PLOS ONE

RESEARCH ARTICLE

# Reproducibility in cytometry: Signals analysis and its connection to uncertainty quantification

**Paul N. Patrone*, Matthew DiSalvo, Anthony J. Kearsley, Geoffrey B. McFadden, Gregory A. Cooksey**

National Institute of Standards and Technology, Gaithersburg, MD, United States of America

* paul.patrone@nist.gov

Check for updates

OPEN ACCESS

**Citation:** Patrone PN, DiSalvo M, Kearsley AJ, McFadden GB, Cooksey GA (2023) Reproducibility in cytometry: Signals analysis and its connection to uncertainty quantification. PLoS ONE 18(12): e0295502. https://doi.org/10.1371/journal.pone.0295502

**Editor:** Ahmed E. Abdel Moneim, Helwan University, EGYPT

**Received:** December 12, 2022

**Accepted:** November 24, 2023

**Published:** December 22, 2023

**Copyright:** This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

**Data Availability Statement:** All relevant data are within the paper and its Supporting information files.

**Funding:** The authors received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

Signals analysis for cytometry remains a challenging task that has a significant impact on uncertainty. Conventional cytometers assume that individual measurements are well characterized by simple properties such as the signal area, width, and height. However, these approaches have difficulty distinguishing inherent biological variability from instrument artifacts and operating conditions. As a result, it is challenging to quantify uncertainty in the properties of individual cells and perform tasks such as doublet deconvolution. We address these problems via signals analysis techniques that use scale transformations to: (I) separate variation in biomarker expression from effects due to flow conditions and particle size; (II) quantify reproducibility associated with a given laser interrogation region; (III) estimate uncertainty in measurement values on a per-event basis; and (IV) extract the singlets that make up a multiplet. The key idea behind this approach is to model how variable operating conditions deform the signal shape and then use constrained optimization to "undo" these deformations for measured signals; residuals to this process characterize reproducibility. Using a recently developed microfluidic cytometer, we demonstrate that these techniques can account for instrument and measurand induced variability with a residual uncertainty of less than 2.5% in the signal shape and less than 1% in integrated area.

## 1 Introduction

In the last 30 years, cytometry has evolved as a powerful technique for clinical diagnostics, drug development, and biotechnology [1, 2]. This success has recently motivated fundamental questions and studies designed to better understand the ultimate capabilities of cytometers, as well as their potential for quantitative, reproduceable, and even traceable measurements [3–6]. However, in-depth uncertainty quantification (UQ) has yet to be fully realized, limiting efforts to refine the metrology aspects of cytometry.

In this context, signals analysis is a challenging and often overlooked task that has a significant impact on uncertainty. For example, conventional flow cytometers assume that individual events are well characterized by simple properties such as the signal area, height, and width [7–10]. However, such approaches discard the vast majority of information in the

measurement. Moreover, cytometers only interrogate a particle once per laser region, so that population variability is inherently convolved with other sources of uncertainty. As a result, it is difficult to determine whether an event corresponds to a valid measurand, characterize reproducibility on a *per-event basis*, and ultimately, quantify confidence in the measurement process.

The goal of this manuscript is to address such problems by incorporating UQ directly into the signals analysis. We propose a collection of techniques that use scale transformations to identify sources of variation in events arising from changes in particle speed, size, and brightness. The main idea behind these analyses is to represent signals in terms of low-order mathematical interpolations and define the relevant transformations in terms of generic physical models. We then use constrained optimization to map all events onto one another, which also yields the physical parameters (e.g. size) associated with the particles. Using a recently developed microfluidic cytometer (see Ref. [11]), we illustrate how this analysis can: (i) quantify phenomena such as noise in flow conditions and sample variability; and (ii) estimate the *per-event* reproducibility associated with all remaining sources of uncertainty. We also demonstrate that this analysis provides a foundation for more advanced signal processing techniques by using it to extract the individual singlets comprising a multiplet signal, i.e. an event composed of multiple overlapping singlets.

The need to incorporate UQ directly into signals analysis arises from several issues unique to cytometers. For example, cells are only measured once per laser region; moreover, it is difficult to ensure identical flow and optical conditions for all measurands. Thus, it is impossible to directly characterize repeatability and reproducibility of any measurement. Signals analysis and mathematical modeling therefore take on new roles, since they allow us to answer the questions, "how does variation in measurement conditions deform the signals, and how can these deformations be undone?" In this way, the theory approximately reconstructs an imaginary scenario in which one could repeat measurements on identical particles, thereby overcoming experimental limitations.

Multiplets provide a similar motivation for our analysis techniques. Because engineering solutions alone cannot prevent multiplets [12–15], common practice discards their information. However, this introduces uncertainty into population variability estimates [12, 14]. Perhaps worse, common data analysis strategies that rely on integrated areas cannot distinguish dim doublets from bright singlets, for example; see Fig 1. This leads to an uncomfortable situation in which even the identity and number of measurands have potentially unquantifiable uncertainties. Again, we demonstrate that signals analysis is an appropriate avenue to address such problems, since it allows us to extract individual singlets by understanding how they combine to make the multiplet. More generally, these examples illustrate that signals analysis is a useful tool for untangling sources of uncertainty that become intertwined by the measurement process.

This example of *multiplet deconvolution* highlights another theme of our work: there is a natural feedback between UQ and signals analysis that must be exploited to realize the full potential of cytometry. While it is obvious that uncertainty estimates arise from data analysis, we also illustrate the reverse: *new data analyses are enabled by UQ*. For example, we recast multiplet deconvolution as the task of finding the most probable singlets that reconstruct the original multiplet. Thus, estimating the uncertainty in the shape of singlets is a prerequisite step. Similar arguments apply to the related but distinct problem of *multiplet detection*. More generally, UQ can inform data analysis by ensuring that results are physically meaningful. These observations clarify our stance that UQ is the broad set of tasks that increase confidence in a measurement.

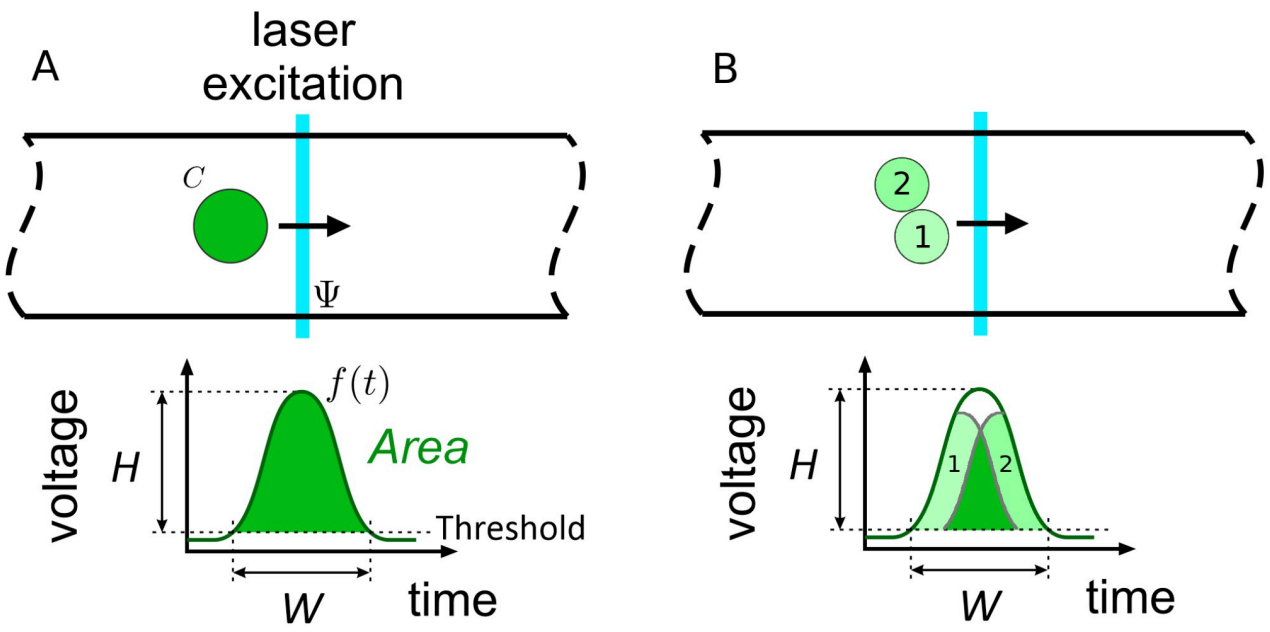



**Fig 1. Schematic of the signal generation process associated with a cytometry event.** *Panel A*. As a bead or cell traverses the laser profile, the concentration of fluorophores and object shape convolve with the laser profile to create fluorescent light. This light is converted to a voltage by a photodetector. See Eq (1) for a mathematical description of this process. Conceptually we can imagine the signal as beginning and ending when the particle reaches an initial and final position as it moves with the fluid. We refer to this as a constant-trajectory signal. In practice, it is easier to acquire a signal over a constant time interval, i.e. by approximately centering the peak in a time window of fixed duration, without regard to the distance traveled. See Eq (1) and the surrounding text for definitions of *C*, Ψ, and *f*. *Panel B*. Simple metrics such as the signal area, height, and width may not be able to distinguish the signals generated by dim doublets and a bright singlet. Compare with Panel A.

https://doi.org/10.1371/journal.pone.0295502.g001

Because the present work develops new analysis tools, it is useful to perform validation measurements on particles with well characterized properties. For this reason, the examples below are restricted to fluorescent microspheres. However, the results herein suggest that extensions to cell-based measurements are viable, and we point to relevant issues throughout. (Such work is also the subject of a manuscript in preparation.) We note also that the current manuscript does not consider issues associated with gating strategies and quantification of population variability per se. Such topics are left for future work.

The reader should also note that while our analysis strives to be as general as possible—for example, we need not specify a form of the laser profile—we make certain assumptions that may not apply to all cytometers or measurands. These relate to uniformity of the laser profile perpendicular to the flow direction and light-collection geometric factors. Importantly, these assumptions allow us to formulate the scale transformations as relatively simple, closed-form expressions, thereby facilitating downstream numerical optimization. An alternative would be to tailor our approach to a specific cytometer through more detailed modeling, but this may introduce significant computational overhead. Such issues are discussed in Sec. 5.

The rest of the manuscript is organized as follows. Section 2 presents the main assumptions of our analysis in the context of a generic model of a cytometry measurement (2.1); derives the associated scale transformations (2.2); and formulates the optimization problem needed to undo deformations associated with variable measurement conditions (2.3). Section 3 validates these methods using experimental data. Section 4 highlights the usefulness of this analysis for UQ (4.1) and multiplet deconvolution (4.2). Section 5 considers this work in the greater context of cytometry and signals analysis thereof. An S1 Appendix presents technical aspects of the multiplet deconvolution.

## 2 Motivation and main ideas

A key observation motivates our work: under generic conditions and for a given cytometer, *the shapes of all signals are identical up to a set of linear transformations that depend only on the particle properties*. As will become clear, these transformations can be interpreted as changes of units that bring the numerical values of different measurements into agreement. This provides a quantitative framework for comparing events, e.g. to determine if they are valid measurands, characterize their properties, and estimate variation about mean behavior. The goal of this section is to specify the conditions under which this observation holds and develop the tools needed to compare signals.

A key challenge in formulating this analysis is that the measurements are given in a form that does not permit direct comparison. That is, we can easily model events over a constant trajectory, but the instrument generally outputs signals having a fixed time-window. Converting between these requires knowledge of the unknown particle speed. Part of our task is therefore to ensure that the data analysis can recast the signals we acquire into ones that can be modeled. The multitude and complexity of steps involved motivates a separation of the analysis into three subsections. In Sec. 2.1, we propose a model that characterizes deformation in signals arising from differences in particle speed and size, assuming they each have the same trajectory. In Sec. 2.2, we develop the machinery to convert signals acquired over the same time interval to ones having the same trajectory length. Section 2.3 combines these results to solve the backwards problem: determine the particle properties by undoing the signal deformations.

### 2.1 Physical effects causing signal variation

Consider the interplay between physical, chemical, and biological processes during a cytometry measurement. We take an event to be the fluorescence signal $f(t)$ collected as a function of time $t$ while a bead or cell passes a laser excitation region; see Fig 1. For particles moving with a constant velocity and parallel to the fluid flow, a model of this process is

$$f(t) = \int_{D_L} \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\, C(x, y, z - vt)\boldsymbol{\Psi}(x, y, z)\boldsymbol{\Phi}(x, y, z), \tag{1}$$

where $C(x, y, z - vt)$ is the concentration of fluorophores on the particle, $\Psi(x, y, z)$ is the laser light intensity, $\Phi(x, y, z)$ is the amount of emitted fluorescent light (per unit laser light and fluorophore number) coming from $(x, y, z)$ and collected by the photodetector, $v$ is the constant advection velocity of the particle, $D_L$ is the spatial domain over which the signal is generated, and $z$ is parallel to the direction of flow, i.e. the axial direction. The goal of our analysis is to quantify how $f(t)$ changes as a function of the bead radius $R$, velocity $v$, and concentration $C$, so that we can then solve the reverse problem of determining these parameters from $f(t)$.

We first simplify Eq (1). Considering, for example, the microfluidic cytometer in Ref. [11], we assume that Eq (1) can be reduced to

$$f(t) = \int_{D_L} \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\, C(x, y, z - vt)\psi(z), \tag{2}$$

where $\psi(z)$ depends only on $z$ and accounts for both the laser and collection optics. In an abuse of terminology, we henceforth refer to $\psi(z)$ as the laser profile.

