

# Uncertainty Quantification of Antibody Measurements: Physical Principles and Implications for Standardization

Paul N. Patrone, Lili Wang, Sheng Lin-Gibson, and Anthony J. Kearsley  
*National Institute of Standards and Technology*  
100 Bureau Drive, Gaithersburg MD, 20899 USA  
(Dated: December 6, 2024)

Harmonizing serology measurements (i.e. rendering them interchangeable) is critical for comparing results across different diagnostics platforms, developing associated reference materials, and thereby informing medical decisions. However, the theoretical foundations of such tasks have yet to be fully explored in terms of antibody thermodynamics and uncertainty quantification (UQ). In the context of SARS-CoV-2, for example, this has restricted the usefulness of standards currently deployed, limited the scope of materials considered as viable standards, and ultimately decreased confidence in serology. To address these problems, we develop rigorous theories of antibody normalization and harmonization. We begin by proposing a mathematical definition of harmonization equipped with structure needed to quantify uncertainty associated with the choice of standard, assay, etc. We then show how a thermodynamic description of serology measurements (i) relates this structure to the Gibbs free-energy of antibody binding, and thereby (ii) induces a regression analysis that directly harmonizes measurements. We supplement this with a novel, optimization-based normalization (not harmonization!) method that validates consistency between the behavior of a reference material and biological samples. *A key result of these analyses is that under physically reasonable conditions, the choice of reference material does not increase uncertainty associated with harmonization.* We validate main ideas via an interlab study that considers monoclonal antibodies as a reference for SARS-CoV-2 serology measurements and discuss connections to correlates of protection.

Keywords: Thermodynamics, SARS-CoV-2, Uncertainty, Antibody, Serology

## I. PREFACE

This manuscript arose from the confused and sometimes murky world of SARS-CoV-2 antibody testing as it existed in late 2021 and 2022 [1, 2]. While it is hard to estimate the total number – hundreds? – of blood and saliva assays that were developed during the pandemic, one thing is clear: the community never agreed upon a scale for comparing the resulting measurements [3–7]. Thus, it was impossible to say, for example, how many anti-SARS-CoV-2 antibodies anyone had, much less determine if someone was immune from infection. The issue, it turns out, was not lack of standards. Rather, we argue in this work that the community required new methods for *interpreting measurements* of those standards. In other words, the problem was one of physics and math, not medicine and biology.

In the exposition that follows, we frame this argument in the context of two questions: (i) what are the physical processes that cause antibody measurements to differ; and (ii) how do we account for this physics to *harmonize* (i.e. render interchangeable) serology measurements? In answering these questions, we decided to follow the series of steps that a data analyst would need to achieve harmonization *in practice*. In part, this is to facilitate reproducibility and promote the perspective that validation steps can (and in our opinion, should) be incorporated into all aspects of the data flow. But it also highlights an important difference between harmonization and the concept of *normalization*, which we feel has been a point of confusion. However, this exposition necessarily makes the manuscript highly interdisciplinary, leveraging ideas

from not only statistical mechanics and thermodynamics, but also applied math, optimization, biology, and immunology. We anticipate that readers will have a typical background in physics and/or applied mathematics, but not necessarily any of the other fields. If needed, readers can consult Appendix A for more information and context on serology before proceeding to Sec. II, which serves as a formal introduction for this manuscript. Where possible, we have also put technical math details in the appendices.

## II. INTRODUCTION

The COVID-19 pandemic highlighted the importance of antibody tests as a means to characterize humoral response, e.g. in high-risk populations such as cancer patients [8]. However, the rapid development of many different SARS-CoV-2 assays led to questions regarding the degree to which such measurements quantify immunity [9–12]. In response, public health and research institutions established reference materials to harmonize antibody scales [8, 13–15]. While these efforts improved correlation between measurements, they did not conclusively achieve harmonization [3–6]. They also led to new questions: how do we compare reference materials, and how do we select the “best” one for a given clinical or research setting?<sup>1</sup> Moreover, the development of antibody

---

<sup>1</sup> Certain commercial products are referenced (directly or indirectly) in this manuscript to clarify our theoretical analysis. Such

standards did not suggest an obvious definition for correlates of protection, which remains an elusive concept to this day [9–12, 16].

