Using Flow Cytometry and Multistage Machine Learning to Discover Label-Free Signatures of Algal Lipid Accumulation

Mohammad Tanhaemami[1], Elaheh Alizadeh[1], Claire Sanders[2], Babetta L. Marrone[2], Brian Munsky[1,3,*]

***Abstract*—Most applications of flow cytometry or cell sorting rely on the conjugation of fluorescent dyes to specific biomarkers. However, labeled biomarkers are not always available, they can be costly, and they may disrupt natural cell behavior. Label-free quantification based upon machine learning approaches could help correct these issues, but label replacement strategies can be very difficult to discover when applied labels or other modifications in measurements inadvertently modify intrinsic cell properties. Here we demonstrate a new, but simple approach based upon feature selection and linear regression analyses to integrate statistical information collected from both labeled and unlabeled cell populations and to identify models for accurate label-free single-cell quantification. We verify the method's accuracy to predict lipid content in algal cells (*Picochlorum soloecismus*) during a nitrogen starvation and lipid accumulation time course. Our general approach is expected to improve label-free single-cell analysis for other organisms or pathways, where biomarkers are inconvenient, expensive, or disruptive to downstream cellular processes.**

***Keywords*—Single cell, flow cytometry, machine learning, label-free quantification, microalgae**

I. Introduction

There are many biological research tasks for which it is important to measure single-cell behavior [1]. These tasks, which include cell counting, cell sorting, and biomarker detection, are widely conducted using flow cytometry (FCM) [1–3]. Flow cytometry is a high throughput analysis technique that performs rapid multiparametric analyses to inspect and quantify large cell populations and subpopulations [2–9]. FCM analysis is usually conducted by first fluorescently labeling cells, and then quantifying fluorescence intensity of individual cells within large populations. Each cell passes through a laser beam to excite fluorophores, and each cell's data is recorded by measuring emitted fluorescence intensity at longer wavelengths [5,7,9]. FCM also provides indirect measurements of cell phenotypes through measurements of intrinsic cellular properties, such as cell size and shape by forward-angle light scatter (FSC), and information about cellular granularity and morphology by side-scattered light intensity (SSC) [8,10]. In addition to quantifying cell populations, the related technique of fluorescence-activated cell sorting (FACS) allows researchers to separate cell populations into different subpopulations with respect to their individual properties [8]. As the name implies, sorting decisions are primarily based upon fluorescent labels [1,11].

Despite broad application of fluorescent labels in flow cytometry measurements [10], application of labels can be costly and may require unnecessary effort [12–14]. Labeling can also alter cell behavior and interfere with cellular processes and downstream analyses by causing activating/inhibitory signal transduction [13,15–19]. Additionally, some stains require cellular fixation or are toxic, which limits downstream processing when sorting [18,20]. A label-free quantification strategy could help prevent these adverse consequences by reducing operation costs and efforts, as well as avoiding side effects of using labels on cells [12,15]. In label-free quantification of FCM measurements, computational methods are used to quantify targeted cellular information based on measurements from other channels, i.e., from features.

Current label-free quantification strategies employ various methods of machine learning within their analyses to make use of large flow cytometry datasets [12,13,15,17,21,22]. However, in these strategies, the best intrinsic cellular features have been selected based solely on information collected from *fluorescently labeled* cells (for instance, see [12,21]). For some biological processes, if labels indirectly affect intrinsic cell properties within training populations, then these interactions could result in unexpectedly poor quantification of cell populations when tested on unlabeled cells. We hypothesize that FCM datasets could be used to develop label-free quantification strategies *even when signatures are weak and are perturbed* during the training process. In this work, we test our hypothesis by combining supervised machine learning algorithms with analysis of the distributions of single-cell data and their corresponding fluctuation fingerprints [23].

To demonstrate our approach, we conduct feature selection and regression analysis to find optimized label-free feature combinations and quantify lipid accumulation in microalgae cells, that can usually produce lipid content of 15% to 35% (potentially up to 80%), depending upon cultivation conditions, growth media, and algal species [24–26]. For such microalgae to become sources of alternative fuels, it will be necessary to monitor and maximize their ability to accumulate lipids [27]. To enable such quantification, we

[1]Department of Chemical and Biological Engineering, Colorado State University; Fort Collins, CO, USA.
[2]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA.
[3]School of Biomedical Engineering, Colorado State University; Fort Collins, CO, USA.
*Correspondence: munsky@colostate.edu

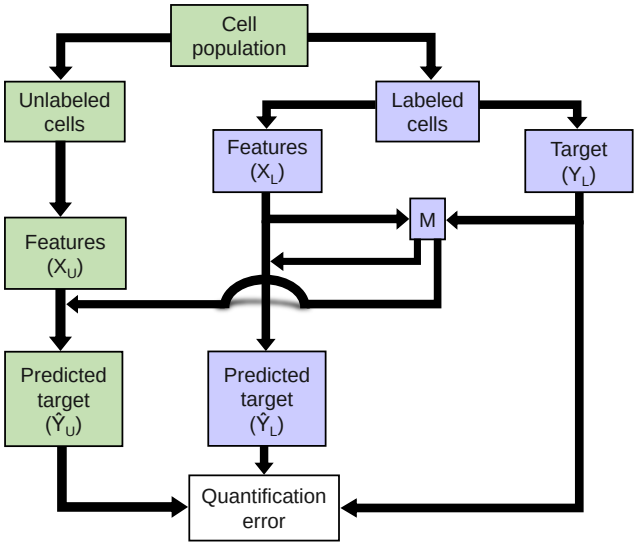

Fig. 1. Flow diagram of preliminary regression analysis to quantify lipid content based using intrinsic (presumably label-free) features. The model is learned using labeled data and then tested on both labeled and unlabeled data.

collect and examine FCM measurements of *Picochlorum soloecismus* under nitrogen replete conditions, and nitrogen deplete conditions that will stress cells and induce them to accumulate lipids. To measure lipid accumulation, we started with a traditional label-based strategy using BODIPY 505/515 fluorescent dye. We measured cell properties with and without the BODIPY stain, and we sought to find signatures in the latter preparation that are capable of reproducing quantities of the former preparation. Using these labeled and unlabeled data, we applied linear and nonlinear supervised machine learning algorithms to select the most informative features and predict lipid content. As opposed to current methods [12,13,15,17,21,22], we show that accurate label-free cell quantification requires rigorous incorporation of statistical information from biological experiments using both labeled and label-free measurements.

II. Results

Figure 1 depicts our initial strategy for label-free quantification. We monitored *P. soloecismus* microalgae for a total of 46 days following nitrogen starvation, and measured data using FCM at 23 different time points. At each time point, we created two identical subsamples as depicted at the top of Fig. 1. To obtain ground truth values for lipid accumulations, we labeled cells in one subsample using BODIPY, and we left the other one unlabeled. We measured the BODIPY signal in the labeled sample using a BD Accuri™ C6 flow cytometer for 10,000 labeled cells per sample. We also collected another set of FCM measurements for 60,000 to 136,000 unlabeled cells. Our FCM analyses recorded 13 features per cell, including the 488 nm excitation, 530/30 nm collection channel (FL1) corresponding to the BODIPY dye. We sought to predict the BODIPY signal intensities using other measured features – flow cytometry measurements of forward scatter (FSC), side scatter (SSC) and other fluorescence wavelengths (FL2 488 nm excitation, 585/40 nm collection, FL3 488 nm excitation, 670LP (long pass) collection, and FL4 640 nm excitation, 675/25 nm collection).

