Using Machine Learning to Identify and Quantify Endotoxins from Different Bacterial Species using Liquid Crystal Droplets

Shengli Jiang[1,2], JungHyun Noh[1,3], Chulsoon Park[1,2], Alexander D. Smith[2],

Nicholas L. Abbott[3*], Victor M. Zavala[2*]

[2]Department of Chemical and Biological Engineering, University of Wisconsin- Madison,
1415 Engineering Dr, Madison, WI 53706, USA.

[3]Smith School of Chemical and Biomolecular Engineering, Cornell University,
113 Ho Plaza, Ithaca, NY 14853, USA.

[1]Equally contributing authors

[*]**Equal corresponding Authors:** victor.zavala@wisc.edu and nla34@cornell.edu

Abstract

Detection and quantification of bacterial endotoxin is important in a range of health-related contexts, including during pharmaceutical manufacturing of therapeutic proteins and vaccines. Here we combine experimental measurements based on nematic liquid crystalline droplets and machine learning methods to show it is possible to classify bacterial source (*Escherichia coli, Pseudomonas aeruginosa, Salmonella minnesota*) and quantify concentration of endotoxin derived from all three bacterial species present in aqueous solution. The approach uses flow cytometry to quantify, in a high-throughput manner, changes in the internal ordering of micrometer-sized droplets of nematic 4-cyano-4'-pentylbiphenyl triggered by the endotoxins. The changes in internal ordering alter the intensities of light side-scattered (SSC, large-angle) and forward-scattered (FSC, small-angle) by the liquid crystal droplets. A convolutional neural network (Endonet) is trained using the large data sets generated by flow cytometry and shown to predict endotoxin source and concentration directly from the FSC/SSC scatter plots. By using saliency maps, we reveal how EndoNet captures subtle differences in scatter fields to enable classification of bacterial source and quantification of endotoxin concentration over a range that spans eight orders of magnitude (0.01 pg/mL to 1 μg/mL). We attribute changes in scatter fields with bacterial origin of endotoxin, as detected by EndoNet, to the distinct molecular structures of the lipid A domains of the endotoxins derived from the three bacteria. Overall, we conclude that the combination of liquid crystal droplets

and EndoNet provides the basis of a promising analytical approach for endotoxins that does not require use of complex biologically-derived reagents (e.g., *Limulus* amoebocyte lysate).

Introduction

Endotoxins are lipopolysaccharides (LPS) found in the outer membranes of Gram-negative bacteria that are responsible for infections that lead to pathophysiological phenomena such as endotoxemia (septic shock caused by severe immune response).[1] Additionally, because the human immune system responds to LPS, all fluids and therapeutics (e.g., saline for rehydration or vaccines against viruses) injected intraveneously into humans must be tested for LPS during manufacturing.[2] The molecular structure of bacterial LPS (Figure 1a) consists of a lipid A group and polysaccharide domains (comprising so-called core and O-antigen regions).[3] LPS from different bacterial organisms possess distinct lipid A structures;[4-5] lipid A from *Pseudomonas aeruginosa* (PA), *Escherichia coli* (EC), and *Salmonella minnesota* (SM) possesses five, six or seven tails, respectively (Figure 1b-d).

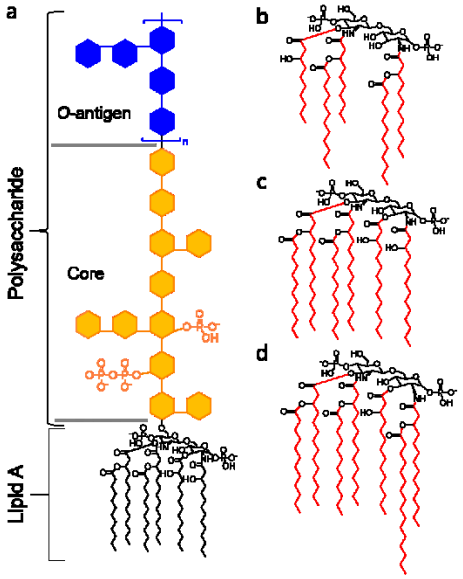

Figure 1. LPS from different bacterial sources. (**a**) Molecular structure of LPS from *Escherichia coli* (EC, O127:B8), showing O-antigens and core polysaccharides, and lipid A groups. (**b-d**) Structures of lipid A portions from *Pseudomonas aeruginosa* (PA), *Escherichia coli* (EC), and *Salmonella minnesota* (SM).

The gold standard for detection of LPS in current use is the *Limulus* amoebocyte lysate (LAL) test,[6] which is based on the blood of the horseshoe crab (*Limulus polyphemus*). The simplest version of this test relies on the formation of a clot upon exposure of LAL to LPS (called the LAL gel clot assay).[6-8] The LAL test can detect LPS at

concentrations as low as 3 pg/mL.[9] The analytical method is, however, expensive (horseshoe crab blood has a market value of 60,000 USD/gal) and not sustainable (hundreds of thousands of crabs are harvested each year to extract their blood and their population is declining). Faster, more accurate, and cheaper methods of analysis of LPS are actively being investigated.[10-14] The most sensitive approach reported so far detects LPS at concentrations as low as 8.7 fg/mL.[1] These analytical methods, however, are not able to identify the bacterial species from which the LPS is derived. Additionally, because they are based on enzymatic processes, they are susceptible to interference by salts and surfactants that are commonly present as part of pharmaceutical formulations.[15] In this paper, we report progress towards development of analytical methods that do not require use of biologically-derived reagents and permit identification of the bacterial source of the LPS.

The approach that we advance in this paper is based on the use of liquid crystals (LCs), which are oils within which molecules exhibit long-range orientational ordering (often referred to as nematic LC phases; Figure 2a-b).[16, 17] The long-range ordering of molecules in LCs generates anisotropic optical and mechanical properties, such as birefringence and elasticity.[18, 19] Past studies have shown that a range of external stimuli and interfacial interactions can change the orientational ordering of LCs, thus generating an optical response to the stimulus.[20, 22] In addition, when LCs are confined within micrometer-sized droplets that are dispersed in water (i.e., LC-in-water emulsions; Figure 2c-h), the geometrical confinement of the LC can generate nanoscopic defects (so-called topological defects) in LC ordering that can serve as hosts for molecular guests.[18, 19] Specifically, we reported previously that LPS from EC triggers changes in the ordering of LCs within micrometer-sized droplets at concentrations of LPS of ~1 pg/mL accumulation of LPS at topological defects induced in the LC by the geometry of the LC droplets.[16-17] When exposed to LPS, LC droplets transition from a so-called bipolar configuration (Figure 2c-e) to a radial internal configuration (Figure 2f-h). In the final radial configuration, a defect is located at the center of the LC droplet and fluorescence imaging reveals accumulation of the LPS at the defects.[16-17] The change in configuration results in a change of the optical appearance of an LC droplet viewed under crossed polars (Figure 2e, h).

