THESIS

USING FLOW CYTOMETRY AND MULTISTAGE MACHINE LEARNING TO DISCOVER LABEL-FREE SIGNATURES OF ALGAL LIPID ACCUMULATION

Submitted by

Mohammad Tanhaemami

Department of Chemical and Biological Engineering

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2020

Master's Committee:

Advisor: Brian Munsky

Ashok Prasad
Hamidreza Chitsaz

ABSTRACT

USING FLOW CYTOMETRY AND MULTISTAGE MACHINE LEARNING TO DISCOVER LABEL-FREE SIGNATURES OF ALGAL LIPID ACCUMULATION

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.

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Brian Munsky of the Department of Chemical and Biological Engineering at Colorado State University. The door to Dr. Munsky's office is always open. He has consistently supported me professionally and mentally, and taught me how to own my research and to think as a professional. He is the best academic advisor one could ever wish for; and I am truly honored to have had the opportunity to work with such an incredible person.

I am indebted to Dr. Ashok Prasad of the Department of Chemical and Biological Engineering, and Dr. Hamidreza Chitsaz of the Department of Computer Science at Colorado State University as the committee members for this thesis. I am grateful for their time and support, and their valuable and challenging comments. It was a great privilege to have them review my work.

I express my deepest gratitude to my beloved parents, who are always supportive – from thousands of miles away. I also thank my wonderful friends, who keep reminding me that laughter and staying positive are the best medicine: Saeid, Bahareh, Behtash, Milad, Alee, Michael, Eric, Amin, Stephanie, and Javad.

Finally, I would like to thank my love, Sogand. Her unwavering support has always helped me push forward, no matter the circumstances. Words cannot describe how grateful I am for having her. I am the luckiest man alive to be with such a wonderful person with the biggest heart.

I am writing my thesis, yet taking my very first steps in this exciting field. As Dr. Munsky once said: *There is no such thing as "done"*.

DEDICATION

To Sogand: my love, my best friend, and my partner in crime.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iii
DEDICATION iv
LIST OF TABLES vi
LIST OF FIGURES vii

Chapter 1 Introduction 1

Chapter 2 Literature Review 3
2.1 Single-Cell Research and Flow Cytometry 3
2.2 Flow Cytometry Analysis 4
2.2.1 Mechanism of a Flow Cytometer 5
2.2.2 Fluorescent Dyes in Flow Cytometry 6
2.2.3 Fluorescence-Activated Cell Sorting (FACS) 8
2.3 Label-Free Quantification Strategies 8
2.4 Motivation 10

Chapter 3 Methods 12
3.1 Cell preparation and flow cytometry measurements 12
3.2 Linear regression analysis 14
3.3 Nonlinear approaches 17
3.4 Feature selection 17
3.5 The Kolmogorov-Smirnov statistic 18
3.6 Weighted model 18

Chapter 4 Results and Discussion 20

Chapter 5 Conclusions and Future Work 45

Bibliography 50

LIST OF TABLES

3.1 Table 3.1: Measured features by the Accuri™ C6 flow cytometer on lipid accumulations of the *P. soloecismus* cells 15

4.1 Table 4.1: Feature selection by the genetic algorithm on linear features 28
4.2 Table 4.2: Feature selection by the genetic algorithm on quadratic features 31
4.3 Table 4.3: Selected linear features for our proposed strategy (weighted model) 33
4.4 Table 4.4: Selected quadratic features for our proposed strategy (weighted model) . . . 34
4.5 Table 4.5: The weight quotient 35
4.6 Table 4.6: Selected final features 39

LIST OF FIGURES

3.1 Figure 3.1: Images of low- and high-lipid cells 13
3.2 Figure 3.2: Flow diagram of preliminary regression analysis 16

4.1 Figure 4.1: Preliminary regression analysis 21
4.2 Figure 4.2: Linear regression on quadratic features 23
4.3 Figure 4.3: Comparison of the features with and without BODIPY stain 25
4.4 Figure 4.4: Regression results after various approaches to feature selection 26
4.5 Figure 4.5: Linear regression on reduced features 27
4.6 Figure 4.6: Linear regression with the genetic algorithm on linear features 29
4.7 Figure 4.7: Linear regression with the genetic algorithm on quadratic features 30
4.8 Figure 4.8: Results of the weighted model 32
4.9 Figure 4.9: Training and validation of the proposed strategy with weighted model . . . 37
4.10 Figure 4.10: Testing the proposed weighted model for all 17 time points 38
4.11 Analysis of the weighted model 40
4.12 Testing the final model on independent cell populations and with a new flow cytometer 42
4.13 Simulation of a typical cell sorting experiment 44

5.1 Figure 5.1: Flow diagram of the final multi-stage label-free quantification strategy . . . 46

Chapter 1

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 data sets [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 data sets 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 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 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.[1]

[1]The Introduction Chapter of this thesis is from self article by Tanhaemami et al. [28].

Chapter 2

Literature Review

In this chapter, we will start with a brief discussion on how single-cell research has changed the view on studying cellular populations. We will then study how a flow cytometry analysis technique is carried out, as well as the broad application of fluorescent labels. Followed by the motivation points of this research, we will discuss about label-free quantification strategies during past few years and their common mistake.[1]

2.1 Single-Cell Research and Flow Cytometry

Before the advent of single-cell research, cell cultures were assumed to act as an ergodic system, i.e., cells are identical and are at equilibrium. Even in a population with growing cells, the assumption of local equilibrium for the cells was accepted [29]. However, in an actually heterogeneous system (which can be can be phenotypic, caused by progression through cell cycles or changes in the local environment, or genotypic, resulting from mutations), assumptions of this nature could lead to incorrect analyses [4,29]. This drawback has led to the development of single-cell studies.

Measurements at a single-cell level provide substantial, otherwise neglected, information about cellular properties [23]. Inclination towards single-cell research, requires the use of technologies that are capable of performing readouts at a single-cell level [9]. One of the most commonly used methods for studying single cells is flow cytometry [2]. Flow cytometry (FCM) is a high-throughput multiparametric analysis technique with the ability to read cells at rates as high as 100,000 cells per second [2, 6]. Multiparametric measurements and multivariate analysis capabilities of FCM, along with its high readout rates, make it a powerful and widespread technique in single-cell research [2–8, 29–31]. Providing insight about physical and chemical characteristics of

[1] Some parts of the Literature Review Chapter of this thesis are from the Curricular Practical Training (CPT) received at the Dana-Farber Cancer Institute and Broad Institute of MIT and Harvard.

individual cells [32], a flow cytometer is mostly used for the purpose of cell sorting, cell counting, and biomarker detection in cellular populations [1–3, 8, 33]. These tasks allow for various studies in single-cell research. For example, FCM can be employed to quantify cytokines, chromosomes, nuclei, DNA and RNA content, and protein accumulation [7, 8]. The ability of FCM in sorting cells has introduced the fluorescence-activated cell sorting (FACS) analysis, which enables distinction of different subpopulations in a larger cell population [1, 11]. FACS analysis can also be used in identification and separation of the cells of interest in a heterogeneous mixture of cellular populations [8].

In the following section, the mechanism of a flow cytometer is described, as well as the common fluorescent dyes, and their advantages and disadvantages. FACS analysis, as a specialized type of flow cytometry that highly depends on use of fluorescent dyes, is also described. In Section 2.3, computational strategies developed to avoid application of fluorescent dyes in FCM are explained, along with their shortcomings and drawbacks. Lastly, in Section 2.4 of this chapter, the motivation for this research is presented.

2.2 Flow Cytometry Analysis

Since its development in the 1950s and the 1960s [29, 34–37], flow cytometry (FCM) has become the main technique for analyzing cellular populations that can quantify multiple characteristics of single cells at a high rate [8, 34, 35, 38]. Differentiating and analyzing cells of interest in a heterogeneous population can be carried out using a flow cytometer. Whether for research, clinical, or industrial purposes, a flow cytometer is used in conducting numerous assays including but not limited to measuring size of the cells, analyzing cellular complexities, quantification of the DNA, RNA, and protein content, cell cycle analysis, and investigating cell membranes [5, 7, 8, 30, 31, 34, 35, 39, 40].

In this section, mechanism of a typical flow cytometer is explained. Application of fluorescent dyes in FCM and FACS analysis technique are also discussed.

2.2.1 Mechanism of a Flow Cytometer

A typical flow cytometer consists of three main parts: (i) Fluidic system, (ii) optical system, and (iii) signal detection and processing [8]. The essential part of a flow cytometer to perform a single-cell analysis is the Fluidic system. The cell suspension is injected into the flow chamber by pressurized air lines. At the same time, a sheath fluid forces the cell sample stream to be at the central core of the flow, resulting in a coaxial flow. Because of the pressure difference between the sample stream and the sheath fluid, the resulting coaxial flow aligns the cells to enter the flow chamber in a single file. The process, known as hydrodynamic focusing, allows the cells to be illuminated uniformly. It is worth noting that the sheath fluid and the cell suspension stream do not mix because the hydrodynamic focusing produces a laminar flow in the flow chamber [8,30,37,41].

As part of the optical system, as the cells pass through the flow chamber, they are illuminated by a focused light source (usually a laser beam) at a certain wavelength. The laser beam hits individual cells at the interrogation point. The process during which a cell traverses the beam is called an event. At each event, the electrons of the fluorochromes attached to the cells are excited by absorbing the energy from the laser beam. As the electrons drop to their lower energy orbitals, they emit energy in the form of light at a higher wavelength. If the cells have autofluorescence capability or are labeled with fluorescent dyes, the excitation-emission procedure can be detected by the flow cytometer. In addition to fluorescence signals, scattered lights deflected from the cells provide a handful of information regarding intrinsic cellular properties. The scattered light can be in the forward angle to the laser beam (typically between 2° and 20°), called the forward scatter (FSC). Alternatively, there is the side scatter (SSC), that is the light collected with an angle of $70° \leqslant \theta \leqslant 110°$. FSC is proportional to the square of the radius of a sphere, hence it provides information regarding size of the cells or their surface area. On the other hand, SSC is related to the light reflected by the nucleus and other contents in a cell and gives information about cell granularity. The optical systems also consists of a collection apparatus. A lens collects emitted light from each event and a group of optical filters and mirrors redirect the light to the detectors for recording and analysis. There are three types of filters used in the optical system: long-pass, short-

pass, and band-pass filters. Long-pass filters allow for passage of lights with wavelengths equal to or higher than a specific value. Short-pass filters transmit light that are equal to or shorter than a fixed wavelength. Band-pass filters, however, only permit light with wavelengths in a certain and narrow range [6, 8, 10, 30, 34, 35].

The third main part of a flow cytometer is the signal detection and processing. Signals from scattered light and fluorescence are detected by the photodetectors in a flow cytometer. Selected based on their sensitivity, photodetectors can be photodiods (PDs) or photomultiplier tubes (PMTs). In general, PMTs, which are more commonly used in FCM, are more sensitive than PDs. While a PD is able to detect stronger light signals generated by FSC, a PMT is preferred when weaker signals generated by SSC and fluorescence are favored. The detectors convert photons from emitted signals to electrical impulses. The resulting electrical current is then directed to an amplifier, where it is converted to either linear or logarithmic analog voltage pulses. By using an analog to digital converter (A-to-D or ADC), the analog signals gathered from amplifiers are converted to digital signals. The data from such digital signals are normally displayed in single-parameter histograms (used for studying fluorescence intensities) or dual-parameter scatter plots (used for assays related to FSC versus SSC) [7, 8, 30, 34, 35, 37, 39, 42].

2.2.2 Fluorescent Dyes in Flow Cytometry

For a fluorescent compound, there is a fixed range of wavelengths at which it can absorb and emit quantum of light energy. When a fluorochrome is exposed to light, namely a laser beam in a flow cytometry assay, the electrons on the fluorescent compound move from their lower orbitals to their higher orbitals as a result of absorbing energy. This process causes the fluorescent compound to be in an excited state. When electrons drop to their lower energy orbits, they emit the excess energy in the form of light at a lower energy frequency, i.e., higher wavelength. Such process of excitation because of absorbing the light energy and then emission of energy in the form of light at a higher wavelength is called fluorescence [8, 34, 37, 39, 43]. The difference between the peak wavelengths of excitation and emission signals in a fluorescence process is called the Stokes shift.

Higher Stokes shift is more desirable since it means the excitation and emission wavelengths can be distinguished from one another more easily [8, 37, 39].

While forward and side scattered light intensities in a flow cytometer offer useful information about intrinsic cellular properties such as cell size, shape, and granularity, fluorescence signals can be investigated to study the structure or functionality of the cells [30]. Although a cell might have autofluorescence capabilities, there are a few intrinsically fluorescent compounds present and they provide limited information about the cell [8]. Hence, fluorescent dyes are widely used in FCM measurements to allow for understanding several characteristics of cellular populations and presence of cell components that would be neglected otherwise [8, 34, 43, 44]. In flow cytometry, a proper fluorochrome may be chosen depending on its compatibility to the wavelength of the laser beam (commonly Argon laser emitting light at 488 nm) and properties of interest in the cells [34, 39, 43]. Moreover, autofluorescence can interfere with detection of target fluorescent signals if they are very dim. Thus, selecting an appropriate fluorochrome in FCM measurements is a vital task and should be conducted with extreme care [30]. A good fluorescent dye should have high quantum yield (able to help detect low concentrations of the stain by emitting high cell-associated fluorescence signal intensity), high Stokes shift, low toxicity, and high photostability. It should also be biologically inert and highly soluble in water [30]. There are numerous fluorescent dyes available and they can be classified into several groups with respect to their applications and mechanism of binding [8]. Some fluorochromes are used to label proteins and antibodies, such as fluorescein isothicyanate (FITC with approximately 492/520-530 nm excitation/emission wavelengths), phycoerythrin (PE with approximately 480-565/575-585 nm excitation/emission wavelengths), and allophycocyanin (APC with approximately 650/660 excitation/emission wavelengths). Some other fluorochromes are used in detecting the DNA content, such as propidium iodide (PI with excitation/emission wavelengths of approximately 305-580/623 nm). Additionally, boron-dipyrromethene (BODIPY with approximate wavelengths of 503-505/512-515 nm of excitation/emission) and Nile Red (with approximate wavelengths of 551/636 nm of excitation/emission wavelengths) are commonly used for identification of lipids [8,18–20,27,30,34,37,39,43]. The de-

sire for higher Stokes shift has led to invention of tandem dyes. In a tandem dye, two fluorochromes are conjugated in a donor-acceptor relationship. The donor fluorochrome gets excited at a wavelength. Being covalently coupled to the acceptor, the donor transforms its excitation energy to the acceptor. The acceptor then emits light at a much higher wavelength. The advent of tandem dyes is an outcome of a process called fluorescence resonance energy transfer (FRET) [8, 34, 43]. Describing the mechanisms of binding and applications for all available fluorescent dyes is beyond the scope of this study, though if interested, the reader is encouraged to read the useful books authored by Macey [34], Givan [37], and Shapiro [36], as well as several other scientific articles cited in this section.

2.2.3 Fluorescence-Activated Cell Sorting (FACS)

A major application of a flow cytometry analysis is to sort a heterogeneous cell population into separate subpopulations. Fluorescence-activated cell sorting (FACS) is a powerful technique developed based on flow cytometry that enables us to sort a population of fluorescently-labeled cells into two or more subsets. In FACS analysis, an operator first gates the cells of interest, i.e., defines the properties of the cells of interest by the flow cytometer software as they are interrogated. Then, as the cells leave the sample chamber, a nozzle containing the stream of the cells – and the sheath fluid are vibrated at a high frequency by an acoustic piezoelectric crystal. The vibration causes the sheath fluid to break into droplets, each of which isolating at most one cell. When droplets leave the nozzle, they are electrically charged according to the operator-defined gates. Later, electrically-charged droplets are deflected by electromagnetic plates which redirect the droplets into their corresponding collection tubes. Alternatively, unwanted cells in the stream will be collected into the waste [1, 8, 11, 34, 35, 40, 45].