As described in the methods section, we sought to identify label-free quantification through several iterative training-validation strategies. First, we conducted a linear regression analysis on FCM measurements of labeled cells (the training step), and then the model was used to predict the lipid content of unlabeled *P. soloecismus* cells. The model was then applied to a different dataset gathered from labeled and unlabeled cells, and we evaluated the prediction accuracy using the Kolmogorov-Smirnov distance.

We performed training on three time points of our data. Time points corresponded to days 1, 14, and 46, which were were selected based on the lowest, the middle, and the highest BODIPY signal intensities. We then validated our model on another three time points corresponding to the second lowest, another middle, and the second highest BODIPY signal intensities (days 0, 15, and 37).

Figure 2 shows the results of applying the simple linear regression analysis using labeled data only. Figure 2(a) shows that at each time point the predicted labeled training data has a strong correlation with the measured data. Figure 2(b) suggests that a preliminary regression analysis provides a strong classification for the labeled training data, which was consistent in Fig. 2(c) for validation on labeled cells (KS distances between predictions and measurements for labeled cells were 0.0480, 0.0527, and 0.0190 for the three validation time points). However, the same regression model failed drastically when it was used to estimate the lipid content in the absence of labels, and Fig. 2(d) shows that the difference between predicted and measured values of the lipid content for unlabeled cells is extreme (KS distances were 0.9737, 0.9460 and 0.9233 for the same validation time points as above). Extended results for the linear regression are provided in supplementary Fig. S1.

To address the possibilities that we were overfitting the data or that linear regression was too simple an analysis to extract the informative label-free features, we also applied three more advanced machine learning approaches to learn lipid content from the intrinsic features: (i) *quadratic*, which corresponds to linear regression applied to linear and second order products of the original features (Methods and Fig. S2); (ii) *gradient boosting machine learning* (GBML) as utilized for label-free classification in Blasi et al. [12] (Fig. S3); and finally a *multilayer perceptron neural network* (MLPNN) [28] as shown in Fig. S4. To reduce effects of over-fitting, the latter two approaches (GBML and MLPNN) both employ cross-validation analysis on random partitions of the labeled training data. However, as shown in Figs. S2-S4, each of these advanced approaches appeared to work very well on the *labeled* training and validation data, but all were insufficient to predict the lipid content for *unlabeled* data.

To explain the failure of the labeled-cell-trained regression model on unlabeled cells, we suspected that some

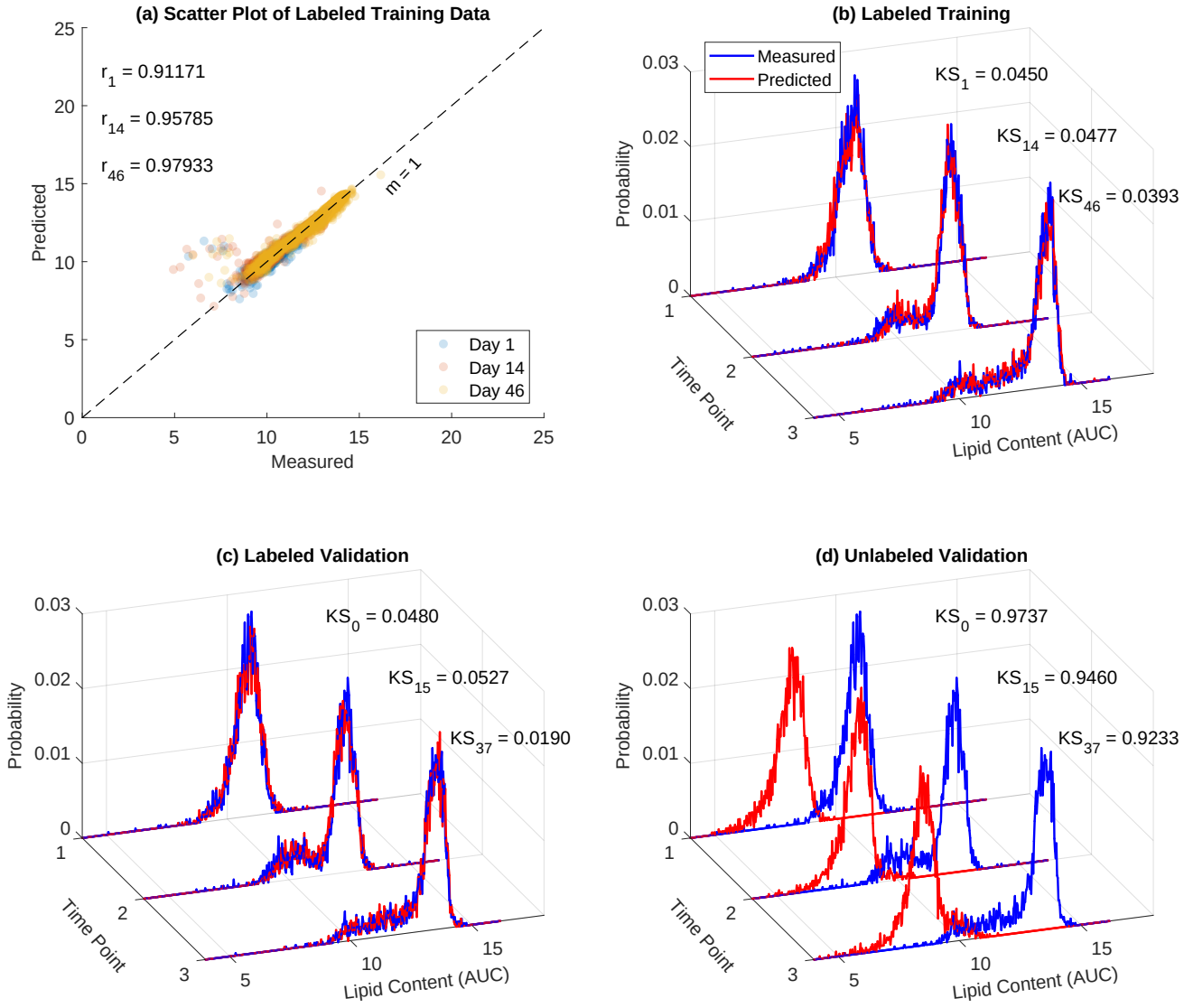

Fig. 2. Preliminary regression analysis. (a) Correlations between measured and predicted values of lipid content for labeled training data. Pearson's correlation coefficients are shown for each time point. (b) Histograms of lipid content for labeled training data. Measured in blue and predicted in red. Kolmogorov-Smirnov distances between the distributions are shown. (c) Histograms of the lipid content for labeled validation data. (d) Histograms of the lipid content for unlabeled validation data. Training data corresponds to days 1, 14, and 46; validation data corresponds to days 0, 15, and 37. All lipid content measurements are in arbitrary units of concentration (AUC). Bin sizes vary logarithmically.

channels in the flow cytometer might be adversely affected by application of the BODIPY stain. Indeed, Fig. 3 shows that some intrinsic features (FL2-A and FL2-H, corresponding to the second channel of the flow cytometer) change substantially when BODIPY is added to the cells. This channel is the closest to the FL1 channel that measures the lipid content, where the BODIPY fluorescent dye is added. Moreover, it is conceivable that the level of this disruption could be correlated with the amount of lipid in the cells, which means that it could be equally present in both training and validation data for the labeled cells. As a result, these changes could disrupt the training and cross-validation procedures and account for prediction failure when tested on unlabeled cells.