In our past studies, by imaging LC droplets between crossed polars, we established that the fraction of droplets in a sample assuming a radial configuration (i.e., observation of characteristic cross as seen in Figure 2h) is strongly correlated to the concentration of LPS from EC, thus providing a basis for quantification of LPS concentration in aqueous solutions.[16] Although other amphiphiles, such as doubled-tailed phospholipids (e.g., 1,2-dilauroyl-sn-glycero-3-phosphatidylcholine (DLPC)) and single-tailed synthetic surfactants (e.g., sodium dodecylsulfate (SDS)) also induce configurational transitions in LC droplets, they are only observed at concentrations of 10 μg/mL or higher (i.e., 6 orders of magnitude higher in concentration than LPS).[16] The high concentration of DLPC and SDS required to achieve a configurational transition is because the internal

configurations of LC droplets induced by DLPC and SDS are driven by adsorption of the amphiphiles at the aqueous interfaces of the LC droplets (leading to surface-driven changes in the orientation of the LC, typically called surface anchoring transitions),[23-25] not accumulation of the amphiphiles at defects. Our past studies with LPS from EC and a synthetic amphiphile designed as a mimic of lipid A of EC (i.e., synthetic amphiphile with six tails) support the hypothesis that it is the lipid A component of LPS that leads to the accumulation of LPS at defects to trigger LC droplet responses as low concentrations.[23-25] In this work, we provide experimental evidence that micrometer-sized LC droplets dispersed in an aqueous solution (i) respond to the LPS of PA and SM, even though the molecular structures of the lipid A domains of LPS from these organisms differ from EC (Figure 1), and (ii) can be used to identify the bacterial source of LPS (from three different bacteria) and quantify LPS concentrations over eight orders of magnitude (0.01 pg/mL to 1 μg/mL). As described below, we achieve these advances by using a machine learning (ML) framework to analyze the optical response of the LC droplets (Figure 2c-h).

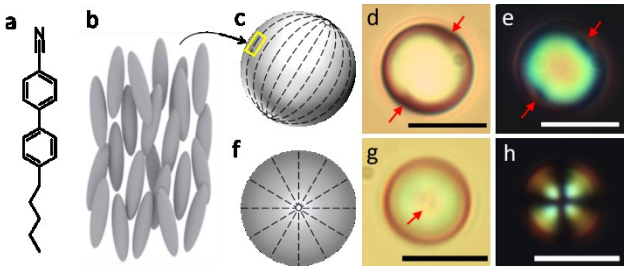

Figure 2. LPS-triggered configurational changes in LC droplets. (a) Molecular structure of 4-cyano-4'-pentylbiphenyl (5CB); (b) schematic illustration of nematic LC. (c-e) Bipolar configuration and (f-h) radial configuration of LC droplets; (c, f) schematic illustration of orientations of LCs, (d, g) bright-field optical micrographs and (e, h) polarized light micrographs between crossed polars. The arrows in d, e and g point to defects in the LC. Scale bars are 5 μm.

Our previous reports showing detection of LPS from EC using LC droplets relied on the use of optical microscopy (and tedious and slow procedures) to image the configuration of individual LC droplets.[16-17] In this paper, we go beyond our past methodology by quantifying the scattering of light from LC droplets, as measured by flow cytometry (Figure 3). This new approach to analysis of LPS provides high throughput characterization of LC droplets and the generation of data sets that are sufficiently large to enable ML approaches. In brief, LC droplets are injected into a fluid stream under laminar flow conditions;[28-30] after flow focusing into a beam of light from a laser, the intensity of light scattered by each LC droplet at both

small (0.5-15 degree, so-called “forward scattering”; FSC) and large scattering angles (15-150 degree, “side scattering”; SSC) is measured (Figure 3a).[23] The FSC signal is approximately proportional to the volume of the droplet and SSC is influenced by the internal ordering of the droplet.[28-30] Previously, we showed that characteristic features of FSC versus SSC plots (Figure 3b) correlate closely with the fraction of LC droplets in a sample that assume a radial configuration.[23] This approach, which we call the radial configuration (RC) method in the remainder of this paper, focused on a characteristic domain of the FSC/SSC scatter field that was observed to depend strongly on LC configurations (additional details are given under Methods), (Figure 3b). While the RC method of analysis of FSC/SSC light scattering during flow cytometry has been applied to characterization of LC droplets in the presence of a range of synthetic amphiphiles that trigger surface anchoring transitions,[23-27] it has not been used to characterize LC droplets in the presence of LPS. As noted above, the mechanism by which LPS changes the ordering of LC droplets differs from conventional surfactants and involves a finer scale of energetics,[16-17] and thus an additional key goal of our study was to determine if flow cytometric methods could be extended to detection of LPS using LC droplets.

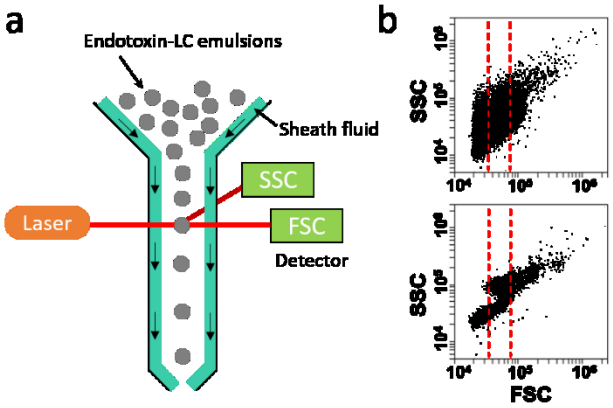

Figure 3. Analysis of LC droplets exposed to LPS using flow cytometry. (**a**) Illustration of LC emulsion dispersed in aqueous LPS being pumped into a flow cytometer in the direction of the sheath fluid flow. Light that is forward and side scattered from the LC droplets is collected at two angles (FSC and SSC). (**b**) Scatter plots for LC droplets in bipolar (top) and radial (bottom) configurations, where the radial configuration is generated by the adsorption of SDS. Two red dashed lines indicate the region within which the FSC intensity lies between 35,000 and 65,000 intensity units. The data shown was obtained using type A microcentrifuge tubes.