The difficulty of addressing these questions arises from a fundamental ambiguity in the information extracted from serology measurements [2, 7]. In theory, the goal is to quantify titers or “concentrations” of antibodies that target a given antigen. Binding assays address this by quantifying the relative number of antibodies in a sample that attach to a substrate. However, the corresponding binding kinetics depend on the assay itself, e.g. through the details of this substrate [17]. Thus, it is more appropriate to assert that such assays characterize the properties of a chemical reaction, for which there is no concept of an absolute and independent “bound antibody number” absent information about the other reactants.<sup>2</sup> This ambiguity becomes *worse* when we recognize that all reference materials suffer the same problem, which leads to the possibility of expressing an ill-defined sample concentration in terms similarly ill-defined standard. Typical single-point reference calibrations, e.g. based on end-point titers [18], also suffer from extrapolation error [19] and thereby further complicate harmonization. Thus, meaningful comparison of antibody measurements and standards cannot be realized without accounting for the physical processes and sources of uncertainty that affect their use.

The present work addresses these problems via a hierarchy of data analyses that marry concepts from statistical mechanics with uncertainty quantification to establish a physics-based foundation for harmonization. We first motivate this hierarchy through a Gibbs free-energy description of antibody binding, which (surprisingly) implies that only the assays, not the reference, control the degree to which harmonization is possible. Importantly, this theory induces a probabilistic model that can be used to validate the thermodynamic assumptions by quantifying – and in some cases removing – measurement variability due to: (i) choice of reference material; (ii) assay; (iii) instrument and operator effects; (iv) uncertainty inherent in samples; and (v) interactions between these elements. This analysis can also be used to determine if and by how much a specific reference material increases uncertainty in harmonization, which becomes a metric for comparing standards. Throughout, we validate these ideas in the context of an interlab study that considers synthetic, monoclonal antibodies (mAbs) as serology standards [20].

A key theme that permeates this work is the need to incorporate uncertainty quantification (UQ) into all aspects of the data analysis. While this entails obvious tasks such as statistical modeling and uncertainty propagation [21], we adopt the broader definition of UQ as

encompassing verification and validation exercises that assess and increase confidence in models and measurement processes *per se*. For example, synthetic mAbs are viewed as being fundamentally different from human antibodies, and thus unsuitable for the purposes of harmonization. To address this stigma, our workflow includes a novel normalization procedure to determine if both types of antibodies exhibit the same behavior in a measurement system. This validation step can thereby indirectly assess whether mAbs are governed by the same physical processes as human antibodies, which is a prerequisite for using the former as a standard for the latter.

A main challenge in this work is the need to revisit and clarify seemingly elementary and resolved ideas in serology. For example, the concept of harmonization has been used in a variety of contexts [22–24], but to the best of our knowledge, it has not been defined with the precision needed to fully ground it in metrology. Thus, a preliminary task in our analysis is to mathematically define this concept and equip it with the structure needed to permit later UQ. The implications of this exercise are not trivial. It uncovers hidden structure in the definition of harmonization that has yet to be exploited and provides a direct connection to the physics of antibody measurements. Moreover, it illustrates why harmonization and normalization are not the same task.

A secondary motivation for this work is the fact that current methods to assess fitness of purpose for serology standards are insufficiently grounded in the physics of harmonization and focus primarily on their *technical performance*, sometimes ignoring issues such as ease of manufacturing, distribution times, etc. For example, there is widespread belief that harmonization via SARS-CoV-2 standards can only be achieved using human-derived, pooled references, although to the best of our knowledge there are no studies validating this conjecture. As a result, the development of current SARS-CoV-2 standards (which are all human derived) took nearly a year [14, 15] despite being needed much sooner. Moreover, such standards are inherently limited stock and must contend with changes between lots [25]. These issues suggest the need to better understand the impacts of using alternatives such as mAbs, which permit better scale-up and quicker development. *Indeed our companion manuscript finds – and the present work justifies – that mAbs and human-derived standards are identical from a performance standpoint when using our physically derived definition of harmonization [26].*

A limitation of our work is that we do not define or fully connect our work with a real-world definitions of immunity. This is due primarily to lack of available data [27, 28]. Studies that address this problem would need to connect information about neutralization measurements to notions of risk and clinical outcomes. To the best of our knowledge, these latter tasks remain open challenges in the SARS-CoV-2 testing community and thus fall beyond the scope of our work [27, 28]. However, we point to extensions of our basic definitions and theory to

reference does not imply endorsement or approval of any kind by NIST.

<sup>2</sup> This mirrors the challenge of estimating binding kinetics for emerging SARS-CoV-2 variants.

neutralization assays, a more direct but complicated tool for characterizing immunity.