To mitigate this effect, we removed features FL2-A and FL2-H from the regression analysis and then repeated the linear regression. Figure 4(a-b) shows quantification results when the above two features are removed. We found that removing corrupted features led to substantial improvement for the quantification of unlabeled data (KS improved from 0.92-0.97 in Fig. 2(d) to 0.11-0.38 in Fig. 4(b)). The supplementary Fig. S5 provides extended plots of the outcomes of regression analyses upon removal of corrupted features. It is interesting to note that removal of disrupted features reduces accuracy of lipid prediction for labeled cells. This occurs because the labeling inadvertently modulates some "intrinsic" features in the labeled cells and introduces extraneous feature-target correlations that are actually detrimental to predictions for unlabeled cells. A troublesome consequence of these correlations between labels and intrinsic features is that these disrupted features are immune to removal when cross-validation analysis is

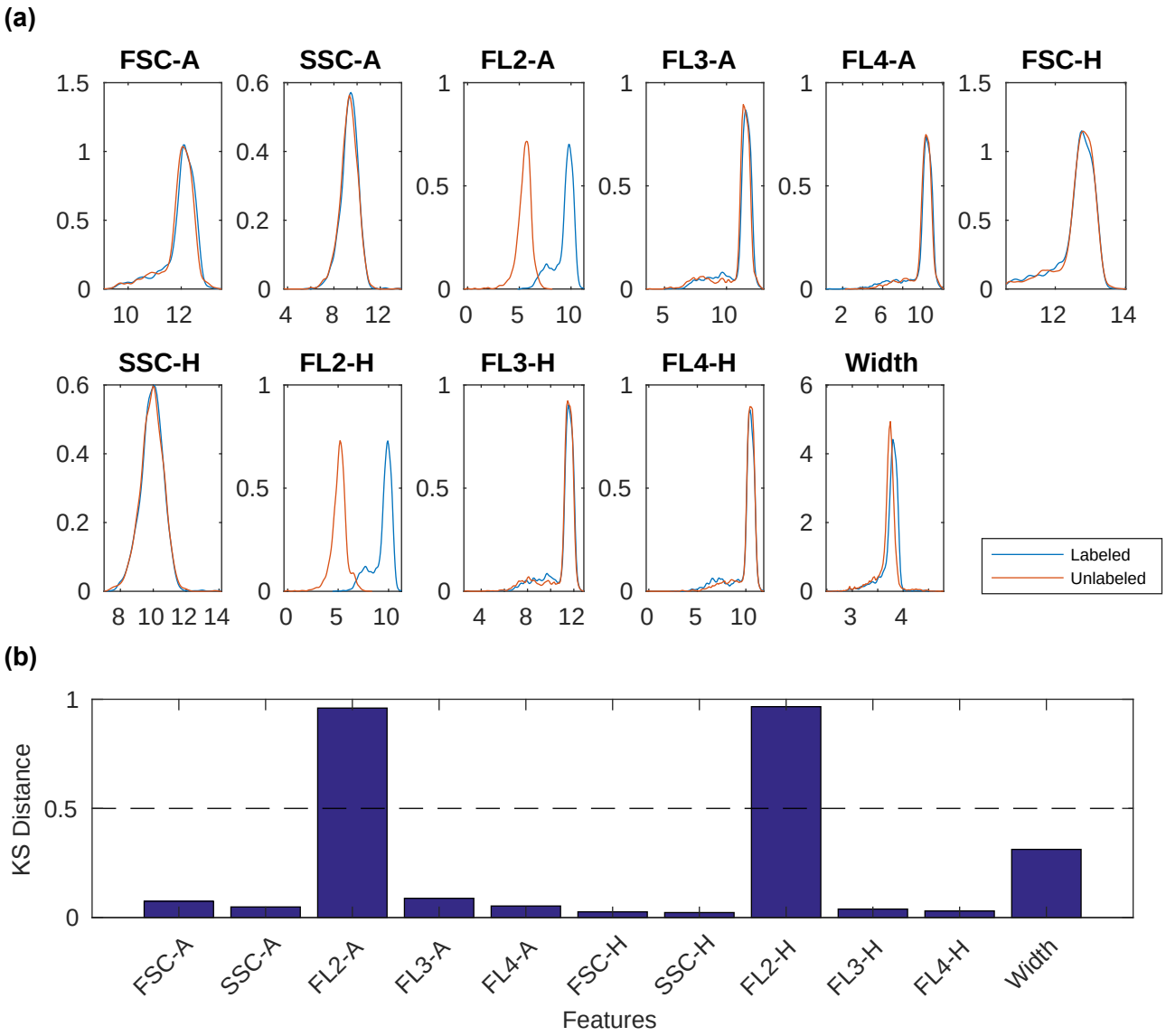

Fig. 3. Comparison of the features with and without BODIPY stain. (a) Kernel densities of features for labeled and unlabeled cells, averaged over all times. Labeled cells are shown in blue, and unlabeled cells are in red. (b) KS distance between labeled and unlabeled features distributions. FL2-A and FL2-H features show clear dependence on the BODIPY stain. Horizontal line denotes threshold used to remove corrupted features.

applied exclusively to labeled cells.

Next, we used the genetic algorithm on combinations of labeled and completely unlabeled data to explore if further feature reduction could enhance label-free classification. Figure 4(c-d) shows the results following the application of the genetic algorithm, which automatically selected FSC-A, SSC-A, FL3-A, FSC-H, and the width of the signal as the most informative features. Down-selecting to these most informative features resulted in a slightly smaller KS distance (0.10 - 0.35) between measured and predicted values of the lipid content for unlabeled cells. Extended results are provided in supplementary Fig. S6.

During automated feature selection for linear regression (Fig. 4(c-d)), we did not incorporate higher order effects (e.g., "interactions") between predictor variables. To enhance our modeling and potentially extract more information from the data, we added an expanded set of products of feature values to the input. As shown in Fig. 4(e,f), expansion of the input matrix of features to include quadratic and first order interaction terms, followed by label-free feature selection via the genetic algorithm, resulted in a slight improvement to label-free predictions for the lipid content. For more detailed results after introducing the quadratic features and application of the genetic algorithm on higher order effects, see Fig. S7 in the supplementary information. In this case, the genetic algorithm identified the product of FSC-A and FL4-H, the square of FSC-H, and the product of FL4-H and signal width as the most informative attributes. Selected features by the genetic algorithm on linear and quadratic features are presented in more detail in supplementary Table S1.

