High-throughput label-free cell classification using machine learning

Claire L. Chen,[1, 2, *] Ata Mahjoubfar,[1, 2] Li-Chia Tai,[2] Ian Blaby,[3] Kayvan R. Niazi,[4, 5] and Bahram Jalali[1, 2, 5, 6]

[1]*Department of Electrical Engineering, University of California, Los Angeles, California 90095, USA*
[2]*California NanoSystems Institute, Los Angeles, California 90095, USA*
[3]*Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, USA*
[4]*NantWorks, LLC, Culver City, California 90232, USA*
[5]*Department of Bioengineering, University of California, Los Angeles, California 90095, USA*
[6]*Department of Surgery, David Geffen School of Medicine,*
University of California, Los Angeles, California 90095, USA
(Dated: June 4, 2015)

Label-free cell analysis is needed for personalized genomics, cancer diagnostics, and drug development as it avoids adverse effects of staining reagents on cellular viability or cell signaling. However, currently available label-free cell assays mostly rely only on a single feature, namely on cell size or dielectric constant. They also have low throughput limiting the sample size that can be analyzed. Here we report that by combining feature extraction and machine learning to high throughput quantitative phase imaging enabled by the photonic time stretch, cells can be classified with high accuracy. Our system captures quantitative phase images in a flow-through microscope and extracts multiple biophysical features such as morphological parameters, optical loss characteristics, and protein concentration indicators of individual cells. These biophysical measurements form a hyperdimensional feature space in which supervised learning is performed for cell classification. As a validation of the enhanced classification sensitivity and specificity of our method, we show binary classification of OT-II white blood T-cells against SW-480 colon cancer cells. Additionally, we show classification of lipid accumulating algal strains for biofuel production.

I. INTRODUCTION

Conventional flow cytometry is a powerful tool for large-scale cell analysis due to its ability to measure anisotropic elastic light scattering of millions of individual cells as well as emission of fluorescent labels conjugated to cells [1, 2]. However each cell is represented with single values per detection channels (forward scatter, side scatter, and emission bands), requiring labeling with specific biomarkers for acceptable classification accuracy [1, 3]. Imaging flow cytometry [4, 5] on the other hand captures images of cells revealing significantly more information about the cells. For example, it can distinguish clusters and debris that would otherwise result in false positive identification in a conventional flow cytometer based on light scattering [?].

In addition to classification accuracy, another critical specification of a flow cytometer is its throughput. Indeed high throughput, typically 100,000 cells per second, is needed to screen a large enough cell population to find rare abnormal cells that are indicative of early stage disease. However there is a fundamental trade-off between throughput and accuracy in any measurement system [6, 7]. Additionally, imaging flow cytometers face a throughput limit imposed by the speed of the CCD or the CMOS cameras, a number that is approximately 2000 cells/s for present systems [8]. Higher flow rate lead to blurred cell images due to the finite camera shutter speed. Many applications of flow analyzers such as cancer diagnostics, drug discovery, biofuel development, and emulsion characterization require classification of large sample sizes with a high-degree of statistical accuracy []. This has fueled research into alternative optical diagnostics techniques for characterization of cells and particles in flow.

Recently, our group has developed a label-free imaging flow-cytometry technique based coherent optical implementation of the photonic time stretch concept [9]. The instrument overcomes the trade-off between sensitivity and speed by using Amplified Time-stretch Dispersive Fourier Transform (ADFT) [10]. In time stretched imaging [11], the object's spatial information is encoded in a spectrum of laser pulses within a pulse duration of sub-nanoseconds (Fig. 1). Each pulse representing one frame of the camera is then stretched in time so that it can be digitized in real-time by an electronic analog-to-digital converter. The ultra-fast pulse illumination freezes the motion of high-speed cells or particles in flow to achieve blur-free imaging. Detection sensitivity is challenged by the low number of photons collected during the ultra-short shutter time (optical pulse width) and the drop in the peak optical power resulting from the time stretch. This issue is solved in time stretch imaging by implementing a low noise-figure Raman amplifier within the dispersive device that performs time stretching [7, 9, 11]. In the coherent version of the instrument utilized here, the time stretch imaging is combined with spectral interferometry to measure quantitative phase images in real-time and at high throughput [12]. Integrated with a microfluidic channel, coherent time stretch imaging system in this work captures 36 million quantitative phase and intensity images of individual cells every second in flow rates as high as 10 meters per second with a throughput

* Email: l.chen@ucla.edu

up to 100,000 cells per second as a high-speed imaging flow cytometer.

On another note, surface markers used to label cells, such as EpCAM, are an indispensable tool for cell classification. However, they are unavailable in some applications; for example, melanoma or pancreatic circulating tumor cells as well as some cancer stem cells are EpCAM-negative and will escape EpCAM-based detection platforms [13]. Furthermore, large-population cell sorting opens the doors to downstream operations where the negative impacts of labels on cellular behavior and viability are often unacceptable []. Cell labels may cause activating/inhibitory signal transduction [], altering the behavior of the desired cellular subtypes, potentially leading to errors in downstream analysis, such as DNA sequencing and subpopulation regrowth. In this way, quantitative phase imaging (QPI) methods [14–16] that categorize unlabeled living cells with high accuracy are needed. Coherent time stretch imaging is a method that combines QPI with high throughput imaging for non-invasive screening of large number of cells in flow.

In this work, image mining and machine learning techniques are applied, for the first time, to label-free time stretch QPI to measure biophysical parameters corresponding to protein concentration, optical loss, and morphological features on single cells with an ultrahigh flow rate and in a label-free fashion. Multiplexed parameters lead to information-rich hyper-dimensional scatter plots for label-free cell classification with high statistical precision. To demonstrate sensitivity, specificity, and accuracy of multi-parameter label-free flow cytometry using our technique, we classified (1) OT-II hybridoma T-lymphocytes and SW480 colon cancer epithelial cells, and (2) Chlamydomonas reinhardtii algal cells (herein referred to as Chlamydomonas) based on their lipid content, which is related to their use in biofuel production. Our preliminary results show that compared to classification by individual biophysical parameters, our label-free hyperdimensional technique improves the detection accuracy from 78.1% to 95.5%, or in other words, reduces the classification inaccuracy by about five times.

II.   TIME STRETCH QUANTITATIVE PHASE IMAGING (TS-QPI)

The application of time stretch quantitative phase imaging (TS-QPI) to imaging flow cytometry has been recently demonstrated in our group [9]. Broadband optical pulses from a mode-locked laser were firstly conditioned in fiber optics and then spatially dispersed in free-space optics with a pair of reflection diffraction gratings creating 1-D "rainbow flashes" (Fig. 1). Each of rainbow flashes was composed of all the wavelength components distributed laterally over the field of view. These flashes illuminated the target as in traditional photography, but in addition, rainbow flashes targeted different spatial points with distinct colors of light, resulting in

space-to-spectrum encoding. Rainbow pulses were then split into the two arms of a Michelson interferometer. Different wavelength components of the rainbow flash traveled parallel to each other but individually focused on the mirror in the reference arm or on the reflective substrate of a microfluidic device in the sample arm. In the sample arm, the cells in the microfluidic channel were hydrodynamically focused [17, 18] into the rainbow's field of view and flowed perpendicular to the rainbow flash. Reflected pulses from the microfluidic device and the reference arm were recombined and coupled back into the fiber, optically amplified and linearly chirped through Raman-amplified TS-DFT system. An ultrafast single-pixel photodetector transformed instantaneous optical power into an electrical signal and subsequently, an analog-to-digital convertor (ADC) samples and quantizes the signal. Acquired data are passed down to processing stages for big data analytics. The interference between time-shifted linearly chirped pulses create a beat (fringe) frequency, which can be adjusted via the interferometer arm length mismatch.

High speed phase contrast imaging was achieved with high sensitivity using an amplified time-stretch system that creates a low-noise distributed Raman amplifier within dispersive fiber with a net optical image gain of approximately 15 dB. The photodetected time-stretched pulses, each representing one line-scan, are converted to analytic signals using Hilbert transformation [19] and the phase component is extracted. Since the phase varies over a wide range (much larger than 2π radians), an unwrapping algorithm is used to obtain the continuous phase profile. The phase profile contains an increasing term with time, corresponding to the fringe (beat) frequency and the phase shift induced by the cell. By eliminating the background phase component, the cell-induced phase shift is extracted, from which a 2D phase image is constructed (Fig. 2).