In this paper, we report that flow cytometric methods can be combined with LC droplets to detect LPS, and that

the methods permit detection of LPS from PA and SM organisms in addition to EC. We use ML techniques to automatically analyze FSC/SSC scatter data to determine LPS concentration (regression analysis) and show that ML techniques can also analyze scatter plots to identify the bacterial species from which the LPS is derived (classification analysis). The ML framework treats FSC/SSC scatter fields as images and uses a convolutional neural network (CNN) to extract features from such images at different resolutions (see Methods for additional discussion of the ML methodologies). Predictions are obtained using a fully connected neural network (FCNN), which uses features from the CNN to predict LPS species type and concentrations. We conduct extensive numerical tests using experimental data obtained with LPS from EC, PA, and SM to demonstrate the effectiveness of the analytical approach.

Methods

Experimental Methods

Materials. The following reagents were purchased from Sigma-Aldrich: LPS from *Escherichia coli* (EC, O127:B8), *Pseudomonas aeruginosa* (PA), *Salmonella minnesota* (SM), phosphate-buffered saline (PBS), and sodium dodecyl sulfate (SDS). The nematic liquid crystal 4-cyano-4'-pentylbiphenyl (5CB) was purchased from HCCH (Jiangsu Hecheng Display Technology). *Limulus* amoebocyte lysate (LAL) reagent water was purchased from Associates of Cape Cod. Ethanol (200 proof) was purchased from Pharmco-AAPER. Neptune pipette tips (validated to have no detectable LPS) were purchased from Continental Lab Product. Disposable glass culture tubes were purchased from VWR (catalog number 47729-570). Polystyrene tubes (catalog number 14-959-2A) and 2.0 mL microcentrifuge tubes (polypropylene) were purchased from Fisher Scientific (catalog numbers 02-681-291 (described below as tube Type A), 14-666-315 (described below as tube Type B), and 022363352 (Eppendorf brand)). Deionization of a distilled water source was performed with a Milli-Q system to provide water with a resistivity of 18.2 MΩ·cm.

Preparation of LC emulsions. The LC-in-water emulsions were prepared by adding 1.5 µL of 5CB and 3 mL of 5 µM SDS solution prepared in PBS buffer (10 mM phosphate-buffered saline at 137 mM NaCl, 0.27 mM KCl, pH 7.4 at 25℃) to a 12×75 mm disposable glass culture tube. Subsequently, three cycles, each comprising 30 seconds of vortex mixing at 3000 rpm, were performed to form a milky emulsion. The LC emulsions were used within 1 hour of preparation. Details regarding the preparation of the SDS solutions in PBS are given in Supporting Information (SI).

Preparation of aqueous dispersions of LPS. 1 mg of powdered LPS was dissolved in 1 mL of LAL reagent

water at room temperature (using a 2 mL Eppendorf-brand microcentrifuge tube). The resulting solution was sonicated for 5 seconds and then mixed by vortexing at 3000 rpm for 30 seconds. The sequence of sonication and vortex mixing were performed once more for 5 and 60 seconds, respectively.

Each run of experiments contained 9 LPS concentrations ranging from 0.01 pg/mL to 1 mg/mL, and two control samples (1 mM SDS to generate reference radial configurations and 5 µM SDS to generate reference bipolar configurations). Prior to dilution to achieve the desired LPS concentrations, each stock solution of LPS (1 mg/mL) was mixed by vortexing at 3000 rpm for 5 seconds. As detailed in SI, we performed the dilutions using two different types of microcentrifuge tubes (type A or B). Type B microcentrifuge tubes were cleaned by immersion in ethanol. Specifically, 300 microcentrifuge tubes were immersed in 3 L of ethanol in a plastic bottle (bottle used by the supplier of the anhydrous ethanol). The ethanol was replaced with fresh ethanol every 24 hours. The total time for cleaning was 72 hours; afterward, the microcentrifuge tubes were rinsed with ethanol and dried with nitrogen. Microcentrifuge tubes of type A were used as purchased. When performing the LPS dilutions using type A microcentrifuge tubes, we found the optimal aqueous diluent phase to be PBS whereas the optimal diluent phase when using type B microcentrifuge tubes was PBS to which was added 27 □M SDS.

After serial dilutions were performed to reach a final LPS concentration of 0.01 pg/mL, each solution was vortexed for 50 seconds, and then 700 µL of each diluted solution was transferred to a polystyrene tube. Finally, 35 µL of LC emulsion (prepared as described above) was added to 700 µL of each LPS solution and the LPS-LC emulsions were incubated for 30 minutes, prior to flow cytometry measurement.

Characterization of LC droplets by Flow Cytometry. LC emulsions were pumped through a BD Accuri C6 flow cytometer at a flow rate of 14 µL/min. Scatter fields were obtained by measuring the intensity of forward light scattering (FSC) and side light scattering (SSC) at detection angles of $0^{\circ} \pm 15^{\circ}$ and $90^{\circ} \pm 15^{\circ}$, respectively (Figure 3a). All flow cytometry measurements were performed at room temperature. SSC/FSC scatter fields were obtained by measuring intensities for 10,000 LC droplets.

Computational Methods

The working hypothesis that motivated this work was that ML algorithms can be used to automatically extract non-obvious sources of information from FSC/SSC fields to predict bacterial species from which LPS is derived and to quantify LPS concentrations. The ML framework described below uses CNNs and FCNNs to simultaneously conduct classification and regression tasks and we use additional techniques (radial configuration counting, support vector machines, and saliency maps) to validate and gain insights from our results.

Radial Configuration (RC) Method. The scatter plots (FSC versus SSC) obtained from flow cytometry are a 2D probability density function that describes the probability (frequency) that a given LC droplet has a certain FSC/SSC value. Figure 3b provides representative examples of scatter plots for LC droplets in bipolar and radial configurations, where the bipolar configuration is generated by using an aqueous solution containing 5 µM SDS and the radial configuration is generated by the presence of 1mM SDS (see Methods for details). As detailed in our prior studies,[23] the so-called radial configuration (RC) method is a technique that seeks to extract dominant features of the FSC/SSC scatter field. This method estimates the percentage of LC droplets that assume a radial configuration in an emulsion. For a given sample, the method measures the number of light scattering events in the FSC/SSC field within the region between 35,000 and 65,000 units of FSC (as shown by the two red lines in Figure 3b; the region between the lines has a higher probability density than the region outside the red dashed lines). This number of events is compared against the number of events in the same region of the FSC/SSC field for a sample of bipolar LC droplets dispersed in 5 µM SDS solution (negative control) (Figure 3b) and a sample of radial LC droplets dispersed in a 1 mM aqueous solution of SDS (positive control) (Figure 3b). From this data, the % radial feature was calculated as

$$\%\,Radial \;=\; 100\% \;\times \frac{\sum_{FSC=35{,}000}^{65{,}000} Count \;-\; \sum_{FSC=35{,}000}^{65{,}000} Count_{negative}}{\sum_{FSC=35{,}000}^{65{,}000} Count_{positive} \;-\; \sum_{FSC=35{,}000}^{65{,}000} Count_{negative}} \qquad (1)$$

The FSC range used to characterize the scatter field is described below as capturing a characteristic "S-region". The % radial feature is a single number (scalar) that has been found to correlate strongly to the fraction of radial LC droplets in a sample (as determined independently by optical microscopy) and surfactant concentration.[23] In this work, we sought to determine if it also correlates closely to LPS concentration. We also highlight that this feature embeds information from reference fields (positive and negative controls). To predict LPS concentration in measurements reported below, we feed the radial configuration percentage (% radial feature) to a FCNN.