Finally, we introduced a new strategy based on weighted

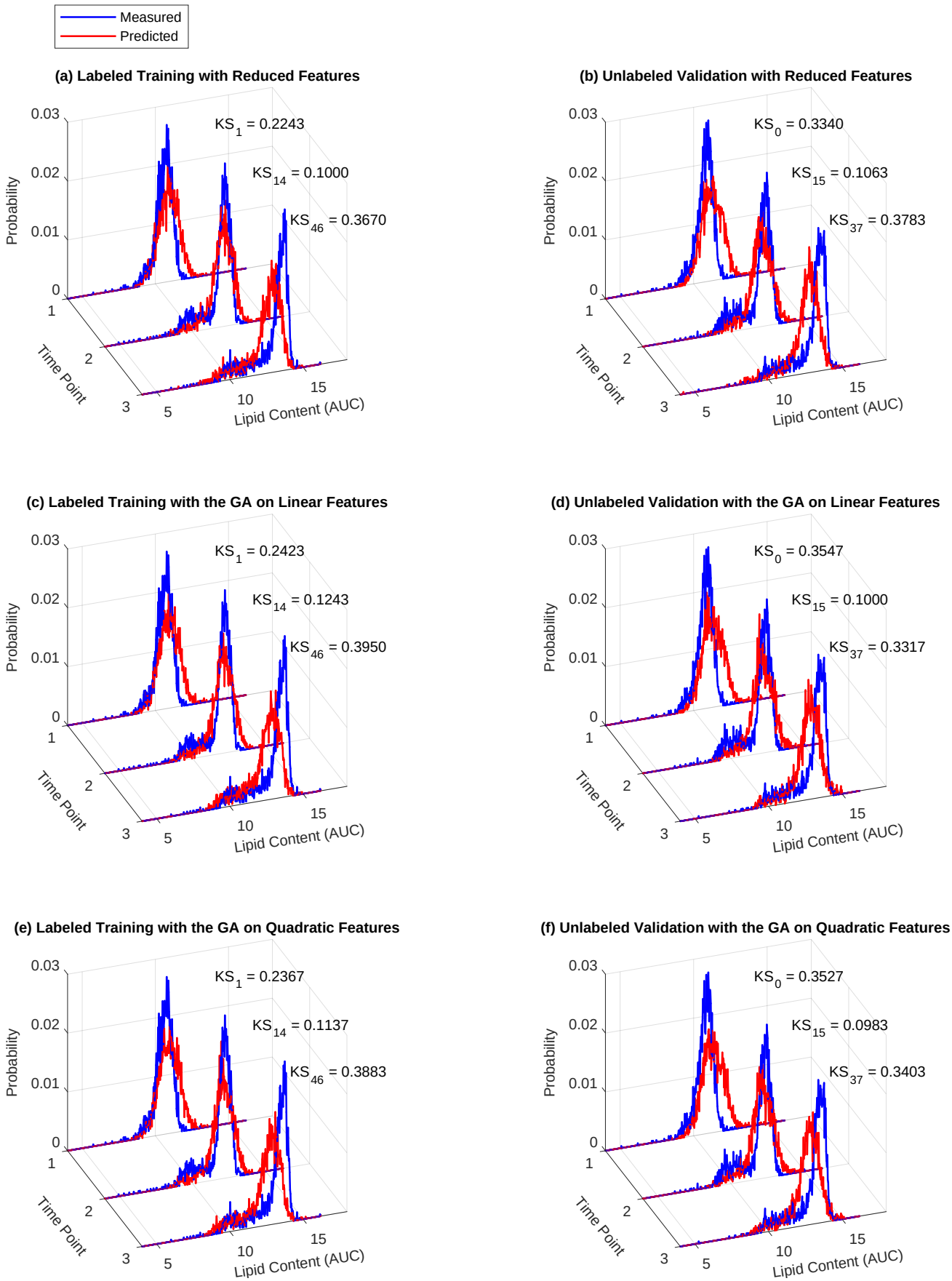

Fig. 4. Regression results after various approaches to feature selection. (a) Training on reduced features. (b) Validation of the model in (a) on unlabeled cells. (c) Training based on the features selected by the GA. (d) Validation of the model in (c) on unlabeled cells. (e) Training based on the features selected by the GA on quadratic features and interactions. (f) Validation of the model in (e) on unlabeled cells. For all cases, measured values are shown in blue and predicted in red. Kolmogorov-Smirnov distances between distributions are shown. Training data corresponds to days 1, 14, and 46; validation data corresponds to days 0, 15, and 37.