A.   Signal Processing for Phase and Loss Extraction

The time stretched temporal waveforms corresponding to each line scan consists of two features. One is a fast oscillating fringe (carrier), caused by the interference of the linearly chirped pulses arriving from the reference and signal arms in the Michelson interferometer. Acting as a radio-frequency (RF) carrier whose frequency is set by the user adjusted arm length mismatch, the fringe frequency is modulated when the optical path length in the sample arm is changed by the arrival of a cell. This frequency shift and the accompanying phase change are used to measure the optical path length of the cell. The second feature in the waveform is a lower frequency envelope corresponding to temporal shape of the optical pulse. The amplitude of this envelope provides information about optical loss caused by surface roughness and inner organelle complexity (Fig. 2). One of the challenges in imaging flow cytometry is the uncertainty in the posi-
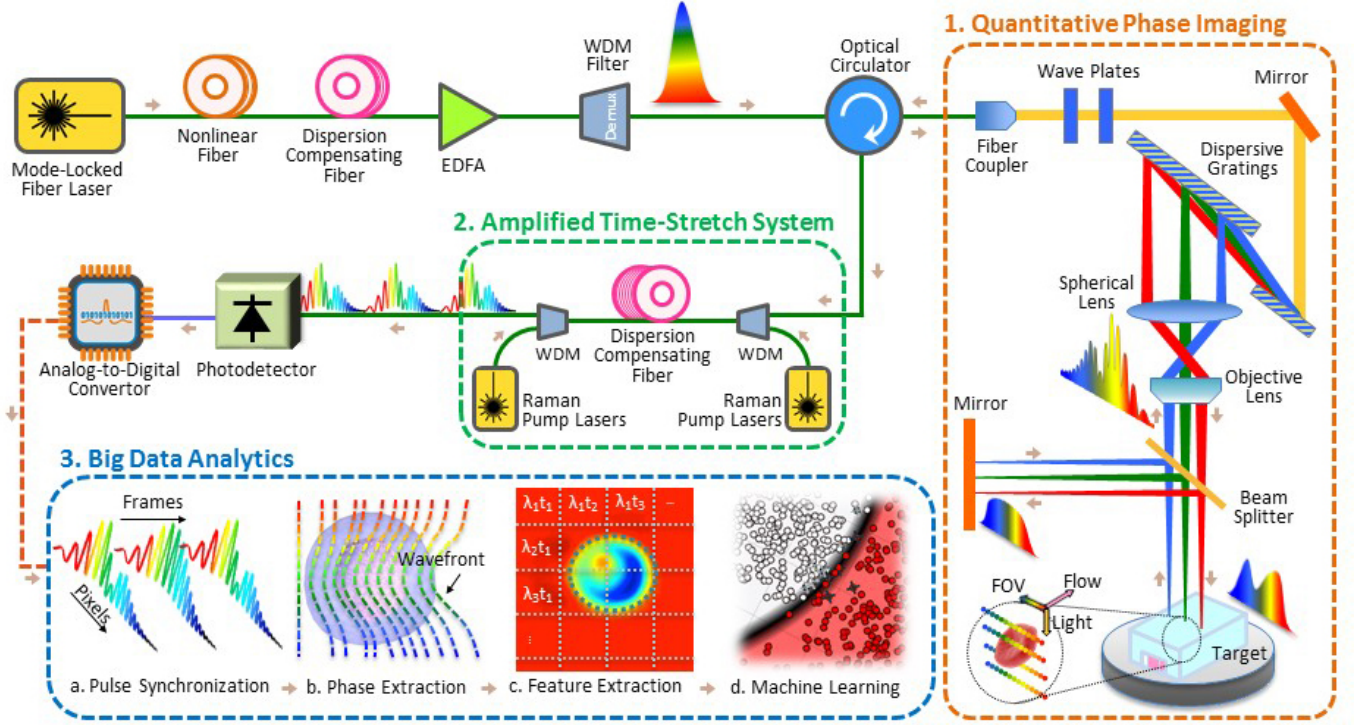
FIG. 1. Time stretch quantitative phase imaging (TS-QPI) system simultaneously acquires brightfield images and quantitative phase images. Box 1 shows a train of broadband optical pulses spatially dispersed into a train of rainbows illuminating the target as line scans. The spatial features of the target are hence encoded into spectrum of broadband optical pulses each representing one 1-D frame. The ultra-short optical pulse illumination freezes the motion of cells during high speed flow to achieve blur-free imaging with a throughput of at least 100,000 cells/s (limited by the flow channel). The phase shift at each location within the field of view is embedded into the interference patterns using by a Michelson interferometer. Box 2: The interferogram pulses were then stretched in time so that spatial information could be mapped into time through time-stretch dispersive Fourier transform (TS-DFT) and then captured by single pixel photodetector. The loss of resolution at high shutter speed is compensated by stimulated Raman amplification during time stretch. Box 3: (1) Pulse synchronization: The envelope and phase of the time-domain signal was firstly mapped into series of spatial information, forming a scanning line brightfield image and phase contrast image. Intensity measurement is calibrated based on the pulses during when cells are absent. (2) The biomass density of a cell leads to a spatially varying optical phase. When a rainbow linescan passes through the cells, the changes in refractive index at difference locations will cause phase walk-off at each interrogation wavelength. Hilbert Transform and unwrapping algorithm are used to generate the spatial information of phase shift. (3) Spatially dispersed rainbow creates multiple interrogation points for cell imaging. Decoding the phase shifts in each pulse at each wavelength and remapping into space reveals information about the distribution about protein concentration within cells. The transparency coefficient of each cell, corresponding to the fluctuation of intensity induced by the cell, is obtained from the amplitude of the slowly varying envelope on interferogram. Thus we achieve both image reconstruction of both brightfield and phase contrast simultaneously.

tion of the cell along the optical axis. Since the optical path length is a quantity that is integrated over the entire depth along the axis, maps of optical path lengths are mostly insensitive to axial position of the cells. The refractive index used as a measure of protein density is obtained by dividing the measured optical path length by the cell thickness. Since cells in suspension relax to a spherical shape (due to surface tension) [20, 21], an independent measure of cell diameter can be obtained from its lateral dimension. Finally, the attenuation coefficient of each cell is obtained from the amplitude of the slowly varying envelope as discussed previously.

III. FEATURE EXTRACTION

Instead of producing a single value from each cell image, we make quantitative image analysis and measure variety of features for large population of cells. The 2D image obtained from sequential linescans provides information about cell. Owing to the high throughput of the time-stretch quantitative phase imager (TS-QPI), multiple parameters of individual cells can be extracted for a large cell population. The measured cell information not only includes morphology, such as diameter, area, shape, aspect ratio, uniformity, etc., but also biophysical features such as protein concentration, transparency, and
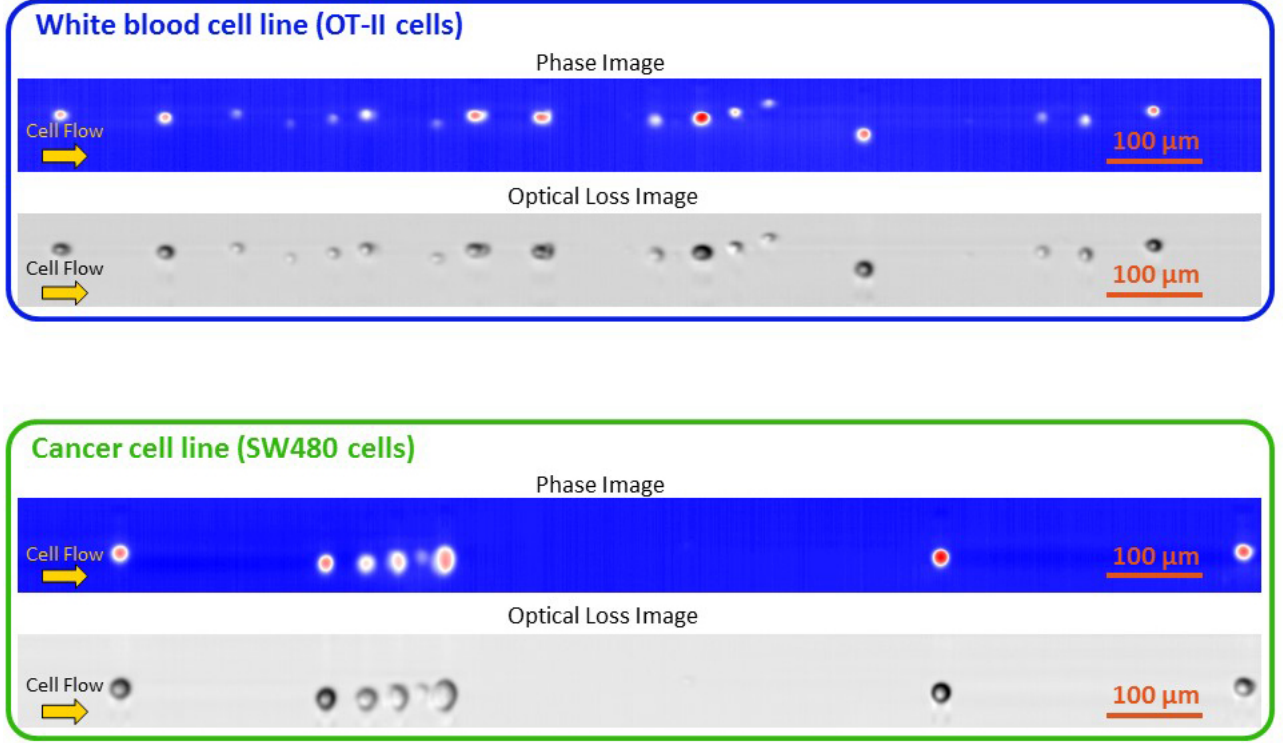
FIG. 2. Phase image (upper) and optical loss image (lower) of SW480 (green) and OT-II cells (blue). The transparency of cells are affected by the attenuation of light going through the cells, as well as the scattering from the surface of the cells and from the internal cell organelles. The intensity is extracted from the lower frequency envelope variation of the waveform and relates to the transparency of individual cells in a flow stream. The phase information is extracted from the analytic forms of each pulse using Hilbert Transformation, followed by a phase unwrapping algorithm. Details can be found in appendix.

attenuation.

It is well known that cells from the same line or tissue exhibit variations in size, structure, and protein expression levels, even when they are genetically identical [?]. These intrinsic variations have important role in analyzing important extrinsic variations caused by drug treatment, cell cycles, labeling, and transcription rate. The large data set captured by the phase contrast time stretch imaging flow cytometer provides statistically relevant biophysical information that is useful in drug development.

We have developed a feature extraction tool that performs segmentation, object detection and feature measurement. The algorithm operates on phase and intensity images simultaneously. The analysis also includes clump identification noise suppression, and smoothing applied before cell identification. In our feature extraction, the morphology of each single cell is described by area, diameter along rainbow, major axis, minor axis, orientation, uniformity, circularity, perimeter, and number of surrounding clumped cells. The change in optical pulse intensity caused by cell scattering and absorption is extracted from the pixel intensities in the attenuation image. Cell protein concentration is extracted from the pixel intensities in the phase image. Overall, 16 features

are extracted for each cell. The relative efficacy of these parameters in cell classification are quantified below.

Among different features, size measurement is particularly important as it is used as a fundamental parameter by itself in many technologies [], but also used to extract the refractive index (and hence density) from measured optical path length here. Size measurement is performed along both the lateral (rainbow flashes) and flow directions. The measurement along the rainbow direction is used for protein concentration estimation since it helps to eliminate measurement inaccuracies caused by fluctuations of flow speed. We also measure the aspect ratio, perimeter, and median radius as well as a host of other parameters listed in Table I. To calibrate the system, we used 5 μm polystyrene beads (from Polysciences, Inc.) with a narrow and known NIST traceable size distribution.