The rest of the manuscript is organized as follows. In Sec. III, we provide physical and mathematical definitions of key concepts used throughout the manuscript. Section IV gives a global overview of our analysis hierarchy by considering normalization in the context of signal generation (IV A) and deriving a thermodynamic theory of antibody binding that induces a probabilistic perspective needed for harmonization (IV B). Section V fully develops the remaining elements of this hierarchy in terms of a regression analysis. Section VI provides historical context for our work, discusses limitations, and considers future directions. Appendices provide additional background on serology testing and summarize technical mathematical arguments.

### III. NOTATION AND TERMINOLOGY

Given the interdisciplinary nature of this work, we begin by considering definitions and conventions that inform our mathematical analysis.

#### A. Definitions

- I. A concentration  $\hat{c}$  is **normalized** if it is given in units of antibodies per volume.
- II. The word **sample** always refers a specimen taken from a human and to which an **assay** (i.e. a serology test) is applied.
- III. The words **reference** and **standard** always refer to a measurand, which can be synthetic or human-derived, used to normalize measurements taken on samples.
- IV. In physical terms, we interpret *harmonization* as the process of determining a mathematical rule that tells one how to modify the numerical value of normalized antibody concentrations for each assay so that their corresponding measurements all agree and can be used interchangeably. *This rule is only considered meaningful if it does not depend on the sample*, but only the assay, reference, and concentration values. In more mathematical terms, let  $s = 1, 2, \dots, S$  and  $n = 1, 2, \dots, N$  index samples and assays for some maximum values  $S$  and  $N$ . Also let  $r$  denote a fixed reference. We say that the assays are **harmonized** by reference  $r$  if for any normalized concentrations  $\hat{c}_{s,n,r}$  and  $\hat{c}_{s,n',r}$  (corresponding to sample  $s$  measured with assays  $n$  and  $n'$  and normalized by  $r$ ), we can find a function  $T(n, c, r)$  such that

$$T(n, \hat{c}_{s,n,r}, r) = T(n', \hat{c}_{s,n',r}, r) = \chi_{s,r}, \quad (1)$$

where  $\chi_{s,r}$  is an assay-independent **consensus value** associated with sample  $s$ . *Consistent with the above physical intuition, this function does not depend directly on the sample index  $s$ , only its normalized concentrations  $\hat{c}_{s,n,r}$ .* See Refs. [22, 23] and the references therein for related ideas. We always assume that  $\chi_{s,r}$  has the same units as  $\hat{c}$ , i.e. antibodies per unit volume.

- V. If we identify parameters  $\epsilon_{s,n}$  (which could be random) such that

$$\begin{aligned} T(n, \hat{c}_{s,n,r}, r)(1 + \epsilon_{s,n}) &= T(n', \hat{c}_{s,n',r}, r)(1 + \epsilon_{s,n'}) \\ &= \chi_{s,r}, \end{aligned} \quad (2)$$

we say that the assays can be **approximately harmonized** with a relative uncertainty quantified by the  $\epsilon$ . In principle we could let  $\epsilon_{s,n}$  also depend on  $r$ , but in later sections we find this assumption unnecessary. In a slight abuse of terminology, we sometimes refer to the concentrations  $T(n, \hat{c}_{s,n,r}, r)$  as harmonized (without the modifier “approximately”) when the meaning is clear from context.

#### B. Notation and Conventions

- For clarity, we reserve certain indices for special purposes. Lowercase  $m$  refers to either a sample or a reference. Lowercase  $s$  and  $r$  refer exclusively to samples and references. Lowercase  $n$  always refers to an assay. Lowercase  $k$  is used generically as an integer index except in any of the previously mentioned cases. *We caution the reader that throughout the manuscript, indices are often the primary (and sometimes only) way we indicate functional dependence between variables.* For example, we use  $F_s$  to denote a fluorescence value that changes with the discrete sample index  $s$ , whereas  $c_{s,n}$  is a concentration depending on both the discrete sample and assay indices  $s$  and  $n$ .
- In certain cases, we need to non-dimensionalize the arguments of transcendental functions by dividing through by the units. In such cases, we use the symbol  $U_\star$  to indicate a quantity whose value is 1 multiplied by the units associated with  $\star$ .
- We treat Gibbs free-energies  $G$  (and differences  $\Delta G$  thereof) as dimensionless, having been divided by the temperature expressed in units of energy.
- We denote a normal random variable with mean  $\mu$  and variance  $\sigma^2$  via  $\mathcal{N}(\mu, \sigma^2)$ .