2.3 Label-Free Quantification Strategies

Application of fluorescent dyes helps investigate multiple cellular properties, functions, and structures [8, 30, 34, 43, 44]. Measurements in a Flow cytometer primarily rely upon fluorescently

labeling the cells for subsequent quantification and analysis [44]. However, labeling with a fluorescent marker has several drawbacks. A fluorescent dye may have a negative impact on natural cellular behavior, interfere with cellular processes, cause unwanted activating/inhibitory signal transduction, and lead to fundamental errors in downstream analyses [12, 13, 15–18, 46, 47]. Furthermore, fluorescent labels may not always be available and require extensive effort to be used in elaborate experiments. Thus, the process of labeling can be complicated and time-consuming [1, 10, 14, 19, 22, 44, 47]. Such adverse consequences has raised needs for methods to measure cellular properties without the labels – or with less amount. Statistical analyses, combined with machine learning methods, has paved the way for opening new windows in label-free quantification of cell populations in FCM measurements. Several label-free quantification strategies have been developed over the past few years to resolve this issue. For instance, Scholtens et al. used an image cytometer combined with a Random Forest algorithm to classify circulating tumor cells (CTCs), apoptotic CTCs, CTC debris, leukocytes, and debris not related to CTCs [48]. Weber et al. employed a parameter-free Fisher's linear discriminant analysis classifier to detect differentiation in a mixed population of proliferating and differentiating cells over time for PC12 cells on the basis of phase contrast images [49]. Some studies applied support vector machines (SVM) for classification purposes. Feng et al. proposed methods to classify Jurkat T cells and Ramos B cells in a polarization diffraction imaging flow cytometry (p-DIFC) analysis using SVM [50]. The technology of p-DIFC was also used in combination with SVM for classification of malignant versus benign cells (PC3 and PCS prostate cell types) by Jiang et al. [51]. Li et al. have also integrated SVM algorithms with complex holographic images to analyze 3-part leukocytes [52]. In another study, Miura et al. discussed advantages of using SVM to quantify large populations of *Euglena gracilis* cells and human cancer cells in light-sheet fluorescence imaging flow cytometry [53]. Other supervised machine-learning based methods have also been introduced for the purpose of label-free quantification. Using an imaging flow cytometer and gradient boosting machine-learning algorithms, Blasi et al. were able to offer label-free strategies for identification of DNA content in fixed and live Jurkat cells [12]. The study was later followed by Hennig et al. to provide a more

user-friendly software [21]. Eulenberg et al. suggested a deep convolutional neural network to reconstruct cell cycle and disease progression analyses [22]. Jiang et al. presented methods to classify aggregated platelets from single platelets and white blood cells by analyzing optofluidic time-stretch microscopy images via logistic regression analysis [54].

The above studies are based on methods to incorporate supervised machine learning algorithms with experimental single-cell measurements for two main goals: (i) to reduce the need for fluorescent labels, and (ii) to compete with (or maintain) high-throughput capabilities of FCM measurements. Nevertheless, they have a fundamental drawback: in their supervised machine learning process, the training-validation steps and feature selection procedures are based on information from fluorescently labeled cells. This key issue is the motivating element of our research, which is discussed in the next section.

2.4 Motivation

As discussed in the previous sections, application of fluorescent labels can be costly, exhaustive, cumbersome, harmful to cells' natural behavior, be intrusive in cellular procedures, and eventually lead to problematic understandings in cell analyses [1, 10, 12–19, 22, 44, 46, 47, 47]. As a response to such important pitfalls in FCM analysis, several studies have focused on integrating various supervised machine-learning algorithms with FCM measurements, hoping to offer a computational label-free quantification strategy [12, 21, 22, 48–54]. Although these studies are doing an excellent job in offering avenues for label-free quantification strategies, they fail to correctly identify and select features in their machine-learning pipelines. Merely concentrating on fluorescently labeled cells to extract the features leads to invalid label-free strategies in investigating single cells. Even though the labels influence intrinsic cell properties only *indirectly*, the corresponding information cannot be reliable in the training step of the used supervised machine-learning algorithm. Strategies developed based upon the above common mistake will yield poor testing results in quantification of cellular populations in a true label-free context.

In this research study, we hypothesize that the training step can be carried out based on label-free measurements. Our approach will produce valid results, *even when signatures are perturbed and are weak*. We show that in employing supervised machine learning, feature-selection iterations should consider *both the labeled and unlabeled* information.

Chapter 3

Methods

We developed our label-free quantification strategy with respect to flow cytometry measurements of lipid accumulations in *Picochlorum soloecismus* cells. *P. soloecismus* cells are marine microalgae with the ability to accumulate lipids of up to 80% under stress, i.e., nitrogen replete and nitrogen deplete conditions [24–26]. We monitored algal lipid accumulations and investigated FCM measurements on *P.soloecismus* cells for several days. We then proceeded to generate a method for label-free quantification of the lipid accumulation with respect to the above FCM measurements.

In this chapter, we will discussed methods to develop our proposed label-free quantification strategy, along with the related theoretical background.[1]

3.1 Cell preparation and flow cytometry measurements

Experiments were conducted by our collaborators at the Los Alamos National Laboratory (LANL). *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 [55, 56]. 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 increased above the set-point. Cells were monitored for a total of 46 days following nitrogen starvation, collected at 23 different days, 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. Figure 3.1 shows representative images of high and low lipid cells. The analysis was performed using a BD Accuri™ C6 flow cytometer with BD CSampler™ (BD Biosciences). Unstained samples

[1] Some parts of the Methods Chapter of this thesis are from self article by Tanhaemami et al. [28], and the Curricular Practical Training (CPT) received at the Dana-Farber Cancer Institute and Broad Institute of MIT and Harvard.

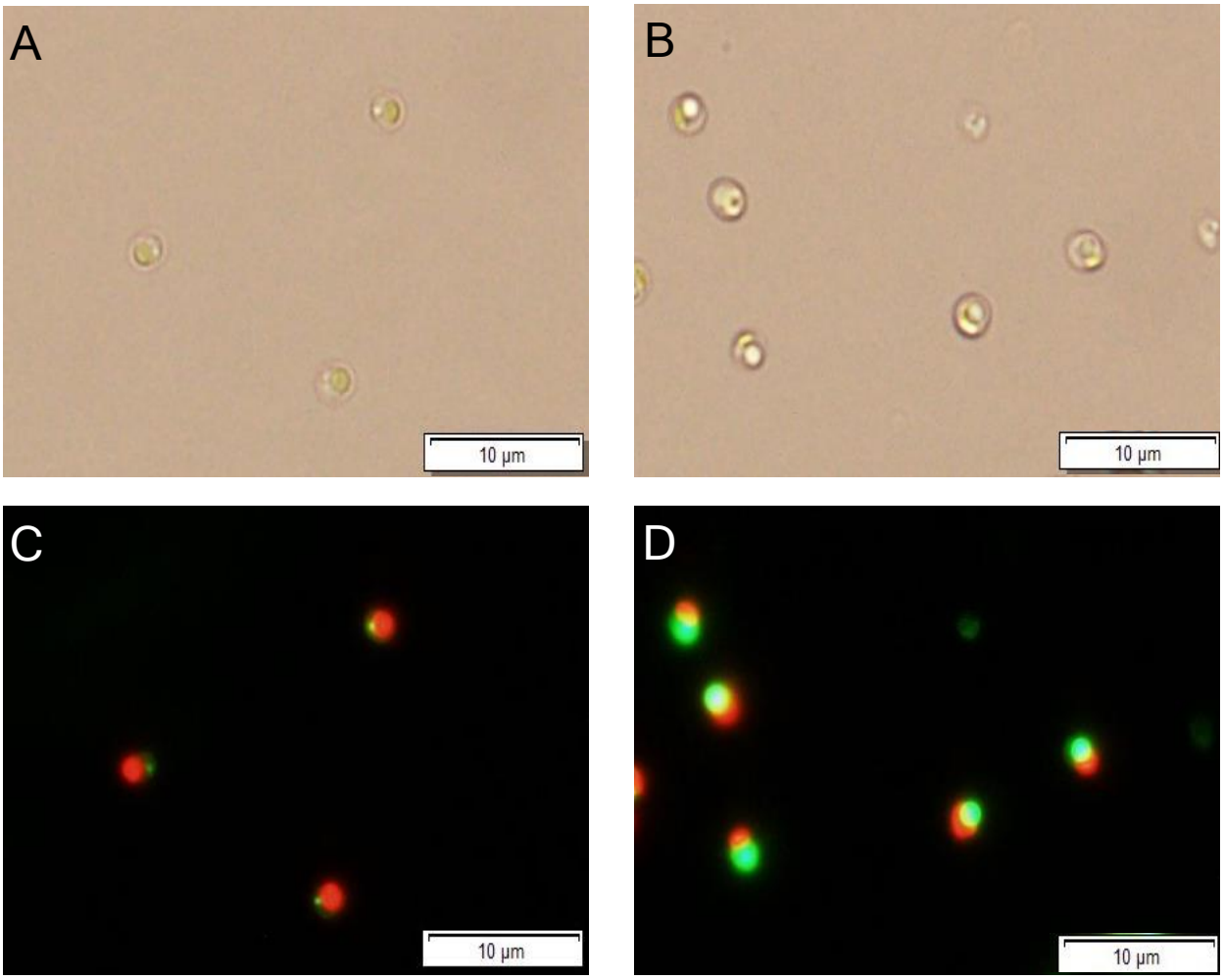

Figure 3.1: Images of low lipid (A,C) and high lipid (B,D) cells. Panels A and B show the bright field and C and D are overlays of BODIPY staining (green) and chlorophyll fluorescence (red).

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 µl/min). Data was exported in .csv format for subsequent analysis.

The Accuri™ C6 flow cytometer collected the data from measurements of *P. soloecismus* cells in the following channels:

- FSC: Forward scatter (low angle scatter, generally related to size)
- SSC: Side scatter (90-degree scatter, generally related to granularity)
- FL1: 488 nm excitation, 530/30 nm collection, the main channel, used to look at BODIPY signals
- FL2: 488 nm excitation, 585/40 nm collection
- FL3: 488 nm excitation, 670 LP (long pass) collection, used to look at auto fluorescence
- FL4: 640 nm excitation, 675/25 nm collection, auto fluorescence is also strong here

For the forward scatter, the side-scatter, and each channel FL1-FL4, the flow cytometer measures a pulse of light as each cell traverses the laser beam. Both the height (-H) and the integrated area (-A) of these pulses were collected, providing two measures per channel, per cell. The "Width" of a measured pulse is also recorded for each cell.

Table 3.1 presents the information provided at each channel in our flow cytometer.

All data was exported in .csv format. With these data, we next examined several iterative training-validation strategies to discover signatures within the label-free data that could reproduce the lipid accumulation at all times. Computations were executed in MATLAB™ R2017b environment.

3.2 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. 3.2). In re-

Table 3.1: Measured features by the Accuri™ C6 flow cytometer on lipid accumulations of the *P. soloecismus* cells

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7
FSC-A	SSC-A	FL1-A	FL2-A	FL3-A	FL4-A	FSC-H
Feature 8	Feature 9	Feature 10	Feature 11	Feature 12	Feature 13	
SSC-H	FL1-H	FL2-H	FL3-H	FL4-H	Width	

gression analysis, there are two main types of variables: the response variable (denoted y) and the explanatory variables (the set of predictors, denoted $\mathbf{x}$) [57]. 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

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \tag{3.1}$$

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 [58]. In linear regression, the response (target) and predictor (feature) variables are assumed to satisfy the linear relationship [58]

$$\mathbf{Y} = \mathbf{X}\mathbf{M}, \tag{3.2}$$

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

$$\mathbf{M} = \mathbf{X}^{-L}\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3.3}$$

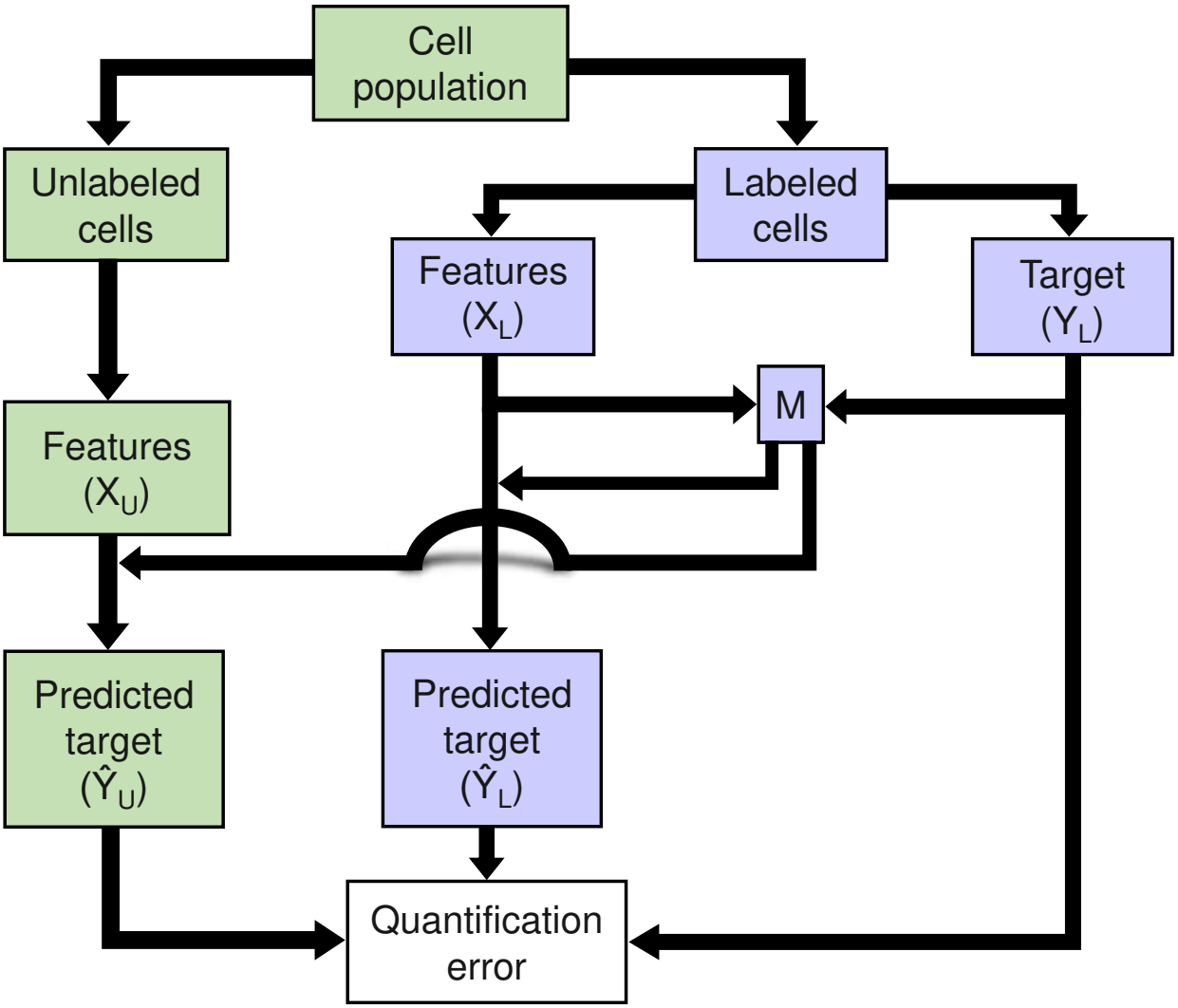

Figure 3.2: Flow diagram of preliminary regression analysis to quantify lipid content based using intrinsic (presumably label-free) features. The model is learned using labeled data and then tested on both labeled and unlabeled data.

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. (3.3) 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 regression coefficient $\mathbf{M}$ computed from training data, we proceeded to predict the lipid content of the labeled and unlabeled validation data sets.

3.3 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:

$$\begin{aligned} y_i = & f(x_1, x_2, ..., x_N, \\ & {x_1}^2, {x_2}^2, ..., {x_{N-1}}^2, {x_N}^2, \\ & x_1 x_2, ..., x_{N-1} x_N) + \varepsilon. \end{aligned} \tag{3.4}$$

This expanded linear regression analysis, which uses all possible quadratic features, is referred to as the *quadratic* regression model.

3.4 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: we used a linear regression analysis – with and without the quadratic terms – to find the regression coefficient $\mathbf{M}$ for a given feature set for the training data. We then applied the genetic algorithm [59] to the select the best combination of features that could predict the target 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 (KS) statistic [60]. The genetic algorithm was used to find the set of features that led to the smallest KS statistic for the unlabeled validation data.

3.5 The Kolmogorov-Smirnov statistic

To evaluate the success of our label-free quantification strategies, we need a metric to compare measured and predicted lipid accumulation. As mentioned in the previous section, lipid content of unlabeled cells cannot be directly measured for lipid accumulations. Hence, instead of individual cells, it is the distributions of the cells that are investigated. For this purpose, the Kolmogorov-Smirnov (KS) statistic is used. Moreover, the KS distance is a useful tool to analyze histogram data and decide whether two sets of data are sampled from the same distribution [61]. The KS distance is defined as the absolute value of the maximum difference between $F_m(x)$ and $F_p(x)$ for all possible values of x [60], where $F_m(x)$ and $F_p(x)$ are defined as the cumulative distributions of the measured and predicted values, respectively.