Eq (2) is justified for a wide range of operating conditions. For example, letting $\boldsymbol{\rho} = (x, y)$ denote the radial direction and $\rho = |\boldsymbol{\rho}|$ be the magnitude of $\boldsymbol{\rho}$, Eq (2) applies when: (i) differently sized, radially symmetric particles whose centers follow streamline $\boldsymbol{\rho}_0$ encounter a laser

profile and geometric factors whose product is linear in $\boldsymbol{\rho}$, so that

$$\boldsymbol{\Psi}(x,y,z)\boldsymbol{\Phi}(x,y,z) = \psi_1(z) + \boldsymbol{\alpha}(z)\cdot(\boldsymbol{\rho}-\boldsymbol{\rho}_0), \tag{3a}$$

$$C = C(\rho, z), \tag{3b}$$

for some vector function $\boldsymbol{\alpha}(z)$; and (ii) arbitrary shaped, non-rotating particles on any streamline encounter a laser profile and geometric factors that are independent of $\boldsymbol{\rho}$, so that

$$\boldsymbol{\Psi}(x,y,z)\boldsymbol{\Phi}(x,y,z) = \psi(z). \tag{4}$$

Either case may happen when the particles are small relative to the laser interrogation region. In particular, case (i) can be shown by symmetry arguments applied to Eq (1).) See Ref. [11] for relevant details associated with the device studied in this manuscript, including in particular the focusing strategy used to ensure that particles do not cross streamlines.

We model $C(x, y, z)$ as a step function of the form

$$C(x,y,z) = c\Theta(R^2 - x^2 - y^2 - z^2), \tag{5}$$

where $c$ is a constant fluorophore concentration and the Heaviside function is defined such that $\Theta(r) = 1$ if $r > 0$ and $\Theta(r) = 0$ otherwise. Eq (5) corresponds to a bead or cell with a uniform concentration of fluorophores throughout its volume. Through appropriate modification of the Heaviside function this expression can describe, for example, surface concentrations. Moreover, Eq (5) can be parameterized to model non-radial deformations associated with deformable cells, although such tasks are beyond the scope of this manuscript.

In light of these simplifications, Eq (1) becomes

$$f(t) = c\pi\int_{-L_0}^{L_0} \mathrm{d}z[R^2 - (z-vt)^2]\Theta[R^2-(z-vt)^2]\psi(z), \tag{6}$$

where where we assume that the laser is fully contained inside the domain $[-L_0, L_0]$ for some length $L_0$. For later convenience, we assume (without loss of generality) that $L_0$ satisfies the inequality

$$0 < a < L_0 - R, \tag{7}$$

where the domain $[-a, a]$ is the smallest set for which $\psi(z) > 0$. Physically, inequality (7) implies an event in which we track a particle starting before it enters the laser and ending after it fully exits; the corresponding function $f(t)$ is padded on the left and right by zeros. (We assume that any constant offset or background has been subtracted from the fluorescence signal.) Because $L_0$ is constant, the locations at which we start and stop tracking the particle are fixed. Taking $t = 0$ as the time when the particle is at the (arbitrary) "center" $z = 0$ of the laser, $f(t)$ is defined on the interval $[-\tau, \tau]$, where $\tau = L_0/v$.

Eq (6) implies that increasing or decreasing $v$ "compresses" or "stretches" the signal in the time domain. Thus, we should be able to solve the reverse problem: estimate the relative velocities by undoing the deformations in such a way that the signals coincide. To achieve this, it is convenient to consider Fourier representations of the signals [16, 17]. We denote the corresponding transforms of $\psi(z)$ and $[R^2 - z^2]\Theta[R^2 - z^2]$ by $\hat{\psi}(k)$ and $\hat{g}(k; R)$, where a simple computation yields

$$\hat{g}(k;R) = \begin{cases} \frac{-4RL_0^2\cos(kR/L_0)}{k^2} + \frac{4L_0^3\sin(kR/L_0)}{k^3} & k \neq 0 \\ \frac{4}{3}R^3 & k = 0 \end{cases} \tag{8}$$

and $k = \pi n$ for any integer $n$. (Note that we use the normalization $\mathfrak{F}[\cdot] = \int_{-L_0}^{L_0} \cdot \exp(-ikz/L_0)\mathrm{d}z$, where $\mathfrak{F}[\cdot]$ is the Fourier transform acting on $\cdot$.) In practice, we assume that $-M \leq n \leq M$, where $M$ is a mode-cutoff associated with the noise floor of the signal; see also Sec. 3. Unless otherwise stated, all sums over $k$ range from $-M\pi$ to $M\pi$. Note that $\hat{\psi}(k)$ is unknown.

In light of these assumptions, Eq (6) becomes

$$f(t) = \frac{c\pi}{2L_0}\sum_k \hat{\psi}(k)\int_{-L_0}^{L_0} \mathrm{d}z\,\left[R^2 - (z-vt)^2\right] \Theta[R^2 - (z-vt)^2]e^{ikz/L_0}, \tag{9}$$

where we have used the fact that

$$\psi(z) = \frac{1}{2L_0}\sum_k \hat{\psi}(k)e^{ikz/L_0}. \tag{10}$$

To further simplify this, use inequality (7) to impose the periodicity relationship $C(x, y, z - vt) = C(x, y, z - vt + 2nL_0)$, where $n \in \mathbb{Z}$ is an integer. This assumption does not change the signal on the domain $-L_0/v \leq t \leq L_0/v$. However, it does allow us to invoke the convolution theorem [18], which yields

$$f(t) \quad = c\pi\sum_k \hat{\psi}(k)\hat{g}(k;R)e^{ikt/\tau}. \tag{11}$$

Eqs (8) and (11) illustrate how the measurand and measurement conditions affect the signal. Varying the concentration $c$ and velocity $v$ (which controls $\tau$) alters the height and width of the signal. Changes in particle size $R$ have more complicated effects as characterized by Eq (8), since a particle is convolved with a larger fraction of the laser profile as $R$ increases. Taking the Fourier transform of Eq (11) with respect to time (over the interval $-\tau$ to $\tau$) yields

$$\hat{f}(k;\tau) = 2c\pi\tau\hat{\psi}(k)\hat{g}(k;R). \tag{12}$$

The inclusion of $\tau$ in the argument of $\hat{f}(k;\tau)$ is to emphasize the time-domain over which the transform is taken, which is important in the following sections. Letting $f(t)$ and $f_0(t)$ denote test and reference signals, the laser profile $\hat{\psi}(k)$ is eliminated by taking the ratio

$$\frac{v\hat{f}(k;\tau)}{v_0\hat{f}_0(k;\tau_0)} = \frac{c\hat{g}(k;R)}{c_0\hat{g}(k;R_0)}, \tag{13}$$

where quantities with 0 subscripts are associated with $f_0(t)$ and the trajectory length $L_0$ is assumed to be the same for both particles. In the remainder of the manuscript, we assume that $\tau_0 = L_0/v_0$, i.e. the reference particle defines the length and time scales against which we compare all events.

Eq (13) is the key result of this section: it defines the shape of one signal in terms of another and as a function of the particle properties, provided the trajectories are identical.

## 2.2 Converting between fixed-time and fixed-trajectory representations

In practice, it is difficult to measure an event over a fixed trajectory length and variable time domain. It is more common for acquisition to occur in fixed time windows, which we assume to be $[-\tau_0, \tau_0]$. However, if $\tau_0$ remains fixed, the corresponding trajectory lengths of each particles must vary, which violates the key assumption of the previous section, so that we cannot use Eq (13). (Allowing trajectory length to vary per-particle would yield varying spectral

representations of $\psi(z)$, making it difficult to eliminate the laser profile via a relationship akin to Eq (13).).

The resolution to this problem is to identify a mapping that converts a fixed-time Fourier representation to one on a fixed spatial domain. Observe first that in Eq (12), the dimensionless frequency $k$ is common to all signals; moreover, $\hat{f}(k;\tau)$ is linear in the particle-dependent time-scale $\tau = L_0/v$. Instead of $\hat{f}(k;\tau)$ we are given a measured signal $\hat{f}_m(k;\tau_0)$, where $\tau_0 = L/v = L_0/v_0$ and $L \neq L_0$ if $v \neq v_0$. The linearity of $\hat{f}(k;\tau)$ in $\tau$ suggests that there is a transformation matrix $\mathbf{X}$ such that

$$\hat{f}(k;\tau) = \sum_{k'} X_{k,k'}(\tau,\tau_0)\hat{f}_m(k';\tau_0), \tag{14}$$

where $\hat{f}(k;\tau)$ and $\hat{f}_m(k';\tau_0)$ are the constant-trajectory and constant-time representations of the event. The subscript $m$ emphasizes that the latter is the measured signal. The time domains associated with the right and left sides of Eq (14) are $D_v = [-\tau, \tau]$ and $D_0 = [-\tau_0, \tau_0]$. (The subscript $v$ on $D_v$ is to emphasize that $v$, not $L_0$, controls $\tau$.) We temporarily assume that $\tau$, $\tau_0$, and $L_0$ are known; in the next section we show how to find these quantities.

Eq (14) does not itself specify the operator $\mathbf{X}$. Its definition depends on how we wish to extend or truncate a signal when mapping it to a constant-trajectory domain. There are two equivalent perspectives one may take: active or passive. In the former, we deform (i.e. expand or contract) the time-series so that its constant-trajectory domain $D_v$ becomes $D_0$. This collapses all time-series onto one another. In the latter perspective, we deform $D_0$ to become $D_v$, keeping the time-series unchanged. While both approaches are equivalent, the corresponding derivations have subtle differences.

To better understand these issues, consider the case in which $v > 1$. Clearly $D_v$ is a subset of $D_0$. In the passive approach, we are given $\hat{f}_m(k;\tau_0)$, i.e. the Fourier modes on $D_0$, and wish to find the corresponding modes on the smaller domain $D_v$. In this sense, $\mathbf{X}$ is a "domain-restriction" operator. Assume that inequality (7) holds, so that $f_m(t) \to 0$ as $t \to \pm L_0/v$. Defining the restriction $f_r(t) = f_m(t)$ for $t \in [-\tau, \tau]$, we take the Fourier transform of $f_r(t)$ on $D_v$ to find

$$\begin{aligned}\hat{f}(k;\tau) &\coloneqq \hat{f}_r(k;\tau) \\ &= \frac{v_0}{v}\sum_{k'} \operatorname{sinc}\left(k - \frac{k'v_0}{v}\right)\hat{f}_m(k';\tau_0),\end{aligned} \tag{15}$$

so that $X_{k,k'} \coloneqq (v_0/v)\operatorname{sinc}(k - k'v_0/v)$ and the $\coloneqq$ symbol means that we *define* the left-hand side to be equal to the right-hand side. We adopt the convention that [19]

$$\operatorname{sinc}(x) = \sin(x)/x. \tag{16}$$

To arrive at this result via the active perspective, we scale the argument of $f_m(t)$ by a factor $v_0/v$, so that $f_m(t) \to f_m(v_0 t/v)$. Then, restricting $f_m(v_0 t/v)$ to the domain $D_0$ and taking the Fourier transform yields an estimate that we denote $\hat{f}_r(k;\tau_0)$ in a slight abuse of notation. (The restriction $f_r(t)$ is only defined on $D_v$, so that by the notation $\hat{f}(k;\tau_0)$ we mean the Fourier transform of the expanded version of $f_r(t)$, which has the constant-trajectory domain $D_0$.) To convert this to $\hat{f}(k;\tau)$ (i.e. the mode-weight on the domain $D_v$), we invoke Eq (13) with $R = R_0$ and $c = c_0$, which yields $\hat{f}_r(k;\tau) = (v_0/v)\hat{f}_r(k;\tau_0)$. This derivation suggests the interpretation that $\mathbf{X}$ is also a "signal-expansion" operator, since $v > v_0$. We leave it as an exercise to the reader to show the the corresponding matrix $\mathbf{X}$ is identical to that given by Eq (15).

When $\nu < 1$, the domain $D_\nu = [-\tau, \tau]$ contains the interval $[-\tau_0, \tau_0]$, so that in the passive-interpretation, $\mathbf{X}$ is a "domain-extension" operator that extends $f_m(t)$ from $D_0$ onto $D_\nu$. While there is no unique extension $f_e(t)$, we assume without loss of generality that $f_m(t) \to 0$ as $t \to \pm\tau_0$. (Signals violating this condition can be treated as multiplets, which are considered in Sec. 4.) This together with inequality (7) suggests the definition

$$f_e(t) := \begin{cases} f_m(t) & t \in [-\tau_0, \tau_0] \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Again, taking the Fourier transform on $D_\nu$, one finds

$$\hat{f}(k;\tau) := \sum_{k'} \operatorname{sinc}\left(k\frac{\nu}{\nu_0} - k'\right)\hat{f}_m(k';\tau_0), \tag{18}$$

so that $X_{k,k'} := \operatorname{sinc}[k(\nu/\nu_0) - k']$. In the active interpretation, $\mathbf{X}$ is a "signal-contraction" operator; we leave derivation of the equivalence with Eq (18) as an exercise for the reader.