Convolutional Neural Networks (CNNs). In our ML approach, we treat the FSC/SSC scatter field directly as a single-channel gray-scale image. The image is generated by counting the number of scattering events in a 2D bin (a pixel). We generate the bins by partitioning the FSC and SSC dimensions into 50 segments (the image has $50 \times 50 = 2{,}500$ pixels). As with the RC method, for each sample, we use reference scatter fields that represent limiting behaviors: bipolar control (negative) and radial control (positive). We use the sample image and the negative and positive reference images to assemble an image with three channels (Figure 4). This approach seeks to magnify differences between the sample images and seeks to provide context. The three-channel image is fed to a CNN, which we call EndoNet. EndoNet was designed to contain a single convolutional layer, a max-pooling layer, and two fully connected layers (Figure 4). This simple architecture achieves high accuracy and

facilitates interpretability. The convolutional layer has 64 filters of size 3×3 and stride 1 and the max-pooling layer has filters of size 2×2 and stride 2. Each of the fully connected layers has 32 nodes and the activation functions between layers are rectified linear units (ReLUs). The CNN structure can conduct both regression and

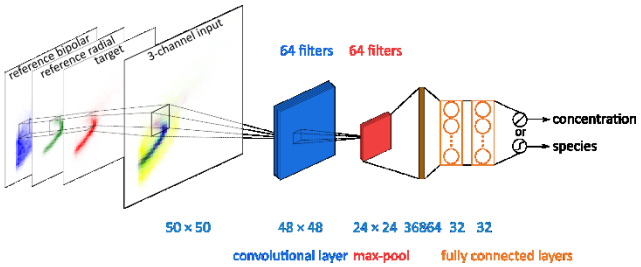

Figure 4. The Architecture of EndoNet. The input to EndoNet is a tensor $\mathbb{R}^{50\times50\times3}$ that consists of three channels: the target sample $\mathbb{R}^{50\times50}$, the reference bipolar sample $\mathbb{R}^{50\times50}$ and reference radial sample $\mathbb{R}^{50\times50}$. The CNN contains a convolutional layer with 64 3×3 filters and a max-pooling layer with 2×2 filters. The feature map generated from the max-pooling layer is a tensor $\mathbb{R}^{24\times24\times64}$ that is flattened into a long vector $\mathbb{R}^{36864}$. This vector is passed to a two-layer fully connected network (each having 32 hidden units). The regression task uses a linear output layer and the classification task uses a softmax output layer.

classification tasks (the final fully connected output layer is changed). In other words, the same features extracted by the CNN are used for both regression and classification. The output for species classification is a vector in $\mathbb{R}^3$ (corresponding to the probability of the LPS being from a specific species) while the output for concentration regression is a scalar in $\mathbb{R}$ (corresponding to concentration). To demonstrate the efficiency of EndoNet, we compared its performance against that of an ML approach that uses a pre-trained CNN (VGG-16) to extract features from the image and perform classification and regression separately. Here, we feed the features of the fifth convolutional layer of VGG-16 to a fully connected network. The fully connected network has the same structure as the one used in EndoNet. We expected this approach to be faster (as it does not require training the CNN) but we sought to determine if the features extracted would be as effective because the CNN was trained using generic images from the internet. In both architectures, we use 20% of the data for validation and conduct a five-cross validation procedure (See SI for details, Figure S1). All the training and testing tasks were performed in PyTorch.

Support Vector Machines. To understand the dominant features of FSC/SSC fields that are important for LPS classification and regression using the ML approach, we extract the 64 trained convolutional filters from EndoNet and apply the filters to each image to calculate the output feature map weight. If a filter captures the pattern in an image perfectly, the output weight value will be large. We feed all the output feature weights (a total of 64) into a support vector machine (SVM) to find the most important filters and associated patterns.

Saliency maps. A saliency map is a tool used to determine the domains of an image that drive predictions.[31] The saliency map is a matrix $M \in \mathbb{R}^{50\times50}$ of the same size as the input target image that is obtained via back propagation. Specifically, we calculate the gradient of the loss function with respect to the actual label via back propagation. The saliency map M is then obtained by finding the dominant elements of the gradient (sensitivity) in each pixel and in each channel. We apply the mask of the saliency map to the original image to highlight the regions important for classification and/or regression.

Results and discussion

Flow Cytometry and LPS Detection

The first goal of our experiments was to test the hypothesis that flow cytometry could be used in combination with LC droplets to detect LPS and to determine LPS concentration. As noted above, prior studies have reported detection of synthetic amphiphiles using LC droplets and flow cytometry,[23-27] but those synthetic amphiphiles trigger ordering transitions in LC droplets via a mechanism that differs from LPS.[16, 17] To address our initial goal, we used LPS from EC, as our past studies based on optical microscopy showed that LC droplets undergo ordering transitions in the presence of LPS from EC.[23] As shown in Figure 5a, we found that the concentration of LPS from EC in a sample influenced FSC/SSC scatter fields generated using LC droplets and flow cytometry in qualitative ways that are similar to those of prior studies based on synthetic surfactants (high LPS concentrations promote radial configurations).[23-27] In contrast to prior studies with synthetic surfactants, however, we found that subtle differences in scatter fields emerge with LPS concentrations that are as low as 0.01 pg/mL (10 fg/mL) (Figure 5a-b). While performing these experiments at low LPS concentrations, we determined also that the type of polypropylene tube used to dilute the LPS had an impact on our flow cytometry results (consistent with the presence of chemical species such as mold release agents on the surfaces of the tubes). Accordingly, we subsequently established that extraction of the polystyrene tubes (see Methods section), along with adjustment of the SDS concentration present in the buffer used to dilute the LPS (see Methods), led to similar responses of LC droplets to LPS concentration. Below, we present the results obtained using these two experimental procedures, which reveal that our methods are robust to variations in procedures.