3.6 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). These models were fixed for all subsequent computations. For any arbitrary population, a new combination model could be formulated as a weighted sum:

$$\mathbf{M} = \alpha_1\mathbf{M}_1 + \alpha_2\mathbf{M}_2 + \alpha_3\mathbf{M}_3, \tag{3.5}$$

where the weights $\mathbf{a} = [\alpha_1, \alpha_2, \alpha_3]$ would be specific to any new population of unlabeled cells.

We then sought to learn a secondary model to estimate $\mathbf{a}$ from populations of unlabeled data. For this task, 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. It is important to note – and will be discussed later in more detail – that because the unlabeled cells are not treated with BODIPY, the statistics contained in $\mathbf{s}_r$ can include the 530/30 nm channel, which allows access to previously unutilized information in the 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

$$\mathbf{a}_r = \mathbf{s}_r \mathbf{Q} + \varepsilon, \tag{3.6}$$

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

$$\mathbf{Q} \approx \mathbf{S}^{-L} \mathbf{A}. \tag{3.7}$$

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 Eq. (3.5), and predict the lipid content using Eq. (3.2).

Chapter 4

Results and Discussion

The proposed label-free quantification strategy in this study opens new windows to correctly identify label-free signatures in flow cytometry measurements. By addressing the common mistake in other label-free quantification approaches (chapter 2), we offer a new method to accurately quantify FCM measurements based on defining an optimized set of *unlabeled* features. In this chapter, the results of our method are demonstrated and discussed in detail.[1]

Figure 3.2 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. 3.2. 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 Chapter 3, 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 data

[1]The Results and Discussion Chapter of this thesis are from self article by Tanhaemami et al. [28].

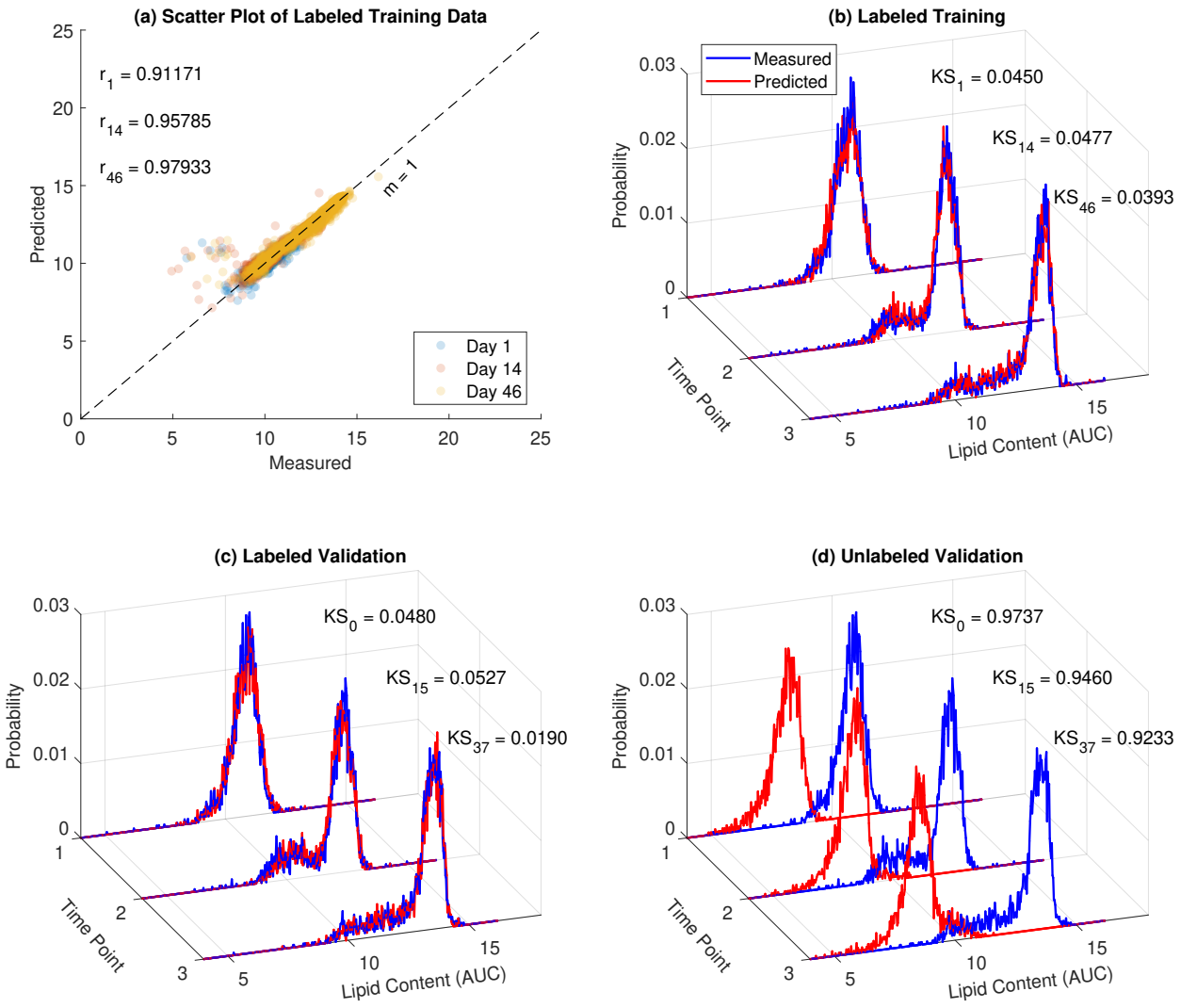

Figure 4.1: Preliminary regression analysis. (a) Correlations between measured and predicted values of lipid content for labeled training data. Pearson's correlation coefficients are shown for each time point. (b) Histograms of lipid content for labeled training data. Measured in blue and predicted in 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.

set gathered from labeled and unlabeled cells, and we evaluated the prediction accuracy using the Kolmogorov-Smirnov statistic.

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 4.1 shows the results of applying the simple linear regression analysis using labeled data only. Figure 4.1(a) shows that at each time point the predicted labeled training data has a strong correlation with the measured data. Figure 4.1(b) suggests that a preliminary regression analysis provides a strong classification for the labeled training data, which was consistent in Fig. 4.1(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. 4.1(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).

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 another more advanced machine learning approach to learn lipid content from the intrinsic features. The *quadratic* regression model, which corresponds to linear regression applied to linear and second order products of the original features, was also used to predict the lipid accumulations. As shown in Fig. 4.2, including second order products of the original features in regression analysis also does a great job on labeled data. Nevertheless, it drastically fails to predict the lipid content for unlabeled cells. From this observation we concluded that it is not the model being too simple that causes the unlabeled predictions to fail.

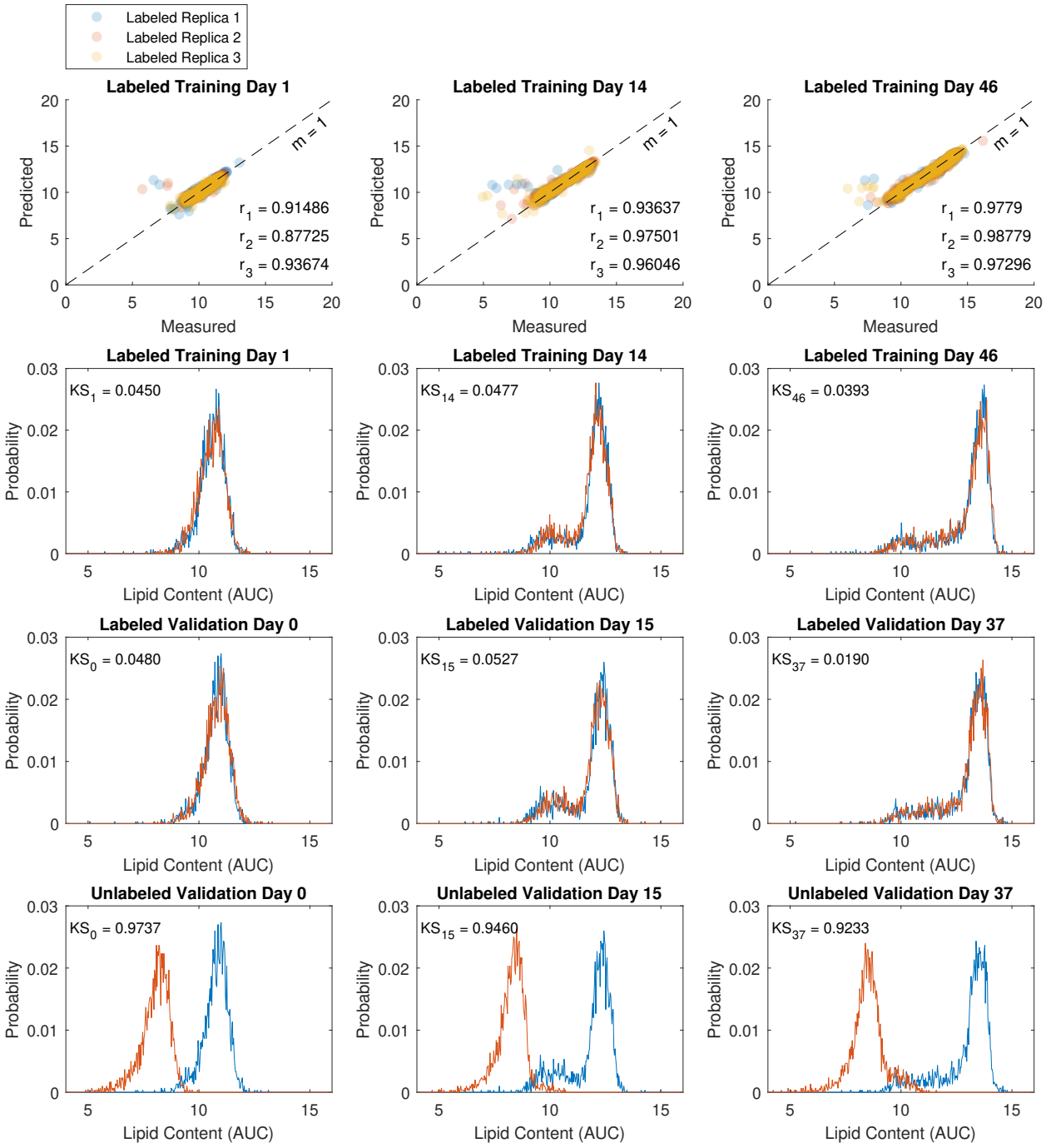

Figure 4.2: The quadratic regression model. The model works well for labeled data (rows 1, 2, and 3). However, it fails to predict the lipid content for unlabeled cells (row 4). The first row represents the correlation between measured and predicted values of the labeled training data. The 3 colors correspond to the 3 measurement replications at each day of FCM analysis (days 1, 14, and 46). Pearson's correlation coefficients are shown for each replication. For validation (rows 2, 3, and 4), we selected days 0, 15, and 37. The histograms show the results of prediction with this model for training (labeled cells) and validation (labeled and unlabeled cells) data. Measured histograms are in blue, predicted are in red. The KS distances between measured and predicted lipid content are shown on each plot. For all histograms, lipid accumulation are shown in arbitrary units of concentration (AUC).

To explain the failure of the labeled-cell-trained regression model on unlabeled cells, we suspected that some channels in the flow cytometer might be adversely affected by application of the BODIPY stain. Indeed, Fig. 4.3 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 4.4(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. 4.1(d) to 0.11-0.38 in Fig. 4.4(b)). 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 applied exclusively to labeled cells. Figure 4.5 provides extended plots of the outcomes of regression analyses upon removal of the corrupted features.

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 4.4(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

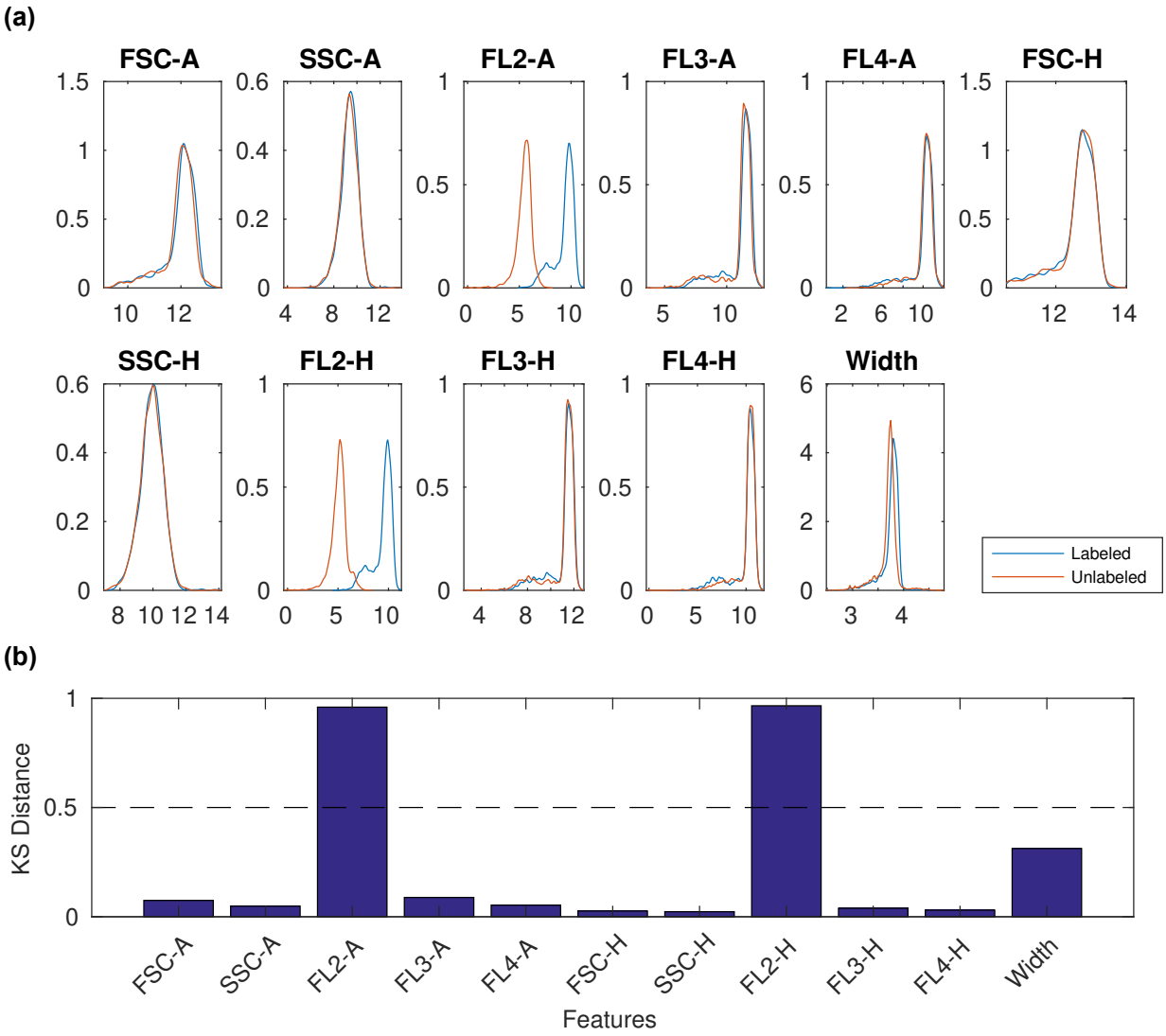

Figure 4.3: Comparison of the features with and without BODIPY stain. (a) Kernel densities of features for labeled and unlabeled cells, averaged over all times. Labeled cells are shown in blue, and unlabeled cells are in red. (b) KS distance between labeled and unlabeled features distributions. FL2-A and FL2-H features show clear dependence on the BODIPY stain. Horizontal line denotes threshold used to remove corrupted features.