Combining these results, we define $f(t)$ on $D_\nu$ to be

$$f(t) := \begin{cases} f_r(t) & \nu \geq \nu_0 \\ f_e(t) & \nu \leq \nu_0, \end{cases} \tag{19}$$

with $\hat{f}(k;\tau)$ being the corresponding Fourier transform. In both Eqs (15) and (18), setting $\nu \to \nu_0$ yields the identity transformation, as expected.

In practical settings, distinct signals $f_m(t)$ may not be identically centered on the domain $D_0$ (i.e. before applying $\mathbf{X}$). However, Eq (6) assumes that the center of any given particle is at $z = 0$ when $t = 0$, and moreover $\mathbf{X}$ deforms the signal about $t = 0$. Thus, it is necessary to consider transformations of the form $f_m(t) \to f_m(t + \Delta t)$. As before, assume that $f_m(t)$ is padded by zeros as $t \to \pm\tau_0$, so that cyclic permutations by sufficiently small $\Delta t$ only wrap zeros around the left and right sides of the peak. Assuming that $f_m(t)$ is periodic with a period of $2\tau_0$, one finds that

$$\begin{aligned} f_m(t) &= \sum_k \hat{f}_m(k;\tau_0)e^{ikt/\tau_0} \\ &\to \sum_k \hat{f}_m(k;\tau_0)e^{ikt/\tau_0 + ik\Delta t/\tau_0}, \end{aligned} \tag{20}$$

so that $\hat{f}_m(k;\tau_0) \to \hat{f}_m(k;\tau_0)\exp(ik\Delta t/\tau_0)$ under time translation. While the padding assumption is technically not necessary for the validity of Eq (20), it ensures that subsequent transformations by $\mathbf{X}$ do not truncate non-zero parts of the signal.

## 2.3 Signal matching

Assume that we have reference and test signals $f_0(t)$ and $f_m(t)$, both mapped to the interval $-\tau_0 \leq t \leq \tau_0$ and sufficiently padded by zeros on left and right. Both Fourier expansions use the same set of frequencies. Let $\mathbf{k}$ denote the vector of frequencies, and $\hat{\mathbf{f}}_0$ and $\hat{\mathbf{f}}_m$ denote the corresponding vectors of Fourier coefficients. Also define the matrix operators $\mathbf{X}$, $\mathbf{T}$, and $\mathbf{S}$

having elements

$$X_{k,k'}(v, v_0) \coloneqq \begin{cases} (v_0/v)\mathrm{sinc}(k - k'v_0/v) & v \geq v_0 \\ \mathrm{sinc}(kv/v_0 - k') & v \leq v_0 \end{cases}, \tag{21}$$

$$T_{k,k'}(\Delta t) \coloneqq \delta_{k,k'} e^{ik\Delta t/\tau_0}, \tag{22}$$

$$S_{k,k'}(R; R_0) \coloneqq \delta_{k,k'} \hat{g}(k; R_0)/\hat{g}(k; R), \tag{23}$$

where $\delta_{k,k'} = 1$ if $k = k'$ and $\delta_{k,k'} = 0$ if $k \neq k'$. Eqs (21) and (22) derive from matrix analogues of Eqs (15), (18) and (20). The quantity $\hat{g}(k; R)$ is given by Eq (8) and defines the relative transformation associated with changing particle radius. The vector analogue of $\hat{f}(k; \tau)$ is denoted $\hat{\mathbf{f}}(\mathbf{k}; \tau)$.

For noise-less signals and a known set of transformation parameters $\Delta t$, $R$, $R_0$, $v$, and $v_0$, Eq (13) implies that

$$\mathbf{\Lambda}\hat{\mathbf{f}}_m(\mathbf{k}; \tau_0) - \hat{\mathbf{f}}_0(\mathbf{k}; \tau_0) = 0, \tag{24}$$

where

$$\mathbf{\Lambda} = \frac{(v/v_0)}{(c/c_0)}\mathbf{S}(R, R_0)\mathbf{X}(v, v_0)\mathbf{T}(\Delta t). \tag{25}$$

The interpretation of Eq (24) is straightforward. The product $\mathbf{\Lambda}\hat{\mathbf{f}}_m(\mathbf{k}; \tau_0)$: (i) centers $f_m$; (ii) scales it to the domain $D_v$; (iii) matches the radius scaling to $R_0$; and (iv) normalizes the amplitude to the reference signal. The order of operations of the matrices mirrors the assumptions of the analysis above and cannot be changed. (Note that the transformations matrices do not all commute with one another. For example, centering a peak and then truncating its edges can yield a different signal than truncating followed by centering.)

In practice, the transformation parameters must be determined from noisy signals. This motivates the objective

$$\mathcal{L} = \left|\left|\mathbf{\Lambda}\hat{\mathbf{f}}_m - \hat{\mathbf{f}}_0\right|\right|^2, \tag{26}$$

where the square is interpreted as the sum over magnitude-squared of the elements of the complex vector argument. Ostensibly minimizing $\mathcal{L}$ as a function of the $\Delta t$, $R$, $R_0$, $v$, $v_0$, $c$, and $c_0$ should yield their numerical values. However, many of these quantities only appear in $\mathbf{\Lambda}$ via the ratios $v/v_0$, $R_0/L_0$, and $c/c_0$. As a result, we can only determine their relative values. This reflects a deeper freedom that we have not yet exploited: the ability to arbitrarily pick the numerical scale of the coordinate system. For convenience, we pick $c_0 = 1$, $L_0 = 1$ (i.e. the reference particle traverses a distance of 2), and $v_0 = 1$, which implies that $\tau_0 = 1$, with all units now being dimensionless. The latter two choices fix the time and length scales of the system, so that $R$, $R_0$, and $v$ are defined relative to the reference trajectory and velocity. We use this convention throughout the remainder of the manuscript.

With these choices, we define the true $c$, $v$, $R$, $R_0$, and $\Delta t$ to be solutions to the optimization problem

$$\{c^*, v^*, R^*, R_0^*, \Delta t^*\} = \underset{\{c,v,R,R_0,\Delta t\}}{\mathrm{argmin}}\, \mathcal{L}(c, v, R, R_0, \Delta t; \hat{\mathrm{f}}_m).$$

The objective $\mathcal{L}$ is highly non-linear and may have local minima. Thus, it is important to identify reasonable initial points as inputs to the optimization. While these issues are discussed at greater length in Sec. 3, we note that simple properties of the signal such as its height, width at half max, etc. yield initial conditions that should be sufficiently close to optimal solutions to ensure convergence of the optimization.

## 3 Validation with experimental data

To validate the analysis presented in Sec. 2, we consider a collection of 1302 events measured in the optofluidic device described in Ref. [11]. The measurands are polystyrene microspheres with dispersed Dragon Green fluorophore and a manufacturer-specified mean diameter of 15.3 $\mu$m (no uncertainties available for the mean). We use a high-speed camera system to manually verify that all events correspond to singlets. The fluorescence signals, denoted $\mathbf{f}_m$, are discretized on a 16-bit data acquisition card that records samples at 2 MHz. Thus, the $\mathbf{f}_m$ are vectors whose entries correspond to the discrete time interval over which the data is collected. The digitizer outputs fluorescence values in units of volts; we divide these measurements by 1 V to non-dimensionlize.

As a preprocessing step, we use a moving average filter to smooth peaks followed by a polynomial fit peak finder to estimate the height $h_m$, width $w_m$, and the time interval of the location of the peak maxima for each event; see Ref. [11, 20, 21]. The first two quantities are used later in the optimization. Using the third, we perform a cyclic permutation of the data (corresponding to temporal shifts) to approximately align the peak maximum with the midpoint of the time interval, which we rescale to be $[-1, 1]$.

Having estimated these quantities, we return to the original (non-smoothed) signal for subsequent analysis. Because the signals contain noise (likely from the photodetector and/or counting statistics of photons), we perform a discrete Fourier transform (DFT) to identify a noise-floor, which begins around the tenth mode, corresponding to $M = 9$ according to our indexing convention; see also Ref. [22]. We take this value to be a cutoff for a low pass filter and also keep the complex conjugate weights, which correspond to the last 9 modes (the case $k = 0$ is its own conjugate mode). (The exact cutoff will depend on application and the noise-floor of the signal.) Given that the DFT yields a spectral representation that is a continuous interpolation of the data, we can directly use the corresponding mode weights and frequencies in calculations involving Eq (26).

To test Eq (26), we pick a reference event $\hat{\mathbf{f}}_0$ at random from among those having a height and width roughly equal to the median value for this population. The specific choice of this event is unimportant, since the goal is to demonstrate that all signals can be collapsed to any chosen reference. Next, we choose $N_s$ additional samples $\hat{\mathbf{f}}_m$ and define the modified objective

$$\mathcal{L}_{N_s} = \sum_{m=1}^{N_s} \mathcal{L}(c_m, v_m, R_m, R_0, \Delta t_m; \hat{\mathbf{f}}_m), \tag{27}$$

where $R_0$ is the unknown radius of the reference particle and $(c_m, v_m, R_m, \Delta t_m)$ are the unknown transformation parameters associated with $\hat{\mathbf{f}}_m$. In Eq (27), $m$ has been upgraded from a subscript to an index. Because the beads have on the order of $10^9$ fluorophores per particle, we anticipate that variations in $c_m$ (e.g. due to shot noise) are negligible. Thus, we set $c_m = 1$ for all values of $m$, so that it does not play a further role in the analysis.

In principle, minimizing $\mathcal{L}_{N_s}$ with respect to the unknown parameters yields estimates of their values, as well as a "consensus" estimate of $R_0$. However, several computational considerations limit our ability to use Eq (27) directly. First, the objective function is nonlinear and

contains on the order of $20N_s$ terms and $3N_s$ parameters. Second, the objective has rapid oscillations due to the forms of $\mathbf{X}(v)$ and $g(k;R)$. Third, in the limit that $R \to 0$ for $R/R_0$ constant, the ratio $g(k;R_0)/g(k;R) \to (R_0/R)^3$, so that $\mathcal{L}_{N_s}$ develops a connected set of solutions. That is, only the ratio of $R_0$ to $R$ can be estimated from $\mathcal{L}_{N_s}$.

A resolution to these problems is to consider a regularized version of the objective

$$\mathcal{L}_r = \mathcal{L}_{N_s} + \epsilon(R_0 - \bar{R}_0)^2, \tag{28}$$

where $\epsilon \ll 1$ is a regularization parameter and $\bar{R}_0$ is an estimate of the particle radius given by outside sources of information. For the beads under consideration, manufacturer specifications indicate that the average radius is approximately 7.625 $\mu$m, which we use as the value for $\bar{R}_0$ in dimensional units. To convert to dimensionless units, we note that: (i) the digitizer samples at 2 MHz; (ii) each peak is 1200 samples long; and (iii) the average velocity as estimated from time-of-flight between two interrogation regions is 337 mm/s. Thus, the characteristic distance $d_c$ traversed by a particle during an event is

$$d_c = \frac{1200\,\text{samples}}{2\text{MHz}} 337\text{mm/s} \approx 202\,\mu\text{m}. \tag{29}$$

The characteristic radius 7.625 $\mu$m normalized by $d_c$ and doubled (to account for mapping to $-1 \leq z \leq 1$) yields $\bar{R}_0 = 0.0755$. We also set $\epsilon = 10^{-2}$ and fix $N_s = 5$.

To estimate transformation parameters associated with the remaining curves, we minimize $\mathcal{L}$ for each of the remaining 1296 curves separately, using the consensus value of $R_0$. The results of this exercise are shown in Fig 2. The top subplot shows the original time-traces, while the bottom subplot shows the data collapse; the inset shows the relative errors. See also Fig 3. We also compute the sample mean $\bar{R}$ and sample variance $\sigma_R^2$ estimates of the particle radius to estimate the coefficient of variation (CV) given by $CV_R = \sigma_R/\bar{R}$ [23]. We find $CV_R = 3.18\%$, which is consistent with the manufacturer specified range of particle sizes (4.4%) and the number of measurements considered; see Fig 4.

Fig 5 shows scatter plots of estimated size and velocity measured using two independent photodetection systems. To record these measurements, waveguides are placed both upstream and downstream of the laser interrogation region in order to collect fluorescence from different perspectives. Each waveguide is connected to a separate photomultiplier and data acquisition hardware, ensuring two independent measurements; see Ref. [11] for more details. The correlation between data analysis results from each detector further supports the conclusion that we have extracted systematic information from each time-series.

## 4 Applications

We now consider two applications of the analysis discussed in Sec. 3: estimating per-event measurement reproducibility and doublet deconvolution. A key theme of these examples is that UQ enables one to extract otherwise inaccessible information from cytometry measurements. In particular, UQ allows us to resolve a *uniqueness problem* of identifying the most probable singlets that generate a multiplet. Our main task is to recast the previous results in a probabilistic framework and demonstrate how this leads to new kinds of signals analyses.