We used hyper-dimensional scatter plots as a tool for high content cell analysis. To demonstrate application in circulating cancer cell (CTC) detection, we used a white blood cell line (OT-II hybridoma T cell) and a cancer cell line (SW480 epithelial colon cancer). Fig. 3 shows the 3D scatter plot based on diameter, protein concentration, and attenuation of a large number of OTII (black dots) and SW480 (red dots) cells. The 2D projections on the

three orthogonal planes are also shown. It is clear that certain pairs of features are better able to differentiate the two populations.

IV. CORRELATION AMONG EXTRACTED FEATURES

As previously mentioned, a total of 16 features are extracted from phase and transparency images of each cell. Parameters that are highly correlated do not provide unique information. Pairwise correlation matrix among 16 parameters extracted from the images is shown as a heat map in Fig. 4. Diagonal elements of the matrix are correlation of each parameter with itself, i.e. the auto-correlation. The set of parameters with the red color are highly correlated. The subset in the lower left shows high correlation because they all measure morphological features. Also, the subset in the upper right are correlated as they all measure optical loss. This suggests that the dataset can be adequately represented by a smaller set that excludes highly correlated parameters.

In order to rank the physical features according to their utility in classification, we performed one dimensional classification using each of the 16 features extracted from phase contrast images. Fig. 5 shows classification accuracy for each feature arranged in descending order. The features are color coded into three categories: morphology, optical phase, and optical loss, to describe the type of information provided by each. The figure provides valuable insight into the relative importance of each category of cell parameters and suggests that morphological information provide the most information about cells but at the same time significant additional information is contained in optical phase and loss measurements. As we will see later in this paper, these conclusions are also supported by results of hyperdimensional classification.

V. CLASSIFICATION

Neural network is a bio-inspired learning algorithm that works well in classification via supervised learning when adequate amount of training data is available. The basic artificial neural network can be described as a series of transformations performed on the input data. Shown in Fig. 6, the network consists of an input layer, output layer producing classification results and hidden layer(s) that map the input to the output.

Each layer performs a linear combination on its inputs and performs soft threshold operation on the sum. The output of the node j in layer $l + 1$, denoted by $z_j^{(l+1)}$ is generated from inputs $x_1, x_2, \ldots, x_N$, as

$$z_j^{(l+1)} = \mathrm{h}(a_j^{(l+1)}) = \mathrm{h}(\sum_{i=0}^{N_l} \omega_{ji}^{(l)} x_i^{(l)}) \quad i = 0, \ldots, N_l; \quad j = 1, \ldots, N_{l+1}; \quad l = 1 \ldots L$$

(1)

Here $a_j^{(l+1)}$ is the linear combination of inputs, $\omega_{ji}^{(l)}$ are the weights of the linear combination. The summation runs over N_l, the total number of nodes in the previous hidden layer, and L is the total number of hidden layers. The function $\mathrm{h}(\cdot)$ is called the activation or hypothesis function and is usually a soft threshold implemented using a logistic sigmoid function $\mathrm{h}(a) = 1/(1 + \exp(-a))$ or a hyperbolic tangent function. The constant term in the summation, $x_0^{(l)}$, is called the bias.

VI. ROC AND AUC

In classification, there is a trade-off between the sensitivity and specificity. Choosing a low detection threshold (the classifier) results in high sensitivity, however this sacrifices the specificity leading to large number of false positives. To obtain the optimum classifier, we use the receiver operating characteristics (ROC) and maximize its area under the curve (AUC) (see inset of Fig. 7). The ROC is obtained by varying the classification threshold while plotting the True Positive Rate on the vertical axis versus the False Positive Rate on horizontal axis. Therefore, ROC is an ideal tool to visualize the impact of the threshold on classification accuracy. A random guess, corresponding to accuracy of 50% in binary classification will have an ROC that is a diagonal line. Conversely, a classifier that accurately separates the classes will have an ROC curve that approaches the upper left corner. The AUC parameter measures the area under the ROC curve and serves as an effective metric for optimizing the threshold.

VII. SINGLE LAYER PERCEPTRON

Single layer perceptron is the name given to neural network with one hidden layer. This simple model works well on linearly separable data. Different values of weights correspond to different linear classification planes. In our work, we use the perceptron with AUC as the cost function and genetic algorithm [ ] for optimization. This approach of opting for the simpler model, perceptron, as opposed to a multilayer neural network, without compromising the accuracy is consistent with the proverbial "simpler is better" known as Occam's Razor in the machine learning community.

Add to neural network section Say it is the bases of deep learning and corical processor

As shown in Fig. 4 and Fig. 5, many of the 16 parameters are correlated and not all measured features in the data set produced by the time stretch quantitative phase imaging have the same amount of information. That result suggests that it may be possible to reduce the 16 dimensional data set to a smaller set of uncorrelated orthogonal dimensions without significantly compromising the classification accuracy. In that spirit, we have
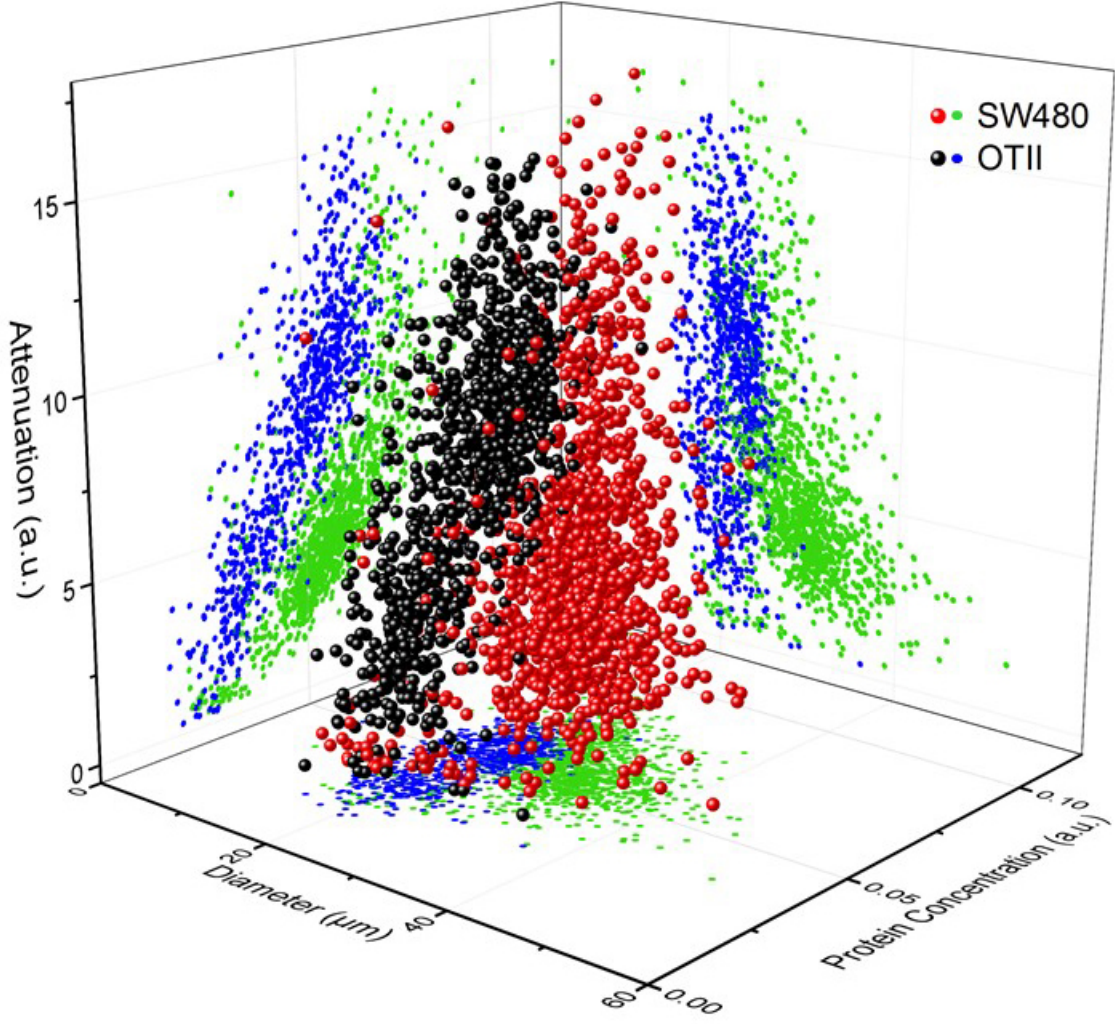
FIG. 3. Three-dimensional scatter plot based on size, protein concentration, and transparency of the cells measured by Time Stretch Quantitative Phase Imaging. The green and blue dots are two-dimensional projections on the plane containing every combination of two parameters. The protein concentration corresponds to the local phase shift within the cell, which was converted into a local optical path difference mapping. The transparency is a parameter describing the intensity fluctuation caused by cell scattering and absorption.

used principle component analysis (PCA) for dimensionality reduction. The PCA algorithm finds an alternative lower dimension space such that variance of data projected onto this subspace is maximized along subspace dimensions. By revealing the structure in data that best explains the variance, PCA achieves data compression via dimensionality reduction.