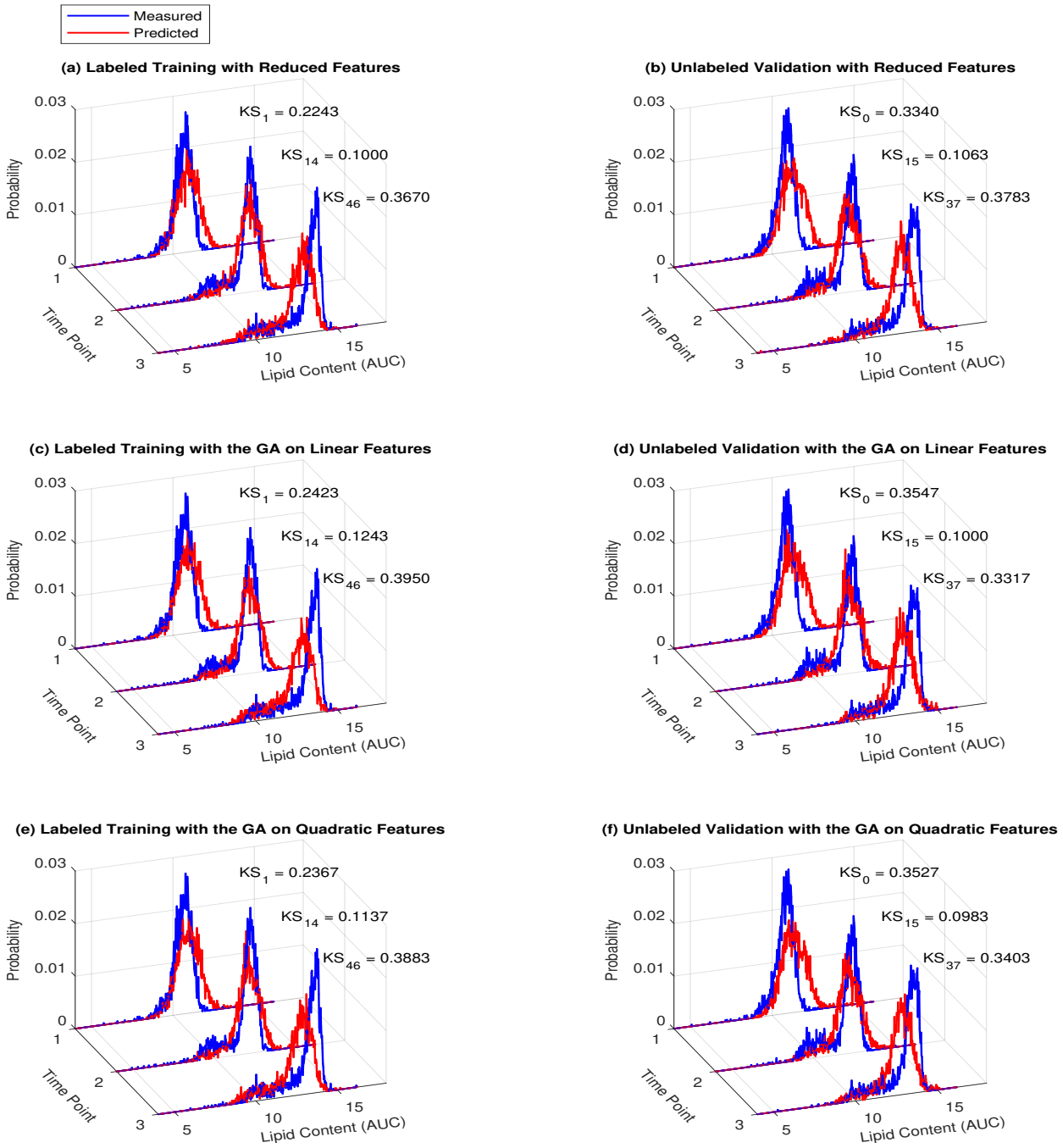

Figure 4.4: Regression results after various approaches to feature selection. (a) Training on reduced features. (b) Validation of the model in (a) on unlabeled cells. (c) Training based on the features selected by the GA. (d) Validation of the model in (c) on unlabeled cells. (e) Training based on the 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.

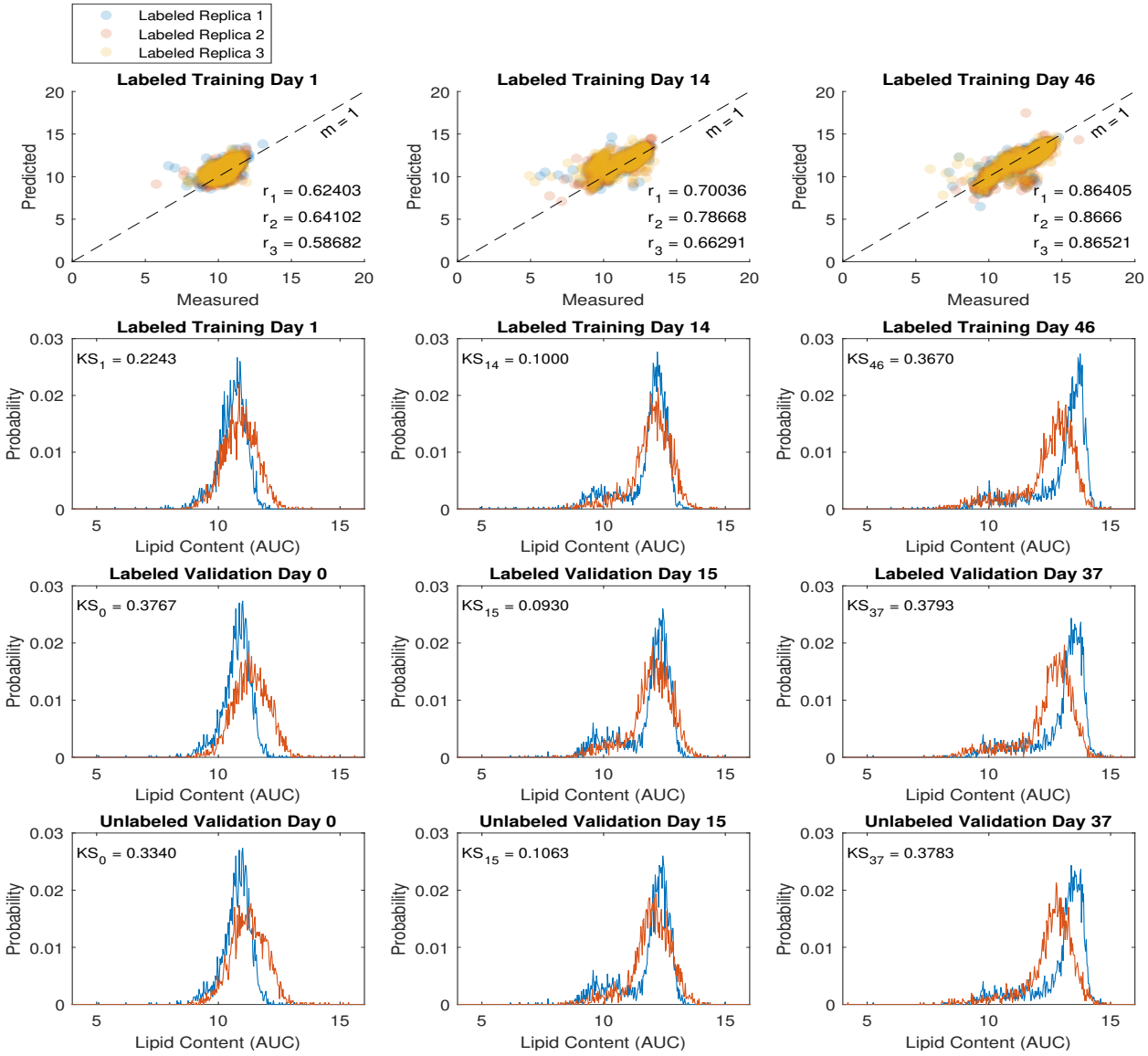

Figure 4.5: Linear regression on reduced features. Accuracy of the model is decreased for labeled data (rows 1, 2, and 3) due to removing the disrupted features. However, we see an important improvement in predicting the lipid content for the unlabeled cells (row 4). The first row represents the correlation between measured and predicted values of the labeled training data. The 3 colors correspond to the 3 measurement replications at each day of FCM analysis (days 1, 14, and 46). Pearson's correlation coefficients are shown for each replication. For validation (rows 2, 3, and 4), we selected days 0, 15, and 37. The histograms show the results of prediction with this model for training (labeled cells) and validation (labeled and unlabeled cells) data. Measured histograms are in blue, predicted are in red. The KS distances between measured and predicted lipid content are shown on each plot. For all histograms, lipid accumulation are shown in arbitrary units of concentration (AUC).

results are provided in Fig. 4.6. Selected features by the genetic algorithm on linear features are also presented in Table 4.1.

Table 4.1: Feature selection by the genetic algorithm on linear features. There are 11 features in our flow cytometry measurements. Selected features are shown in green. Features FL1-A and FL1-H correspond to the main fluorescence channel (targets to be predicted). Hence, they are not considered in the application of the genetic algorithm. Regression coefficient values for each selected feature are shown.

'FSC-A'	6.3395
'SSC-A'	0.0548
'FL2-A'	0
'FL3-A'	-0.1703
'FL4-A'	0
'FSC-H'	-2.9008
'SSC-H'	0
'FL2-H'	0
'FL3-H'	0
'FL4-H'	0
'Width'	-6.8632

During the automated feature selection for linear regression on linear features (Fig. 4.4(c-d) and Fig. 4.6), 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. 4.4(e,f) and in Fig. 4.7, 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. 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 quadratic features are also presented in Table 4.2.

After cross-validation and feature reduction, the predictions using label-free data had improved substantially, but we noticed that there remained some substantial systematic prediction errors. In particular, predictions using a single regression model appeared to be biased toward the average lipid levels and led to over-prediction of low lipid populations (early time points) and under-

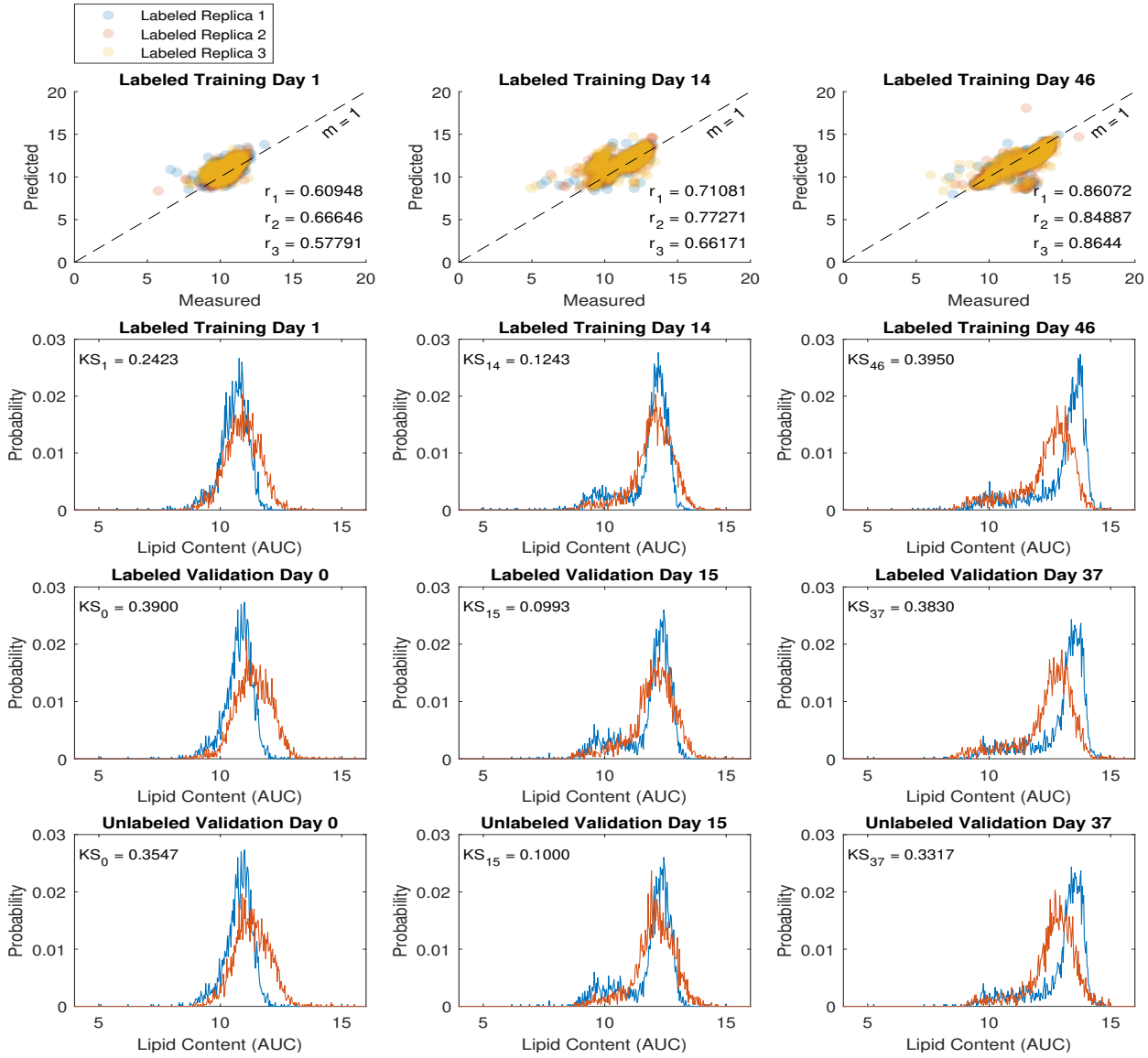

Figure 4.6: Regression analysis after performing automated feature selection by the genetic algorithm on linear features. Better prediction accuracy for the unlabeled cells. The first row represents the correlation between measured and predicted values of the labeled training data. The 3 colors correspond to the 3 measurement replications at each day of FCM analysis (days 1, 14, and 46). Pearson's correlation coefficients are shown for each replication. For validation (rows 2, 3, and 4), we selected days 0, 15, and 37. The histograms show the results of prediction with this model for training (labeled cells) and validation (labeled and unlabeled cells) data. Measured histograms are in blue, predicted are in red. The KS distances between measured and predicted lipid content are shown on each plot. For all histograms, lipid accumulation are shown in arbitrary units of concentration (AUC).

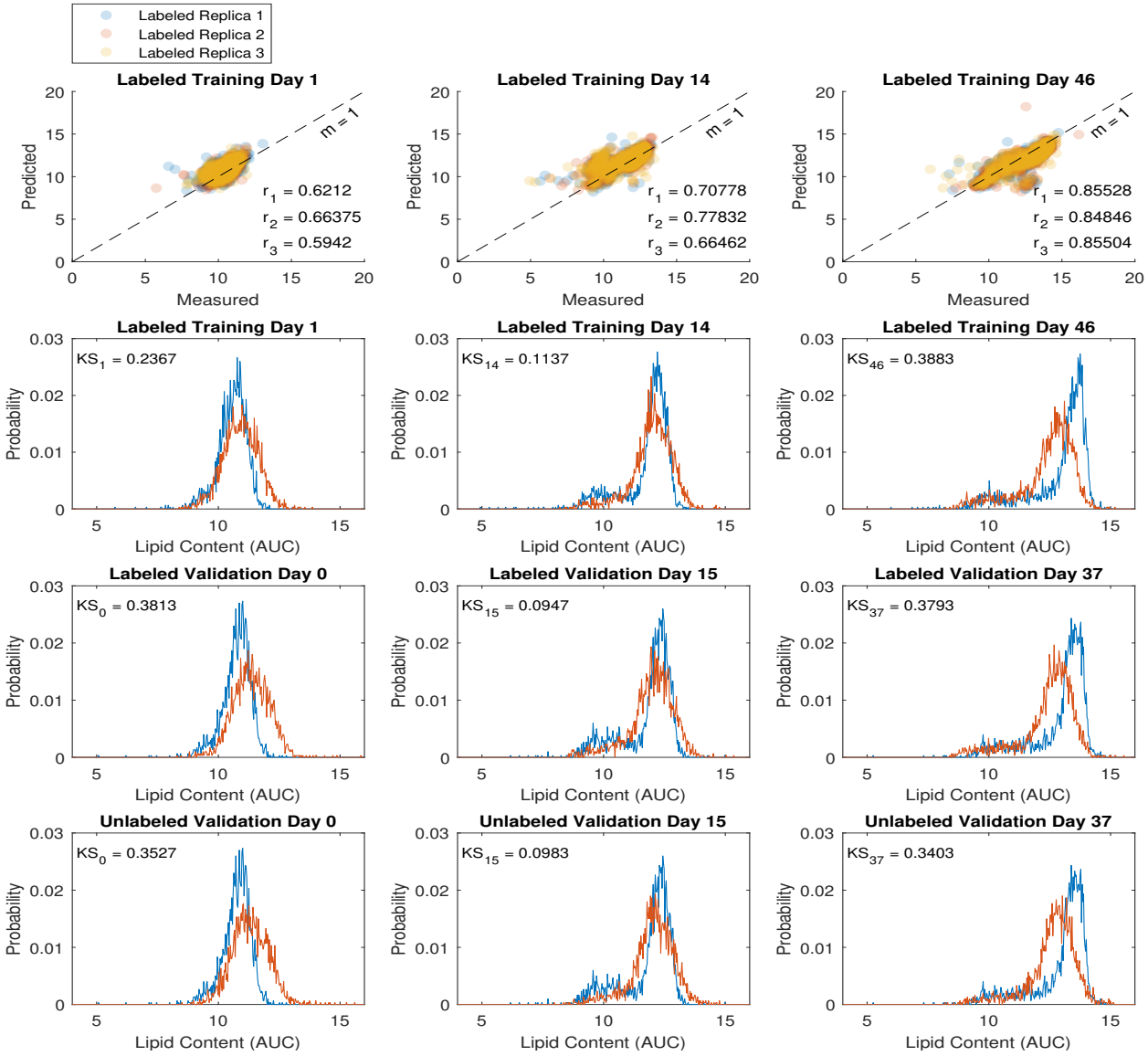

Figure 4.7: Regression analysis after performing automated feature selection by the genetic algorithm on quadratic features. Slight improvement is observed for prediction of the unlabeled cells' lipid content. The first row represents the correlation between measured and predicted values of the labeled training data. The 3 colors correspond to the 3 measurement replications at each day of FCM analysis (days 1, 14, and 46). Pearson's correlation coefficients are shown for each replication. For validation (rows 2, 3, and 4), we selected days 0, 15, and 37. The histograms show the results of prediction with this model for training (labeled cells) and validation (labeled and unlabeled cells) data. Measured histograms are in blue, predicted are in red. The KS distances between measured and predicted lipid content are shown on each plot. For all histograms, lipid accumulation are shown in arbitrary units of concentration (AUC).