### 4.1 Uncertainty quantification

Eq (24) implies that the optimal transformation parameters undo signal deformation due to effects such as particle speed, size, brightness, etc. Thus, each transformed signal $\mathbf{\Lambda}\hat{\mathbf{f}}_m$ can be viewed as a realization of a measurement of identical particles, i.e. a version of $\hat{\mathbf{f}}_0$. Treating

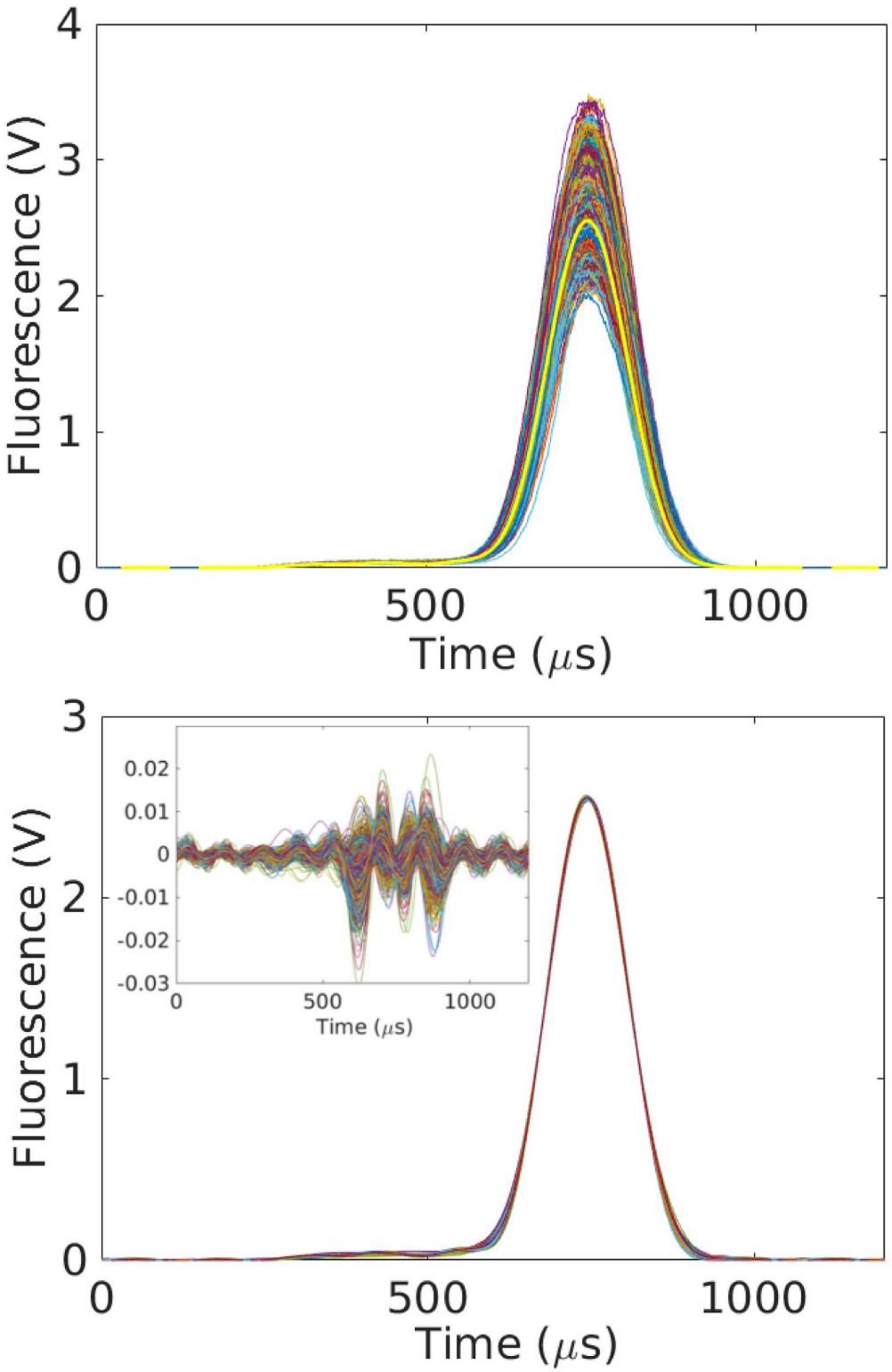



**Fig 2. Example of data collapse for 1302 time-series.** *Top*: The original curves to which we apply the scale transformations described in the main text. The bold-yellow time-trace is used as the reference. The time and fluorescence values are given in their original units, although subsequent analysis is done after non-dimensionalizing. *Bottom*: All 1302 curves after data collapse and conversion back to the original units. To achieve, collapse, we use Eq (24) to turn each measured curve into a realization of $f_0(t)$. For the purposes of defining units in this and other figures,

we adopt this active interpretation of the transformations. The inset shows the point-wise residuals between the transformed curves and reference. The difference is normalized by the maximum value of the reference. Note that the residuals have the characteristic frequency of the mode cutoff.

https://doi.org/10.1371/journal.pone.0295502.g002

these as independent and identically distributed random variables, we can estimate the per-event reproducibility in terms of corresponding statistical estimators. For example, if $A_j$ denotes the integrated area of the $j$th transformed signal, then reproducibility in area measurements is given in terms of a standard deviation $\sigma_A$ computed via

$$\sigma_A^2 = \frac{1}{N_e - 1}\sum_{j=1}^{N_e}(A_j - \bar{A})^2, \tag{30}$$

$$\bar{A} = \frac{1}{N_e}\sum_{j=1}^{N_e} A_j, \tag{31}$$

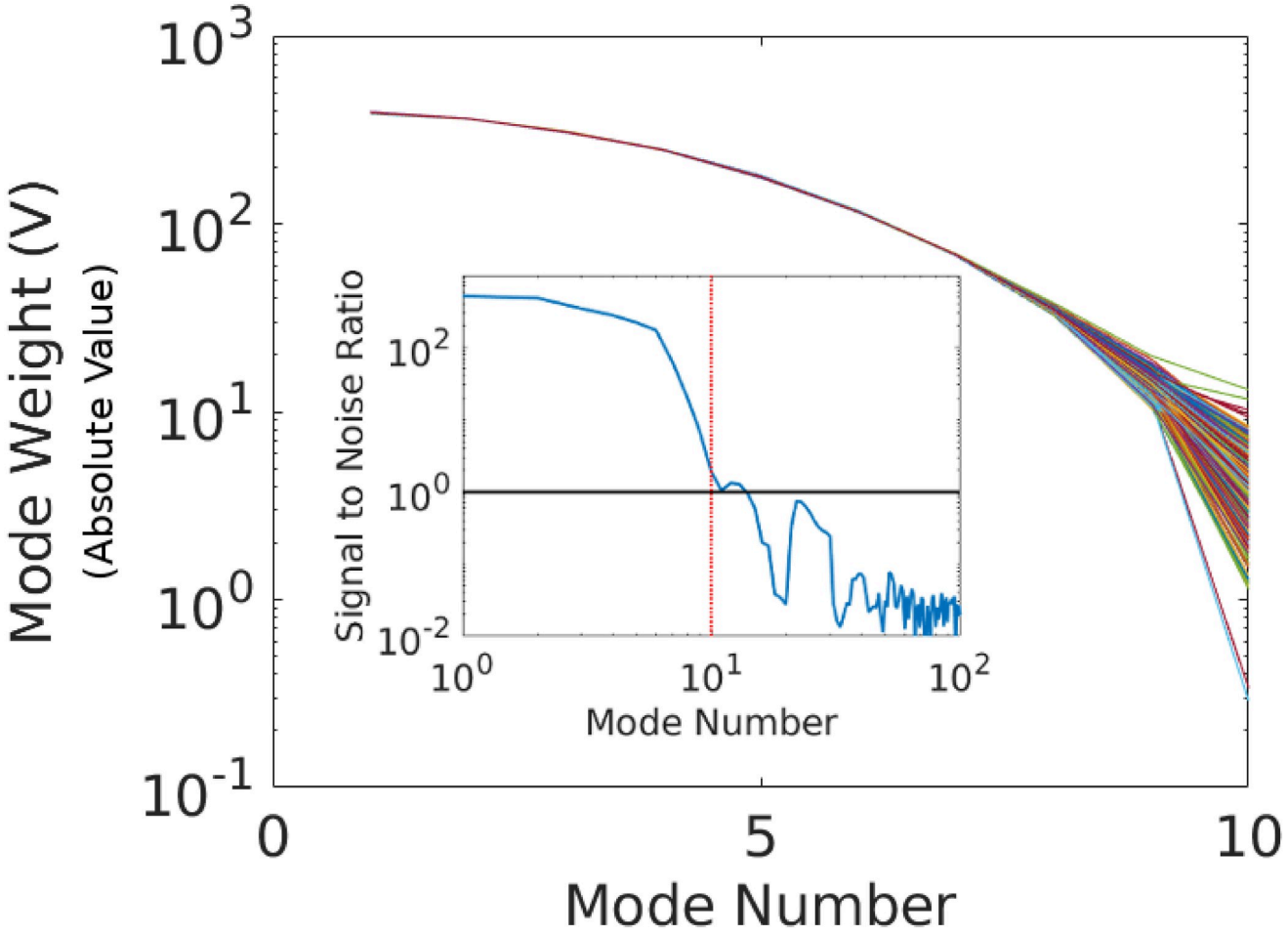



**Fig 3. Absolute values of the mode-weights associated with the 1302 curves in Fig 2 after transformation and collapse via Eq (24).** The inset shows the signal-to-noise ratio, which was computed by: (i) using the first 10 modes to determine the transformation parameters; (ii) applying the transformation matrices to the first 100 modes of each signal; and (iii) computing the ratio of the absolute value of the mean to standard deviation of the resulting mode-weights. The dotted-red vertical line shows the mode cutoff, which occurs when the mode-weights approach a signal-to-noise ratio of unity.

https://doi.org/10.1371/journal.pone.0295502.g003

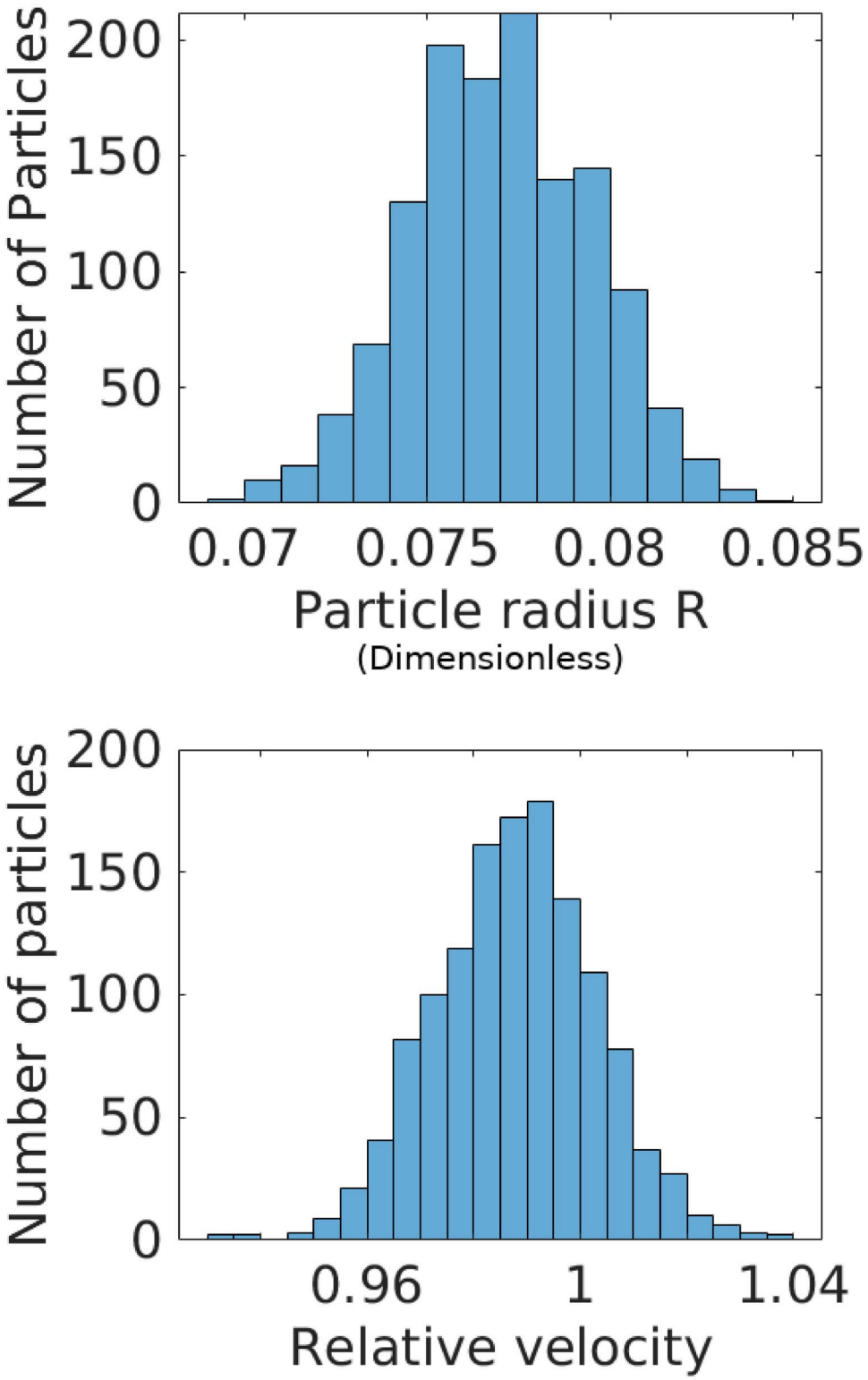



**Fig 4. Histograms of relative particle sizes (top) and velocities (bottom).** The coefficient of variation (CV) in particle radii is 3.18%, which matches well with the manufacturer specification of 4.4% (Bangs Laboratories, FSDG009).