A second key goal of our experiments was to determine if LPS derived from different bacterial organisms triggered ordering transitions in LC droplets (the previous experiments were performed only with LPS from EC)[16, 17]. Our past studies establish that the LC droplets are responding to the lipid A portion of LPS and, as shown in Figure 1b-d, the structure of the lipid A portion of LPS is dependent on the bacterial source (the number of tails

varies with bacterial species). Figure 5c shows experimental results revealing that LPS obtained from all three sources of bacteria (from PA, SM and EC) do trigger ordering transitions in LC droplets, generating qualitatively similar scatter plots (FSC versus SSC). Below we show that ML techniques applied to these scatter fields permits accurate classification of the bacterial origin of the LPS as well as concentration of the LPS.

Radial Configuration (RC) Method

Figure 5d shows the relationship between the radial configuration percentage (% Radial) and the concentration of LPS derived from EC, PA and SM. Inspection of Figure 5d (lower panels) reveals that the average fraction of radial droplets increases with concentration of LPS for all three sources of LPS but that the level of scatter in the individual measurements (top panels; and as reflected in the standard deviations indicated by error bars in the lower panels) prevents identification of qualitative differences between the three plots (corresponding to the three different bacterial sources). As detailed in the Methods section, we used a FCNN to predict LPS concentration from the radial configuration percentage; Figure 6a and Table 1 show prediction accuracy measured in terms of the coefficient of determination (R^2), which quantifies the strength of the linear relationship between two variables. The R^2 for the RC method is estimated to have an average value of 0.68 (R^2 = 1 means perfect prediction). The best prediction is obtained for the LPS from PA while the worst prediction is obtained for the LPS from SM. Figure 6a also reveals that the predictions degrade significantly at high LPS concentrations. This arises because the radial configuration percentage saturates at high concentrations (it is difficult to distinguish

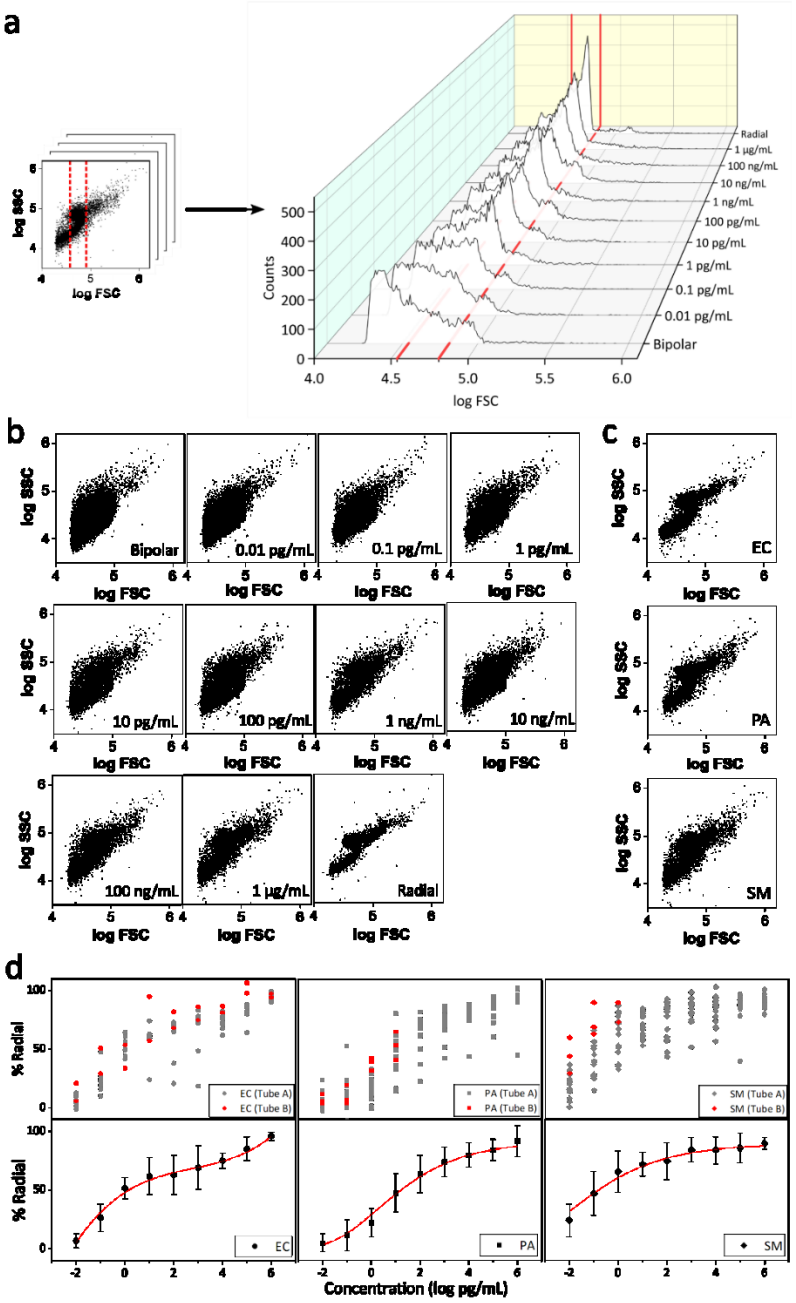

Figure 5. The effect of LPS concentration and bacterial source on FSC/SSC scatter fields. (**a**) Scatter field (left) generated by LC droplets exposed to 100 pg/mL of LPS from SM and (right) the marginal probability densities of FSC light intensity plotted as a function of LPS concentration (right). The region between two red lines is the characteristic "S-region". The number of points in the characteristic "S-region" increases with endotoxin concentration. (**b**) Scatter fields obtained using LC droplets exposed to various concentrations of LPS from SM. With increasing endotoxin concentration, the LC droplet population shifts from a bipolar to a radial configuration. The negative control (bipolar) was prepared using LC droplets dispersed in 5 µM SDS

solution and the positive control (radial) was prepared using LC droplets dispersed in 1 mM SDS solution. (**c**) Scatter fields obtained using LC droplets exposed to LPS obtained from different bacterial species at the same concentration (1 µg/mL). Type A microcentrifuge tubes were used to obtain data shown in a-c. (**d**) % Radial output obtained using LC droplets exposed to various concentrations of LPS derived from three different bacterial organisms: EC (left, N_A = 9, N_B = 2), PA (middle, N_A = 14, N_B = 3), and SM (right, N_A = 16, N_B =3). The top row shows scatter plots of repeat experiments (grey and red data points were obtained with type A and type B microcentrifuge tubes, respectively) and the bottom row shows average values of all data with error bars. Error bars are the standard deviations.

concentrations from the % Radial data alone). Overall, these results highlight limitations of the RC method. Because the scatter fields show perceptible differences outside the "S-region" with changing LPS concentrations (Figure 5b), as we discuss below, we conclude that the RC approach misses relevant patterns of FSC/SSC scatter fields that change with LPS concentration and bacterial origin. This point is illustrated by data in Figure 5c, which shows LC droplet responses to LPS from different species of bacteria. The CNN approach described below was pursued to address this limitation of the RC method.