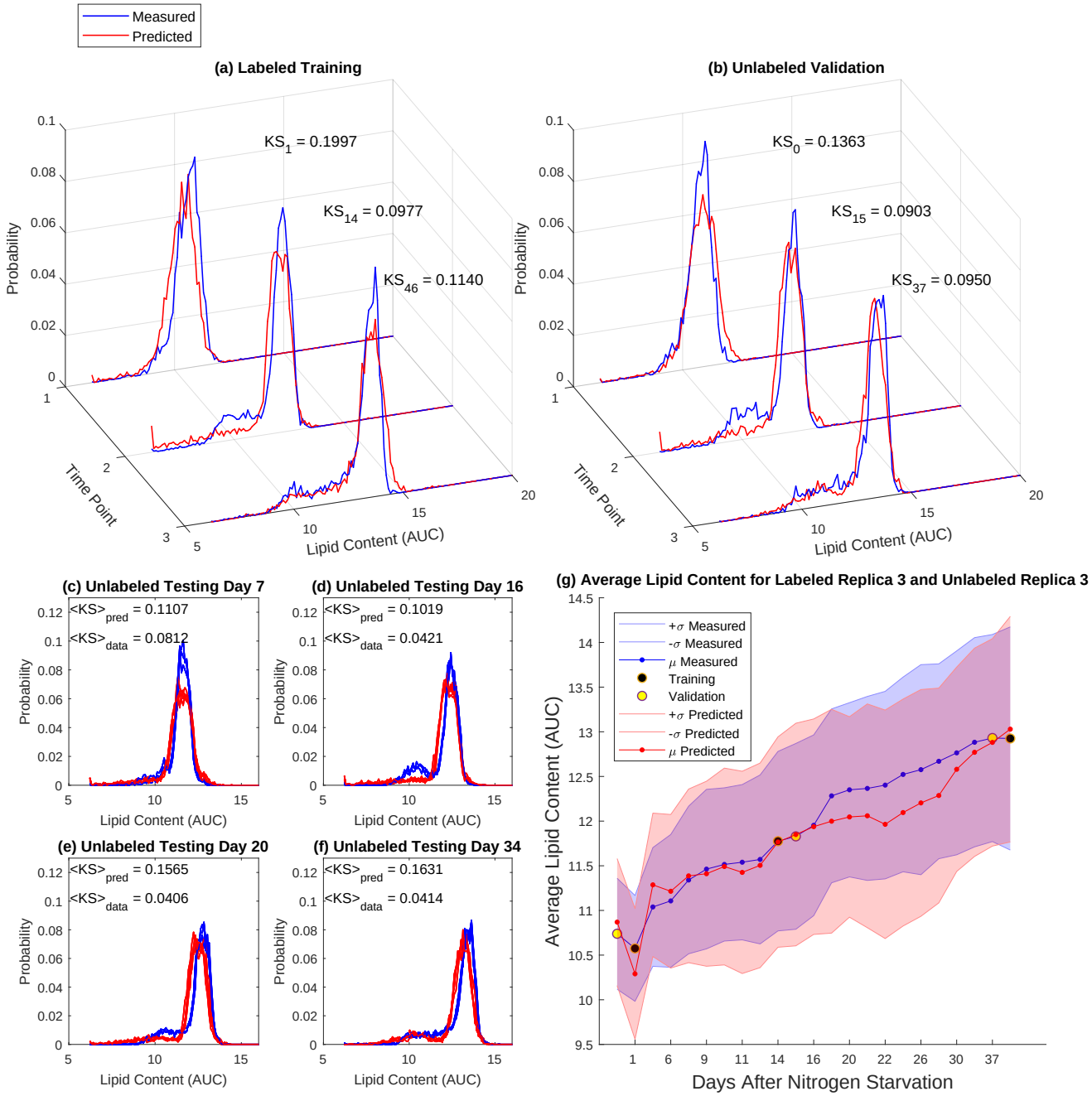

Fig. 5. Results of analysis. Distributions of lipid content for (a) labeled training data. and (b) unlabeled validation data. KS distances between distributions are shown. (c-f) Testing the final strategy on four unlabeled testing time points: Days 7, 16, 20, and 34. See Fig. S8 for corresponding results for all 17 testing time points. "KS data" is the average KS distance between measured lipid distributions. (g) Average lipid content at each day after nitrogen starvation. The blue and red shaded areas show the standard deviation as measured and predicted, respectively.

models (see Methods section). Our weighted model was formed by a linear combination of three models, each learned from labeled and unlabeled data at three training time points. The weights applied to these three models were estimated (using a secondary regression analysis) from measured statistics of the *unlabeled features*. Importantly, the re-weighting of the models allows incorporation of the 530/30 nm FCM channel, which was previously discarded due to the fact that it was needed for the measurement of BODIPY in the labeled cells.

Figure 5 shows the results of our new label-free quantification strategy for labeled cells (Fig. 5(a)) and unlabeled cells (Fig. 5(b-g)). It can be seen here that using a weighted modeling strategy based on statistics of unlabeled features enables the model to predict the BODIPY signal with a remarkably high accuracy. The expanded weighted model analysis allows for a substantially improved ability to quantify lipid content for both labeled and unlabeled cells. The very small KS distance (0.14, 0.09, and 0.09) on the three validation time points represent an exceptional

success in predicting the BODIPY signals based on label-free measurements.

For the final machine learning model, the genetic algorithm selected the product of SSC-A and SSC-H, the square of FL3-A, the product of FL4-A and SSC-H, and square of FL3-H as the most informative features for the construction of the regression analyses at the three training time points. Table S1 of the supplementary information presents these selected features in detail. For the secondary regression analysis used to define the weights of the regression analyses, the optimum found by the genetic algorithm relied on statistical information from all fluorescence channels (including the 530/30 nm channels that was previously discarded during labeled cells measurements). The selected columns of the test statistic are presented in supplementary Table S2.

After we validated the final label-free lipid estimation model, we fixed all parameters and sought to test it for label-free quantification on a much larger set of time points. The final model yielded exceptional prediction accuracy of the BODIPY signal for this previously unseen testing data, as can be seen in the predicted distribution of lipid content at specific time points (Fig. 5(c-f) and supplementary Fig. S8). Figure 5(g) also shows that the trained model correctly quantified average and standard deviation of lipid accumulation (in log scale) at each day following nitrogen starvation.

III. Conclusions

Single-cell quantification and classification are crucial tasks in many biological and biomedical applications, and flow cytometry (FCM) is one of the most common tools used for these tasks. Computational strategies have substantial potential to identify label-free markers and mitigate the expense or disruptive effects of traditional FCM analyses. In this article, we have demonstrated the use of mathematical tools and statistical methods, including regression analysis and machine learning to extract quantitative information from intrinsic properties of unlabeled cell populations. We discovered that computational classifiers that are learned using intrinsic features measured in labeled cell populations may appear to be highly predictive when compared to other labeled cells, but these same models may then fail dramatically when tested on truly label-free data (Figs.2 and S2-S4).

The key to our integrated strategy is careful consideration of the variations within heterogeneous single-cell populations. Drawing inspiration from our past work to identify gene regulation models from single-cell distributions [23,29,30], we reasoned that distributions of labeled and unlabeled cell populations should have shared statistics that could help to circumvent the issue of data corruption due to label applications. Under that inspiration, we developed a multi-stage regression approach that incorporates collections of both labeled and unlabeled data in the same conditions. From these data sets, we learn which features' statistics are conserved, which features vary between different treatments, and which features are most valuable to predict lipid content in unlabeled cells when trained using labeled cells. Figure 6 depicts a flow diagram of our new approach and its three main components of (i) linear regression applied to features and feature products to discover the correlations between intrinsic features and lipid content within labeled cells; (ii) genetic algorithms to automatically select features that contain useful information, but which avoid misleading or distracting artifacts contained within large FCM datasets; and (iii) a new model-weighting strategy to allow application of different statistical models in different situations.

The combination of regression analyses, genetic algorithms and model weighting approaches yields a final set of models and weights that are uniquely determined from the statistical properties of unlabeled cell population measurements. Using this approach, we can then extract sufficient information to provide efficient label-free quantification of lipid content in *Picochlorum soloecismus* over time during nitrogen starvation. Our final model accurately estimates lipid content distributions over time that span several orders of magnitude (Figs. 5 and S8). Moreover, although direct verification of lipid content for unlabeled single-cells is not possible, our final regression models preserved single-cell prediction accuracy for lipid content in labeled cells, especially at later time points when lipid content is highest (Pearson's correlation coefficient of R = 0.74-0.87; see Fig. S8).

Together, the proposed computational tools could help circumvent the need for biochemical labels to reduce expense and open new avenues for single-cell research. For example, label-free quantification will be instrumental to sort cells into different subpopulations, without the (potentially terminal) cellular disruptions associated with standard biochemical markers. Once trained through several rounds of regression and genetic algorithms, our final model for algal lipid quantification reduces down to a simple linear operation applied to a handful of 7 second-order products of features of the unlabeled cells. Such operations are easily computed in less than a microsecond per cell, making the label-free analysis ideal for use in gating and sorting applications as a stand-in for fluorescence in fluorescence-activated cells sorting (FACS) analyses. Such populations could then be instrumental in future advanced studies such as analysis with subsequent growth assays, application to directed evolution to improve productivity or yield, exploration of additional perturbation responses, and other assays that require live, unmodified cells for subsequent analyses.