#### IV. PHYSICAL PRINCIPLES OF OUR ANALYSIS

The task of harmonizing measurements necessarily requires that they first be put on some scale, i.e. via normalization. Thus, it stands to reason that harmonization can be impacted in a detrimental way if we do not first consider normalization *per se*.

In the subsections that follow, we consider both tasks from a physics-based perspective, as this suggests validation exercises and uncertainty models that can be used to interrogate the quality of data. Subsection IV A analyzes normalization in the generic context of signal generation and yields a test for answering the question, “when does a reference material behave like a typical sample?” Subsection IV B unravels the underlying thermodynamic processes and answers the question, “what causes measurements to differ between instruments?”

The reader should keep in mind that for both normalization and harmonization, the technical exposition is always as follows. We first solve the “forward problem” by formulating and analyzing a model of the measurement (whose output values are known) as a function of physical parameters, which are often unknown in practice. This model then motivates a regression analysis that solves the “reverse problem” by extracting the unknown parameters from the data.

##### A. Generic Aspects of Signal Generation: Implications for Antibody Normalization and Units

Normalization aims to quantify the concentration  $c$  of antibodies that bind to a substrate. However, this concentration is never measured directly; neither is it possible to measure the *total concentration* of antibodies except as part of the manufacturing process for certain reference materials. Instead, the instrument outputs some numerical value such as median fluorescence intensity (MFI)  $F$  expressed, for example, in units of voltage. This  $F$  is typically assumed to be proportional to the bound concentration. The constant of proportionality, which we denote  $\Gamma$ , has units such as voltage (i.e. some proxy for MFI) times volume per number of bound antibodies; viz.

$$F = \Gamma c. \quad (3)$$

In turn,  $c$  is assumed to be proportional to the total concentration  $y$  of antibodies of a fixed type via the theory described in Sec. IV B; see also Refs. [18, 19]. We denote the corresponding proportionality constant by  $K$ , which is dimensionless but should be thought of as having units of bound antibody number per total antibody number. Thus, one trivially finds

$$c = Ky \implies F = \Gamma Ky, \quad (4)$$

where the product  $\Gamma K$  has units of voltage per concentration of total antibodies.

Normalization is typically performed by computing a ratio of the form

$$\hat{y}_{s,r} = \frac{F_s}{F_r} \quad (5a)$$

$$= \frac{c_s}{c_r} = \frac{K_s y_s}{K_r y_r} \quad (5b)$$

where subscripts  $s$  and  $r$  denote corresponding quantities for a sample and reference [18, 29]. While this  $\hat{y}_{s,r}$  is ostensibly a normalized antibody value, it is dimensionless; i.e. it is the number of bound sample antibodies per bound reference antibody. To make  $\hat{y}_{s,r}$  consistent with Definition I, we multiply by  $y_r$ , which is assumed to be known, yielding

$$\hat{c}_{s,r} = \hat{y}_{s,r} y_r. \quad (6)$$

The left-hand side (LHS) of Equation (6) has units of total antibodies per volume.

Four comments are in order.

First, recognize that many of the above quantities ( $\hat{c}_{s,r}$ ,  $\hat{y}_{s,r}$ ,  $c_r$ ,  $c_s$ , etc.) depend on the choice of assay. In subsequent sections we make this explicit by including the subscript  $n$ . Here, we have suppressed the assay dependence since it is not central to a discussion of units.

Second, a more appropriate definition of  $\hat{c}_{s,r}$  would be  $\hat{y}_{s,r} c_r$ , since this has the biologically relevant units of bound antibodies per volume [14, 15, 24].<sup>3</sup> However,  $K_r$  is generally unknown, which implies the same for  $c_r$ . Thus, one can only normalize bound antibody concentration up to an unknown scale factor associated with the  $K_r$ , which turns out to be an equilibrium constant; see Sec. IV B.

Third, while  $\hat{y}_{s,r}$  and  $\hat{c}_{s,r}$  are nominally different, they are interchangeable from a theoretical standpoint because of Eq. (6). The quantity  $\hat{y}_{s,r}$  is mathematically more convenient because it is dimensionless, and we often use this form of the normalized measurement. We refer to  $\hat{y}_{s,r}$  as a *scaled* antibody concentration and take  $\chi_{s,r} = \chi_{s,r}/y_r$  to be the corresponding consensus value; see Definition IV.