EndoNet and VGG-16

As detailed in the Methods section, in contrast to the RC method, EndoNet analyzes the full range of data in the scatter plots of samples, along with their controls (bipolar and radial), as three-color images. Regression results for EndoNet are presented in Figure 6a and Table 1, along with comparisons to the pre-trained CNN VGG-16 and the RC method. The results reveal that EndoNet has a smaller error, which is estimated in terms of root mean squared error (RMSE), and higher R^2 compared to VGG-16 and the RC method. Specifically, EndoNet reduces the RMSE value of the RC method by 50% and decreases the RMSE value of VGG-16 by 27%. The regression results for VGG-16 are better than those of the RC method but worse than those of EndoNet. This indicates that VGG-16 captures patterns in the scatter fields that the RC method misses but also indicates that the filters of VGG-16 are not optimal (these have been obtained from images that are not related to our actual application).

Table 1. Summary of regression prediction accuracy for EndoNet, VGG-16, and RC method, expressed in terms of RMSE and R^2 (the latter is indicated within parentheses).

	RMSE(R^2)			
	EC	PA	SM	Average
EndoNet	0.83(0.87)	0.62(0.93)	0.76(0.91)	0.73(0.91)
VGG-16	1.20(0.73)	0.85(0.83)	1.03(0.86)	1.01(0.81)
RC	1.54(0.70)	1.09(0.82)	1.68(0.54)	1.43(0.68)

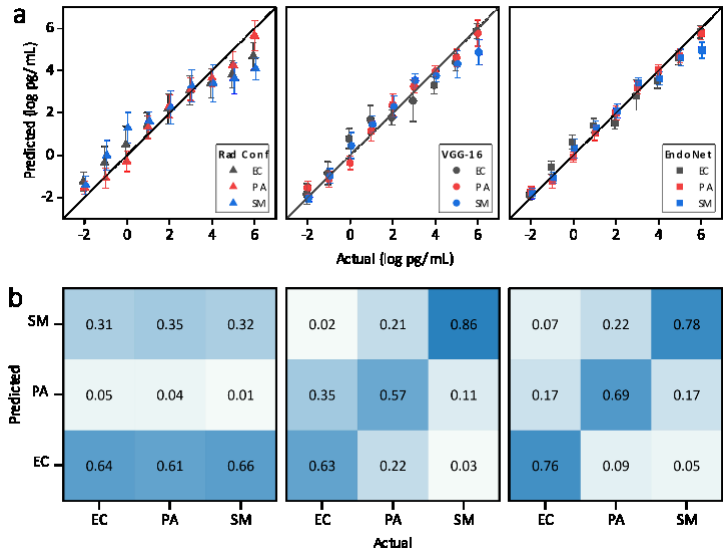

Figure 6. Prediction accuracies for RC method, VGG-16, and EndoNet. (**a**) Predicted and actual concentrations using RC method (left), VGG-16 (middle), and EndoNet (right). EndoNet predictions are closer to perfect predictions (black diagonal line). Within each individual prediction point (after 5-fold cross-validation) EndoNet has the smallest variance (the predictions are more stable). (**b**) Average confusion matrix for LPS classification using RC method (left), VGG-16 (middle), and EndoNet (right). Perfect agreement between predicted species and true species would be indicated by diagonal blocks containing the number one. For example, the first column in the EndoNet confusion matrix indicates that 76% of EC is accurately classified as EC but 17% is classified as PA and 7% as SM. EndoNet classifies species more accurately than VGG-16 and RC. The analysis is based on data obtained using type A and type B microcentrifuge tubes.

We interpret these results to indicate that EndoNet can extract pattern information within each channel and between channels of the input image (which include negative and positive controls). Capturing differences between channels provides context to the CNN and has the effect of highlighting domains in

the scatter field containing the most information. To validate this hypothesis, we conducted predictions for EndoNet using only one channel as input (the target image) and we ignored the positive and negative channels (obtained using bipolar and radial droplets). For this approach, the RMSE values for EC, PA and SM were 0.91, 0.80 and 0.82, while the RMSE values for the three-channel approach were 0.83, 0.62 and 0.76. In Figure S2, we present cumulative probability plots for the prediction error for the three species under study. EndoNet is consistently better than VGG-16 and the RC method. Specifically, over the entire concentration range, EndoNet has a higher probability of giving a smaller prediction error. We also see that the largest improvement is obtained for SM and the smallest improvement is obtained for PA.

To conduct classification tasks with EndoNet, we only changed the fully connected network layer (the input data and the convolutional layers are the same as those used for regression). We found that the classification accuracy (measured in terms of the f_1-score) remains high across the entire concentration range. After obtaining the precision and recalls for each group (bacterial species), we use a weighted average based on the number of samples to calculate the average precision and recall. From the average precision and recall, the final f_1-score is derived. Classification prediction results are presented in Figure 6b and Figure S3. The average classification f_1-score for EndoNet is 0.75; in contrast, the RC method achieves an average f_1-score of 0.2 (and improvement of 275%). This indicates that the characteristic "S-region" used in the RC method does not provide sufficient information to distinguish between different bacterial sources of LPS. For the RC method, we also observed a lower classification accuracy at higher LPS concentrations. This highlights that, at high concentrations, the FSC/SSC fields are more difficult to distinguish (reinforcing our observations obtained for regression). The confusion matrix for EndoNet indicates that EC is correctly predicted 76% of the time, PA is correctly predicted 69% of the time, and SM is correctly predicted 78% of the time. In comparison, the RC method can only correctly predict LPS type 64%, 4%, and 32% of the time, respectively. The confusion matrix for RC also reveals that the predicted results are highly biased towards EC.

Analysis of EndoNet Features

Figure 7a-b present saliency maps for EndoNet for regression and classification, respectively. Here, we make several observations regarding patterns evident in saliency maps. First, the saliency maps cluster (highlighted in red) around the characteristic "S-region" used in the RC method (the region between the red lines). This indicates that EndoNet searches for information in the same region as the RC method; however, the saliency maps also indicate that there is a clear diagonal pattern that is not considered in the RC method. The saliency map reveals that this diagonal pattern becomes more dominant at high concentrations. Specifically, from Figure 4a, we can see that the FSC/SSC field becomes more clustered as the LPS concentration increases. In Figure 7a, we also

see that the S region and the diagonal pattern are relevant in regression tasks.

