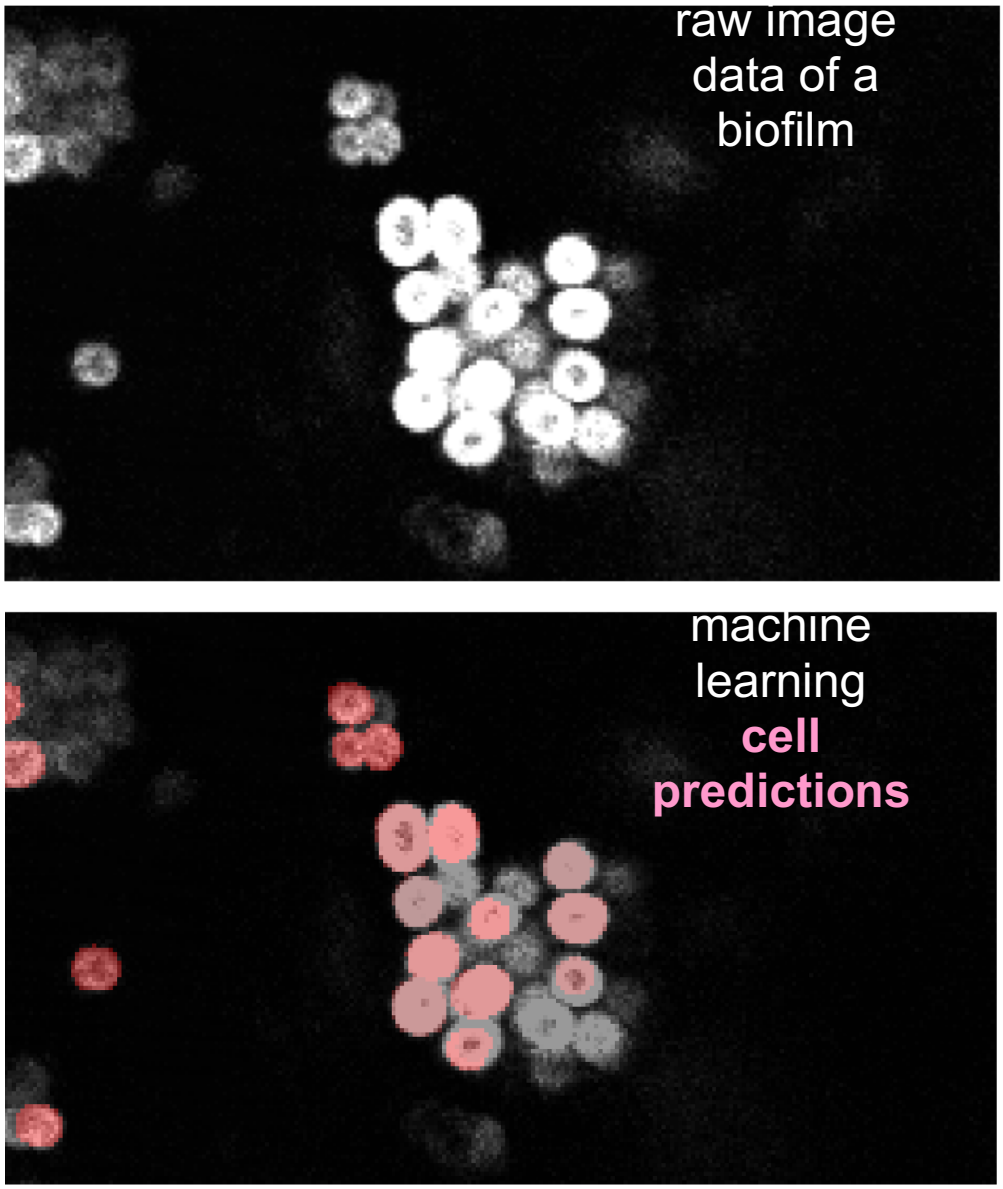
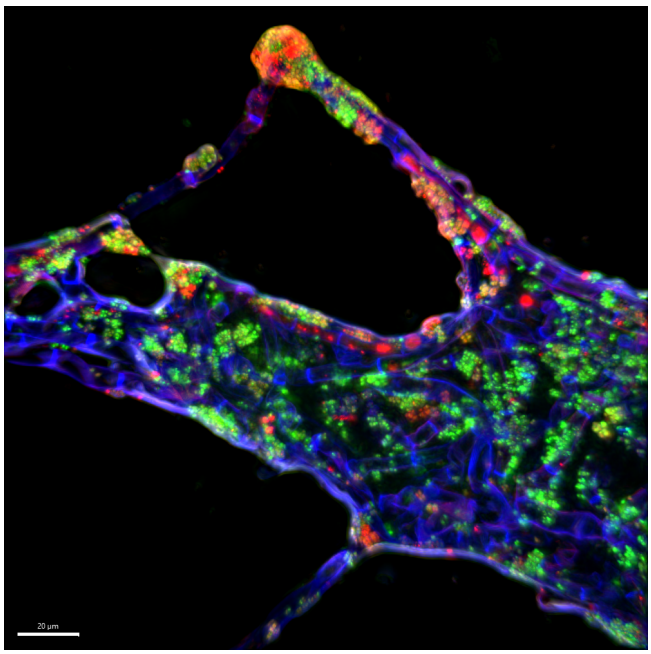
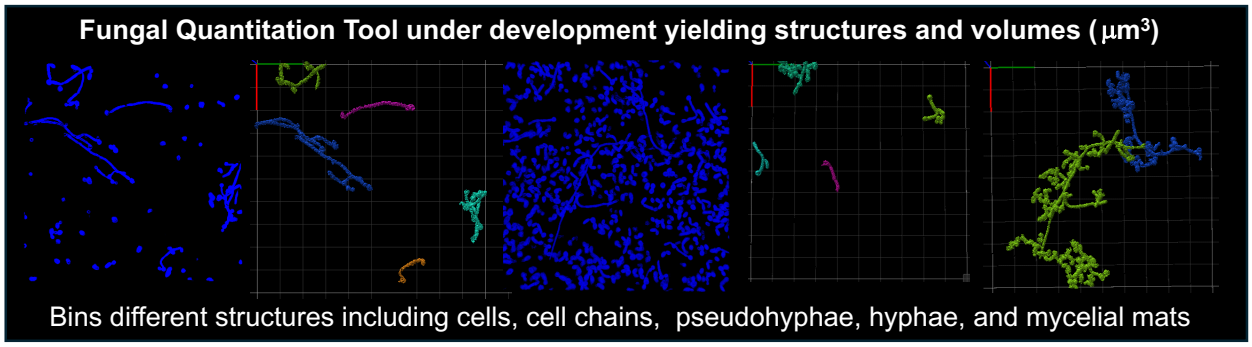