IV. Methods

A. Cell preparation and flow cytometry measurements

P. soloecismus was grown in f/2 media containing half the recipe nitrogen and using Instant Ocean sea salt (Blacksburg, VA) at 38 g/L [31,32]. Cultures were grown at room temperature on a 16 hour light/8 hour dark cycle and mixed by stirring. PH was maintained at 8.25 with on-demand CO_2 injection when the pH

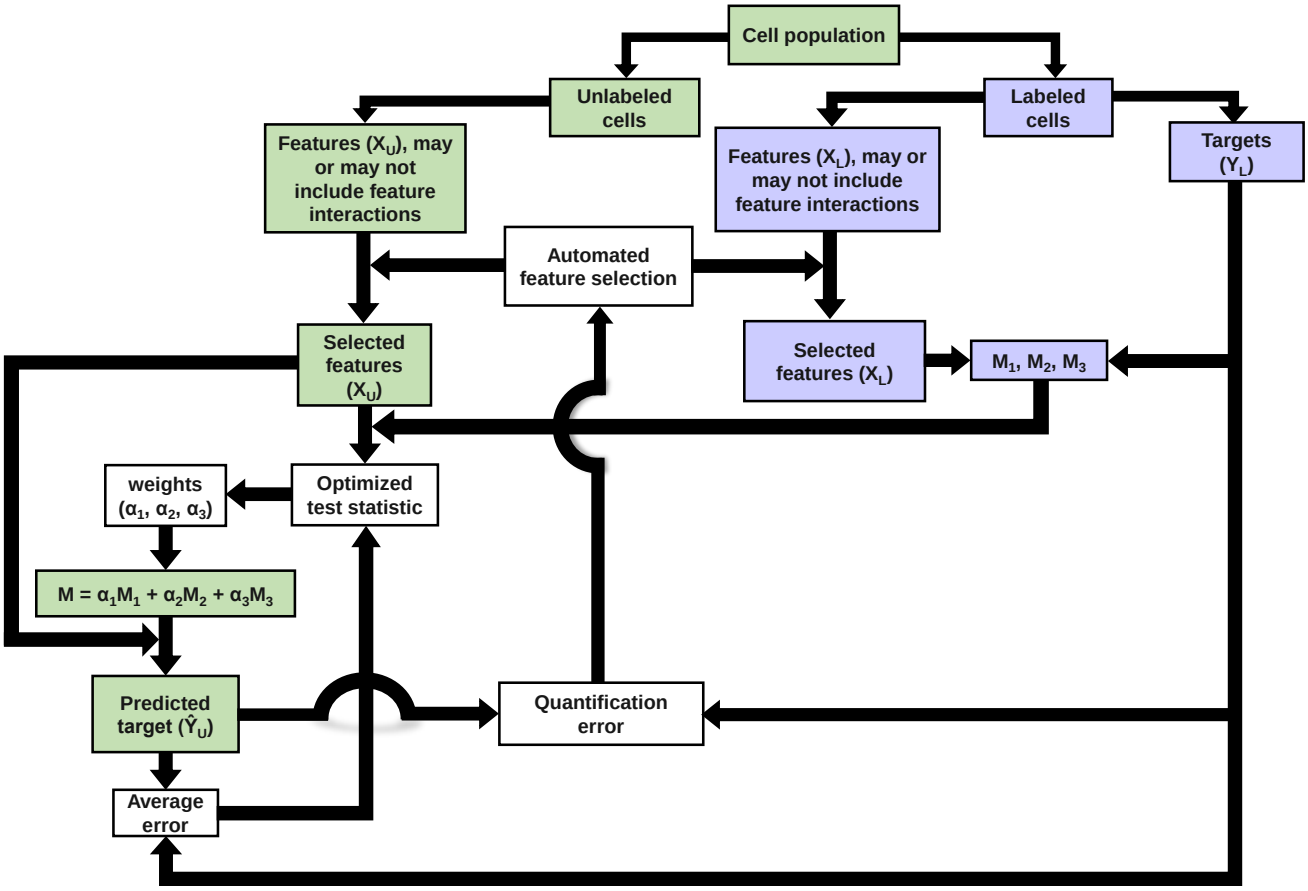

Fig. 6. Flow diagram of the final multi-stage label-free quantification strategy.

increased above the set-point. Cells were collected and stored at 4 °C prior to analysis.

Stained populations of cells were incubated with 22.6 µM BODIPY 505/515 (Thermo Fisher Scientific) with 2.8% DMSO in media for 30 minutes at room temperature prior to analysis. Analysis was conducted using a BD Accuri™ C6 flow cytometer with BD CSampler™ (BD Biosciences). Unstained samples were collected with a set volume of 10 µl on a high flow rate (66 µl/min), for stained samples 10,000 events were collected on a low flow rate (14 l/min). Data was exported in .csv format for subsequent analysis.

B. Linear regression analysis

In an initial attempt to identify label-free signatures of lipid content, we considered linear regression applied to match intrinsic features of labeled cells to lipid content (Fig. 1). In regression analysis, there are two main types of variables: the response variable (denoted y) and the explanatory variables (the set of predictors, denoted $\mathbf{x}$) [33]. In this study, the response vector is the accumulation of the lipid content for each cell (called the target) and the predictor is a matrix containing the data for intrinsic cellular properties measured by FSC, SSC, and other fluorescence wavelengths (called the features). In regression analysis, the response is approximated as a function of the predictors as

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \tag{1}$$

where $\mathbf{x}_i = (x_1, \ldots, x_N)_i$ is the vector of N intrinsic features for the i^{th} cell, and ε_i is a random measurement error for that cell [34]. In linear regression, the response (target) and predictor (feature) variables are assumed to satisfy the linear relationship [34]

$$\mathbf{Y} = \mathbf{X}\mathbf{M}, \tag{2}$$

where the vector $\mathbf{Y} = [y_1, \ldots, y_{N_c}]^T$ is the vector of targets for N_c training cells; $\mathbf{X} = [\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_{N_c}^{\mathrm{T}}]^{\mathrm{T}}$ is the corresponding matrix of features for the same cells; and $\mathbf{M}$ is the regression parameter or regression coefficient.

Linear regression provides a preliminary insight about potential relationships between the predictor and the response variables. After defining the features and the target, the regression coefficient that minimizes the sum of squared difference of $|\mathbf{Y} - \mathbf{X}\mathbf{M}|_2^2$ can be calculated as

$$\mathbf{M} = \mathbf{X}^{-L}\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3}$$

To perform a preliminary regression analysis, we first selected three *training* time points, corresponding to the lowest, the middle, and the highest BODIPY fluorescence intensities (in this experiment, days 1, 14, and 46, respectively). We chose these days to capture the greatest possible range of lipid accumulation phenotypes. For each time point, we considered FCM measurements from a random set of 3000 labeled cells. We computed the regression coefficient, $\mathbf{M}$, by Eq. (3) using the labeled data sets $\mathbf{X}_L^{(\text{train})}$ and $\mathbf{Y}_L^{(\text{train})}$. Next, we selected another three *validation* time points, corresponding to the second lowest, another middle, and the second highest BODIPY fluorescence intensities (in this experiment, days 0, 15, and 37, respectively). This time, we extracted information for both labeled, $\mathbf{X}_L^{(\text{valid})}$ and $\mathbf{Y}_L^{(\text{valid})}$, and unlabeled cells, $\mathbf{X}_U^{(\text{valid})}$. Using the $\mathbf{M}$ computed from training data, we proceeded to predict the lipid content of the labeled

and unlabeled validation data sets by the regression coefficient computed previously.