Additionally, we analyzed the convolutional filters of EndoNet using a linear support vector machine (SVM). Specifically, we used a linear SVM to identify the four most important CNN filters (from a total of 64 filters). Figure S4 presents activated images associated with the dominant filters (these are obtained by applying the most important filters to the target image) for SM with a concentration of 1 pg/mL. We determined that the filter associated with the highest weight searches in the characteristic "S-region" and the second most dominant weight searches for the diagonal pattern. These results correlate closely with results obtained with saliency maps. The third and fourth dominant filters search for contrasts in the boundaries of the scattered field. We found that the

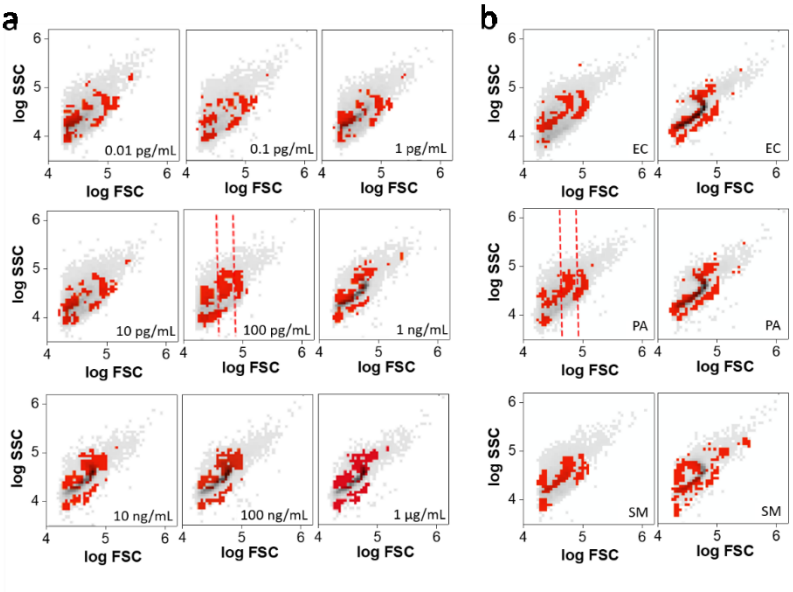

Figure 7. Saliency maps revealing dominant features of FSC/SSC scatter fields. (**a**) Saliency maps for concentration regression of LPS from SM. The region between the dashed lines is the characteristic "S-region" analyzed by the RC method. The saliency maps show that the characteristic "S-region" contains significant information but also indicate that other sources of information are present in the scatter fields. (**b**) Saliency maps for LPS classification under EndoNet. The left column is at 0.01 pg/mL and the right column is at 1 μg/mL. The region between the two red dashed lines is the characteristic "S-region". The features of the scattered fields that drive regression and classification are similar.

most dominant filter of EndoNet is given by the 3×3 matrix:

$$\begin{bmatrix} -0.02 & 0.14 & -0.22 \\ -0.16 & 0.10 & -0.20 \\ -0.16 & 0.10 & 0.07 \end{bmatrix}.$$

This filter corresponds to a difference operator along the x-axis (as evidenced by adjacent columns with positive and negative values), indicating that the dominant filter used by EndoNet to perform regression and classification is an edge detection filter for vertical edges (similar to a Sobel filter). In other words, the structure of this learned filter (the structure of the matrix) reveals that the CNN is picking out gradients (contrasts) in the scatter field along the x direction. This result is important because techniques typically used to characterize scatter fields from flow cytometry look only at the density of points in different regions of the field but do not look at the gradients (how quickly the density decays along the x or y direction). If the scatter field is viewed as a probability distribution in 3 dimensions (where the third dimension is the density), the gradient characterizes how quickly the distribution function is growing along the x axis. That is, the CNN is characterizing the shape of the probability distribution function. In summary, whereas the RC method involves counting the number of points along a certain vertical column in the S region (relies on frequency information), the dominant filter of EndoNet indicates that it searches for gradients in the number of points.

Conclusions

Our study reveals that aqueous dispersions of micrometer-sized LC droplets, when characterized using flow cytometry (light scattering mode) and analyzed using a ML approach, can be used to identify LPS from three bacterial organisms and predict their concentration in aqueous solution. The ML approach uses a CNN (that we call EndoNet) to process FSC/SSC scatter fields obtained from flow cytometry of the LC droplets. We show that this strategy leads to significant improvements in prediction accuracy compared to approaches that characterize local features of scatter fields (e.g., so-called RC method). Importantly, EndoNet can also perform classification tasks (i.e., identify from which bacterial species the LPS was derived). EndoNet achieves high accuracy over a concentration range that spans eight orders of magnitude and is found to be robust to changes in experimental procedures. Importantly, we note that concentrations that are as low as 10 fg/mL can be identified using EndoNet. This indicates that the approach is capable of recognizing subtle difference in scatter fields.

By analyzing saliency maps and the convolutional filters of EndoNet, we gained insight into the most important hidden features in the flow cytometry scatter fields that are analyzed by EndoNet. Specifically, we found that a characteristic "S-region" analyzed by the RC method embeds high information, but we also conclude that other dominant regions emerge at high concentrations of LPS or for particular bacterial species. The presence

of multiple dominant patterns explains why the RC method fails at high concentrations and is unable to perform accurate classification. The saliency maps also reveal regions that do not provide information for classification and regression (e.g., regions of high FSC and SSC values).

Our results generate several open questions for future investigation. Although we have succeeded in identifying important hidden features of the scatter plots using saliency maps and the convolutional filters of Endonet, we do not yet know which of these hidden features is used by Endonet to classify the bacterial source of the LPS. Additionally, while past work establishes both that the lipid A portion of LPS plays a central role in triggering the response of LCs droplets and that the lipid A portion of LPS differs in structure for EC, PA and SM, it is also possible that differences in the polysaccharide domains of the LPS may contribute to changes in scatter plots of LC droplets detected by Endonet.