Fourth, normalization is typically accomplished by computing the ratio given by Eq. (5a), i.e. directly in terms of the measurement outputs at a single concentration. However, nothing restricts us from iteratively diluting a sample and measuring the associated range of fluorescence values.<sup>4</sup> Mathematically, this amounts to multiplying  $y_m$  by a *dilution factor*  $d$ , which can take any positive value less than one ( $0 < d < 1$ ). Equation

<sup>3</sup> When  $y_r$  is unknown, as is the case for human-derived, pooled samples, its value is assigned arbitrarily [14, 15, 24]. In this case, it is equally valid to fix  $c_r$  instead. This is purely a semantic choice that does not play a role in our analysis.

<sup>4</sup> In fact, serology measurements are often performed on such a *dilution series*, although often only one dilution is used for data analysis.

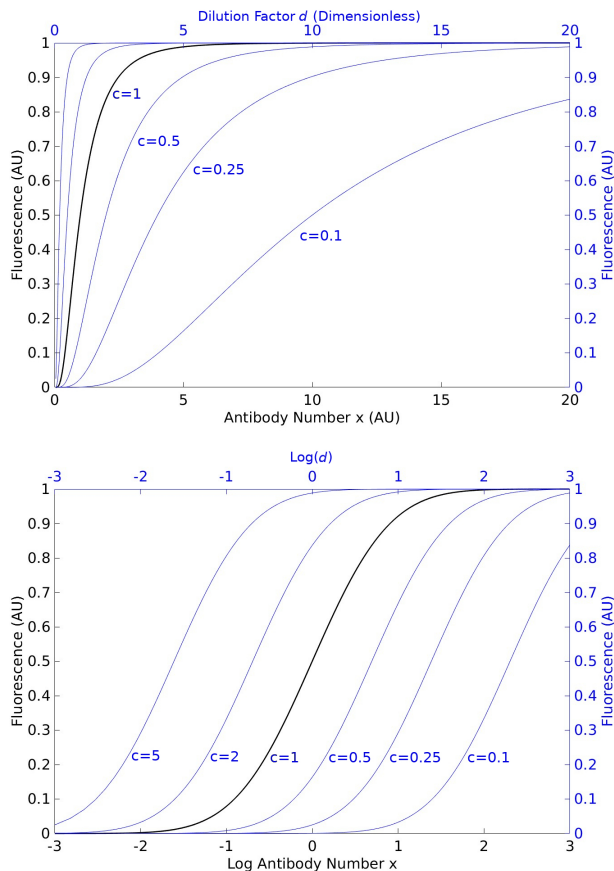


FIG. 1. Plots of synthetic data motivating Eqs. (7) and (8). Note that colors are matched to axes and do not have interchangeable interpretations. Our normalization algorithm prescribes a mathematical method for reconciling their differences. In both plots,  $d$  corresponds to a dilution factor. In the top plot, the left-most blue curves correspond to  $c = 5$  and  $c = 2$  from left to right. In all plots, the black dilution curve corresponds to a reference material. Changing the concentration of antibodies in the reference traces out the dilution curve. Equivalently, if we fix the concentration (say to  $c = 1$ ), diluting ( $d < 1$ ) or concentrating ( $d > 1$ ) the reference yields the same curve. *Importantly, this dual interpretation only applies to the reference.* Taking a sample for which  $c > 1$  implies that the sample must be diluted *more* relative to the reference to generate the same fluorescence signal. In order to collapse the blue curve onto the reference, we must scale the dilution by a factor of  $c$  (i.e.  $c_s/c_r$  for  $c_r = 1$  and  $c_s = c$ ), which corresponds to a horizontal shift on a log scale (bottom plot). A similar interpretation applies to the case  $c < 1$ .

(5b) then implies that the scaled antibody concentration is in fact an invariant quantity *if* both the reference and sample are diluted by the same amount.

This simple observation leads to a normalization method that estimates  $\hat{y}_{s,r}$  and validates whether a reference material exhibits the same behavior as a typical sample.