https://doi.org/10.1371/journal.pone.0295502.g004

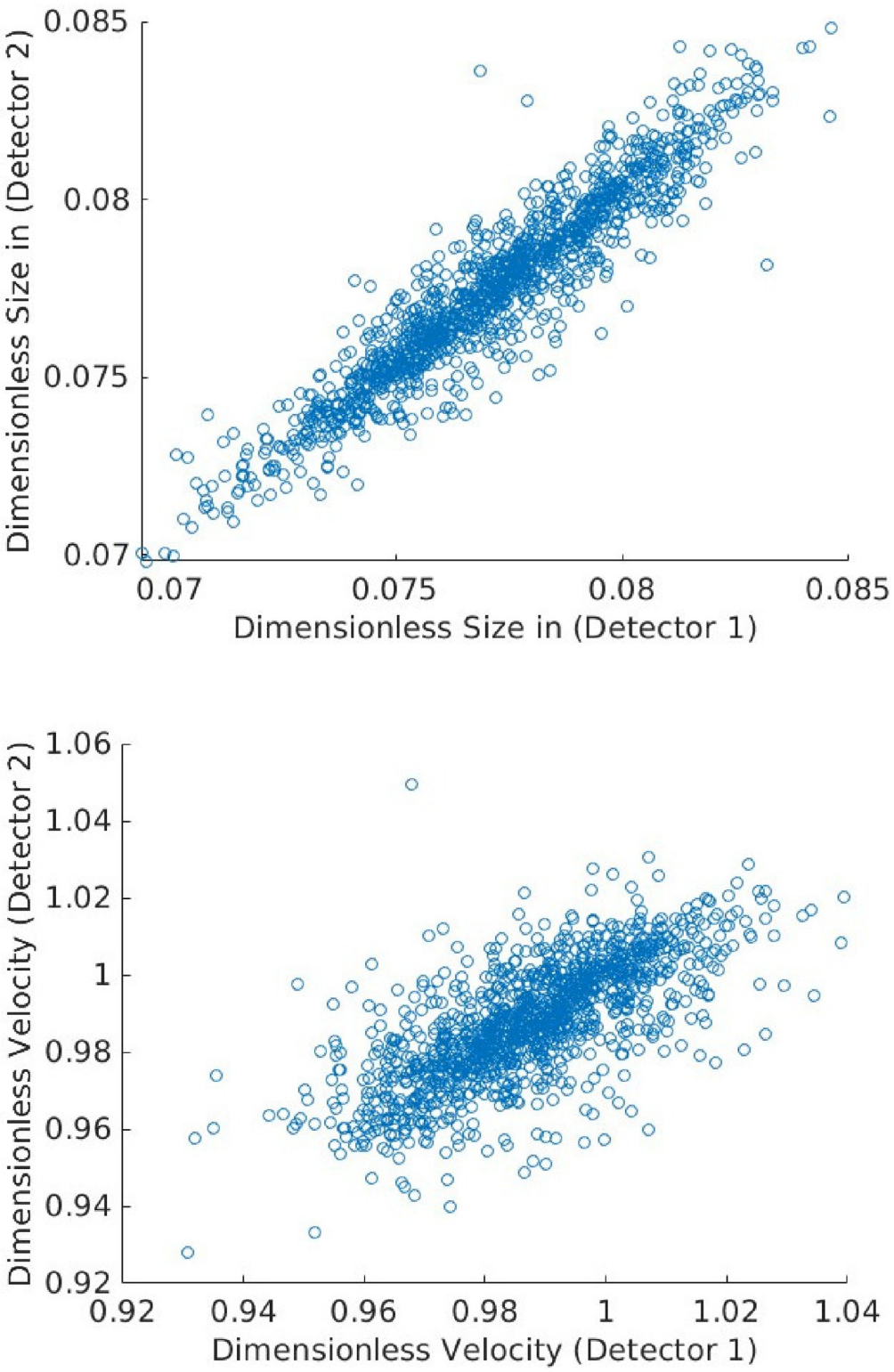



**Fig 5. Scatter plots of estimated size and velocity of the same events measured using two independent photodetection systems.** Velocities and sizes are correlated to within roughly 1% relative uncertainty, which is consistent with the uncertainty in the shape shape.

https://doi.org/10.1371/journal.pone.0295502.g005

where $N_e$ is the number of events. Note that in Sec. 2 we use the subscript $m$ to denote a measured time-series, which we implicitly index from $1 \leq m \leq N_e$. Unless otherwise specified, in this section we use the subscript $j$ when referring to transformed signals. (The number of transformed signals is still denoted $N_e$.) Fig 6 shows the results of this analysis for the 1302 curves considered in Fig 2. The inset to the bottom plot of Fig 2 indicates that the point-wise (in time and relative to the reference amplitude) uncertainty in shape of the filtered signals is on the order of 2% or less.

We make this last observation more precise by considering the Fourier modes of each realization of $\hat{\mathbf{f}}_0$, which we denote by $\hat{\mathbf{f}}_{0,j}$. As a practical matter, it is easier to consider the real and imaginary parts of $\hat{f}_{0,j}(k;\tau_0)$ separately, since their fluctuations should be independent. Dropping explicit reference to $\tau_0$, we decompose $\hat{f}_{0,j}(k) = \hat{f}_{j,\mathrm{Re}}(k) + i\hat{f}_{j,\mathrm{Im}}(k)$, where

$$\mathrm{Re}[\hat{f}_{0,j}(k)] = \hat{f}_{j,\mathrm{Re}}(k) = \bar{f}_{\mathrm{Re}}(k) + \Delta\hat{f}_{j,\mathrm{Re}}(k), \tag{32a}$$

$$\mathrm{Im}[\hat{f}_{0,j}(k)] = \hat{f}_{j,\mathrm{Im}}(k) = \bar{f}_{\mathrm{Im}}(k) + \Delta\hat{f}_{j,\mathrm{Im}}(k), \tag{32b}$$

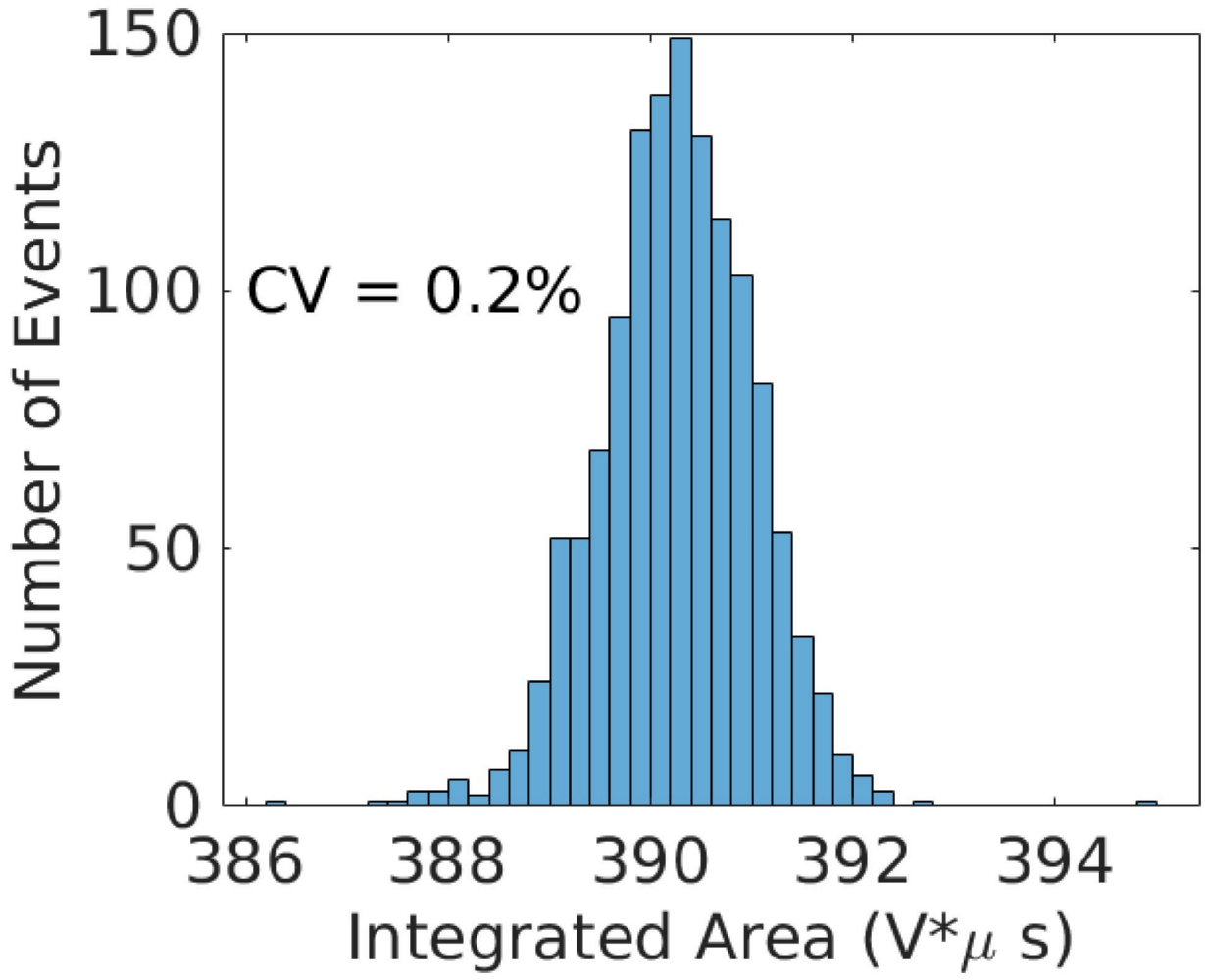



**Fig 6. Histogram of integrated areas of the transformed signals in Fig 2.**

https://doi.org/10.1371/journal.pone.0295502.g006

$\bar{f}_{\mathrm{Re}}(k)$ [$\bar{f}_{\mathrm{Im}}(k)$] are average real [imaginary] Fourier modes, and $\Delta\hat{f}_{j,\mathrm{Re}}(k)$ [$\Delta\hat{f}_{j,\mathrm{Im}}(k)$] are random variables associated with uncertainty in each mode. The average mode weights are constructed by analogy to Eq (31). As a bookkeeping step, it is useful to define a vector $\hat{\mathfrak{f}}_j$ whose first $M + 1$ elements (where $M$ is the mode cutoff) are the real parts of modes $k = 0, \pi, 2\pi, \ldots, M\pi$ of $\hat{\mathbf{f}}_{0,j}$ and whose next $M$ elements are the corresponding imaginary parts for $k \neq 0$; see S1 Appendix for more detailed motivation of this decomposition. Note that $\hat{\mathfrak{f}}$ contains all of the same information as $\hat{\mathbf{f}}$ because $f(t)$ is real. There are $N_e$ such realizations of $\hat{\mathfrak{f}}_j$, where $1 \leq j \leq N_e$. The corresponding average (sample mean) weights and perturbations are $\bar{\mathfrak{f}}$ and $\Delta\hat{\mathfrak{f}}_j$. We also use the $\hat{\mathfrak{f}}_j$ to construct a covariance matrix $\Xi$, as well as the probability of a deviation $\Delta\hat{\mathfrak{f}}$

$$P_1(\Delta\hat{\mathfrak{f}}) \propto \exp\left[-\frac{1}{2}\Delta\hat{\mathfrak{f}}^{\mathrm{T}}\Xi^{-1}\Delta\hat{\mathfrak{f}}\right], \tag{33}$$

where the proportionality factor depends on the dimensionality and determinant of $\Xi$. Eq (33) characterizes uncertainty in the shape of a signal. Note that the real basis underlying $\hat{\mathfrak{f}}$ in Eq (33) allows us to express correlations between modes in terms of the usual covariance matrix. Likewise, we construct a probability density

$$Q_1(\Delta\chi) \propto \exp\left[-\frac{1}{2}\Delta\chi^{\mathrm{T}}\Upsilon^{-1}\Delta\chi\right], \tag{34}$$

where $\Delta\chi^{\mathrm{T}} = (R - \bar{R}, v - \bar{v})$ characterizes deviation from the average radius and velocity, and Y is the corresponding covariance matrix. In the next section, we show how these probability densities play a fundamental role in the tasks of doublet deconvolution.

## 4.2 Doublet deconvolution

As an illustration of how UQ can inform downstream analysis, consider the task of doublet deconvolution. In typical cytometry protocols, such data is identified indirectly via gating procedures and subsequently rejected. We propose an alternate strategy based on constrained optimization.