Figure 5. Example of 3D image data from Cyanobacteria biofilms as shown in Figure 2 with predictions of individual cyanobacteria cell locations by an ML algorithm.



(A)



(B)

Figure 6. Quantifying fungi within biofilms presents additional challenges: (A) an example of multi-domain biofilm consisting of *Aspergillus aculeatus* and *Pseudomonas aeruginosa* (green = all bacteria, red = membrane compromised bacterial cells, and blue = fungal hyphae); (B) an example of a fungal quantitation tool under development.

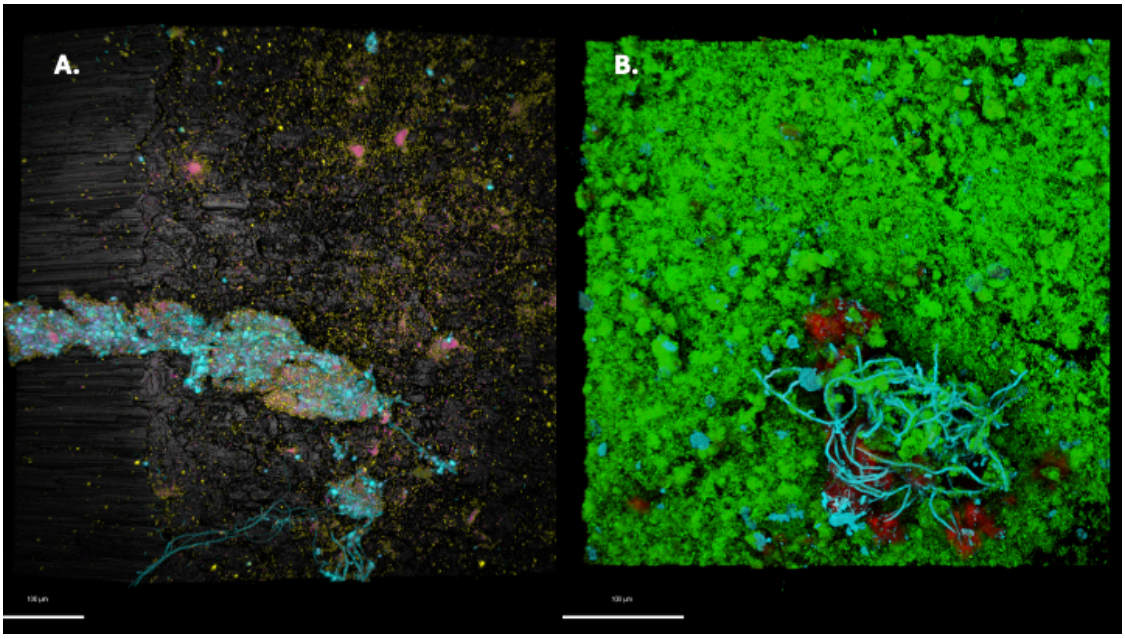


Figure 7. Images of two MDB from NASA-funded research at the CBE: (A) shows the persistence of an MDB following 252 days of dormancy following a water flush (yellow = all bacteria, magenta = membrane compromised bacterial cells, and cyan = fungal cells); (B) shows the complex 3D structure of fungal hyphae and bacterial associations (green = all bacteria, red = membrane compromised bacterial cells, and blue = fungal hyphae).