We have performed PCA on the 16 dimensional data set acquired by our time stretch QPI system. This maps the data from the physical feature space to the PCA space. Fig. 8 shows the percent of the variance in data explained by each component (lower chart). The key observation is that most of the variance can be accounted for by the first two principle components. The upper portion of the plot shows the AUC for binary classification using each of the principle components. Interestingly,

the first component with the highest explained variance is not necessarily the most important component for classification. This is not surprising since PCA components are optimized to the variance of data but not to the data labels. Therefore, apriori intuition about the physical significance of the features in the case here, is superior to PCA in eliminating dimensions that do not provide high value in classification. Nevertheless, in situations where physical intuition about meaning of features does not exist, PCA is a valuable tool. A second observation is that principle components that have low explained variance do have some value in classification.
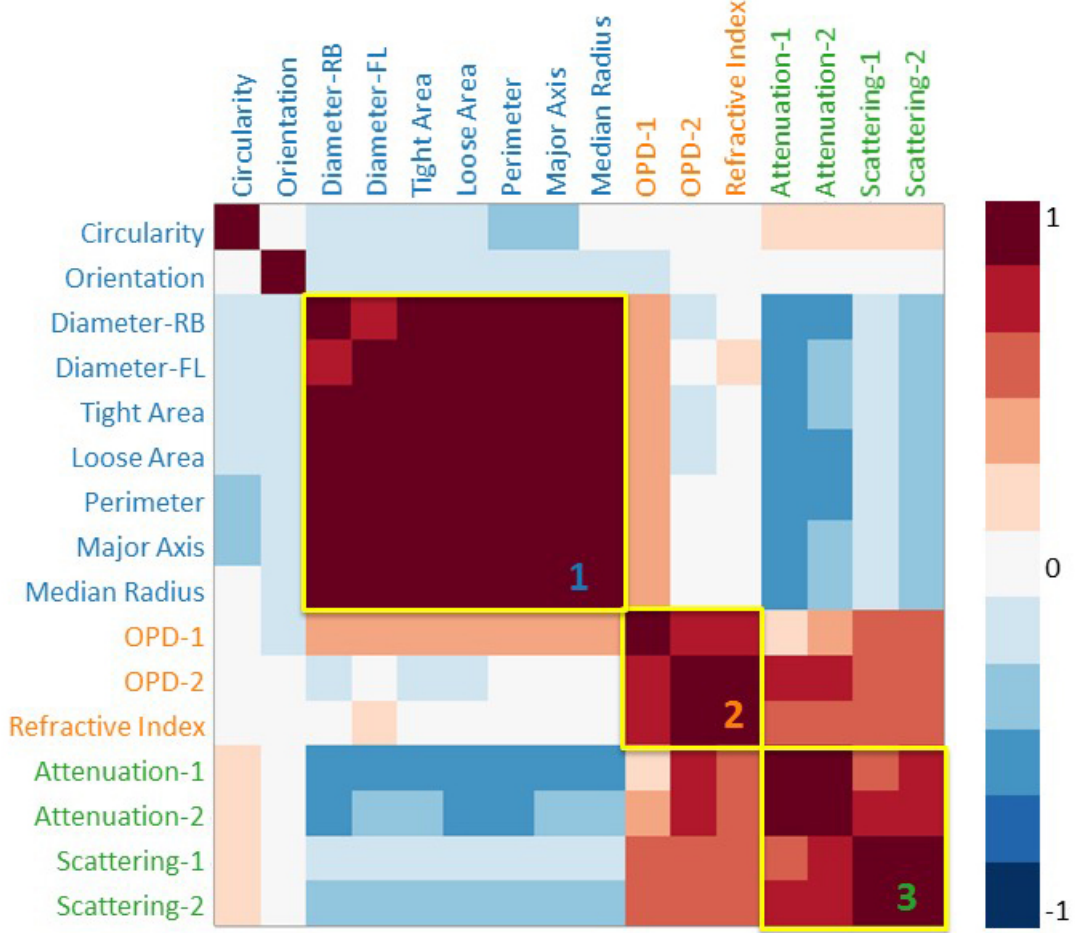
FIG. 4. Pairwise correlation plot visualized as heat map showing all major 16 features extracted from the images. Diagonal elements of the matrix are correlation of each parameter with itself, i.e. the autocorrelation. The set of parameter with the red color are highly correlated. The subset in the lower left show high correlation because they all measure morphological features. Also the subset in the upper right are correlated as they all measure attenuation and light scattering. This suggest that the dataset can be adequately represented by a smaller set that excludes highly correlated parameters.

A. Single layer perceptron optimized via AUC

To optimize the classifier performance in neural networks, cross entropy and mean squared error are typical cost functions used. In this work, we use the ROC as the cost function to train the neural network because it considers the balance between the sensitivity and specificity of the classification [Hanley 1982]. AUC has been proven to be advantageous to the mean square error for evaluating learning algorithms [22] and minimizing the error rate may not lead to the best possible AUC values [Yan 2003, Cortes AUC optimization, Huang AUC].Here we use single layer perceptron but trained by globally maximizingthe area under the ROC curve (AUC). By varying the bias and the weights, the offset and slope of the classifier is scanned, respectively. To speed up the computation for high dimensionality analysis, we employ genetic algorithm.

VIII. DEMONSTRATION IN CLASSIFICATION OF OTII AND SW480 CELLS

Classification of cells based on size alone in a label-free manner has been previously reported by several groups [Vona et al, others]. However, in contrast to these single parameter (size) approaches, our multi-parameter classification, enabled by the time stretch QPI, offers considerable improvements in detection sensitivity and accuracy (see Fig. 7 and Fig. 9). To demonstrate the effectiveness of the machine learning plus time stretch QPI imaging system for cancer cell detection, we used OT-II hybridoma T cells as a model for normal white blood cells and SW-480 colon cancer cells. A 16 feature data set was measured using the imaging system and feature
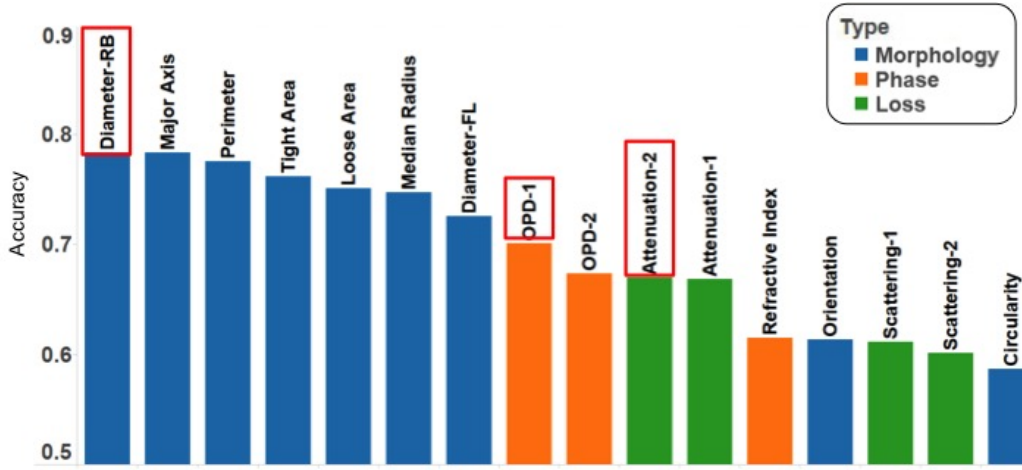
FIG. 5. Ranking of feature performance based on the accuracy in 1D classification. The value is the area under ROC curve of each individual parameter. Blue color shows morphological parameters includes diameter along the interrogation rainbow, diameter along the flow direction, tight cell area, loose cell area, perimeter, circularity, major axis length, orientation, and median radius. Orange bars show phase shift information include optical path length difference, refractive index difference and refractive index difference after noise reducing. Green bars show light scattering information. Red box shows the best performed parameters in these three categories respectively. Morphology contains most information.
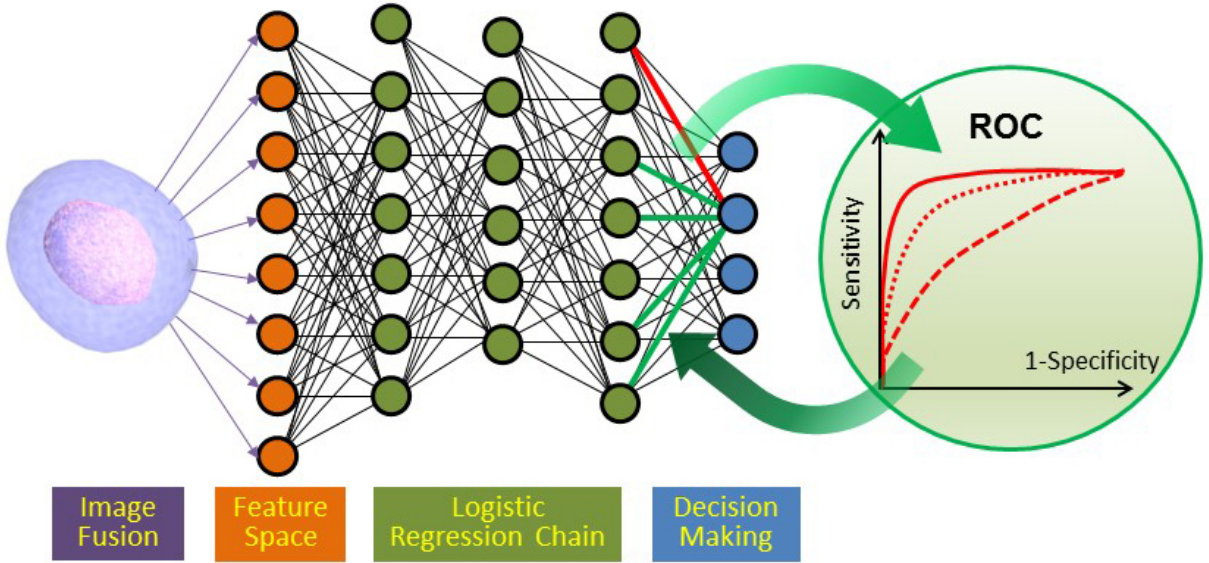

FIG. 6. Neural network architecture and training via Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC). Multivariate features of each measurement (cell) are feed into the neural network and the output shows classification labels. We train the weights of the last layer perceptron by optimizing the area under ROC curve. Each ROC curve corresponds to a set of weighting matrix to one of the output unit generated by scanning the bias weight. AUC is proven to be statistically consistent and more discriminating than accuracy.

extraction technique described previously. By sliding the detection limit along the direction perpendicular to the optimum classification plane, a receiver operating characteristic (ROC) curve for multi-dimensional data is generated (Fig. 7).