First assume that a doublet $d(t)$ is a linear combination singlet signals. In Fourier space this implies

$$\begin{aligned}\hat{\mathfrak{d}} &= \boldsymbol{\Theta}^{-1}(c_1, v_1, R_1, R_0, \Delta t_1)[\bar{\mathfrak{f}} + \Delta\hat{\mathfrak{f}}_1] \\ &+\boldsymbol{\Theta}^{-1}(c_2, v_2, R_2, R_0, \Delta t_2)[\bar{\mathfrak{f}} + \Delta\hat{\mathfrak{f}}_2],\end{aligned} \tag{35}$$

where the parameters $c_j$, $v_j$, $R_j$, and $\Delta t_j$ transform the reference signal into the corresponding singlets that comprise the doublet, and $\hat{\mathfrak{d}}$ and $\boldsymbol{\Theta}$ are the representations of $\hat{\mathbf{d}}$ and $\boldsymbol{\Lambda}$ in the same basis as $\Delta\hat{\mathfrak{f}}$ (cf. the S1 Appendix). We use the inverse $\boldsymbol{\Theta}^{-1}$, since we are transforming *from* the reference signal to the measured signal. Generalizing Eq (35) to an arbitrary multiplet $m(t)$, yields

$$\hat{\mathfrak{m}} = \sum_j \boldsymbol{\Theta}^{-1}(c_j, v_j, R_j, R_0, \Delta t_j)[\bar{\mathfrak{f}} + \Delta\hat{\mathfrak{f}}_j]. \tag{36}$$

Assuming that a sample is known to be a multiplet comprised of $\mathcal{M}$ singlets, the probability that the singlets have a given set of deformations $\Delta\mathfrak{f}_1, \Delta\mathfrak{f}_2, \ldots, \Delta\mathfrak{f}_{\mathcal{M}}$ *before transformation to* $\hat{\mathfrak{m}}$

can be expressed as

$$P_{\mathcal{M}}(\Delta\hat{\mathrm{f}}_1, \Delta\hat{\mathrm{f}}_2, ..., \Delta\hat{\mathrm{f}}_{\mathcal{M}}) = \prod_{j=1}^{\mathcal{M}} P_1(\Delta\hat{\mathrm{f}}_j). \tag{37}$$

Likewise, the probability of a set of transformation parameters is given by

$$Q_{\mathcal{M}}(\Delta\chi_1, \Delta\chi_2, ..., \Delta\chi_{\mathcal{M}}) = \prod_{j=1}^{\mathcal{M}} Q_1(\Delta\chi_j). \tag{38}$$

We assume that the true values of $\Delta\hat{\mathrm{f}}_j$ and $\Delta\chi_j$ are those that maximize the joint probability $P_{\mathcal{M}}Q_{\mathcal{M}}$. However, the singlets must reconstruct the multiplet exactly. Together this suggests the objective

$$\mathcal{L}_{\mathcal{M}} = \sum_j \Delta\hat{\mathrm{f}}_j^{\mathrm{T}} \boldsymbol{\Xi}^{-1} \Delta\hat{\mathrm{f}}_j + \Delta\chi_j^{\mathrm{T}} \Upsilon^{-1} \Delta\chi_j, \tag{39}$$

which we minimize with respect to the $\Delta\hat{\mathrm{f}}_j$ and $\Delta\chi_j$, subject to the constraint given by Eq (36). Note that Eqs (36) and (39) are valid for an arbitrary multiplet composed of $\mathcal{M}$ singlets, although $\mathcal{M}$ is assumed known; see Sec. 5.2 for additional considerations on determining $\mathcal{M}$.

This optimization problem is performed over $\mathcal{M}(2M + 1) + 4\mathcal{M}$ parameters, where $\mathcal{M} \geq 2$ is the number of singlets comprising the signal, and $M$ is the mode cutoff. In practice, however, this can be reduced to a $4\mathcal{M}$ dimensional problem, since the constraint is linear in the $\Delta\hat{\mathrm{f}}_j$. We leave details of the optimization to the S1 Appendix. We also set $c_j = 1$, consistent with Sec. 3.

Fig 7 illustrates the results of this analysis applied to a synthetic doublet constructed from two singlets taken from Fig 2. The analysis recovers the original signals to within roughly 5% accuracy or better point-wise in time. The actual transformation parameters associated with each particle are recovered to within approximately 2%. Fig 8 shows the relative errors (with doublets inset) associated with different peak spacings between the singlets. All reconstructions have point-wise errors of less than 10%, even for closely spaced singlets.

Fig 9 shows the results of this analysis applied to two measured signals that were visually verified to be doublets. In the second event (bottom plot), one of the beads was likely out-of-focus of the camera, leading to a reduced visual signal. See Ref. [11] for more details. In both cases the analysis recovers singlets that are consistent with the synthetic examples shown in Fig 7. While more work is needed to make this analysis fully robust for commercial settings, the examples provided herein indicate that the constrained optimization formulation is a useful tool for doublet deconvolution.

## 5 Discussion

### 5.1 The role of modeling in cytometry

From a metrology perspective, the operation of a cytometer is at odds with the fundamental assumptions used to characterize reproducibility and uncertainty. The inherent separation of scales—thousands of distinct, micron-sized cells traveling meters per second in a complicated fluid-dynamic system—coupled with the native variability of biological systems means that it is challenging to repeat an independent measurement on the same particle in the same optical region. Moreover, as Fig 2 demonstrates, even small variations in *reference materials* can lead to dramatic changes in raw measurement signals. Thus, a key challenge in characterizing the accuracy of a cytometer arises from an inability to design operating conditions that isolate

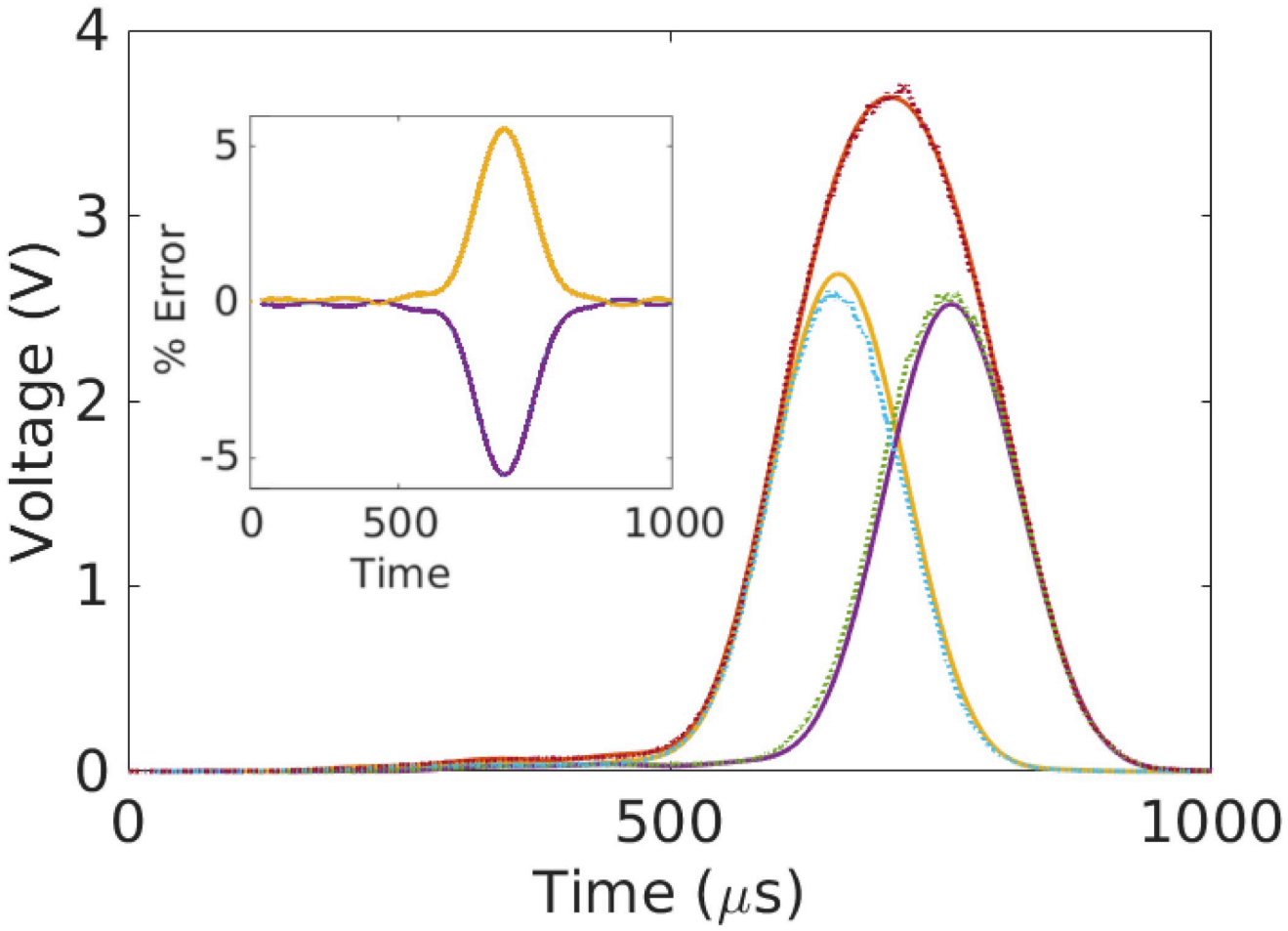



**Fig 7. Example of doublet deconvolution.** Signals associated with two known singlets were added together to make a synthetic doublet. Original signals are shown in dotted lines. The solid lines show the recovered signals after deconvolution (purple and yellow) as well as the filtered doublet (red). The inset shows the error in the recovered singlets relative to the original filtered singlets, normalized by the maximum values of the latter.

https://doi.org/10.1371/journal.pone.0295502.g007

individual sources of uncertainty. In other words, experimental *analysis* of uncertainty is difficult.

The modeling herein provides a distinct approach to this problem through *synthesis, i.e. building up a prediction of the experimental result*, taking into account the cumulative effects of multiple sources of variation. Under ideal circumstances, comparing this synthetic result with reality leads to a unique quantification of the physical phenomena (e.g. particle size, speed, etc.) that generate each signal. Any remaining variation that we cannot account for is then treated as a reasonable proxy for reproducibility.

This distinction between analysis and synthesis highlights both the potential roles and challenges of using mathematical modeling for device characterization. In particular, the model permits us to infer the relative magnitudes of coupled physical effects, thereby compensating for experimental limitations. But critically, the accuracy of these estimates relies on the validity of the underlying theoretical assumptions. For example, the assumption that the particles are spherical may not be appropriate for deformable cells, requiring revision of Eq (9). Likewise, the properties of laser profiles considered herein may not apply to all cytometers.

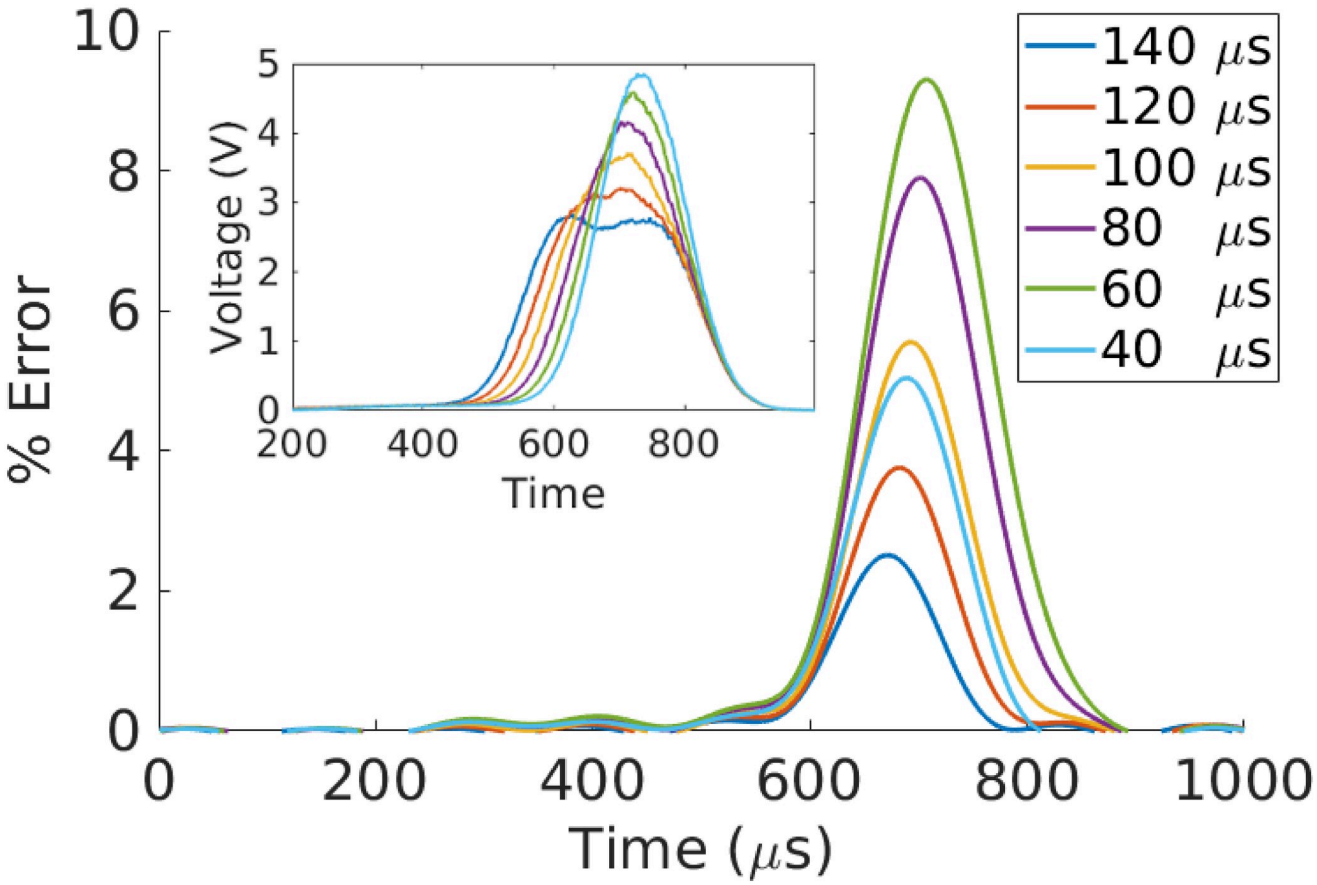



**Fig 8. Impact of peak separation on doublet deconvolution.** The inset shows synthetic doublets constructed by varying the peak separation distance of the singlets in Fig 7. The main figure shows the time-dependent relative difference between one of the recovered singlets and its true (filtered) signal. Colors have the same interpretation in both the main plot and inset.

https://doi.org/10.1371/journal.pone.0295502.g008

These observations suggest a need for deeper coordination between UQ, signals analysis, and design of cytometers. While models can often be revised to account for increasingly complex phenomena (e.g. deformable cells, non-uniform lasers), computations invariably become too expensive to be useful. In such cases, theory can instead inform design changes that may be experimentally achievable and lead to improved accuracy through consistency with modeling assumptions. See also Ref. [11] for related ideas.