To realize this in practice, we solve the forward modeling problem via a few simple observations. First let

$x = cd$  be a diluted bound concentration, and define  $F(x)$  be the fluorescence as a function of  $x$ . It is reasonable to assume that  $F(x)$  is strictly monotone increasing in  $x$ ; i.e. more bound antibodies increase the measurement signal. Moreover, even when  $c_r$  and  $c_s$  are unknown, we can always measure the associated fluorescence curves  $F_r(d)$  and  $F_s(d)$  by varying the dilution factor. It is not necessary that  $F(x)$ ,  $F_r(d)$ , or  $F_s(d)$  be linear, as photodetectors may saturate if there is too much measurand in an instrument. But we do always assume that concentrations  $c$  are linear in  $d$ .

In this context, the fundamental assumption of serology measurements can be stated as the joint requirements

$$F_r(d) = F(c_r d), \quad F_s(d) = F(c_s d), \quad (7)$$

and

$$F(x_s) = F(x_r) \iff x_s = x_r, \quad (8)$$

where the latter is guaranteed by the monotonicity of  $F(x)$ . To make use of this, assume that we find a quantity  $\alpha_s$  such that  $F_r(d) = F_s(\alpha_s d)$ . Then Eqs. (7) and (8) imply that this  $\alpha_s$  is in fact the inverse of the scaled concentration; viz.

$$\frac{c_s}{c_r} = \frac{1}{\alpha_s} = \hat{y}_{s,r}, \quad (9)$$

Moreover, we may define

$$\hat{c}_{s,r} \equiv \frac{c_r}{\alpha_s} \quad (10)$$

to be the normalized concentration.

These observations imply that the reverse problem, i.e. normalization, is tantamount of finding a per-sample scale factor  $\alpha_s$  that, when applied to the argument of  $F_s(d)$ , collapses this dilution curve onto the reference dilution curve  $F_r(d)$ ; see Fig. 1.<sup>5</sup> This has several benefits over normalization via Eq. (5a): by leveraging more data one reduces susceptibility to noise, and it is not necessary to identify a linear regime for the function  $F(x)$ . In practice, however, finding scale factors  $\alpha_s$  is more computationally challenging. For example, typical dilutions series only sample a few points from  $F_r(d)$  and  $F_s(d)$ , which can make it difficult to test for data collapse. Moreover, the underlying  $F(x)$  is generally not known *a priori*. In Appendix B, we discuss a constrained optimization approach that overcomes both issues, leveraging only generic information about the structure of  $F(x)$ .

Figure 2 shows an example of this analysis applied to 34 SARS-CoV-2 positive samples that were normalized on the scale of the mAb reference material. We emphasize that while the reference material is a synthetic antibody, the points sampled from its dilution curve fall on

<sup>5</sup> Throughout we use the phrase **data collapse** to refer to a transformation that maps a function or dataset onto another function or dataset.