Blue curves in Fig. 7 show the ROC for classification using all sixteen biophysical features extracted from the time-stretch QPI images. The grey diagonal line shows a random guess classifier with no discrimination skill. For generating the ROC curves, we employ the normal unit vector of a decision boundary in n-sphere and use genetic algorithm for optimization. At step k+1 a trial set of weights (children) is randomly generated in the neighborhood of the present set (parents). Each generation produces new ROC curves. The genetic algorithm randomly selects from this population and uses them as par-
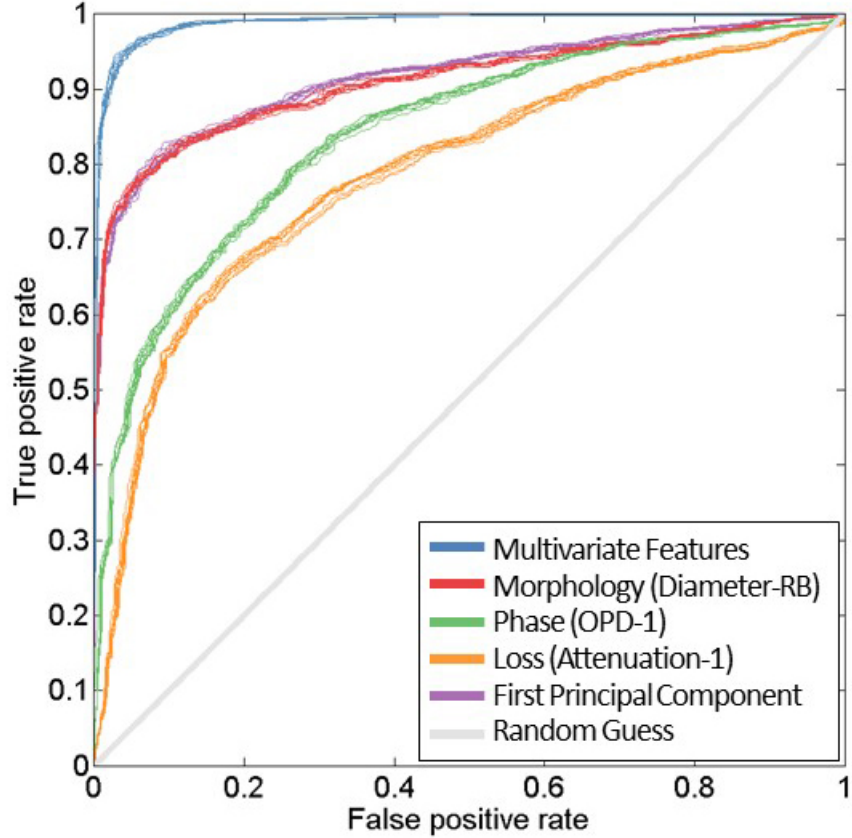
FIG. 7. Receiver Operating Characteristics (ROC) curves for binary classification of white blood cell (OTII) and cancer cells (SW480). The grey diagonal line shows results of random guess. Blue curves show the classifier performance using all 16 biophysical features extracted from the time-stretch QPI images. We show ROC for classification based on single parameters. Red, green and orange curves show the classifier decision made using only the three major physical features: diameter, optical path length, and attenuation, respectively. Multivariate analysis based on time-stretch QPI shows significant improvement in classification. For each group we show the ROC for 10 fold cross validation data set. Additionally purple curves shows the classifier optimized with the first principal component in PCA feature space after data cleaning. Each kind of classifiers is shown with ten find curves demonstrating the results with ten-fold cross validation. Here we take 90% data for training and the other 10% for test. In the ten-fold cross validation, the total data set is split into 10 equal subsets and each time the test is performed for one part while the other nine parts are used for training. The fact that the classifiers remain almost unchanged during the ten runs of cross validation shows consistency and robustness of the data set.

ents to produce the children for the next generation by mutation and crossover. Over successive generations, the population evolves toward an optimal solution that maximizes the AUC. The area under the ROC curve (AUC) was calculated using a non-parametric method based on constructing trapezoids under the curve.

As a way to visualize the improvement in classification by hyperdimensional feature space of TS-QPI, we also show the ROC curves of classification based on several single features: diameter along the interrogation rainbow, integral of cell's optical path length difference, and cell attenuation. These three biophysical parameters individually perform the highest accuracy among morphology, optical phase, and optical loss parameter groups respectively. But multivariate analysis based on TS-QPI shows significant improvement in classification. For each type of classification we show the ROC for 10 fold cross

validation data set. As the curves approach random guess line, the 10 fold cross validation curves deviate more from each other. Additionally purple curves shows the classifier optimized with the first principal component in PCA feature space.

IX. DEMONSTRATION IN ALGAE LIPID CONTENT CLASSIFICATION

Microalgae are considered one of the most promising feedstock for biofuels. The productivity of these photosynthetic microorganisms in converting carbon dioxide into useful carbon-rich lipids greatly exceeds that of agricultural crops [insert reference]. Worldwide, research and demonstration programs are being carried out to develop the technology needed to expand algal lipid production
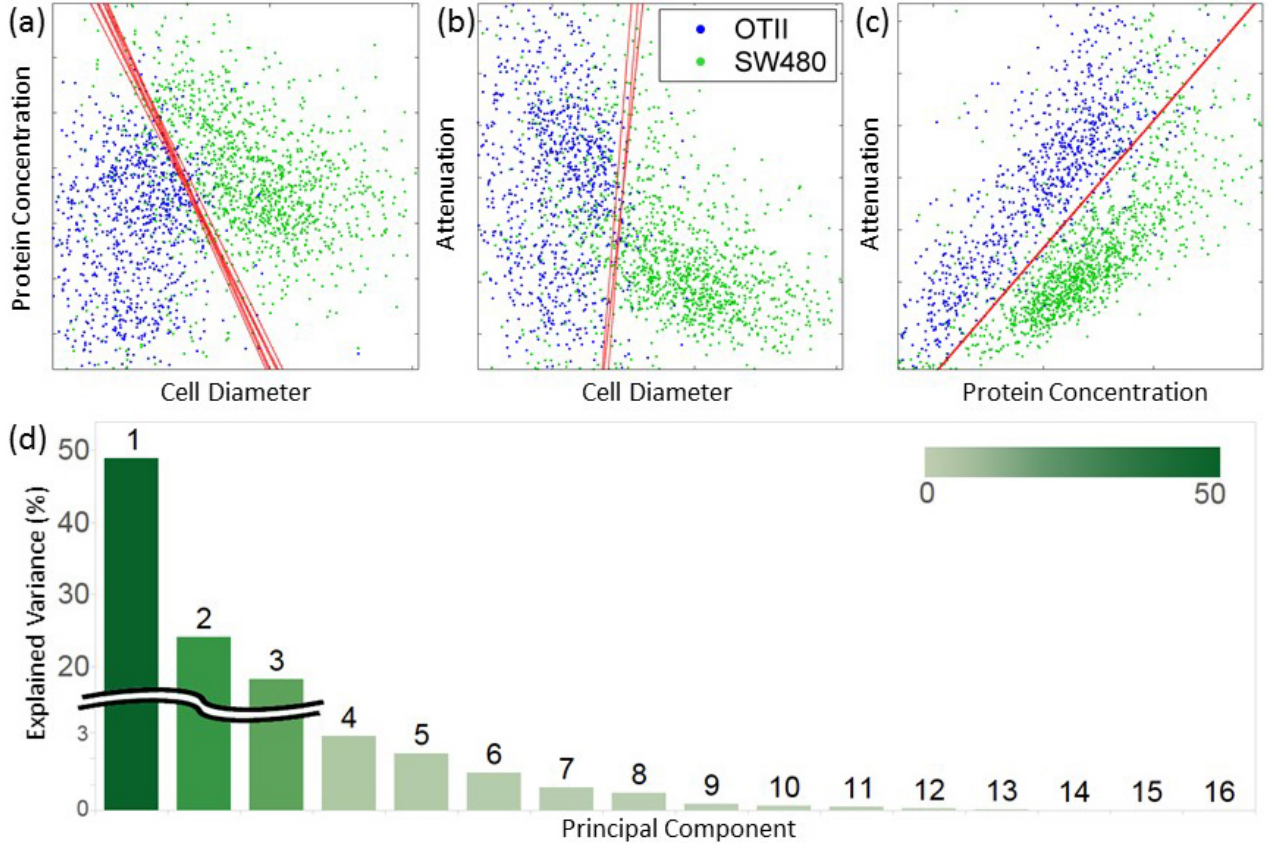
FIG. 8. (a-c) Decision boundary (red lines) in simultaneous two-parameter scatter plots using ten-fold cross validation. It shows the consistency and robustness of supervised classification. (a) diameter and protein concentration (b) diameter and scattering measurement (c) protein concentration and scattering. (d) Performance ranking of each individual principal components in PCA space. The values of each bar show the percentage of the total variance explained by each principal components, accounting for the variability expressed in the data. Principal components with larger variability don't necessarily give high accuracy in classification.

as a major industrial process. Selecting high yield microalgae with fast growth factors are essential in biofuel production industry. Because algae differ greatly in size and structure, cell size alone provides insufficient information for cell classification. Here we show that adding refractive index and transparency data, obtained by the phase contrast time stretch imaging flow cytometer, to size data enables cells to be distinguished on the basis of lipid content.

To test our apparatus for its ability to separate high and low-lipid containing cells we exploited the starch-null sta6 strain of Chlamydomonas. This strain is deleted for STA6 (encoding the small subunit of ADP-glucose-pyrophosphorylase), and when nitrogen-deprived accumulates more lipid than wild-type. Comparison of the two strains therefore provides an ideal setup to test our ability to distinguish lipid-content phenotypes.