Table 4.2: Feature selection by the genetic algorithm on quadratic features. There are 11 features in our flow cytometry measurements. Selected features are shown in green. Features FL1-A and FL1-H correspond to the main fluorescence channel (targets to be predicted). Hence, they are not considered in the application of the genetic algorithm. Regression coefficient values for each selected feature are shown.

			Feature										
			'FSC-A'	'SSC-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature	'FSC-A'	0	0	0	0	0	0	0	0	0	0	6.7021	0
	'SSC-A'	0		0	0	0	0	0	0	0	0	0	0
	'FL2-A'	0			0	0	0	0	0	0	0	0	0
	'FL3-A'	0				0	0	0	0	0	0	0	0
	'FL4-A'	0					0	0	0	0	0	0	0
	'FSC-H'	0						-1.5845	0	0	0	0	0
	'SSC-H'	0							0	0	0	0	0
	'FL2-H'	0								0	0	0	0
	'FL3-H'	0									0	0	0
	'FL4-H'	0										0	-6.9141
	'Width'	0											0

prediction of high lipid content populations (late time points). We hypothesized that this bias to the middle could be corrected by allowing the model itself to adapt in accordance with signatures in the label-free data.

To test this idea, we introduced a new strategy based on weighted models that could be learned from all measurement of unlabeled features. 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* (see Chapter 3 for more details on the related methods). 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 4.8 shows the results of our new label-free quantification strategy for labeled cells (Fig. 4.8(a)) and unlabeled cells (Fig. 4.8(b-g)). It can be seen here that using model weights, which are based only 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.

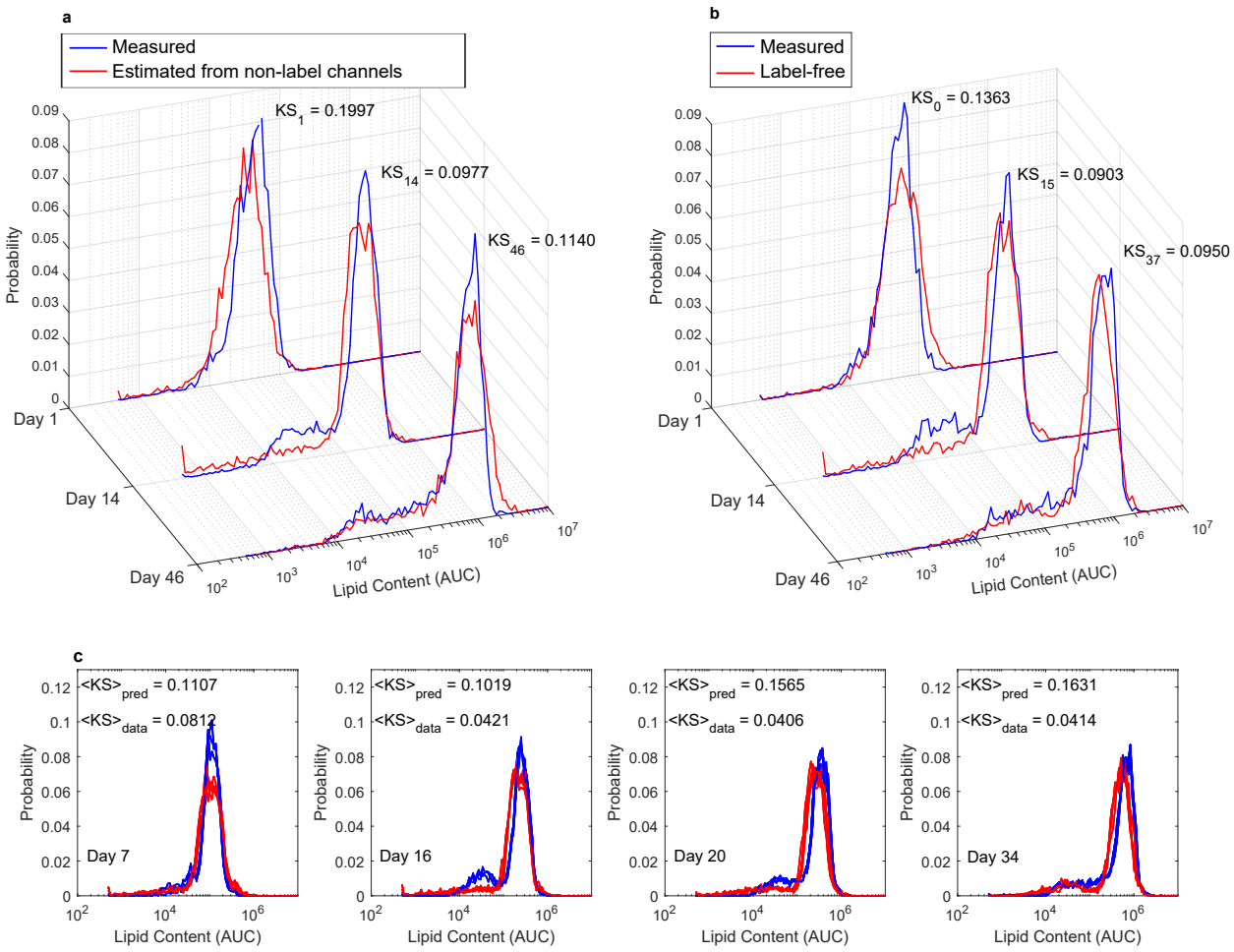

Figure 4.8: Results of the weighted model. Distributions of lipid content for (a) labeled training data, and (b) unlabeled validation data. KS distances between distributions are shown. (c) Testing the final strategy on four unlabeled testing time points: Days 7, 16, 20, and 34. See supplementary figure S10 for corresponding results for all 17 testing time points. “KS data" is the average KS distance between measured lipid distributions. All lipid contents are in arbitrary units of concentration (AUC).

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 4.3 presents the selected features based on our proposed strategy, when applied on *linear* features. Alternatively, Table 4.4 presents the selected features based on our proposed strategy, when applied on *quadratic* features. 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 (based on *quadratic* features) are presented in Table 4.5.

Table 4.3: Our proposed strategy (weighted model). Application of the genetic algorithm on *linear* features (selected features are shown in green). The 3 trained models will have different weights. The genetic algorithm is performed to select the most informative features of each model separately. Regression coefficient values for each selected feature are shown.

'FSC-A'	0	0	0
'SSC-A'	0.3116	0	0.5483
'FL2-A'	0	0	0
'FL3-A'	0	0	-0.0983
'FL4-A'	-0.0754	1.1708	-0.1790
'FSC-H'	0	0	0
'SSC-H'	0	0	0
'FL2-H'	0	0	0
'FL3-H'	0	0	0.9724
'FL4-H'	0.9466	0	0
'Width'	0	0	0

Table 4.4: Our proposed strategy (weighted model). Application of the genetic algorithm on *quadratic* features (selected features are shown in green). The genetic algorithm is applied to the 3 models M1, M2, and M3 separately.

			Feature (Model 1)										
			'FSC-A'	'SSC-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature (Model 1)	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		0	0	-0.5060	0	0	0	0	0.9099	0	0
	'FL2-A'	0			0	0	0	0	0	0	0	0	0
	'FL3-A'	0				0	0	0	0	0	0	0	0
	'FL4-A'	0					0.1716	0	0	0	0	0	0
	'FSC-H'	0						0	0	0	0	0	0
	'SSC-H'	0							0	0	0	0	0
	'FL2-H'	0								0	0	0	0
	'FL3-H'	0									0	0	0
	'FL4-H'	0										0	0
	'Width'	0											0

			Feature (Model 2)										
			'FSC-A'	'SSC-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature (Model 2)	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		0	0	0	0	0	0	0	0	0	0
	'FL2-A'	0			0	0	0	0	0	0	0	0	0
	'FL3-A'	0				0	0	0	0	0	0	0	0
	'FL4-A'	0					0.5854	0	0	0	0	0	0
	'FSC-H'	0						0	0	0	0	0	0
	'SSC-H'	0							0	0	0	0	0
	'FL2-H'	0								0	0	0	0
	'FL3-H'	0									0	0	0
	'FL4-H'	0										0	0
	'Width'	0											0

			Feature (Model 3)										
			'FSC-A'	'SSC-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature (Model 3)	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		0	0	0	0	0	0.3494	0	0	0	0
	'FL2-A'	0			0	0	0	0	0	0	0	0	0
	'FL3-A'	0				-0.0629	0	0	0	0	0	0	0
	'FL4-A'	0					0	0	-0.1393	0	0	0	0
	'FSC-H'	0						0	0	0	0	0	0
	'SSC-H'	0							0	0	0	0	0
	'FL2-H'	0								0	0	0	0
	'FL3-H'	0									0.4719	0	0
	'FL4-H'	0										0	0
	'Width'	0											0

Table 4.5: After performing a secondary regression analysis, our strategy yielded a weight quotient (containing information related to the means and standard deviations of the label-free measurements – see Methods section) used to calculate each corresponding weight (α_1, α_2, and α_3). As a result of using the genetic algorithm, the most important columns of the test statistic (or rows of the weight quotient) are selected. Selected information by the genetic algorithm are indicated by green boxes.

The means of the features, their quadratic values, and their pairwise interactions.

			Feature Means (Model 1)												
			'FSC-A'	'SSC-A'	'FL1-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL1-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature Means for α_1	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		-0.0002	0	0	0	0	0	0	0	0	0	0	0
	'FL1-A'	0			0	0	0	0	0	0	0	0	0	0	0
	'FL2-A'	0				0	0	0	0	0	0	0	0	0	0
	'FL3-A'	0					0	0	0	0	0	0	0	0	0
	'FL4-A'	0						0.7151	0	0	0	0	0	0	0
	'FSC-H'	0							0.1017	0	0	0	0	0	0
	'SSC-H'	0								0	0	0	0	0	0
	'FL1-H'	0									-1.6985	0	0	0	0
	'FL2-H'	0										0	0	0	0
	'FL3-H'	0											0	0	0
	'FL4-H'	0												0	0
	'Width'	0													0

			Feature Means (Model 2)												
			'FSC-A'	'SSC-A'	'FL1-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL1-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature Means for α_2	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		-0.5351	0	0	0	0	0	0	0	0	0	0	0
	'FL1-A'	0			0	0	0	0	0	0	0	0	0	0	0
	'FL2-A'	0				0	0	0	0	0	0	0	0	0	0
	'FL3-A'	0					0	0	0	0	0	0	0	0	0
	'FL4-A'	0						-0.2651	0	0	0	0	0	0	0
	'FSC-H'	0							0.2427	0	0	0	0	0	0
	'SSC-H'	0								0	0	0	0	0	0
	'FL1-H'	0									1.2247	0	0	0	0
	'FL2-H'	0										0	0	0	0
	'FL3-H'	0											0	0	0
	'FL4-H'	0												0	0
	'Width'	0													0

			Feature Means (Model 3)												
			'FSC-A'	'SSC-A'	'FL1-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL1-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature Means for α_3	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		0.5597	0	0	0	0	0	0	0	0	0	0	0
	'FL1-A'	0			0	0	0	0	0	0	0	0	0	0	0
	'FL2-A'	0				0	0	0	0	0	0	0	0	0	0
	'FL3-A'	0					0	0	0	0	0	0	0	0	0
	'FL4-A'	0						-0.4344	0	0	0	0	0	0	0
	'FSC-H'	0							-0.2852	0	0	0	0	0	0
	'SSC-H'	0								0	0	0	0	0	0
	'FL1-H'	0									0.3706	0	0	0	0
	'FL2-H'	0										0	0	0	0
	'FL3-H'	0											0	0	0
	'FL4-H'	0												0	0
	'Width'	0													0

Table 4.5 (continued)

The standard deviations of the features, their quadratic values, and their pairwise interactions.

Feature Standard Deviations for α_1														
		'FSC-A'	'SSC-A'	'FL1-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL1-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature Standard Deviations for α_1	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		0	0	0	0	0	0	0	0	0	0	0
	'FL1-A'	0			0	0	0	0	0	0	0	0	0	0
	'FL2-A'	0				0	0	0	0	0	0	0	0	0
	'FL3-A'	0					0	0	0	0	0	0	0	0
	'FL4-A'	0						0	0	0	0	0	0	0
	'FSC-H'	0							0	0	0	0	0	0
	'SSC-H'	0								0	0	0	0	0
	'FL1-H'	0									0	0	0	0
	'FL2-H'	0										2.1244	0	0
	'FL3-H'	0											0.4623	0
	'FL4-H'	0												0
	'Width'	0												

Feature Standard Deviations for α_2														
		'FSC-A'	'SSC-A'	'FL1-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL1-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature Standard Deviations for α_2	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		0	0	0	0	0	0	0	0	0	0	0
	'FL1-A'	0			0	0	0	0	0	0	0	0	0	0
	'FL2-A'	0				0	0	0	0	0	0	0	0	0
	'FL3-A'	0					0	0	0	0	0	0	0	0
	'FL4-A'	0						0	0	0	0	0	0	0
	'FSC-H'	0							0	0	0	0	0	0
	'SSC-H'	0								0	0	0	0	0
	'FL1-H'	0									0	0	0	0
	'FL2-H'	0										-2.7642	0	0
	'FL3-H'	0											-0.6221	0
	'FL4-H'	0												0
	'Width'	0												

Feature Standard Deviations for α_3														
		'FSC-A'	'SSC-A'	'FL1-A'	'FL2-A'	'FL3-A'	'FL4-A'	'FSC-H'	'SSC-H'	'FL1-H'	'FL2-H'	'FL3-H'	'FL4-H'	'Width'
Feature Standard Deviations for α_3	'FSC-A'	0	0	0	0	0	0	0	0	0	0	0	0	0
	'SSC-A'	0		0	0	0	0	0	0	0	0	0	0	0
	'FL1-A'	0			0	0	0	0	0	0	0	0	0	0
	'FL2-A'	0				0	0	0	0	0	0	0	0	0
	'FL3-A'	0					0	0	0	0	0	0	0	0
	'FL4-A'	0						0	0	0	0	0	0	0
	'FSC-H'	0							0	0	0	0	0	0
	'SSC-H'	0								0	0	0	0	0
	'FL1-H'	0									0	0	0	0
	'FL2-H'	0										0.6488	0	0
	'FL3-H'	0											0.1523	0
	'FL4-H'	0												0
	'Width'	0												

After we validated the final label-free lipid estimation model (Figs. 4.8(a) and 4.9), we fixed all parameters and sought to test it for label-free quantification in new circumstances. Figure 4.8(c) shows that the final model yielded exceptional prediction accuracy of the BODIPY signal for this previously unseen testing data at time points corresponding to days 7, 16, 20 and 34, and Fig. 4.10 shows the corresponding predictions at the remaining 17 time points not used in the training data set.