More broadly, the results presented in this paper are particularly encouraging in terms of robustness, as predictions made by Endonet are tolerant to the presence of contaminants present on plastic tubes, and proteins do not cause changes in LC droplet configurations of the type triggered by LPS. Additionally, our past studies have reported that lipids directly shed by bacteria can also trigger changes in LC droplet configurations of the type reported in this paper.[32] These observations open new exciting avenues for identification of LPS from specific bacterial organisms and quantification of their concentration. Notably, this analytical endpoint can be achieved without requiring complex biological reagents (as in the LAL test) in a high-throughput format that leverages advances in confined soft matter and artificial intelligence. LC-based methods for validation of the absence of LPS contamination in simple process fluids (e.g., water), as commonly performed during pharmaceutical manufacturing, would be an impactful initial application. In the long term, the approach has the potential to impact clean water technologies and environmental monitoring.

Acknowledgments

NLA acknowledges support from the Army Research Office (W911NF-19-1-0071 and W911NF-15-1-0568) and National Science Foundation (CBET-1803409). VZ acknowledges support from the Wisconsin MRSEC.

Conflicts of Interest

There are no competing interests.

Notes and references

[1] Proctor, R. A. *Handbook of Endotoxins*, Vol. 2.; Elsevier: Amsterdam, 1984.

[2] Matthews, K. A.; Taylor, D. K. *J. Am. Assoc. Lab. Anim. Sci.,* **2011**, *50*, 708-712.

[3] Gutsmann, T.; Schromm, A.; Brandenburg, K. *Int. J. Med. Microbiol.*, **2007**, *297*, 341–352.

[4] Khan, M. M.; Ernst, O.; Sun, J.; Fraser, I. D.C.; Ernst, R. K.; Goodlett, D. R.; Nita-Lazar, A. *J. Mol. Biol.*, **2018**, *430*, 2641–2660.

[5] Proctor, R.A. *Handbook of Endotoxins*, Vol. 1.; Elsevier: Amsterdam, 1984.

[6] Cooper, J. F.; Levin, J.; Wagner, H. N. *J. Lab. Clin. Med.,* **1971**, *78*, 138–148.

[7] Muta, T.; Oda, T.; Iwanaga., S. *J. Biol. Chem.*, **1993**, *268*, 21384–21388.

[8] Ong, K. G.; Leland, J. M.; Zeng, K.; Barrett, G.; Zourob, M.; Grimes, C. A. *Biosens. Bioelectron.*, **2006**, *21*, 2270–2274.

[9] Roslansky, P. F.; Novitsky, T. J. *J. Clin. Microbiol*., **1991**, 29, 2477–2483.

[10] de Haas, C. J. C.; Haas, P.-J.; van Kessel, K. P. M.; van Strijp, J. A. G. *Biochem. Biophys. Res. Commun.*, **1998**, *252*, 492–496.

[11] Yeo, T. Y.; Choi, J. S.; Lee, B. K.; Kim, B. S.; Yoon, H. I.; Lee, H. Y.; Cho, and Y. W. *Biosens. Bioelectron.,* **2011**, *28*, 139–145.

[12] da Silva, J. S. L.; Oliveira, M. D. L.; de Melo, C. P.; Andrade, C. A. S. *Colloids Surf. B: Biointerfaces*, **2014**, *117*, 549–554.

[13] Oliveira, M. D. L.; Andrade, C. A. S.; Correia, M. T. S.; Coelho, L. C. B. B.; Singh, P. R.; Zeng, X. *J. Colloid Interface Sci.*, **2011**, *362*, 194–201.

[14] Limbut, W.; Hedström, M.; Thavarungkul, P.; Kanatharana, P.; Mattiasson, B *Anal. Bioanal. Chem.,* **2007**, *389*, 517–525.

[15] Wang, C.; Nelson, T.; Chen, D.; Ellis, J. C.; Abbott, N. L *J. Colloid Interface Sci.*, **2019**, *552*, 540–553.

[16] Lin, I.-H.; Miller, D. S.; Bertics, P.J.; Murphy, C. J.; de Pablo, J. J.; Abbott, N. L. *Science*, **2011**, *332*, 1297–1300.

[17] Miller, D. S.; Abbott, N. L. *Soft Matter*, **2013**, *9*, 374–382.

[18] de Gennes, P. G.; Prost, J. *The Physics of Liquid Crystals*, 2nd ed.; Clarendon Press; Oxford University Press: Oxford, New York, 1993.

[19] Kleman, M.; Laverentovich, O. D. *Soft Matter Physics: An Introduction*; Springer: New York, 2003.

[20] Bukusoglu, E.; Pantoja, M. B.; Mushenheim, P. C.; Wang, X.; Abbott, N. L. *Annu. Rev. Chem. Biomol. Eng.*, **2016**, *7*, 163-196.

[21] Nakayama, M.; Kajiyama, S.; Kumamoto, A.; Nishimura, T.; Ikuhara, Y.; Yamato, M.; Kato, T. *Nat. Comm.*, **2018**, *9*, 1-9.

[22] Zola, R. S.; Bisoyi, H. K.; Wang, H.; Urbas, A. M.; Bunning, T. J.; Li, Q. *Adv. Mater.*, **2019**, *31*, 1806172.

[23] Miller, D. S.; Wang, X.; Buchen, J.; Lavrentovich, O. D.; Abbott, N. L. *Anal. Chem.*, **2013**, *85*, 10296–10303.

[24] Tan, L. N.; Wiepz, G. J.; Miller, D. S.; Shusta, E. V.; Abbott, N. L. *Analyst*, **2014**, *139*, 2386-2396.

[25] Carter, M.; Miller, D. S.; Jennings, J.; Wang, X.; Mahanthappa, M. K.; Abbott, N. L.; Lynn, D. M. *Langmuir,* **2015***, 31*, 12850-12855.

[26] Adamiak, L.; Pendery, J.; Sun, J.; Iwabata, K.; Gianneschi, N. C.; Abbott, N. L. *Macromolecules,* **2018**, *51*, 1978-1985.

[27] Kim, Y.-K.; Huang, Y.; Tsuei, M.; Wang, X.; Gianneschi, N. C.; Abbott, N. L. *ChemPhysChem,* **2018***, 19*, 2037-2045.

[28] Givan, A. L. *Flow cytometry: First principles,* 2nd ed.; John Wiley & Sons, Inc.: New York, 2013.

[29] Melamed, M. R.; Lindmo, T.; Mendelsohn, M. L.; Bigler, R. D. *Am. J. Clin. Oncol.*, **1991**, *14*, 90.

[30] Shapiro, H. M. *Practical flow cytometry*, 4th ed.; John Wiley & Sons, Inc.: New Jersey, 2005.

[31] Simonyan, K.; Vedaldi, A.; Zisserman, A. arXiv:1312.6034, **2013**.

[32] Sivakumar, S.; Wark, K.L.; Gupta, J.K.; Abbott, N.L.; Caruso, F. *Advanced Functional Materials*, **2009**, 19 (14), 2260-2265.