## 5.2 Further applications and open directions

The analysis presented herein offers routes to solving several outstanding problems in signals analysis for cytometry. In particular, the ability to quantify reproducibility of a measurement suggests a task wherein one seeks to *minimize* this uncertainty as a function of operating conditions, e.g. flow velocity and degree of focusing, particle density, etc. Moreover, reproducibility estimates suggest the possibility for propagating uncertainty into populations studies so as to inform best gating and classification strategies. In this spirit, the work presented in Ref. [24] may be relevant.

Recent publications have also suggested the importance of cell deformability during the measurement process. Eq (6) makes a key simplifying assumption that requires modification

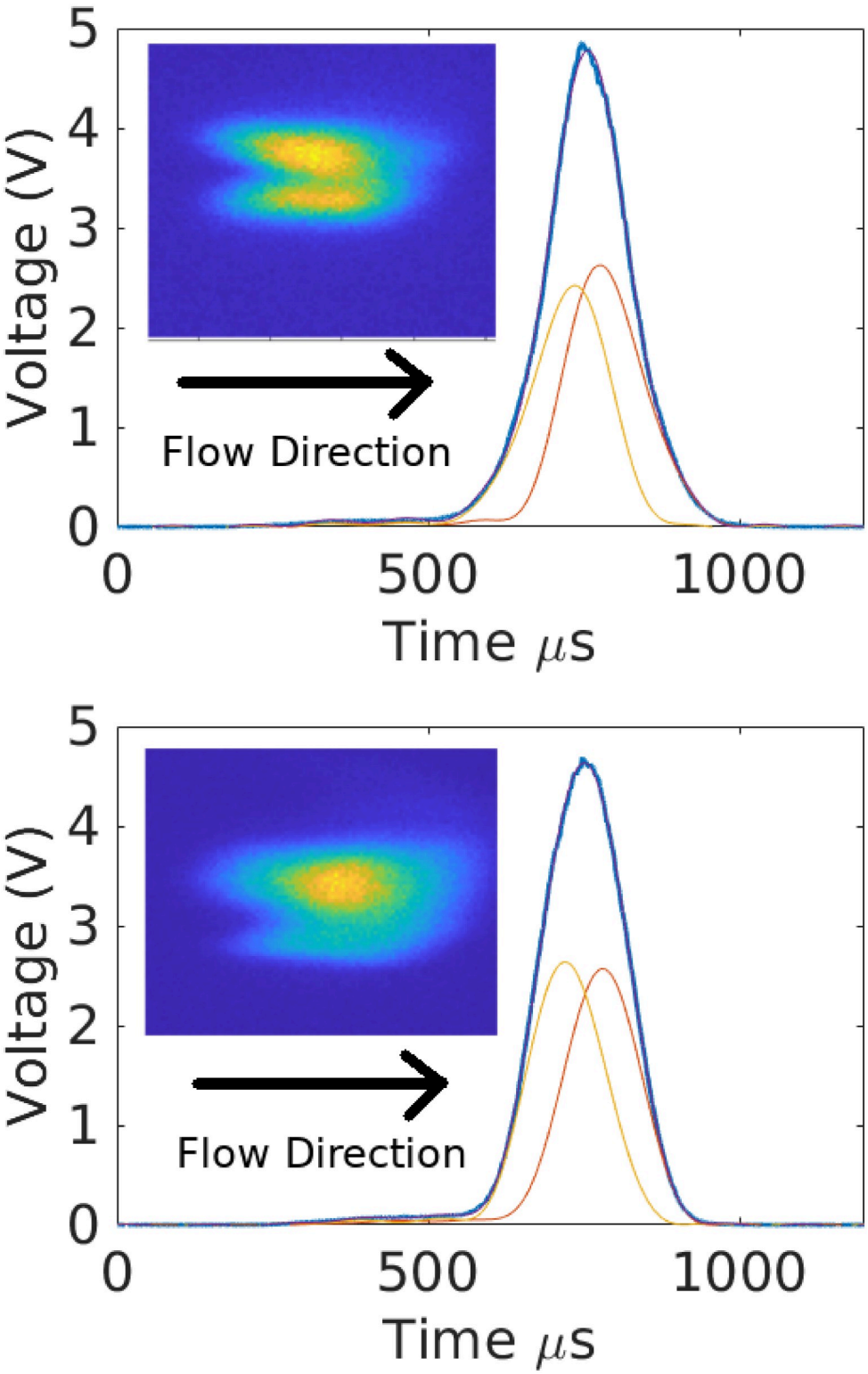



**Fig 9. Deconvolution of two visually verified doublets.** The inset shows false color images of the particles traversing the laser interrogation region. In the inset, the horizontal axis is parallel to the flow direction, and the vertical direction is parallel to the laser. The horizontal elongation is due to blurring. In both plots, two lobes are visible. The main figure shows the original signal (blue), as well as the reconstructed singlets (orange and yellow). The sum of singlets is superimposed in purple over the blue curve but is visually indistinguishable. In the bottom plot, one particle is likely

rotated out of the focal plane, which would account for the differences in visual intensities relative to signal peaks. This assumes the geometric factors associated with illumination and light collection are constant for both particles; see Ref. [11] for the validity of this assumption.

https://doi.org/10.1371/journal.pone.0295502.g009

in this case. To the extent that it is possible to parameterize deformation modes (e.g. in terms of spherical harmonics or empirically defined shapes), the transformations associated with Eq (8) can be generalized to account for non-spherical particles. This ability to extract shape information from a relatively simple time-series could yield significant improvements in throughput relative to imaging cytometers while still characterizing more nuanced information about the cell status.

The problem of *doublet identification* also requires further investigation. We define this as the task of estimating the number of singlets in a signal, irrespective of their relative contributions to the measurement. Contrast this with multiplet deconvolution as addressed in Sec. 4.2. The latter assumes a number $\mathcal{M}$ of singlets and extracts their individual signals from the measurement. To distinguish these problems, it is useful to think of deconvolution in terms of conditional probability, e.g. as being conditioned on $\mathcal{M}$. Thus, the identification problem must be solved prior to deconvolution.

With this in mind, we note that traditional strategies address identification through subjective gating of populations, although several recent works has proposed methods closer in spirit to conditional probability described above [25, 26]. Using methods from the present work, the distribution of Fourier spectra may provide more objective, probability-based methods. The left plot of Fig 10 shows the envelope of spectra (red region) associated with the distribution of

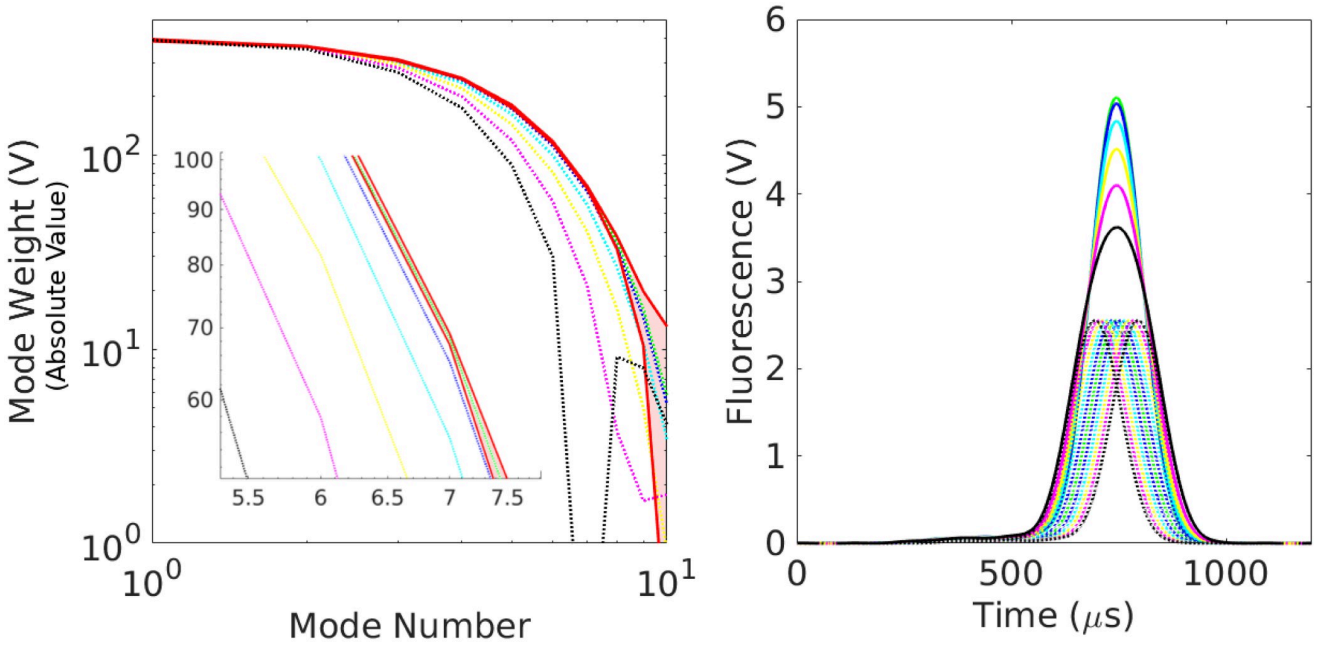



**Fig 10. Fourier spectra of doublets.** *Left*: Fourier spectra of synthetic doublets (dotted) compared with the uncertainty window for singlets (solid red). For a fixed mode number, the uncertainty window for singlets is the set of all mode weights between the minimum and maximum values computed in Fig 2. Synthetic doublets are composed of two copies of a singlet (taken at random from Fig 2) whose peaks are separated by 0 (green), 10 (blue), 20 (cyan), 30 (yellow), 40 (magenta), and 50 (black) microseconds. The doublet mode weights are divided by 2 to normalize the them to scale of the singlets. Note that after a separation of only 10 microseconds the doublet spectra fall outside the uncertainty window of the singlets. *Right*: The doublets whose spectra are shown on the left plot. Solid lines are the doublets, whereas dotted lines of the same color are the corresponding singlets. The colors have the same interpretation as on the left plot. Note that all doublets appear as a single peak and are visually difficult to distinguish from singlets.

https://doi.org/10.1371/journal.pone.0295502.g010

transformed singlets in Fig 2. The right plot shows a collection of synthetic doublets; their corresponding spectra (normalized to the same area as the mean singlet) are shown relative to the singlet envelope on the left. Note that for almost any peak separation at least one mode associated with the doublet falls outside the admissible window for singlets. This suggests that signal *shape* may be an exquisitely sensitive tool for identifying events that are candidates for our doublet deconvolution algorithm. Such questions, however, are left for future work.

The task of generalizing our analysis to scattering channels is likely straightforward, although we leave this for future work. In a related vein, determining when particles have non-spherical geometries and/or fluorescence staining is likely possible but may introduce additional complexity. For example, we speculate that one can parameterize shape in terms of spherical harmonics whose corresponding weights are determined through the data collapse process. It is unclear how many such harmonics can be recovered, but it is plausible that one could distinguish spherical and ellipsoidal particles, for example. Dimensional analysis also suggests that additional features of particles could be extracted from signals by changing the width and/or sharpness of the laser profile. Such generalizations would be necessary to address tasks such as determining particle size when the underlying distribution is multimodal. Such issues remain open questions. Finally, we do not consider convergence analyses of the signal reconstruction and optimization methods herein. Such issues are left for future work.

## 5.3 Metrics in the context of past work

Signals analysis has been integral to cytometry since its inception [7–10, 22]. However, virtually all techniques characterize cells in terms of properties such as the signal area, height, width, etc., and the justifications for such analyses are not universally valid. For example, common practice dictates that forward-scatter (FSC) vs area or FSC-height measurements are appropriate for detecting doublets, as cells should nominally follow one another single file [12]. But as Fig 9 illustrates, cells may pass the laser interrogation region side-by-side or even at a diagonal to the flow direction. A recent study of inertial effects also suggests that typical flow focusing strategies are ineffective [11]. Furthermore, biological processes of interest may interfere with measurements typically used to distinguish doublets [27].

Many of these problems arise from the fact that quantities such as signal area, width, and height are functionals of high-dimensional data (i.e. the full event time-trace) that produce a scalar. Thus, much of the underlying information that could be used to characterize measurands is lost before signals are actually compared. This suggests a need to revisit the order of operations, e.g. by directly comparing full signals *before* quantifying their differences in terms of simple descriptors.