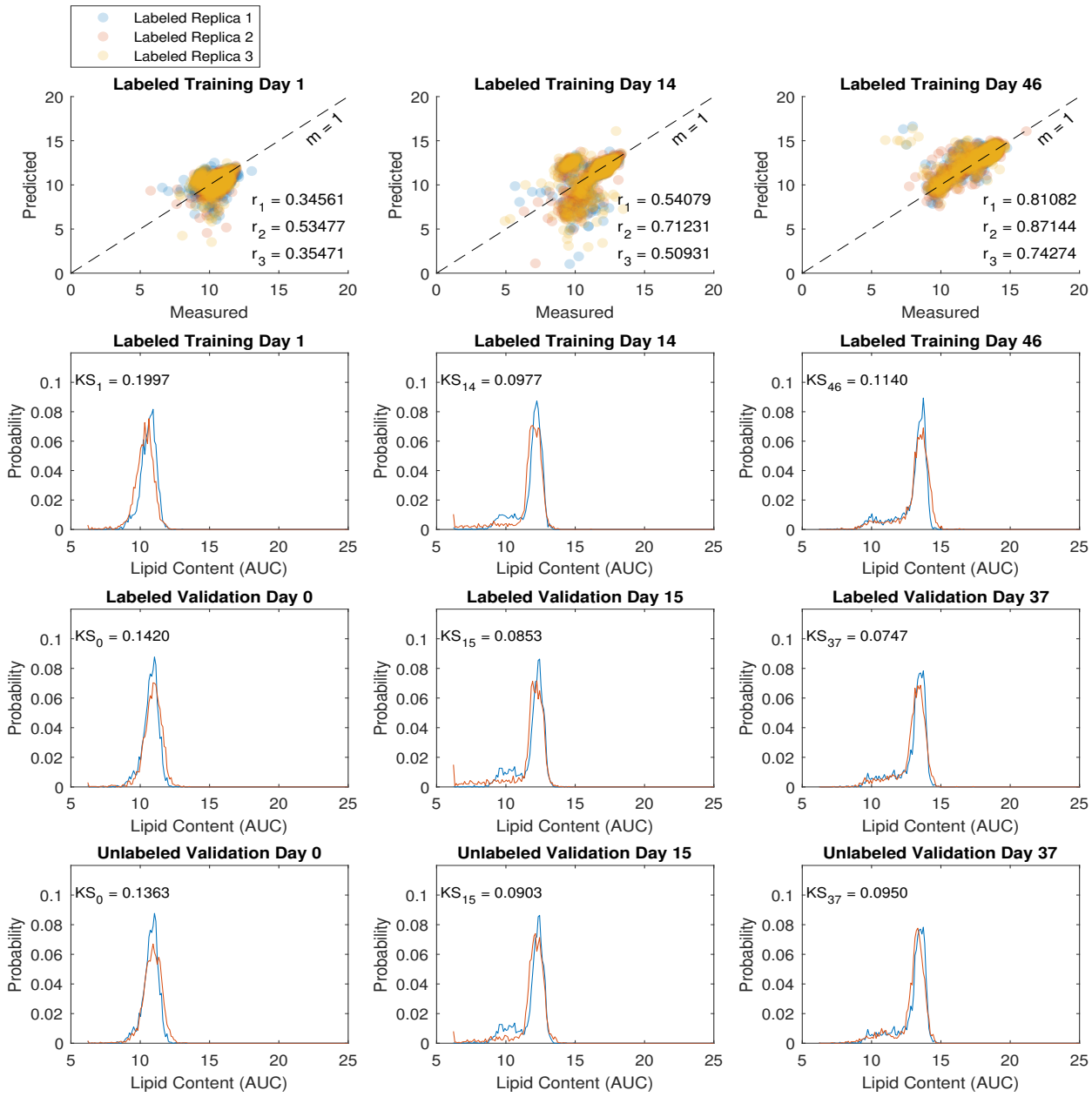

Figure 4.9: Our optimized label-free quantification strategy based on weighted models. The weights applied to the 3 trained models were estimated (using a secondary regression analysis) by measured tests statistics of the unlabeled features. The model was able to predict the lipid content of unlabeled cells with a remarkable high accuracy. The first row represents the correlation between measured and predicted values of the labeled training data. The 3 colors correspond to the 3 measurement replications at each day of FCM analysis (days 1, 14, and 46). Pearson's correlation coefficients are shown for each replication. For validation (rows 2, 3, and 4), we selected days 0, 15, and 37. The histograms show the results of prediction with this model for training (labeled cells) and validation (labeled and unlabeled cells) data. Measured histograms are in blue, predicted are in red. The KS distances between measured and predicted lipid content are shown on each plot. For all histograms, lipid accumulation are shown in arbitrary units of concentration (AUC).

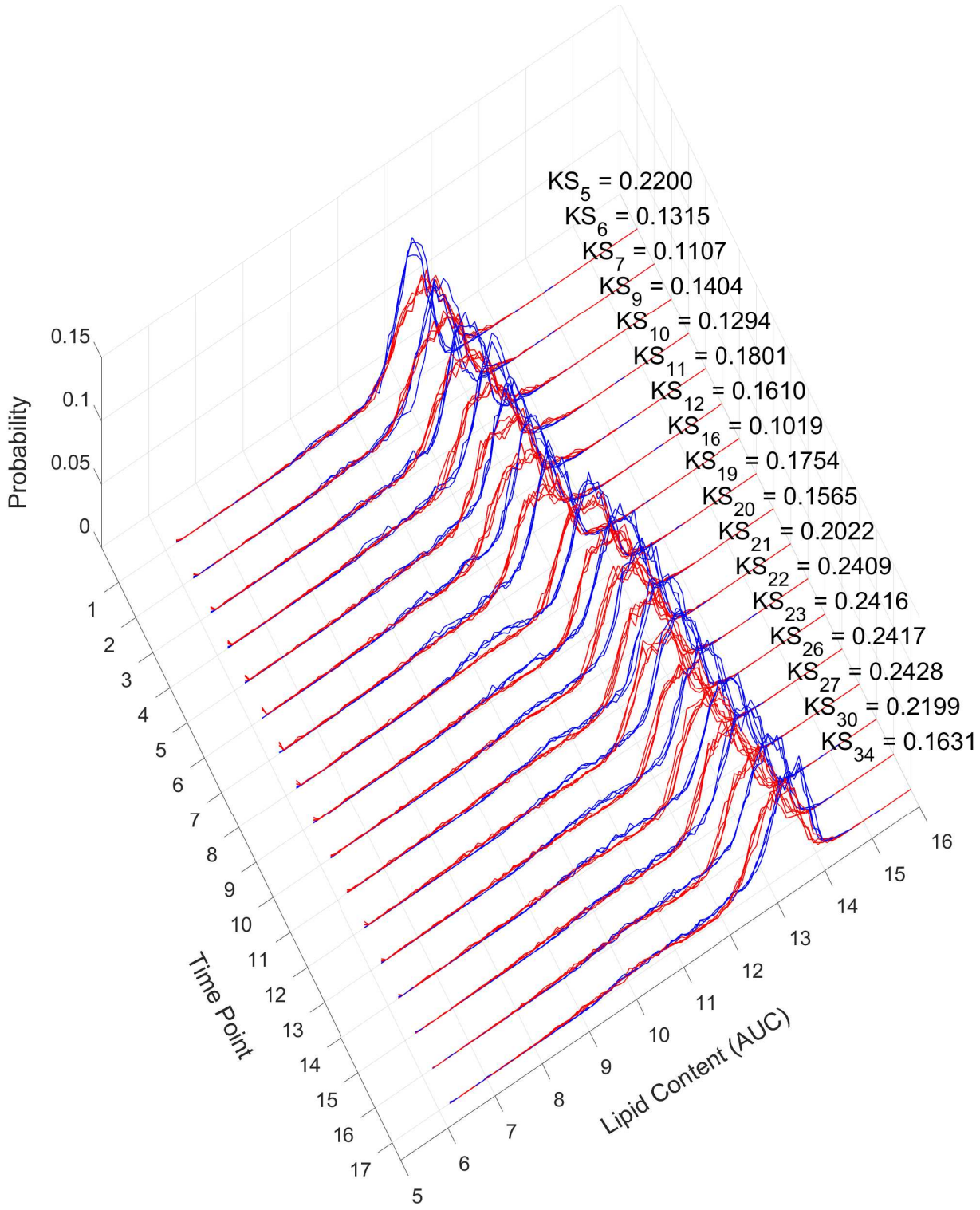

Figure 4.10: Testing the weighted model on all 17 testing time points. The data were not seen previously by the model. Measured histograms are in blue, predicted are in red. The average KS distances between measured and predicted are shown. Lipid accumulation are shown in arbitrary units of concentration (AUC).

Figure 4.11(a) also shows that the trained model correctly quantified average and standard deviation of lipid accumulation (in log scale) at each day following nitrogen starvation. We note that the training and validation data were taken at only three time points each, yet the model sufficed to predict the lipid levels for all of the remaining 17 time points. Figure 4.11(b) shows the changes in model weights, $\mathbf{a}$, which were estimated solely from the statistics of the unlabeled data (three biological replicas per time point) and without any information about the time of measurement. The figure demonstrates that the secondary model correctly adapts these weights from a domination of α_1 at early times, α_2 at middle times and α_3 at late times.

Figure 4.11(c) compares how much the lipid distributions changed over the course of the nitrogen starvation experiment as quantified using labeled (blue) or label-free (red) strategies. We found that the KS difference between the initial and final time points found for the label-based and label-free measurements were in excellent agreement of 0.83 and 0.82 respectively. Using the KS distance, we can now compare the dependence of the population distribution on changes to underlying variables, using analyses similar to those demonstrated in [62] to quantify population responses to external regulatory factors. In our case, figure 4.11(d) shows the KS distance between distributions at variable time t compared to the initial or final times and calculated from the direct lipid measurements (blue, gold) or label-free estimates (red, purple). Once again, we find that the comparisons using label-free measurements are in excellent agreement with the label-based measurements for all time points throughout the nitrogen starvation process.

Table 4.6: Selected features at the three training time points corresponding to $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$. The genetic algorithm selects feature means, and feature standard deviations for the weight quotient $\mathbf{Q}$.

M1	SSC-A*FL3-A , SSC-A*FL3-H , FL4-A*FL4-A
M2	FL4-A*FL4-A
M3	SSC-A*SSC-H , FL3-A*FL3-A , FL4-A*SSC-H , FL3-H*FL3-H
Q (means)	SSC-A*SSC-A , FL4-A*FL4-A , FSC-H*FSC-H , FL1-H*FL1-H
Q (standard deviations)	FL2-H*FL2-H , FL3-H*FL3-H

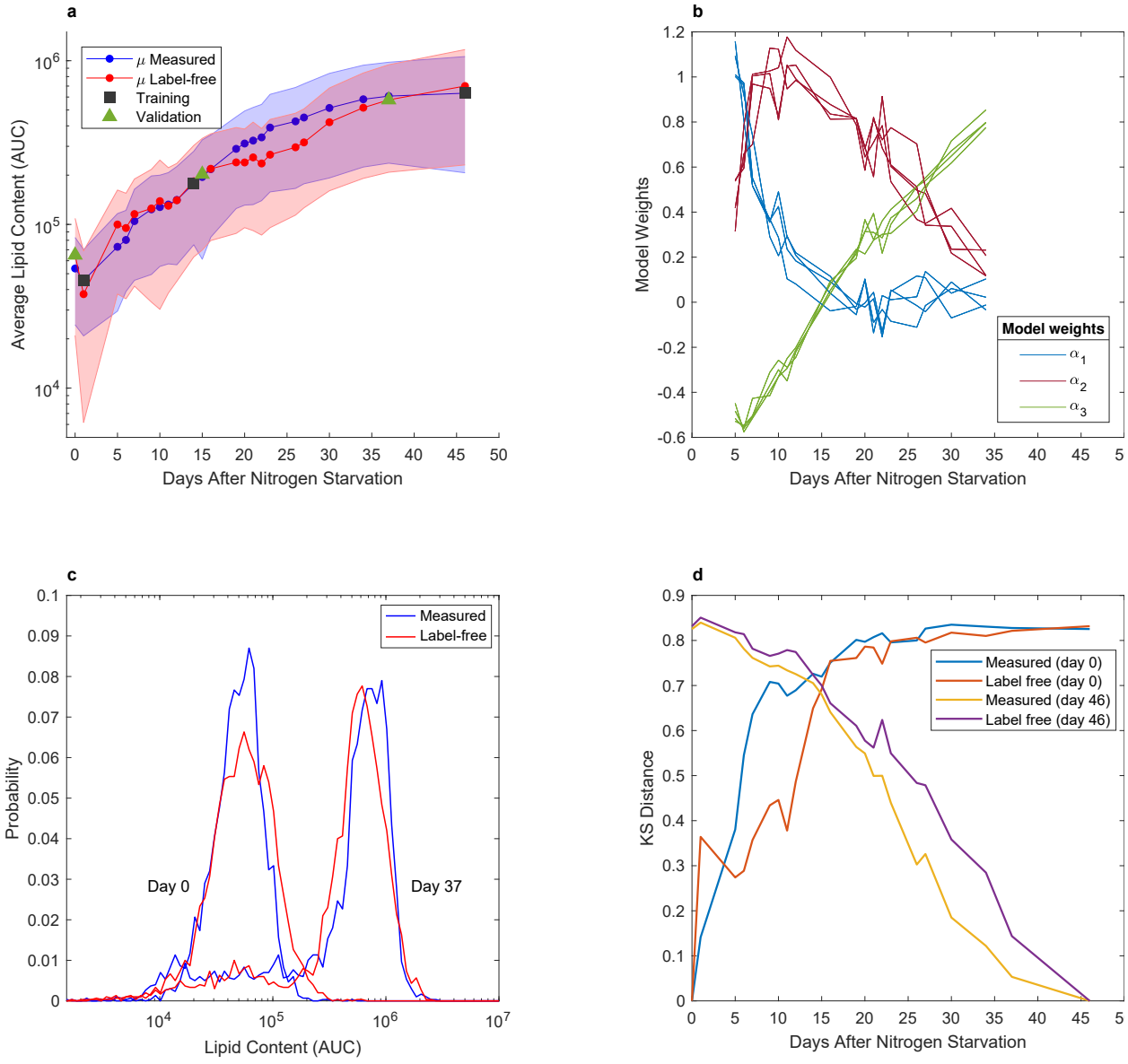

Figure 4.11: Analysis of the weighted model. (A) Average lipid content at each day after nitrogen starvation. The blue and red shaded areas show the standard deviation as measured and predicted, respectively. Training, validation, and testing time points are shown in black rectangles, green triangles, and red and blue circles, respectively. (B) Model weights calculated based on label-free information of the testing data at each day after nitrogen starvation. α_1,α_2, and α_3 correspond to the weights for model 1, model 2, and model 3, respectively. (C) Comparison of the distributions of the lipid content between days 0 and 37 after nitrogen starvation. The KS distance between the measured values is 0.8277. The KS distance between the label-free values is 0.8213. (D) Change in the KS distance with respect to time. Blue and red show the changes in KS distance between day 0 and other days after nitrogen starvation for measured and label-free data, respectively. Yellow and purple show the changes in KS distance between day 46 and other days after nitrogen starvation for measured and label-free data, respectively. All lipid contents are in arbitrary units of concentration (AUC).

Table 4.6 presents the most informative features selected by the genetic algorithm for the construction of the regression analyses at the three training times and for the multi-model weighting coefficients. Tables 4.3 - 4.5 provide the specific numerical values for all regression coefficients. In this case, the feature selection results can be interpreted in terms of known biology.To aid in this interpretation, figure 4.12(A) shows the median levels of three of the most important label-free features (SSC, FL3-A, and FL4-A) and the median measured lipid content versus time after nitrogen starvation. First, we note that the label-free SSC-A measurement is positively correlated with lipid content throughout the time course. This is easily explained by noting that SSC is indicative of granularity of the cells, and as lipids generally accumulate in lipid bodies as shown in figure 3.1, these bodies are likely to account for the increased scattering measurements in the flow cytometer. Second, we note that the fluorescence channels FL3 and FL4 exhibit weak negative correlations to lipid content at later times. Much of the fluorescence measured in these channels is likely to originate from chlorophyll. Our analysis suggests that nutrient deprived cells, which are accumulating lipids as a stress response, slowly deplete their levels of chlorophyll over time, an observation that is consistent with previous studies applying bulk cell culture analyses to other species of algae [63, 64]. For the secondary regression analysis used to define the weights of the regression analyses, the optimum found by the genetic algorithm relied primarily on these same features, but were supplemented by statistical information from other fluorescence channels, including the 530/30 nm channel that was discarded to conduct training on labeled cells.