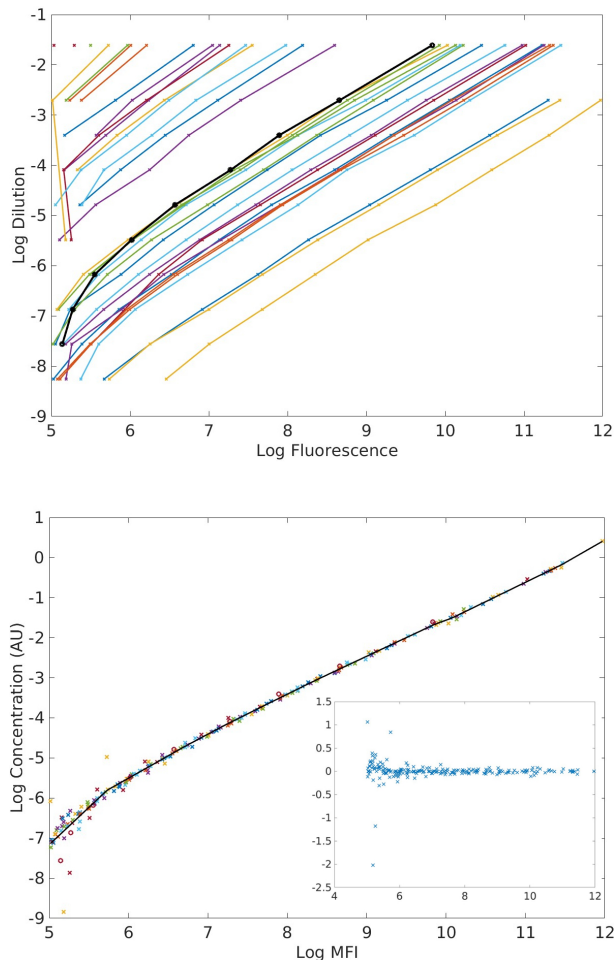


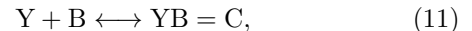
FIG. 2. Raw data and data collapse associated with 34 SARS-CoV-2 positives samples and a mAb reference material, all measured via a ligand binding assay. See Ref. [20] for experimental details. *Top*: Raw data associated with the samples. The reference is labeled with the black circles. Lines are guides for the eye, whereas the overlaid discrete data points are the dilution-fluorescence pairs measured by the instrument. The axes are flipped relative to Fig. 1. By eye, it is plausible that a vertical shift applied to each dilution curve is sufficient to collapse them. *Bottom*: Data collapse associated with minimizing Eq. (B8). The reference material is assumed to have a dimensionless concentration of 1. The solid black curve is a reconstruction of the function  $f(x)$ ; see Eq. (B1) and surrounding text. Note that the reference material has the average (weakly) non-linear behavior of the sample data after transformation. The inset shows the residuals defined here as the difference between the transformed raw data and the estimated dilution curve in black.

the composite function  $F(x)$  constructed from all of the data. The degree of data collapse serves as validation that the mAb is not inherently different from a human sample in regards to its behavior under dilution. While beyond the scope of this manuscript, one could establish a cutoff criterion so that a reference is only considered viable if data collapse relative to  $F(x)$  can only be achieved to within a certain level.

## B. Thermodynamics of Antibody Measurements

Having analyzed serology at the level of signal generation, we now turn to thermodynamic considerations that (i) explain why different measurement systems yield distinct normalized concentrations and (ii) clarify the assumptions underlying Eq. (4). In this section, we focus primarily on the physics of harmonization *per se*. In Sec. V we consider the forward and reverse modeling problems associated with this task.

A binding antibody assay can be viewed as a chemical reaction [17, 30]. A free antibody  $Y$  attaches to a substrate  $B$  to create a bound pair  $YB$ ; viz.



where  $C$  stands for the antibody-substrate complex. This reaction is assumed to be reversible, so that the system reaches an equilibrium described by detailed balance [30, 31]. For a fixed antibody type  $Y_m$  (associated with the  $m$ th sample or reference) and substrate  $B_n$  associated with an the  $n$ th assay or antigen epitope, the equilibrium constant  $\hat{K}$  can depend on *both* the antibody and the substrate. That is,

$$\frac{c_{m,n}}{(y_m - c_{m,n})(b_n - c_{m,n})} = \hat{K}_{m,n}, \quad (12)$$

where  $c_{m,n}$ ,  $y_m$ , and  $b_n$  are the concentrations of  $C_{m,n}$ , all antibodies (free and bound), and all substrates  $B_n$ . The  $n$ -dependence of  $\hat{K}_{m,n}$  reflects the fact that changing the substrate can alter the equilibrium concentration, and hence number of bound antibodies. The physical intuition for this dependence arises from the definition

$$\frac{\hat{K}_{m,n}}{U_{\hat{K}}} = e^{-\Delta G_{m,n}} \rightarrow \Delta G_{m,n} = -\ln(\hat{K}_{m,n}/U_{\hat{K}}), \quad (13)$$

where  $\Delta G_{m,n}$  is the Gibbs free-energy change associated with Eq. (11) [32]. That is, changing the substrate and/or antibody alters the free-energy landscape, and thus the equilibrium constant. Note that Eq. (12) only models antibody affinity, but not *avidity* (i.e. capacity for multivalent binding); see Ref. [33] for justification in the context of the examples considered herein.

From a measurement standpoint, it is desirable for  $c_{m,n}$  to be independent of the substrate concentration; doing so ensures that the former increases linearly with the total antibody concentration [7, 29]. A straightforward Taylor expansion of Eq. (12) reveals that this condition is approximately satisfied if either (i)  $b_n \gg y_m$  and  $\hat{K}_{m,n} \gtrsim \mathcal{O}(1)$  unit volume, or (ii)  $\hat{K}_{m,n} \ll 1$  unit volume for  $b_n$  and  $y_m$  order one concentrations. We henceforth assume that either (i) or (ii) is true,<sup>6</sup> which yields the

<sup>6</sup> It is necessary to distinguish nonlinearity in the measurement due to Eq. (12) from nonlinear effects due to detection equipment such as photodetectors. This distinction is important for the analysis in Sec. IV A.