Fig. 9a shows the 3D scatter plot showing the three principle physical features for the two algae populations. Also shown is the scatter plot obtained using a conventional flow cytometer. It is readily apparent that the TS-QPI capture much more information about individual cells and is better suited for differentiation of low- and high-lipid content algae species. In Fig. 9b, we show ROC curves for binary classification of these populations. Blue curve shows the classifier performance using all 16 physical features extracted from the time-stretch QPI images. Red, green, and orange curves show the classifier decision made using only the three major biophysical features diameter, optical path length, and light attenuation. Additionally we show ROC curves in purple for classification based on the first principal components in PCA space. We observe that while biophysical parameters, in particular cell diameter, is an effective feature for classification, hyperdimensional classification using all 16 parameters offers the best accuracy.

X. SUMMARY

Time-stretch quantitative phase imaging (TS-QPI) is capable of capturing images of flowing cells with minimal motion distortion at unprecedented rates of 100,000 cells/s. TS-QPI relies on spectral multiplexing to cap-
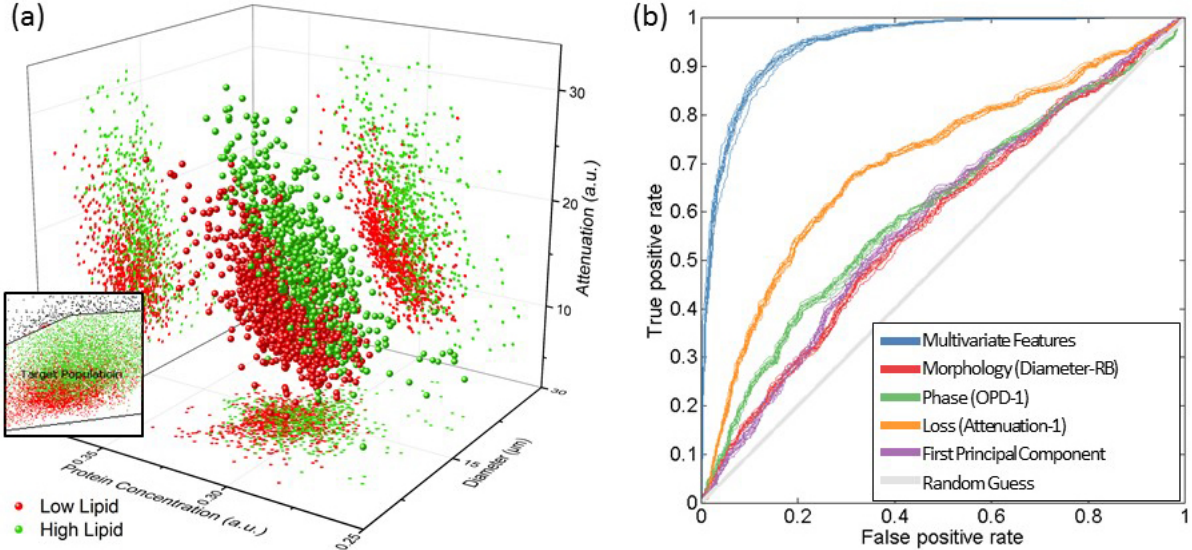
FIG. 9. (a) Study of algae cells (Chlamydomonas reinhardtii) time-stretch QPI. Three-dimensional scatter plot based on size, protein concentration, and transparency of the cells measured by time-stretch QPI, with 2D projections for every combination of two parameters. Insert: Conventional flow cytometry using forward scattering and side scattering is not enough to distinguish the difference between high-lipid content and low-lipid content algae cells. Time-stretch QPI is much more effective in separating the two algae populations. (b) Receiver Operating Characteristics (ROC) curves for binary classification of normal and lipid-rich algae species. Blue curve shows the classifier performance using all 16 physical features extracted from the time-stretch QPI images. Red, green and orange curves show the classifier decision made using only the three major physical features, rainbow diameter, optical path length difference, and attenuation, respectively. Additionally we show ROC curves in purple for classification based on the first principal components in PCA space. The diagonal line shows results of random guess.

ture simultaneously both phase and intensity quantitative images in a single measurement, generating a wealth of information of each individual cell eliminating the need for labeling with undesirable biomarkers. Here, we summarized the information content of these images in a set of 16 parameters for each cell, and performed classification in the hyperdimensional space of this dataset. Our classifier learning method is based on evolutional optimization of the area under the receiver operating characteristic curve for a single-layer binary-classifier neural network. The results from two experimental demonstrations, one on detection of cancerous cells among white blood cells, and another one on identification of lipid-rich algae, shows that classification accuracy by using the TS-QPI hyperdimensional space is more than 17% better than the conventional size-based techniques.

XI. METHODS

A. Time Stretch Quantitative Phase Imaging (TS-QPI) System

Broadband optical pulses from a mode-locked laser (center wavelength = 1565 nm, repetition rate = 36.128 MHz, pulse width ~ 100 fs) are further broadened using a highly nonlinear fiber to approximately 100 nm with a spectral range up to 1605 nm. These broadband pulses

are then linearly chirped to 1.2 ns pulse width by a short dispersion compensating fiber (DCF) of 60 ps/nm, so that an erbium doped fiber amplifier (EDFA) can amplify them with minimal distortion. A coarse wavelength division multiplexer (WDM) filters the pulses from 1581 nm to 1601 nm, where the spectrum is reasonably flat. Therefore, the total bandwidth of the pulses interrogating the cells in our setup is less than 20 nm centered at 1591 nm, giving a negligible fractional bandwidth of 1.3%. These filtered pulses then pass through an optical circulator and are coupled to free-space with a fiber collimator.

Free-space laser pulses were linearly polarized with quarter- and half-wave plates, and then spatially dispersed with a pair of reflection diffraction gratings, so that each wavelength component of the collimated beam was positioned at a different lateral point similar to a line flash rainbow. A beam reducer shrank the rainbow beam 6 times with a pair of 90 degree off-axis parabolic gold-coated mirrors with reflected focal lengths as 152.4 mm and 25.4 mm, respectively. Next, a 15 degree off-axis parabolic gold-coated mirror with 635 mm reflected focal length and a long working-distance objective lens with 0.4 numerical aperture further shrank the rainbow to about 130 μm in width, i.e. field of view. Reflective optics with parabolic gold-coated mirrors is used in our experimental demonstration to minimize loss, aberration, and polarization sensitivity. The rainbow flashes

were then split into the two arms of a Michelson interferometer by a beam splitter. In the sample arm, the rainbow pulses pass through the cells and are reflected by the reflective substrate of the microfluidic device. In the reference arm, a dielectric mirror reflected the rainbow with a length mismatch with the sample arm causing 5 GHz interference fringes. Cells are hydrodynamically focused at the center of the channel flow at a velocity of 1.3 m/s. The reflected pulses from reference and sample arms were recombined at the beam splitter, compressed by the two diffraction gratings and coupled back into the fiber. These return pulses were spectrally encoded by the spatial information of the interrogation field of view. Then they were redirected by the optical circulator to a Raman-amplified time-stretch dispersive Fourier Transform (TS-DFT) system followed by a 10 Gb/s photodetector (Discovery Semiconductors DSC-402APD). An analog-to-digital convertor (Tektronix DPO72004C) with a sampling rate of 50 GS/s and 20 GHz bandwidth is used to acquire the output signal of the photodetector.

B. Coherent Detection and Phase Extraction

Unlike in conventional heterodyne detection, which uses a narrow band continous-wave beam as the local oscillator, the coherent detection in our time stretch system uses an unmodulated copy of the original optical input, which is a broadband chirped optical pulse.

The input optical signal is split into two arms of the Michaelson Interferometry at the beam splitter. The one in sample arm, $E_s(t) \exp(i\omega_s t)$, passes through the object, and then, is combined with the other one from the reference arm $E_{LO}(t) \exp(i\omega_{LO}t)$. $E_s(t)$ and $E_LO(t)$ are the electric field amplitudes of the sample and reference sources, and ω_s and ω_{LO} are the optical frequencies of sample and reference fields, respectively. Not only the electric field amplitude after passing through semitransparent objects, $E'_s(t)$, will be modulated by the optical attenuation and scattering in the sample arm, but also the optical path length difference will lead to a phase shift, $\Delta\varphi(x,t)$, inducted by refractive index change along the interrogation beam. Thus the combined electrical field is

$$E_{total}(t) = E'_s(t) \exp\{i[\omega_s t + \Delta\varphi(x,t)]\} + E_{LO}(t) \exp(i\omega_{LO}t) \tag{2}$$

In our case, $E_{LO}(t)$, as a local oscillator (LO), is same as $E_s(t)$ but with reduced power. After time stretch, the frequency components in ω_s and ω_{LO}, are mapped into time. That is, electric field detected at time t is one-to-one mapped from frequency domain,

$$E_{total}(t) = E'_s(t_p,\omega_s) \exp\{i[\omega_s t + \Delta\varphi(x,t)]\} + E_{LO}(t_p,\omega_{LO}) \exp(i\omega_{LO}t) \tag{3}$$

where t_p corresponds to the p-th incoming pulse. Note that $E_s(t_p,\omega_s)$ can be simplified as $E(\omega_s)$ when pulse

shape (magnitude of each wavelength) is stable from pulse to pulse.