A substantial impediment to the development of label-free strategies for flow cytometry analysis is that collective dynamics can cause one cell population to behave differently from another, even when they are prepared in similar environmental conditions. Moreover, it is not uncommon that two flow cytometers, with different settings or containing different optical components, could yield different measurements, even when used to measure the same cell populations. With these issues in mind, we next sought to test the generality of our approach when applied to a new preparation of *P. soloecismus* over time during nitrogen starvation and quantified using a different flow cytometer (BD Accuri™ C6 Plus, which has a different fluidics system) with matching detectors.

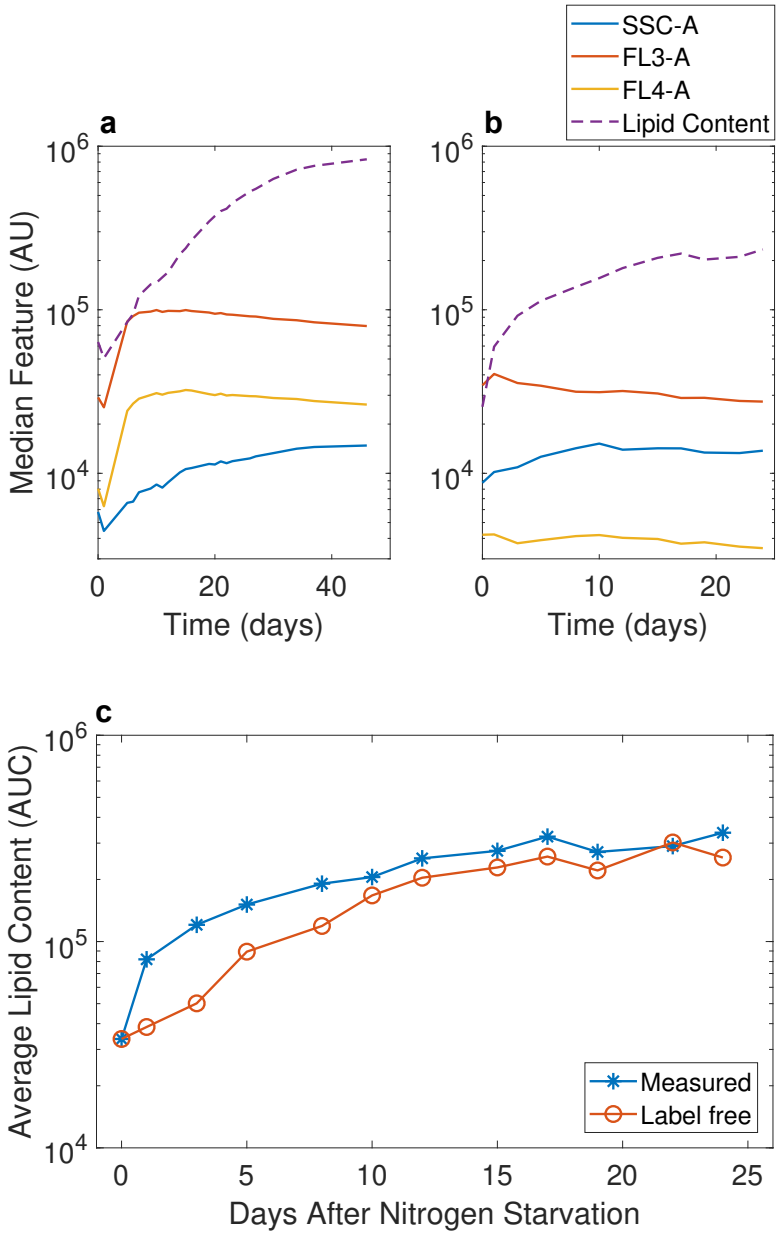

Figure 4.12: Testing the final model on independent cell populations and with a new flow cytometer. (A,B) The median values of the most informative label-free features (solid lines) and the label-based lipid measurement (dashed lines) versus time after nitrogen depletion using (A) the original cell preparation and (B) for a subsequent independent cell preparation measured with a different flow cytometer. (C) Evaluation of the final weighted model's estimates of lipid content when applied to new cells with the new flow cytometer. Blue shows the lipid accumulation for the measured values of the lipid content at each day after nitrogen starvation. Red shows the label-free predictions of the lipid content for the same days. Both curves have been normalized relative to the measure lipid level on Day 0. Lipid content are in arbitrary units of concentration (AUC).

Without any re-training of our previous model (i.e., using the same features and model parameters identified above), we sought to quantify the lipid accumulations over time for the new data set. Owing to variation in the flow cytometer and its settings, the quantitative values for the measurements changed considerably, as can be seen in figures 4.12(A) and 4.12(B), which show the median measurements for the label-free features and lipid measurement for the old and new data sets, respectively. Despite substantial differences, figure 4.12(C) demonstrates that our previous model still correctly captures the trend of increasing lipid accumulations over time based on the label-free information collected using the new cell preparation and new flow cytometer.

Finally, we used our weighted model to explore the possibility that it could be used to sort unlabeled cells according to the lipid content within those cells. To simulate this situation, we generated mixed cell populations by combining 2500 cells randomly chosen from the initial (low lipid) time point and 2500 cells randomly chosen from the final (high lipid) time point. We then applied the previously identified model ($\{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{Q}\}$) on the entire mixed population to predict the distribution of lipids from the label-free features. Figure 4.13 shows the results of this mixture for the quantification based upon labels (panel A) and based upon label-free measurements (panel B). In each case, the green lines correspond to the sub-population taken from the early time points, the purple lines correspond to the label-free measurements, and the black lines correspond to the full mixed distributions. We assumed an optimal gate (vertical dashed line in figure 4.13), and we asked what fraction of cells from the green/purple sub-populations would be correctly assigned to the low/high populations. As a benchmark, we found that the label-based sorting accuracy was 72.84% to classify low lipids cells and 94.32% to classify high lipid cells (figure 4.13(A)). The label-free sorting accuracy performed equally well at 79.64% correct classification of low lipids and 92.2% correct classification of high lipids and (figure 4.13(B)). These results suggest that label-free classification could be used in principle for sorting applications, but full evaluation of this hypothesis, as well as strategies to optimize label-free sorting gates, remain to be validated through future experimental investigations.

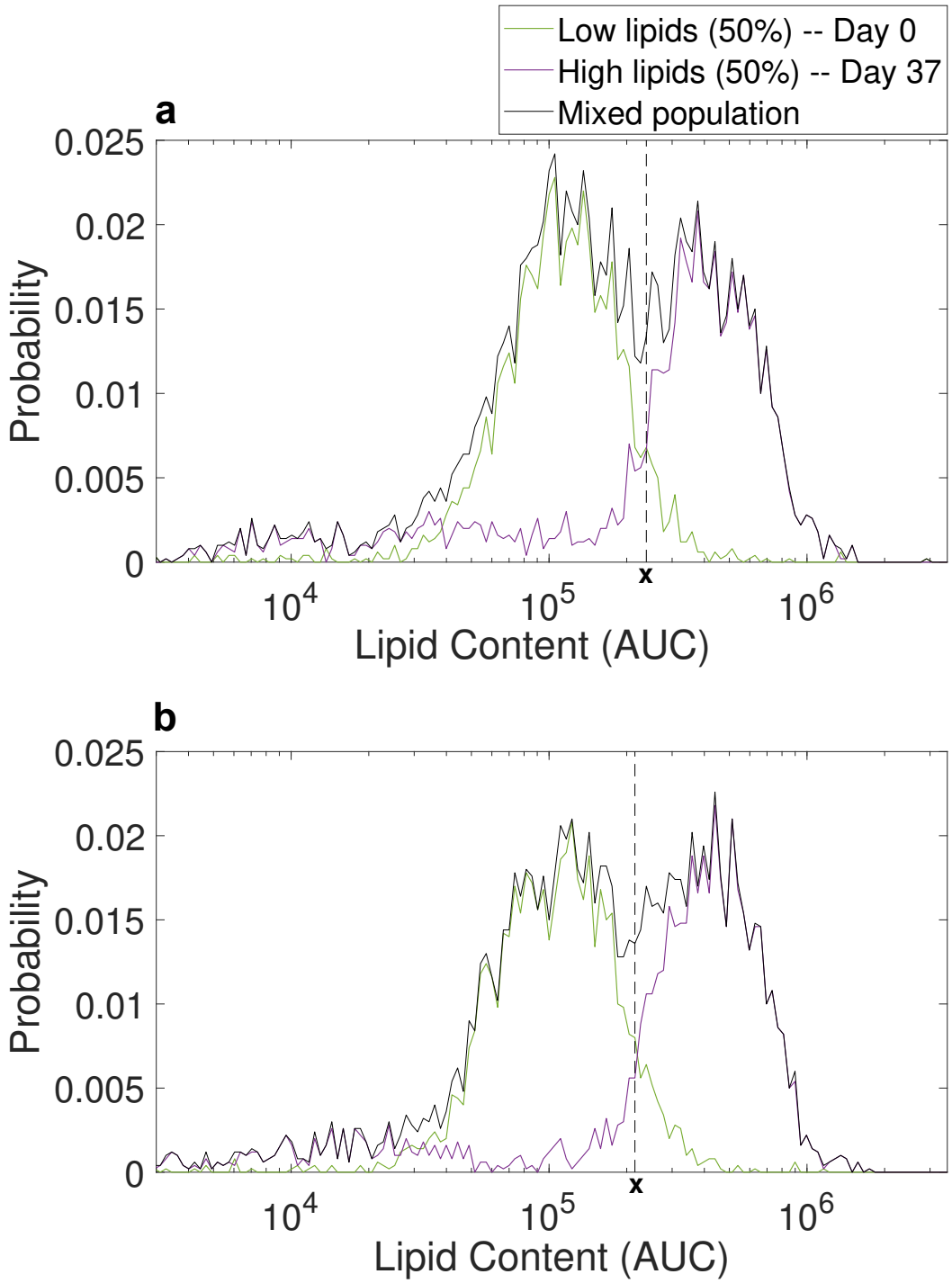

Figure 4.13: Simulation of a typical cell sorting experiment. (A) Sorting the labeled cells based on our optimized weighted model into high and low lipid content. Label-based sorting accuracy of 94.32%. (B) Sorting the unlabeled cells based on our optimized weighted model into high and low lipid content. Label-free sorting accuracy of 92.2%. For all panels, the lipid content are in arbitrary units of concentration (AUC). Simulated populations were generated by combining experimental data from 2500 cells from Day 0 and 2500 cells from Day 37.

Chapter 5

Conclusions and Future Work

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 study, 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. 4.1 and 4.2).

The key to our integrated strategy is careful consideration of the variations within heterogeneous single-cell populations. 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 5.1 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 data sets; and (iii) a new model-weighting strategy to allow application of different statistical models in different situations.

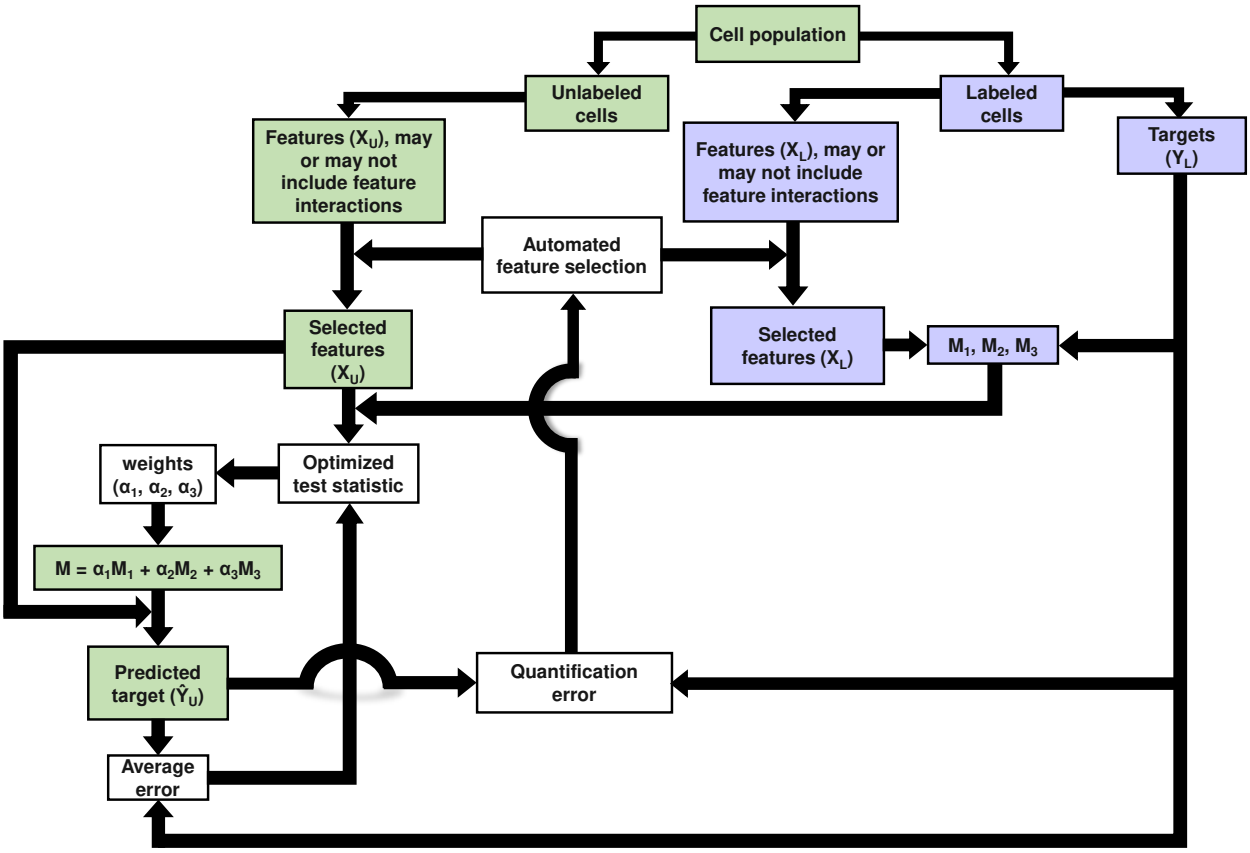

Figure 5.1: Flow diagram of the final multi-stage label-free quantification strategy.

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. 4.8, 4.9, and 4.10). 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; shown on Fig. 4.9).

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.

Applied to algae cultivation for biofuels and bioproducts (food and feed ingredients), real time monitoring of cultures can provide information on the health and productivity of the algal cells. This allows for harvesting when the cells are at maximum yield, or prior to being overtaken by pests or predators. Moreover, the ability to monitor without the addition of dyes increases the speed of analysis and decreases costs. Additionally, the ability to sort cells of interest without labels would enable selection of subpopulations with a desired phenotype of interest, such as higher content of lipid or other value added products, such as specialty oils or cannabinoids. This type of selection would allow for directed improvement of strains without direct genetic modification.

Furthermore, the ability to analyze and sort cells without labels is broadly useful as a strategy to improve productivity of cell cultivation operations for a variety of applications in the medical or pharmaceutical industry. While our machine learning approach will not substitute for the use of labels in every application in flow cytometry, it is applicable in cases where there are subtle morphological features that accompany a biological change in a cell. This approach allows identification of those morphological changes that would be normally detected with a label, and in this case we were able to identify them with machine learning and allow detection in the absence of a dye. For example in analyzing Ewing sarcoma, sorting a heterogeneous population of cells into normal and cancerous cells would require an excessive dedication of time and energy. The researchers would normally profile the cell population on a single-cell level using a microfluidic-enabled platform. The expression of the EWS-FLI1 fusion gene (which is the main cause for Ewing sarcoma) is then detected via a marker of interest. Although, finding the correct and useful biomarker can be challenging. Marker detection can be highly dependable on the flow cytometer and a marker of interest that is shown in one flow cytometer may not be detected in another. Such obstacle can lead to serious bias towards identification of the marker of interest for EWS-FLI fusion gene expression. The entire process, as a result, would be exhaustive and exposed to various errors. A correctly trained label-free computational strategy could remove such errors and save an enormous amount of time and energy.

Our analysis and development of our label-free quantification strategy was performed on *Picochlorum soloecismus* microalgae cells, which are known to have high autofluorescence (collected in the the FL3 channel, see Chapter 3). Although the signals collected from unlabeled cells are weak for those cells compared to the labeled cells, autofluorescence could be strong enough to help us correctly identify unlabeled signatures. As a follow-up study, to further expand our strategy, the same analysis approach should be taken for cells with very low or no autofluorescence capabilities. The result of such analysis could be either a more generalized label-free quantification method, or to define a specific range for autofluorescence of the investigated cells.