C. Nonlinear approaches

To generalize our initial simple linear regression approach, we then added new features corresponding to all possible products of the individual features as follows:

$$\begin{aligned} y_i = & f(x_1, x_2, ..., x_N, \\ & {x_1}^2, {x_2}^2, ..., {x_{N-1}}^2, {x_N}^2, \\ & x_1 x_2, ..., x_{N-1} x_N) + \varepsilon. \end{aligned} \tag{4}$$

This expanded linear regression analysis, which uses all possible quadratic features, is referred to as the *quadratic* regression model. To further generalize the analysis, we also formulated a multilayer perceptron neural network (MLPNN) [28] and also applied the gradient boosting machine learning (GBML) method presented by Blasi et al. [12] to predict the BODIPY signals in our FCM measurements (see Figs. S2-S4 in the supplementary information for details).

D. Feature selection

To select the optimal features, we applied iterative training-validation strategies, in which we applied a fitness function based on *label-free measurements* to select the most informative features. To select the best combination of features we employed a supervised learning strategy, in which we used linear regression analysis with and without quadratic interaction terms to find $\mathbf{M}$ for a given feature set for training data, and we applied the genetic algorithm [35] to the select the best combination of features to predict the validation data.

Direct measurement of lipid content is unavailable for unlabeled cells, so direct validation of label-free lipid predictions is not possible. However, since the labeled and unlabeled cells were sampled from the same original population and at the same time, we reasoned that the labeled and unlabeled populations should have the same distributions or statistics for their single-cell lipid levels. Therefore, to validate label-free predictions, we compare label-free distribution predictions to the labeled measurement distributions using the Kolmorogorov-Smirnov statistic (KS), [36]. The genetic algorithm was used to find the set of features that led to the smallest KS statistic for the unlabeled validation data.

We conducted all linear regression and genetic algorithm computations in MATLAB™ R2017b environment. For the MLPNN, computations were performed in Python 2.7 (see supplementary information for the MLPNN).

E. Weighted model

To further improve predictions of BODIPY signals for unlabeled cells, we considered a weighted model that could be learned from all measurement of unlabeled features, including the fluorescent channel in which BODIPY was measured in the labeled cells. To achieve this weighted model, we first learned three separate regression coefficients $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ based on the three training time points (days 1, 14, and 46). While these models were fixed for all subsequent computations, we defined a combination model that could be formulated as a weighted sum:

$$\mathbf{M} = \alpha_1 \mathbf{M}_1 + \alpha_2 \mathbf{M}_2 + \alpha_3 \mathbf{M}_3. \tag{5}$$

In the above equation, $\mathbf{a} = [\alpha_1, \alpha_2, \alpha_3]$ contains the weights applied to their corresponding $\mathbf{M}_i$'s with respect to the measured unlabeled features. Hence, at each given time point, there is a unique weighted model $\mathbf{M}$ based on fixed regression coefficients $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ and unlabeled features.

We then sought to learn a secondary model to estimate $\mathbf{a}$ from populations of unlabeled data. We defined $\mathbf{s}_r = [\mu_1^{(r)}, \ldots, \mu_n^{(r)}, \sigma_1^{(r)}, \ldots, \sigma_n^{(r)}]$ as a vector that contains the population means and standard deviations of each feature (including quadratic features) in any population of unlabeled cells. We then constructed the population sample statistics matrix $\mathbf{S} = [\mathbf{s}_1^T, \ldots, \mathbf{s}_R^T]$ using R different randomly sampled sub-population from the original training and validation data. For each r^{th} random population, we also performed a computational search to find an optimized model scaling factor $\mathbf{a}_r$ that yields the best possible comparison between measured and predicted targets in the training and validation data, and we collected these into the matrix $\mathbf{A} = [\mathbf{a}_1^T, \ldots, \mathbf{a}_R^T]^T$. With these definitions, we formulated a secondary regression analysis for $\mathbf{a}_r$ as a function of $\mathbf{s}_r$ with the assumed linear form

$$\mathbf{a}_r = \mathbf{s}_r \mathbf{Q} + \varepsilon, \tag{6}$$

for which we could estimate the weight quotient $\mathbf{Q}$ as

$$\mathbf{Q} \approx \mathbf{S}^{-L} \mathbf{A}. \tag{7}$$

In this expression, $\mathbf{Q}$ defines a relationship between the unlabeled features (from computing $\mathbf{s}$) and the weights ($\mathbf{a}$). To prevent overfitting in the determination of the weights, we generated another set of random population samples from our training and validation data, and we used the genetic algorithm to down select among the best columns of $\mathbf{S}$ (or rows of $\mathbf{Q}$) to utilize for the estimate of $\mathbf{a}$.

Once fixed using the training and validation data, the multi-scale regression operators $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$ and $\mathbf{Q}$ could be applied to any new data sets $\mathbf{X}_U$ and their summary statistics $\mathbf{s}$ to calculate $\mathbf{a} = \mathbf{s}\mathbf{Q}$, estimate $\mathbf{M}$ using Eqn. 5, and predict the lipid content using Eqn. 2.

V. Acknowledgments

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM124747.

References

[1] D. R. Gossett, W. M. Weaver, A. J. Mach, S. C. Hur, H. T. K. Tse, W. Lee, H. Amini, and D. Di Carlo, "Label-free cell separation and sorting in microfluidic systems," *Analytical and bioanalytical chemistry*, vol. 397, no. 8, pp. 3249–3267, 2010.

[2] Y. Han, Y. Gu, A. C. Zhang, and Y.-H. Lo, "Review: imaging technologies for flow cytometry," *Lab on a Chip*, vol. 16, no. 24, pp. 4639–4647, 2016.

[3] Y. Saeys, S. Van Gassen, and B. N. Lambrecht, "Computational flow cytometry: helping to make sense of high-dimensional immunology data," *Nature Reviews Immunology*, vol. 16, no. 7, pp. 449–462, 2016.

[4] D. D. Carlo and L. P. Lee, "Dynamic single-cell analysis for quantitative biology," 2006.

[5] N. Aghaeepour, G. Finak, H. Hoos, T. R. Mosmann, R. Brinkman, R. Gottardo, R. H. Scheuermann, F. Consortium, D. Consortium *et al.*, "Critical assessment of automated flow cytometry data analysis techniques," *Nature methods*, vol. 10, no. 3, p. 228, 2013.

[6] G. Lee, W. Finn, and C. Scott, "Statistical file matching of flow cytometry data," *Journal of biomedical informatics*, vol. 44, no. 4, pp. 663–676, 2011.

[7] M. Brown and C. Wittwer, "Flow cytometry: principles and clinical applications in hematology," *Clinical chemistry*, vol. 46, no. 8, pp. 1221–1229, 2000.

[8] A. Adan, G. Alizada, Y. Kiraz, Y. Baran, and A. Nalbant, "Flow cytometry: basic principles and applications," *Critical reviews in biotechnology*, vol. 37, no. 2, pp. 163–176, 2017.