ω_s detected at t, is mixed with local oscillator ω_{LO} at t, which is a copy of $\omega_s(t)$ within the same pulse but delayed by $\Delta\tau$,

$$\omega_{LO}(t) = \omega_s(t - \Delta\tau) \tag{4}$$

where $\Delta\tau = 2\Delta L/c$ is the delay mismatch in Michelson interferometry and ΔL is the length mismatch in the two arms. Define t_i as the relative time within each pulse,

$$E_{LO}(t_p,\omega_{LO}) = E_{LO}(t_p,\omega_{LO}(t_i)) = E_s(t_p,\omega_s(t_i - \Delta\tau)) \tag{5}$$

Since the difference in arrival times at the photodetector after time stretch between each frequency component of reference and sample signals is same as the delay difference in interferometry arms, $\Delta\tau$, the beating frequency or modulation frequency of the interference fringes is

$$\omega_f = \omega_s(t_i) - \omega_{LO}(t_i) = \omega_s(t_i) - \omega_s(t_i - \Delta\tau)$$
$$\approx \frac{2\pi c}{\lambda_c^2}(\lambda(t_i) - \lambda(t_i - \Delta\tau)) = \frac{2\pi c}{\lambda_c^2} \cdot \frac{\Delta\tau}{\beta L_f} = \frac{4\pi\Delta L}{\lambda_c^2 \beta L_f}$$

where λ_c is the central wavelength, β is the group velocity dispersion, and L_f the length of dispersive fiber. Thus the energy flux absorbed at the photodetector, is proportional to the product of the sum of the electric fields

$$I_{PD}(t) \propto |E'_s(t_p,t_i)|^2 + |E_s(t_p,t_i - \Delta\tau)|^2$$
$$+2\,\text{Re}\{E'_s(t_p,t_i) \cdot E_s^*(t_p,t_i - \Delta\tau) \cdot \exp[i(\omega_f t + \Delta\varphi(x,t))]\} \tag{6}$$

where the asterisk signifies complex conjugation, and Re is the real part. Note that high frequency terms are averaged over time as it is far beyond the photodetector response time.

Thus the time stretched temporal waveforms corresponding to each line scan image consists of two features [12]. One is $|E'_s(t_p,t_i)|^2 + |E_s(t_p,t_i - \Delta\tau)|^2$, a temporal envelope of the time-stretched optical pulse usually at baseband frequencies. The amplitude of this envelope corresponds to the temporal shape of the optical pulse and its deviations caused by the object as in brightfield microscopy. It provides information about optical loss, i.e. light attenuation, absorption, and scattering caused by surface roughness, granularity, and inner cell organelle complexity. The other feature is $I_PD^{(f)} = 2\,\text{Re}\{E'_s(t_p,t_i) \cdot E_s^*(t_p,t_i - \Delta\tau) \cdot \exp[i(\omega_f t + \Delta\varphi(x,t))]\}$, a fast oscillating fringe, caused by the interference of the linearly chirped pulses multiplexed between the sample and reference arms in the Michelson interferometer.

The second feature, $I_{PD}^{(f)}$, containing only spectral features modulated at ω_f is separated by a band-pass filter to extract the object phase shift, $\Delta\varphi(x,t)$. The instantaneous phase of this feature can be readily retrieved from its analytic representation given by Hilbert transform, $\mathcal{H}$:

$$\angle I_{PD}^{(f)} = \omega_f t + \Delta\varphi(x,t) = \arg[I_{PD}^{(f)}(t_p,t_i) + j \cdot \mathcal{H}(I_{PD}^{(f)}(t_p,t_i))] \tag{7}$$

Here arg means the argument of a complex number. A one-dimensional phase unwrapping algorithm followed by background phase removal gives the object phase shift,

$$\Delta\varphi(x,t) = \text{unwrap}(\angle I_{PD}^{(f)}) - \omega_f t \quad (8)$$

The unwrapping algorithm used in our processing acts when the absolute phase difference between two consecutive samples of the signal is greater than or equal to π radians, and adds multiples of 2π to the following samples in order to bring the consecutive samples phase difference in the acceptable range of $-\pi$ to π.

C.   Image Reconstruction

We reconstruct both quantitative brightfield and phase-contrast images simultaneously from single-shot frequency-multiplexed interferometric measurements. The envelope and phase of the time-domain signal $I_PD(t_i, t_d)$ was firstly mapped into series of spatial information $I_{PD}(t_i, x_j)$, forming a line scanning brightfield image and phase contrast image, illuminated by the optical pulse at time t_i (Fig. 2b). This is because within each optical pulse, the spatial information is mapped one-to-one into spectral domain, $x_j \rightarrow \lambda_j$, and spectrum is stretch in time, $\lambda_j \rightarrow t_d^{(i)}$, where $t_d^{(i)}$ is the relative group delay time of each frequency component within a pulse with respect to the central wavelength. These line-scan images based on $I_{PD}(t_1, x)$, $I_{PD}(t_2, x)$, $I_{PD}(t_3, x)$, . . . were then cascaded into a two dimensional image corresponding to $I_{PD}(y, x)$, where the second dimension y is the spatial mapping of time elapse based on object flow speed.

The optical path length difference image can be formed by the phase shift line scans as

$$OPD(x,y) = \frac{\lambda(t_i, t_d)}{2\pi}\Delta\varphi(x,t) = \frac{\lambda(x,y)}{2\pi}\Delta\varphi(x,y) \quad (9)$$

On the other hand, if the axial thickness of the cell at reconstructed image pixel (x, y) is $d(x, y)$,

$$OPD(x,y) = 2[n_{cell}(x,y) - n_{solution}(x,y)] \cdot d(x,y) \quad (10)$$

in which n_{cell} and $n_{soluiton}$ are the refractive indices of the cell and the surrounding buffer solution, respectively. The factor 2 is to account for the fact that each wavelength component passes the cell twice in Michelson interferometer.

If we integrate Eq. (5) over the area of the cell, we can derive an average refractive index contrast for the cell, which corresponds to the average protein concentration of the cell:

$$\overline{\Delta n_{cell}} = \overline{n_{cell} - n_{solution}} = \frac{\iint_{cell} OPD(x,y)dxdy}{2V} \quad (11)$$

where $\iint_{cell} t(x,y)dxdy$ is the volume of the cell obtained from its diameter, d, as $V \approx \pi d^3/6$.

The combined scattering and attenuation coefficient of each cell, corresponding to the fluctuation of intensity induced by the cell, or optical loss, is obtained from the amplitude of the slowly varying envelope feature of the interferogram as

$$\Delta I = \frac{I_{PD}(t_i, t_d)}{I_{PD}(t_{background}, t_d)} \quad (12)$$

where $t_{background}$ corresponds to a measurement (an interrogation time) that there is no cell in the field of view, and only the background pulse shape is captured.

D.   Big data analytics pipeline

The high-content image analysis and cell screening pipeline is implemented by combining multiple informatics tools, namely CellProfiler for image processing [23, 24], MySQL/MangoDB for database, Matlab for machine learning, and Javascript for interactive visualization. Feature extraction from quantitative phase and intensity images leads to multiplexed hyperdimensional analysis. When monitoring a large amount of cells, it is not initially trivial which features are statistically significant for cell classification. However, once the learning is performed, a ranking of the significance of the parameters can be helpful for future tasks on the same type of samples.

In our pipeline first of all, image noise reduction and smoothing have been performed. Smoothing can remove artifacts that are smaller than optical resolution limit. Then for object segmentation, which is one of the most challenging steps in image processing and highly influences the downstream classification, we use the Otsu's thresholding method [ ]. Since the ratio of the number of foreground (objects or cells) pixels to that of the background pixels is small, and it normally does not vary substantially from image to image), the histogram of pixel gray levels is fitted into two classes, corresponding to the background and foreground pixels. The optimum threshold is calculated so that the total weighted variance within both classes is minimized or intra-class variance is maximized []. Once objects are identified in the image, morphology of each single cell can be described by area, diameter, uniformity, aspect ratio, perimeter, number of surrounding clumped cells, etc.

The capability to identify clumped cells from single large cells greatly reduces the misclassification rate in imaging flow cytometry compared to traditional flow cytometry. Intensity peaks of pixel brightness within each object are used to distinguish clumped objects. The object centers are defined as local intensity maxima in the smoothed image. Retaining outlines of the identified objects helps validate and visualize the algorithm. In the next step, we discard the objects touching the borders of the image, i.e., the edges of the field of view and data acquisition time window. However, the chance of cells showing up at the edges is very low due to hydrodynamic

focusing. We are also capable of excluding dust, noise, and debris by neglecting the objects that are too small or their aspect ratio is too extreme to be a cell. Size measurements are performed along both the rainbow direction and the cell flow direction. Measurement of the diameter along the rainbow direction is primarily used for cell classification since it helps to eliminate measurement inaccuracies caused by fluctuations of flow speed.

The intensity fluctuation caused by cell scattering and absorption can be extracted from the pixel intensities in the brightfield images ,and the phase shift caused by the cell protein concentration can be extracted from the pixel intensities in the phase contrast images. Due to limited optical resolution of the setup, the edges of bead or cell are blurred, generating the tails of point spread functions in phase and intensity signals outside of the diameter boundaries. In order to maximize the accuracy in morphological, phase, and loss measurements, after object segmentation, we expanded the object boundaries (Fig. 2, orange circles) by 2.5 um (optical resolution of the setup measured by knife-edge method), which serve as loose boundaries, indicating the area within which the pixel intensities are measured and integrated in phase and transmission images (Fig. 2, green circles).

To calculate the refractive index, the axial diameter along the interrogation optical beam is assumed to be equal to the lateral diameter along the rainbow direction, since cells in suspension relax to a spherical shape due to surface tension.

E. Calibration of size measurement

To calibrate the imaging system and image processing pipelines for size measurement, 5 μm polystyrene beads (from Polysciences, Inc.) with NIST traceable particle size standards were analyzed. Size measurement of the polystyrene beads had a distribution with 5.06 μm expected mean and 0.5 μm standard deviation. The broadened standard deviation was within the range of optical resolution limit and was caused mainly by performing object recognition on resolution limited images.

In comparison, phase shift images are robust and insensitive to axial defocusing caused by the limited depth-of-focus of the objective lens and alignment of the cells determined by the microfluidic channel structure. Any unscattered photon that passed through the cellshas the chance to be influential in phase images. This makes the size measurement in phase shift images relatively accurate and consistent.

F. Data cleaning

Data cleaning includes two steps. Firstly, Hotelling's T-squared distribution is calculated [ ] and top 2% of the extreme data was set as outliers due to experimental errors or object recognition errors. Secondly, debris

discrimination is performed; any data point with negative phase shift was considered as either air bubble, flow turbulence, or object recognition errors.