Although we used regression for modeling and the genetic algorithm for feature selection, our strategy is not specific to them. These two steps can be replaced and examined by any other method to model and select the most informative features. However, computational costs may vary. Our developed label-free quantification strategy can go beyond flow cytometry. We can use our method for any large data set to find features that best express a targeted information. We can utilize our current quantification strategy to examine large data sets with more features, e.g., imaging flow cytometry.

Lastly, it should be mentioned that utilizing a computational strategy can be exhaustive for a biologist. Our strategy, while successful in quantifying the target signals in unlabeled cells, cannot have the chance to be more developed if it is pure code. Hence, there is an essential need to provide a graphical user interface (GUI) for our label-free quantification tools. Developing a GUI can be done in MATLAB, as well as several other programming platforms. Moreover, since such modeling approaches and developments are computationally expensive, use of cloud-based environments such as the Google Cloud Platform™, Terra™ (formerly known as the Google FireCloud™), and Docker™ could be advantageous. [1]

[1]Some parts of the Conclusions and Future Work Chapter of this thesis is from self article by Tanhaemami et al. [28], and the Curricular Practical Training (CPT) obtained at the Dana-Farber Cancer Institute and Broad Institute of MIT and Harvard.

Bibliography

[1] Daniel R Gossett, Westbrook M Weaver, Albert J Mach, Soojung Claire Hur, Henry Tat Kwong Tse, Wonhee Lee, Hamed Amini, and Dino Di Carlo. Label-free cell separation and sorting in microfluidic systems. *Analytical and bioanalytical chemistry*, 397(8):3249–3267, 2010.

[2] Yuanyuan Han, Yi Gu, Alex Ce Zhang, and Yu-Hwa Lo. Review: imaging technologies for flow cytometry. *Lab on a Chip*, 16(24):4639–4647, 2016.

[3] Yvan Saeys, Sofie Van Gassen, and Bart N Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7):449–462, 2016.

[4] Dino Di Carlo and Luke P Lee. Dynamic single-cell analysis for quantitative biology, 2006.

[5] Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, Richard H Scheuermann, FlowCAP Consortium, DREAM Consortium, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228, 2013.

[6] Gyemin Lee, William Finn, and Clayton Scott. Statistical file matching of flow cytometry data. *Journal of biomedical informatics*, 44(4):663–676, 2011.

[7] Michael Brown and Carl Wittwer. Flow cytometry: principles and clinical applications in hematology. *Clinical chemistry*, 46(8):1221–1229, 2000.

[8] Aysun Adan, Günel Alizada, Yağmur Kiraz, Yusuf Baran, and Ayten Nalbant. Flow cytometry: basic principles and applications. *Critical reviews in biotechnology*, 37(2):163–176, 2017.

[9] Natasha S Barteneva, Elizaveta Fasler-Kan, and Ivan A Vorobjev. Imaging flow cytometry: coping with heterogeneity in biological systems. *Journal of Histochemistry & Cytochemistry*, 60(10):723–733, 2012.

[10] Bartek Rajwa, Murugesan Venkatapathi, Kathy Ragheb, Padmapriya P Banada, E Daniel Hirleman, Todd Lary, and J Paul Robinson. Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier. *Cytometry Part A*, 73(4):369–379, 2008.

[11] Karen Cheung, Shady Gawad, and Philippe Renaud. Impedance spectroscopy flow cytometry: on-chip label-free cell differentiation. *Cytometry Part A*, 65(2):124–132, 2005.

[12] Thomas Blasi, Holger Hennig, Huw D Summers, Fabian J Theis, Joana Cerveira, James O Patterson, Derek Davies, Andrew Filby, Anne E Carpenter, and Paul Rees. Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nature communications*, 7:10256, 2016.

[13] Jonghee Yoon, YoungJu Jo, Min-hyeok Kim, Kyoohyun Kim, SangYun Lee, Suk-Jo Kang, and YongKeun Park. Identification of non-activated lymphocytes using three-dimensional refractive index tomography and machine learning. *Scientific reports*, 7(1):6654, 2017.

[14] Bernd Wollscheid, Damaris Bausch-Fluck, Christine Henderson, Robert O'brien, Miriam Bibel, Ralph Schiess, Ruedi Aebersold, and Julian D Watts. Mass-spectrometric identification and relative quantification of n-linked cell surface glycoproteins. *Nature biotechnology*, 27(4):378, 2009.

[15] Claire Lifan Chen, Ata Mahjoubfar, Li-Chia Tai, Ian K Blaby, Allen Huang, Kayvan Reza Niazi, and Bahram Jalali. Deep learning in label-free cell classification. *Scientific reports*, 6:21471, 2016.

[16] Sophie E Boddington, Elizabeth J Sutton, Tobias D Henning, Alexander J Nedopil, Barbara Sennino, Anne Kim, and Heike E Daldrup-Link. Labeling human mesenchymal stem cells

with fluorescent contrast agents: the biological impact. *Molecular Imaging and Biology*, 13(1):3–9, 2011.

[17] Baoshan Guo, Cheng Lei, Hirofumi Kobayashi, Takuro Ito, Yaxiaer Yalikun, Yiyue Jiang, Yo Tanaka, Yasuyuki Ozeki, and Keisuke Goda. High-throughput, label-free, single-cell, microalgal lipid screening by machine-learning-equipped optofluidic time-stretch quantitative phase microscopy. *Cytometry Part A*, 91(5):494–502, 2017.

[18] Judith Rumin, Hubert Bonnefond, Bruno Saint-Jean, Catherine Rouxel, Antoine Sciandra, Olivier Bernard, Jean-Paul Cadoret, and Gaël Bougaran. The use of fluorescent nile red and bodipy for lipid measurement in microalgae. *Biotechnology for biofuels*, 8(1):42, 2015.

[19] Judith T Cirulis, Bridget C Strasser, John A Scott, and Gregory M Ross. Optimization of staining conditions for microalgae with three lipophilic dyes to reduce precipitation and fluorescence variability. *Cytometry Part A*, 81(7):618–626, 2012.

[20] Raphael Alford, Haley M Simpson, Josh Duberman, G Craig Hill, Mikako Ogawa, Celeste Regino, Hisataka Kobayashi, and Peter L Choyke. Toxicity of organic fluorophores used in molecular imaging: literature review. *Molecular imaging*, 8(6):7290–2009, 2009.

[21] Holger Hennig, Paul Rees, Thomas Blasi, Lee Kamentsky, Jane Hung, David Dao, Anne E Carpenter, and Andrew Filby. An open-source solution for advanced imaging flow cytometry data analysis using machine learning. *Methods*, 112:201–210, 2017.

[22] Philipp Eulenberg, Niklas Köhler, Thomas Blasi, Andrew Filby, Anne E Carpenter, Paul Rees, Fabian J Theis, and F Alexander Wolf. Reconstructing cell cycle and disease progression using deep learning. *Nature communications*, 8(1):463, 2017.

[23] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

[24] P Biller, R Riley, and AB Ross. Catalytic hydrothermal processing of microalgae: decomposition and upgrading of lipids. *Bioresource technology*, 102(7):4841–4848, 2011.

[25] Harvind K Reddy, Tapaswy Muppaneni, Jalal Rastegary, Saeid A Shirazi, Abbas Ghassemi, and Shuguang Deng. Asi: Hydrothermal extraction and characterization of bio-crude oils from wet chlorella sorokiniana and dunaliella tertiolecta. *Environmental Progress & Sustainable Energy*, 32(4):910–915, 2013.

[26] Jalal Rastegary, Saeid Aghahossein Shirazi, Tracy Fernandez, and Abbas Ghassemi. Water resources for algae-based biofuels. *Journal of Contemporary Water Research & Education*, 151(1):117–122, 2013.

[27] Clifford J Unkefer, Richard T Sayre, Jon K Magnuson, Daniel B Anderson, Ivan Baxter, Ian K Blaby, Judith K Brown, Michael Carleton, Rose Ann Cattolico, Taraka Dale, et al. Review of the algal biology program within the national alliance for advanced biofuels and bioproducts. *Algal Research*, 22:187–215, 2017.

[28] Mohammad Tanhaemami, Elaheh Alizadeh, Claire Sanders, Babetta Marrone, and Brian Munsky. Using flow cytometry and multistage machine learning to discover label-free signatures of algal lipid accumulation. *bioRxiv*, page 497834, 2018.

[29] Hazel M Davey and Douglas B Kell. Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses. *Microbiological reviews*, 60(4):641–696, 1996.

[30] Mario Díaz, Mónica Herrero, Luis A García, and Covadonga Quirós. Application of flow cytometry to industrial microbial bioprocesses. *Biochemical engineering journal*, 48(3):385–407, 2010.

[31] Chantal Jayat and Marie-Hélène Ratinaud. Cell cycle analysis by flow cytometry: principles and applications. *Biology of the Cell*, 78(1-2):15–25, 1993.

[32] Rudolf I Amann, Brian J Binder, Robert J Olson, Sallie W Chisholm, Richard Devereux, and David A Stahl. Combination of 16s rrna-targeted oligonucleotide probes with flow cytom-

etry for analyzing mixed microbial populations. *Applied and environmental microbiology*, 56(6):1919–1925, 1990.

[33] Paul K Horan and Leon L Wheeless. Quantitative single cell analysis and sorting. *Science*, 198(4313):149–157, 1977.

[34] Marion G Macey and Marion G Macey. *Flow Cytometry*. Springer, 2007.

[35] Marion G Macey. Principles of flow cytometry. In *Flow Cytometry*, pages 1–15. Springer, 2007.

[36] Howard M Shapiro. *Practical flow cytometry*. John Wiley & Sons, 2005.

[37] Alice Longobardi Givan. *Flow cytometry: first principles*. John Wiley & Sons, 2013.

[38] Katherine M McKinnon. Flow cytometry: An overview. *Current protocols in immunology*, 120(1):5–1, 2018.

[39] Pranab Dey. Flow cytometry: Basic principles, procedure and applications in pathology. In *Basic and Advanced Laboratory Techniques in Histopathology and Cytology*, pages 171–183. Springer, 2018.

[40] Robert-Jan Bleichrodt and Nick D Read. Flow cytometry and facs applied to filamentous fungi. *Fungal Biology Reviews*, 2018.

[41] Anna Porwit and Marie Christine Béné. *Multiparameter Flow cytometry in the diagnosis of hematologic malignancies*. Cambridge University Press, 2018.

[42] Steven J Kussick. Flow cytometric principles in hematopathology. In *Hematopathology (Third Edition)*, pages 686–711. Elsevier, 2019.

[43] Desmond A McCarthy. Fluorochromes and fluorescence. In *Flow Cytometry*, pages 59–112. Springer, 2007.

[44] Kotaro Hiramatsu, Takuro Ideguchi, Yusuke Yonamine, SangWook Lee, Yizhi Luo, Kazuki Hashimoto, Takuro Ito, Misa Hase, Jee-Woong Park, Yusuke Kasai, et al. High-throughput label-free molecular fingerprinting flow cytometry. *Science Advances*, 5(1):eaau0241, 2019.

[45] Hugo Pereira, Peter SC Schulze, Lisa Maylin Schüler, Tamára Santos, Luísa Barreira, and João Varela. Fluorescence activated cell-sorting principles and applications in microalgal biotechnology. *Algal research*, 30:113–120, 2018.

[46] Ata Mahjoubfar, Claire Chen, Kayvan R Niazi, Shahrooz Rabizadeh, and Bahram Jalali. Label-free high-throughput cell screening in flow. *Biomedical optics express*, 4(9):1618–1625, 2013.

[47] Alex Ce Zhang, Yi Gu, Yuanyuan Han, Zhe Mei, Yu-Jui Chiu, Lina Geng, Sung Hwan Cho, and Yu-Hwa Lo. Computational cell analysis for label-free detection of cell properties in a microfluidic laminar flow. *Analyst*, 141(13):4142–4150, 2016.

[48] Tycho M Scholtens, Frederik Schreuder, Sjoerd T Ligthart, Joost F Swennenhuis, Jan Greve, and Leon WMM Terstappen. Automated identification of circulating tumor cells by image cytometry. *Cytometry Part A*, 81(2):138–148, 2012.

[49] Sebastian Weber, María L Fernández-Cachón, Juliana M Nascimento, Steffen Knauer, Barbara Offermann, Robert F Murphy, Melanie Boerries, and Hauke Busch. Label-free detection of neuronal differentiation in cell populations using high-throughput live-cell imaging of pc12 cells. *PloS one*, 8(2):e56690, 2013.

[50] Yuanming Feng, Ning Zhang, Kenneth M Jacobs, Wenhuan Jiang, Li V Yang, Zhigang Li, Jun Zhang, Jun Q Lu, and Xin-Hua Hu. Polarization imaging and classification of j urkat t and r amos b cells using a flow cytometer. *Cytometry Part A*, 85(9):817–826, 2014.

[51] Wenhuan Jiang, Jun Qing Lu, Li V Yang, Yu Sa, Yuanming Feng, Junhua Ding, and Xin-Hua Hu. Comparison study of distinguishing cancerous and normal prostate epithelial cells by

confocal and polarization diffraction imaging. *Journal of biomedical optics*, 21(7):071102, 2015.

[52] Yuqian Li, Bruno Cornelis, Alexandra Dusa, Geert Vanmeerbeeck, Dries Vercruysse, Erik Sohn, Kamil Blaszkiewicz, Dimiter Prodanov, Peter Schelkens, and Liesbet Lagae. Accurate label-free 3-part leukocyte recognition with single cell lens-free imaging flow cytometry. *Computers in biology and medicine*, 96:147–156, 2018.

[53] Taichi Miura, Hideharu Mikami, Akihiro Isozaki, Takuro Ito, Yasuyuki Ozeki, and Keisuke Goda. On-chip light-sheet fluorescence imaging flow cytometry at a high flow speed of 1 m/s. *Biomedical optics express*, 9(7):3424–3433, 2018.

[54] Yiyue Jiang, Cheng Lei, Atsushi Yasumoto, Hirofumi Kobayashi, Yuri Aisaka, Takuro Ito, Baoshan Guo, Nao Nitta, Natsumaro Kutsuna, Yasuyuki Ozeki, et al. Label-free detection of aggregated platelets in blood by machine-learning-aided optofluidic time-stretch microscopy. *Lab on a Chip*, 17(14):2426–2434, 2017.

[55] Robert RL Guillard. Culture of phytoplankton for feeding marine invertebrates. In *Culture of marine invertebrate animals*, pages 29–60. Springer, 1975.

[56] Robert RL Guillard and John H Ryther. Studies of marine planktonic diatoms: I. cyclotella nana hustedt, and detonula confervacea (cleve) gran. *Canadian journal of microbiology*, 8(2):229–239, 1962.

[57] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[58] Samprit Chatterjee and Ali S Hadi. *Regression analysis by example*. John Wiley & Sons, 2015.

[59] Melanie Mitchell. *An Introduction to Genetic Algorithms*. Complex Adaptive Systems. MIT Press, 2014.

[60] Raul HC Lopes. Kolmogorov-smirnov test. In *International Encyclopedia of Statistical Science*, pages 718–720. Springer, 2011.

[61] Ian T Young. Proof without prejudice: use of the kolmogorov-smirnov test for the analysis of histograms from flow systems and other sources. *Journal of Histochemistry & Cytochemistry*, 25(7):935–941, 1977.

[62] Karolina Tudelska, Joanna Markiewicz, Marek Kochańczyk, Maciej Czerkies, Wiktor Prus, Zbigniew Korwek, Ali Abdi, Sławomir Błoński, Bogdan Kaźmierczak, and Tomasz Lipniacki. Information processing in the nf-κb pathway. *Scientific reports*, 7(1):15926, 2017.

[63] Sangeeta Negi, Amanda N Barry, Natalia Friedland, Nilusha Sudasinghe, Sowmya Subramanian, Shayani Pieris, F Omar Holguin, Barry Dungan, Tanner Schaub, and Richard Sayre. Impact of nitrogen limitation on biomass, photosynthesis, and lipid accumulation in chlorella sorokiniana. *Journal of applied phycology*, 28(2):803–812, 2016.

[64] Yanqun Li, Mark Horsman, Bei Wang, Nan Wu, and Christopher Q Lan. Effects of nitrogen sources on cell growth and lipid accumulation of green alga neochloris oleoabundans. *Applied microbiology and biotechnology*, 81(4):629–636, 2008.