[9] N. S. Barteneva, E. Fasler-Kan, and I. A. Vorobjev, "Imaging flow cytometry: coping with heterogeneity in biological systems," *Journal of Histochemistry & Cytochemistry*, vol. 60, no. 10, pp. 723–733, 2012.

[10] B. Rajwa, M. Venkatapathi, K. Ragheb, P. P. Banada, E. D. Hirleman, T. Lary, and J. P. Robinson, "Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier," *Cytometry Part A*, vol. 73, no. 4, pp. 369–379, 2008.

[11] K. Cheung, S. Gawad, and P. Renaud, "Impedance spectroscopy flow cytometry: on-chip label-free cell differentiation," *Cytometry Part A*, vol. 65, no. 2, pp. 124–132, 2005.

[12] T. Blasi, H. Hennig, H. D. Summers, F. J. Theis, J. Cerveira, J. O. Patterson, D. Davies, A. Filby, A. E. Carpenter, and P. Rees, "Label-free cell cycle analysis for high-throughput imaging flow cytometry," *Nature communications*, vol. 7, p. 10256, 2016.

[13] J. Yoon, Y. Jo, M.-h. Kim, K. Kim, S. Lee, S.-J. Kang, and Y. Park, "Identification of non-activated lymphocytes using three-dimensional refractive index tomography and machine learning," *Scientific reports*, vol. 7, no. 1, p. 6654, 2017.

[14] B. Wollscheid, D. Bausch-Fluck, C. Henderson, R. O'brien, M. Bibel, R. Schiess, R. Aebersold, and J. D. Watts, "Mass-spectrometric identification and relative quantification of n-linked cell surface glycoproteins," *Nature biotechnology*, vol. 27, no. 4, p. 378, 2009.

[15] C. L. Chen, A. Mahjoubfar, L.-C. Tai, I. K. Blaby, A. Huang, K. R. Niazi, and B. Jalali, "Deep learning in label-free cell classification," *Scientific reports*, vol. 6, p. 21471, 2016.

[16] S. E. Boddington, E. J. Sutton, T. D. Henning, A. J. Nedopil, B. Sennino, A. Kim, and H. E. Daldrup-Link, "Labeling human mesenchymal stem cells with fluorescent contrast agents: the biological impact," *Molecular Imaging and Biology*, vol. 13, no. 1, pp. 3–9, 2011.

[17] B. Guo, C. Lei, H. Kobayashi, T. Ito, Y. Yalikun, Y. Jiang, Y. Tanaka, Y. Ozeki, and K. Goda, "High-throughput, label-free, single-cell, microalgal lipid screening by machine-learning-equipped optofluidic time-stretch quantitative phase microscopy," *Cytometry Part A*, vol. 91, no. 5, pp. 494–502, 2017.

[18] J. Rumin, H. Bonnefond, B. Saint-Jean, C. Rouxel, A. Sciandra, O. Bernard, J.-P. Cadoret, and G. Bougaran, "The use of fluorescent nile red and bodipy for lipid measurement in microalgae," *Biotechnology for biofuels*, vol. 8, no. 1, p. 42, 2015.

[19] J. T. Cirulis, B. C. Strasser, J. A. Scott, and G. M. Ross, "Optimization of staining conditions for microalgae with three lipophilic dyes to reduce precipitation and fluorescence variability," *Cytometry Part A*, vol. 81, no. 7, pp. 618–626, 2012.

[20] R. Alford, H. M. Simpson, J. Duberman, G. C. Hill, M. Ogawa, C. Regino, H. Kobayashi, and P. L. Choyke, "Toxicity of organic fluorophores used in molecular imaging: literature review," *Molecular imaging*, vol. 8, no. 6, pp. 7290–2009, 2009.

[21] H. Hennig, P. Rees, T. Blasi, L. Kamentsky, J. Hung, D. Dao, A. E. Carpenter, and A. Filby, "An open-source solution for advanced imaging flow cytometry data analysis using machine learning," *Methods*, vol. 112, pp. 201–210, 2017.

[22] P. Eulenberg, N. Köhler, T. Blasi, A. Filby, A. E. Carpenter, P. Rees, F. J. Theis, and F. A. Wolf, "Reconstructing cell cycle and disease progression using deep learning," *Nature communications*, vol. 8, no. 1, p. 463, 2017.

[23] B. Munsky, G. Neuert, and A. van Oudenaarden, "Using gene expression noise to understand gene regulation," *Science*, vol. 336, no. 6078, pp. 183–187, 2012.

[24] P. Biller, R. Riley, and A. Ross, "Catalytic hydrothermal processing of microalgae: decomposition and upgrading of lipids," *Bioresource technology*, vol. 102, no. 7, pp. 4841–4848, 2011.

[25] H. K. Reddy, T. Muppaneni, J. Rastegary, S. A. Shirazi, A. Ghassemi, and S. Deng, "Asi: Hydrothermal extraction and characterization of bio-crude oils from wet chlorella sorokiniana and dunaliella tertiolecta," *Environmental Progress & Sustainable Energy*, vol. 32, no. 4, pp. 910–915, 2013.

[26] J. Rastegary, S. A. Shirazi, T. Fernandez, and A. Ghassemi, "Water resources for algae-based biofuels," *Journal of Contemporary Water Research & Education*, vol. 151, no. 1, pp. 117–122, 2013.

[27] C. J. Unkefer, R. T. Sayre, J. K. Magnuson, D. B. Anderson, I. Baxter, I. K. Blaby, J. K. Brown, M. Carleton, R. A. Cattolico, T. Dale *et al.*, "Review of the algal biology program within the national alliance for advanced biofuels and bioproducts," *Algal Research*, vol. 22, pp. 187–215, 2017.

[28] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[29] B. Munsky, G. Li, Z. R. Fox, D. P. Shepherd, and G. Neuert, "Distribution shapes govern the discovery of predictive models for gene regulation," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. 7533–7538, 2018.

[30] G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden, "Systematic identification of signal-activated stochastic gene regulation," *Science*, vol. 339, no. 6119, pp. 584–587, 2013.

[31] R. R. Guillard, "Culture of phytoplankton for feeding marine invertebrates," in *Culture of marine invertebrate animals*. Springer, 1975, pp. 29–60.

[32] R. R. Guillard and J. H. Ryther, "Studies of marine planktonic diatoms: I. cyclotella nana hustedt, and detonula confervacea (cleve) gran." *Canadian journal of microbiology*, vol. 8, no. 2, pp. 229–239, 1962.

[33] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012, vol. 329.

[34] S. Chatterjee and A. S. Hadi, *Regression analysis by example*. John Wiley & Sons, 2015.

[35] M. Mitchell, *An Introduction to Genetic Algorithms*, ser. Complex Adaptive Systems. MIT Press, 2014. [Online]. Available: https://books.google.com/books?id=3ezAoQEACAAJ

[36] R. H. Lopes, "Kolmogorov-smirnov test," in *International Encyclopedia of Statistical Science*. Springer, 2011, pp. 718–720.