G. Microfluidic channel design and fabrication

The PDMS microfluidic channel is custom-designed so that it could fit into the reflective optics design. Cells are hydrodynamically focused [17, 18] at the center of the channel flowing at a velocity of 1.3 m/s. The microfluidic device consists of a hydrodynamic focusing region and an imaging region targeted by the interrogation rainbow flashes in TS-QPI system. At the hydrodynamic focusing region, the sheath pressure focused the sample at the center of the channel by narrowing its flow width from 200 μm to about 40 μm with a sheath to sample volume ratio of 3:1. The dimension of the channel was chosen as 200 μm (width) × 25 μm (height) so that the cells will be imaged within depth of view with a narrow lateral distribution. The size of the entire PDMS channel is optimized for fitting on a 2 inch diameter dielectric mirror with sufficient space at the edges to achieve strong bonding. The thickness of the channel top layer is optimized for stabilizing peek tubes performance reliability while accommodate the working distance of the objective lens.

The PDMS microfluidic channel is fabricated using standard soft lithography. The mask was designed in AutoCAD and printed with a resolution down to 1 um. Then a 4-inch silicon wafer was spin-coated with 75um thickness of a negative photoresist (SU-8 from MicroChem) and was exposed under the mask using an aligner. After post-exposure baking, the wafer was developed at room temperature, rinsed with isopropyl alcohol (IPA), and placed in a petri dish. A Polydimethylsiloxane (PDMS) mixture (Sylgard 184 Silicone Elastomer, Dow Corning) was poured onto the patterned wafer, degassed in a vacuum chamber for 30 min and cured at 80 °C for one hour. Once cured, the PDMS channel was cut out and peeled off from the master wafer. We used 1.25 um diameter hollow needle to punch the inlet and outlet holes. The punched PDMS channel was then cleaned with nitrogen gun and magic tape (3M), treated with oxygen plasma, and bonded to a 2-inch broadband dielectric mirror (Thorlabs BB2-E04) for obtaining high reflectance from channel substrate at near infrared spectral window. Finally microtubes (PE-50 tubing, .023x.038in) with steel catheter couplers (Instech, 22 ga x 15 mm) are connected to the inlet and outlet punctures.

H. Preparation of algae cell lines

Chlamydomonas reinhardtii strains used were *cw15* (*nit1 NIT2 mt⁺⁻*) and *sta6* (*cw15 nit1 NIT2 arg7-7 sta6-1::ARG7 mt⁺*), available as CC-4568, CC-4348

respectively from the Chlamydomonas resource center (CRC)[25].

Cells were grown in tris-acetate-phosphate (TAP) medium supplemented with arginine ($100\,\mu g\,mL^{-1}$). Cultures were grown in Innova incubators (New Brunswick Scientific, Edison, NJ) at $24\,°C$, agitated at 180 rpm with continuous light ($95\,\mu mol\,m^{-2}\,s^{-1}$, 6 cool white fluorescent bulbs at $4100\,K$ and 3 warm white fluorescent bulbs at $3000\,K$ per incubator). To induce lipid production, cells were cultured to mid-log phase in regular TAP prior to deprivation of N by transfer to ammonium-free (i.e. nitrogen-free) TAP medium, as described previously [26]. Briefly, cells subjected to nitrogen deprivation were grown to $4 \times 10^6\,cells\,mL^{-1}$ and collected by centrifugation at 1006 xg for 5 min at room temperature. The supernatant was discarded, and the cells were washed in nitrogen-free TAP. Cells were then resuspended in nitrogen-free TAP to a final cell count of $2 \times 10^6\,cells\,mL^{-1}$. Cell densities were determined using a hemocytometer.

[1] H. M. Shapiro, *Practical flow cytometry* (John Wiley & Sons, 2005).

[2] J. V. Watson, *Introduction to flow cytometry* (Cambridge University Press, 2004).

[3] S. P. Perfetto, P. K. Chattopadhyay, and M. Roederer, Nature Reviews Immunology **4**, 648 (2004).

[4] D. A. Basiji, W. E. Ortyn, L. Liang, V. Venkatachalam, and P. Morrissey, Clinics in laboratory medicine **27**, 653 (2007).

[5] D. A. Basiji and W. E. Ortyn, "Imaging and analyzing parameters of small moving objects such as cells," (2001), uS Patent 6,249,341.

[6] B. Razavi, *Principles of data conversion system design*, Vol. 126 (IEEE press New York, 1995).

[7] A. Mahjoubfar, K. Goda, G. Betts, and B. Jalali, JOSA A **30**, 2124 (2013).

[8] K. Goda, A. Ayazi, D. R. Gossett, J. Sadasivam, C. K. Lonappan, E. Sollier, A. M. Fard, S. C. Hur, J. Adam, C. Murray, *et al.*, Proceedings of the National Academy of Sciences **109**, 11630 (2012).

[9] A. Mahjoubfar, C. Chen, K. R. Niazi, S. Rabizadeh, and B. Jalali, Biomedical optics express **4**, 1618 (2013).

[10] K. Goda and B. Jalali, Nature Photonics **7**, 102 (2013).

[11] K. Goda, K. Tsia, and B. Jalali, Nature **458**, 1145 (2009).

[12] A. Mahjoubfar, C. Chen, K. Niazi, S. Rabizadeh, and B. Jalali, in *SPIE LASE* (International Society for Optics and Photonics, 2014) pp. 89720F–89720F.

[13] J. Kling, Nature biotechnology **30**, 578 (2012).

[14] T. Ikeda, G. Popescu, R. R. Dasari, and M. S. Feld, Optics letters **30**, 1165 (2005).

[15] G. Popescu, *Quantitative phase imaging of cells and tissues* (McGraw Hill Professional, 2011).

[16] H. V. Pham, B. Bhaduri, K. Tangella, C. Best-Popescu, and G. Popescu, PloS one **8**, e55676 (2013).

[17] J. B. Knight, A. Vishwanath, J. P. Brody, and R. H. Austin, Physical Review Letters **80**, 3863 (1998).

[18] G.-B. Lee, C.-C. Chang, S.-B. Huang, and R.-J. Yang, Journal of Micromechanics and Microengineering **16**, 1024 (2006).

[19] F. W. King, *Hilbert transforms*, Vol. 2 (Cambridge University Press Cambridge, UK, 2009).

[20] J. Revel, P. Hoch, and D. Ho, Experimental cell research **84**, 207 (1974).

[21] P. Whur, H. Koppel, C. Urquhart, and D. Williams, Journal of cell science **24**, 265 (1977).

[22] H. Verrelst, Y. Moreau, J. Vandewalle, and D. Timmerman, Advances in Neural Information Processing Systems , 978 (1998).

[23] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, *et al.*, Genome biology **7**, R100 (2006).

[24] L. Kamentsky, T. R. Jones, A. Fraser, M.-A. Bray, D. J. Logan, K. L. Madden, V. Ljosa, C. Rueden, K. W. Eliceiri, and A. E. Carpenter, Bioinformatics **27**, 1179 (2011).

[25] "Chlamydomonas Resource Center, University of Minnesota," http://chlamycollection.org/.

[26] I. K. Blaby, A. G. Glaesener, T. Mettler, S. T. Fitz-Gibbon, S. D. Gallaher, B. Liu, N. R. Boyle, J. Kropat, M. Stitt, S. Johnson, *et al.*, The Plant Cell Online **25**, 4305 (2013).

TABLE I. List of extracted features

Feature Name	Description	Category
Diameter-RB	Diameter along the interrogation rainbow. It is insensitive to flow rate fluctuation. For higher accuracy, it is calibrated by the spatial nonuniform distribution of rainbow wavelengths.	Morphology
Diameter-FL	Diameter along the flow direction. It is sensitive to flow rate fluctuation but can be a candidate parameter for monitoring flow speed and channel condition.	Morphology
Tight Area	Total number of pixels in the segmented region in the phase shift image	Morphology
Perimeter	Total number of pixels around the boundary of each segmented region	Morphology
Circularity	$4\pi\mathrm{Area}/\mathrm{Perimeter}^2$	Morphology
Major Axis	Considering the cell as elliptical in lateral imaging plane, the length of the major axis of the ellipse with a normalized second central moment same as the cell.	Morphology
Orientation	Angle between the flow direction and the major axis of the elliptical shape	Morphology
Loose Area	Total number of pixels in the expanded segmented region for measurement of the pixel intensities	Morphology
Median Radius	The median distance of any pixel in the object to the closest pixel outside of the object. [MeasureObjectSizeShape]	Morphology
OPD-1	Integrated optical path length shift within the entire segmented area, namely the cell. Calibrated by the power distribution within different wavelength components of the incident laser pulses.	Phase
OPD-2	Integrated optical path length within the entire segmented area, namely the cell. On top of OPD-1, calibrated by the waveforms when cell is absent within the 1 µs detection window.	Phase
Refractive index	The refractive index difference compared to buffer solution, which is calculated based on OPD and size measurement (see detail in appendix). Refractive index difference in cell is proportional to protein concentration.	Phase
Attenuation-1	Integrated attenuation coefficient assuming the detection is DC-coupled and the attenuation is mainly caused by transmitted light gets absorbed within the 20nm bandwidth. Calibrated by the waveforms when cell is absent within the 1 µs detection window.	Optical Loss
Attenuation-2	Integrated opacity assuming the detection is AC-coupled and the attenuation is mainly caused by transmitted light gets absorbed within the 20nm bandwidth. Calibrated by the waveforms when cell is absent within the 1 µs detection window.	Optical Loss
Scattering-1	Integrated scattering coefficient assuming the detection is DC-coupled and the attenuation is mainly caused by transmitted light scatters from the surface of the cell and granular cellular components. Calibrated by the waveforms when cell is absent within the 1 µs detection window.	Optical Loss
Scattering-2	Integrated opacity assuming the detection is AC-coupled and the attenuation is mainly caused by transmitted light scatters from the surface of the cell and granular cellular components. Calibrated by the waveforms when cell is absent within the 1 µs detection window.	Optical Loss