To better understand this point, it is useful to consider the concept of a *metric*, which addresses the question: how "far apart" are the generic objects $h$ and $g$? In other words, the metric, often denoted $d(h, g)$, defines a notion of distance appropriate to the structure of $h$ and $g$. While a complete treatment is beyond the scope of this work (see Ref. [28]), we note two fundamental properties: (i) $d(h, g) \geq 0$ for any objects $h$ and $g$, i.e. all distances are non-negative; and (ii) $d(h, g) = 0$ implies $h = g$.

While seemingly abstract, these observations have important ramifications for UQ of cytometry. Conventional analyses implicitly use the absolute value metric

$$d(\mathcal{P}[f_j(t)], \mathcal{P}[f_k(t)]) = |\mathcal{P}[f_j(t)] - \mathcal{P}[f_k(t)]| \tag{40}$$

and related notions of Euclidean distance as the basis for comparing measurements, where $\mathcal{P}[f_j(t)]$ is a functional that returns a scalar value (e.g. area, height) associated with the time-series $f_j(t)$. Since the dimensionality of $\mathcal{P}[f_j(t)]$ is low, the statement $d(h, g) = 0$ applied to Eq

(40) implies that two distinct time-series may still be treated as equivalent. For example, a large, dim cell may yield the same integrated area as a small bright one, despite the objects being very different. Yet according to Eq (40) both objects would be the same based on intensity alone. Moreover, such approaches make it impossible to separate effects (e.g. flow rate) that may change the signal shape while keeping scalar properties such as the height constant. Such shortcomings have limited UQ studies to those effects due to photodetectors and total uncertainties as characterized by population histograms [29–32].

The objectives considered herein address such problems by defining the metric in terms of the time-series explicitly; that is, Eq (26) considers the situation $d(\hat{\mathbf{f}}_j, \hat{\mathbf{f}}_k) = ||\hat{\mathbf{f}}_j - \hat{\mathbf{f}}_k||$ in which the notion of distance $||\cdot||$ is applied directly to the full time-series (expressed as its Fourier transform). Eq (26) in particular is a variation on the commonly used $\ell_2$ or "sum-of-squares" metric [28]. A benefit of this approach is that it leverages the full information content of each signal to characterize multiple sources of variation. However, the resulting analysis is more computationally expensive and does not have a simple interpretation. Nonetheless, the example of doublet deconvolution highlights the usefulness of such techniques, and we speculate that more advanced signals analyses in cytometry will require further development of appropriate metrics.

## 5.4 Limitations

The modeling framework described Sec. 2.1 sets forth the minimum assumptions required for the validity of our analysis. A key goal of the theory is to avoid the need for detailed device characterization, which can be costly. However, our main assumptions on the laser profile and geometric factors may not be applicable to all cytometers. In such cases, using our analysis would decrease confidence in results by introducing additional model-form uncertainty.

We acknowledge that extending the method to particles on different trajectories is fundamentally difficult and ill-posed if the laser profile and geometric factors do not satisfy Eq (4). The signal generated by the instrument is a one-dimensional time-series, whereas the underlying physical phenomena take place in three-dimensions. Loosely speaking, the physics permits more degrees of freedom than can be measured by the instrument. Eq (2) is useful insofar as it reduces the spatial model to one dimension, thereby rendering the data analysis problem (more) well-posed. That is, it yields a one-to-one correspondence between the spatial variable $z$ in the model and temporal variable $t$ associated with a time-series. In the event that particles are allowed to take different trajectories in a way that violates Eq (4), a signal generated at a given time could correspond to a particle located at many positions.

While this seems like a fundamental limitation of our analysis, in practice it is not. At flow rates typical of cytometers, the Reynolds number is $\mathcal{O}(1)$, so that inertial effects matter. Under such conditions, spherical-like particles in a flow-channel have a tendency to focus to inertial nodes that, surprisingly, do not coincide with the channel center [11]. The cytometer used in this work takes advantage of this inertial effect to ensure that all of the particles travel on the same streamline [11]; i.e. they all have the same trajectory.

In principle, when it is necessary to measure particles that may be on different streamlines, the aforementioned limitations can potentially be overcome by more detailed modeling of the experimental system. The general structure of Eq (26) can be maintained, although the transformation matrix $\mathbf{\Lambda}$ may not exist in closed form. Rather, it may be necessary to evaluate it on-the-fly in terms of a simulation or other computational model of the measurement process. In this case, a key challenge will be to formulate an optimization routine that can minimize Eq (26) in a reasonable amount of time. Reduced-order modeling or computationally inexpensive approximations are possible routes for addressing such problems.

## Supporting information

**S1 Appendix. Optimization for multiplet deconvolution.**
(PDF)

**S1 Data.**
(ZIP)

## Acknowledgments

The authors thank Drs. Bradley Alpert and Mark Stiles for helpful feedback during prepration of this manuscript. This work is a contribution of the National Institute of Standards and Technology, and as such, is not subject to copyright in the United States of America. Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

## Author Contributions

**Conceptualization:** Paul N. Patrone, Gregory A. Cooksey.

**Data curation:** Matthew DiSalvo.

**Formal analysis:** Paul N. Patrone, Anthony J. Kearsley.

**Investigation:** Paul N. Patrone.

**Methodology:** Paul N. Patrone, Geoffrey B. McFadden.

**Project administration:** Anthony J. Kearsley, Gregory A. Cooksey.

**Validation:** Paul N. Patrone.

**Writing – original draft:** Paul N. Patrone.

**Writing – review & editing:** Paul N. Patrone, Matthew DiSalvo, Anthony J. Kearsley, Geoffrey B. McFadden, Gregory A. Cooksey.

## References

1. Adan A, Alizada G, Kiraz Y, Baran Y, Nalbant A. Flow cytometry: basic principles and applications. Critical Reviews in Biotechnology. 2017; 37(2):163–176. https://doi.org/10.3109/07388551.2015.1128876 PMID: 26767547
2. McKinnon KM. Flow Cytometry: An Overview. Current Protocols in Immunology. 2018; 120(1):5.1.1–5.1.11. https://doi.org/10.1002/cpim.40 PMID: 29512141
3. Wang L, Bhardwaj R, Mostowski H, Patrone PN, Kearsley AJ, Watson J, et al. Establishing CD19 B-cell reference control materials for comparable and quantitative cytometric expression analysis. PLOS ONE. 2021; 16(3):1–18. https://doi.org/10.1371/journal.pone.0248118 PMID: 33740004
4. Laskowski TJ, Hazen AL, Collazo RS, Haviland D. Rigor and Reproducibility of Cytometry Practices for Immuno-Oncology: A multifaceted challenge. Cytometry Part A. 2020; 97(2):116–125. https://doi.org/10.1002/cyto.a.23882 PMID: 31454153
5. Tighe RM, Redente EF, Yu YR, Herold S, Sperling AI, Curtis JL, et al. Improving the Quality and Reproducibility of Flow Cytometry in the Lung. An Official American Thoracic Society Workshop Report. American Journal of Respiratory Cell and Molecular Biology. 2019; 61(2):150–161. https://doi.org/10.1165/rcmb.2019-0191ST PMID: 31368812
6. Kalina T. Reproducibility of Flow Cytometry Through Standardization: Opportunities and Challenges. Cytometry Part A. 2020; 97(2):137–147. https://doi.org/10.1002/cyto.a.23901 PMID: 31593368

7. Xiong F, Tanik MM. Deciphering flow cytometry electronics. In: SoutheastCon 2015; 2015. p. 1–5.
8. Zilmer NA, Godavarti M, Rodriguez JJ, Yopp TA, Lambert GM, Galbraith DW. Flow cytometric analysis using digital signal processing. Cytometry. 1995; 20(2):102–117. https://doi.org/10.1002/cyto.990200203 PMID: 7664621
9. Grimaldi D, Palumbo A, Rapuano S. Hardware and software improvements in flow-cytometry measurements. Measurement. 2004; 36(2):111–119. https://doi.org/10.1016/j.measurement.2004.02.002
10. Godavarti M, Rodriguez JJ, Yopp TA, Lambert GM, Galbraith DW. Automated particle classification based on digital acquisition and analysis of flow cytometric pulse waveforms. Cytometry. 1996; 24 (4):330–339. https://doi.org/10.1002/(SICI)1097-0320(19960801)24:4%3C330::AID-CYTO4%3E3.0.CO;2-J PMID: 8866217
11. DiSalvo M, Patrone PN, Kearsley AJ, Cooksey GA. Serial flow cytometry in an inertial focusing optofluidic microchip for direct assessment of measurement variations. Lab Chip. 2022; 22:3217–3228. https://doi.org/10.1039/D1LC01169C PMID: 35856829
12. Stadinski BD, Huseby ES. How to Prevent yourself from Seeing Double. Cytometry Part A. 2020; 97 (11):1102–1104. https://doi.org/10.1002/cyto.a.24045 PMID: 32573091
13. Donnenberg VS, Donnenberg AD. Coping with artifact in the analysis of flow cytometric data. Methods. 2015; 82:3–11. https://doi.org/10.1016/j.ymeth.2015.03.012 PMID: 25823849
14. Ahmed R, Omidian Z, Giwa A, Cornwell B, Majety N, Bell DR, et al. A Public BCR Present in a Unique Dual-Receptor-Expressing Lymphocyte from Type 1 Diabetes Patients Encodes a Potent T Cell Autoantigen. Cell. 2019; 177(6):1583–1599.e16. https://doi.org/10.1016/j.cell.2019.05.007 PMID: 31150624
15. Burel JG, Pomaznoy M, Lindestam Arlehamn CS, Seumois G, Vijayanand P, Sette A, et al. The Challenge of Distinguishing Cell–Cell Complexes from Singlet Cells in Non-Imaging Flow Cytometry and Single-Cell Sorting. Cytometry Part A. 2020; 97(11):1127–1135. https://doi.org/10.1002/cyto.a.24027 PMID: 32400942
16. Grafakos L. Classical Fourier Analysis. Graduate Texts in Mathematics. Springer New York; 2008.
17. Bloomfield P, Collection KMR. Fourier Analysis of Time Series: An Introduction. Wiley Series in Probability and Statistics—Applied Probability and Statistics Section. Wiley; 1976.
18. Oppenheim AV. Discrete-Time Signal Processing. Pearson education signal processing series. Pearson Education; 1999.
19. Olver FWJ, of Standards NI, ) TUS. NIST Handbook of Mathematical Functions Paperback and CD-ROM. Cambridge University Press; 2010.
20. Wallace WE, Kearsley AJ, Guttman CM. An Operator-Independent Approach to Mass Spectral Peak Identification and Integration. Analytical Chemistry. 2004; 76(9):2446–2452. https://doi.org/10.1021/ac0354701 PMID: 15117182
21. Kearsley AJ, Wallace WE, Bernal J, Guttman CM. A numerical method for mass spectral data analysis. Applied Mathematics Letters. 2005; 18(12):1412–1417. https://doi.org/10.1016/j.aml.2005.02.033
22. Sloot PMA, Tensen P, Figdor CG. Spectral analysis of flow cytometric data: Design of a special-purpose low-pass digital filter. Cytometry. 1987; 8(6):545–551. https://doi.org/10.1002/cyto.990080603 PMID: 3428039
23. JCGM. JCGM 100: Evaluation of measurement data—Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology. 2008.
24. Patrone PN, Kearsley AJ. Classification Under Uncertainty: Data Analysis for Diagnostic Antibody Testing. 2021.
25. Baur M, Reisbeck M, Hayden O, Utschick W. Joint Particle Detection and Analysis by a CNN and Adaptive Norm Minimization Approach. IEEE Transactions on Biomedical Engineering. 2022; 69(8):2468–2479. https://doi.org/10.1109/TBME.2022.3147701 PMID: 35104207
26. Caselli F, De Ninno A, Reale R, Businaro L, Bisegna P. A Bayesian Approach for Coincidence Resolution in Microfluidic Impedance Cytometry. IEEE Transactions on Biomedical Engineering. 2021; 68 (1):340–349. https://doi.org/10.1109/TBME.2020.2995364 PMID: 32746004
27. Bohmer RM, Bandala-Sanchez E, Harrison LC. Forward light scatter is a simple measure of T-cell activation and proliferation but is not universally suited for doublet discrimination. Cytometry Part A. 2011; 79A(8):646–652. https://doi.org/10.1002/cyto.a.21096
28. Kreyszig E. Introductory Functional Analysis with Applications. Wiley Classics Library. Wiley; 1989.
29. Hoffman RA, Wood JCS. Characterization of Flow Cytometer Instrument Sensitivity. Current Protocols in Cytometry. 2007; 40(1):1.20.1–1.20.18. https://doi.org/10.1002/0471142956.cy0120s40 PMID: 18770846

30. Ubezio P, Andreoni A. Linearity and noise sources in flow cytometry. Cytometry. 1985; 6(2):109–115. https://doi.org/10.1002/cyto.990060205 PMID: 3979215

31. Steen H. Noise, sensitivity, and resolution of flow cytometers. Cytometry. 1992; 13(8):822–830. https://doi.org/10.1002/cyto.990130804 PMID: 1458999

32. Chase ES, Hoffman RA. Resolution of dimly fluorescent particles: A practical measure of fluorescence sensitivity. Cytometry. 1998; 33(2):267–279. https://doi.org/10.1002/(SICI)1097-0320(19981001)33:2%3C267::AID-CYTO24%3E3.0.CO;2-R PMID: 9773890