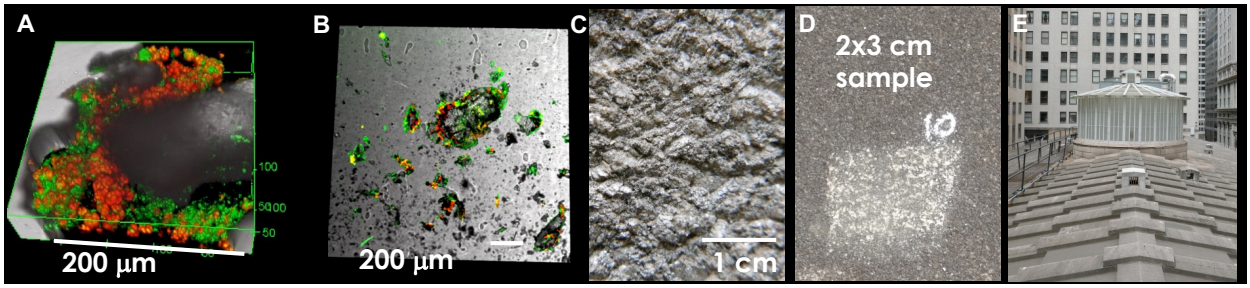


Figure 8. An example of a multiscale modeling project being undertaken by the authors to estimate the size of Cyanobacteria and heterotroph communities on the roof of Federal Hall in New York City: (A) high magnification image of the community on a particle; (B) lower magnification image of the communities on particles; (C) blackened regions associated with biofilm presence; (D) size of the scraped area for imaged particles; and (E) the roof of Federal Hall with darkened areas containing biofilms.

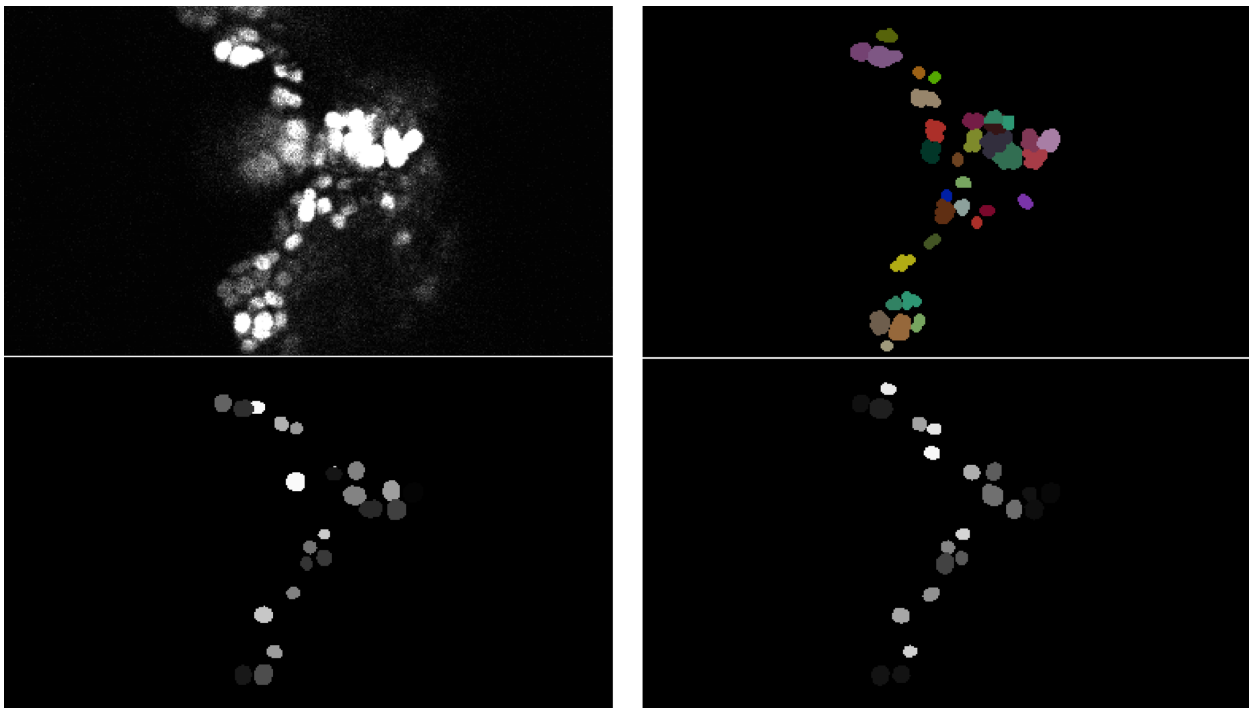


Figure 9. An example of the annotation process. The original image appears on the upper left. A manually annotated image appears on the upper right. An automatically annotated image appears on the lower left that uses a model trained on fully annotated data. A similar automatically annotated image appears on the lower right which is the result of a model trained on partially annotated images.