approximate model

$$c_{m,n} = K_{m,n} y_m, \quad (14)$$

where  $K_{m,n}$  is an appropriately rescaled equilibrium constant.<sup>7</sup>

An interesting question that motivates our harmonization analysis is whether the equilibrium constant is *separable*, meaning it can be expressed in the form

$$K_{m,n} = \kappa_{Y_m} \kappa_{B_n}, \quad (15)$$

for some constants  $\kappa_{Y_m}$  and  $\kappa_{B_n}$  depending only on the sample or assay, but not both [34].<sup>8</sup> Physically, Eq. (15) is interpreted as the condition in which the antibodies' contribution to the equilibrium constant is independent of the substrate contribution.<sup>9</sup> In the context of Eq. (13), this implies that the change in free energy can be expressed as a sum

$$\Delta G_{m,n} = \Delta \hat{G}_m + \Delta \hat{G}_n, \quad (16)$$

where  $\Delta \hat{G}_m$  and  $\Delta \hat{G}_n$  are distinct quantities depending on either the sample or substrate, but not both. Heuristically Eq. (16) is plausible if *any antibody* that binds to a fixed antigen (corresponding to constant  $n$ ) always changes the latter's conformation in the same way, so that  $\Delta G_{m,n}$  only varies with the internal energy and entropy differences between the antibodies themselves.

To understand the usefulness of separability, recall that antibody normalization amounts to determining the ratio  $c_{s,n}/c_{r,n}$  for sample  $s$ , reference  $r$ , and a fixed assay  $n$ . Separability then amounts to the condition that

$$\frac{c_{s,n}}{c_{r,n}} = \frac{\kappa_{Y_s} y_s}{\kappa_{Y_r} y_r} \quad (17)$$

which is equivalent to

$$\ln \left( \frac{c_{s,n}}{c_{r,n}} \right) = \ln \left( \frac{y_s}{y_r} \right) - \Delta \hat{G}_s + \Delta \hat{G}_r. \quad (18)$$

*In other words, relative concentrations of bound antibodies are independent of the assay being used for the measurement, since the right-hand side (RHS) has no dependence on  $n$ . Separability therefore implies that normalization automatically harmonizes assays in the sense of Definition I, and we can simply set the function  $T(n, c) = c$ .*

<sup>7</sup> When  $\hat{K}_{m,n} \ll 1$  unit volume,  $K_{m,n} = \hat{K}_{m,n} b_n$ . When  $b_n \gg y_m$ , one finds  $K_{m,n} = 1$ .

<sup>8</sup> This property is also called *rank factorization*, since it amounts to representing a matrix in terms of objects with lower rank.

<sup>9</sup> Note that the  $\kappa_{Y_m}$  and  $\kappa_{B_n}$  are not determined uniquely by Eq. (15). We can always define new constants  $\hat{\kappa}_{Y_m} = \kappa_{Y_m}/\alpha$  and  $\hat{\kappa}_{B_n} = \kappa_{B_n} \alpha$  for any positive constant  $\alpha$  such that the product  $K_{m,n} = \hat{\kappa}_{Y_m} \hat{\kappa}_{B_n} = \kappa_{Y_m} \kappa_{B_n}$  is unchanged.

If we relax the separability assumption, harmonization is no longer guaranteed.<sup>10</sup> To see this, assume that

$$K_{m,n} = \kappa_{Y_m} \kappa_{B_n} \exp(-\Delta g_{m,n}), \quad (19)$$

where  $\Delta g_{m,n}$  is a relative free-energy deviation from separability. We require that  $\Delta g_{m,n}$  depend non-trivially on both of its indices. More precisely, when viewed as a matrix with elements  $s, n$ , we require  $K_{s,n}$  to have rank greater than one. Taking the logarithm of  $c_{s,n}/c_{r,n}$  yields

$$\ln \left( \frac{c_{s,n}}{c_{r,n}} \right) = \ln \left( \frac{\kappa_{Y_s} y_s}{\kappa_{Y_r} y_r} \right) - \Delta g_{s,n} + \Delta g_{r,n}. \quad (20)$$

The term  $\Delta g_{s,n}$  is problematic; it implies that the normalized concentration depends on the free-energy of the specific sample-assay pair, which is nominally inconsistent with harmonization.

In practice, we expect that Eq. (20) is a more realistic description of antibody measurements; biological variation between human samples and differences in substrate epitopes will cause some antibodies to bind differently to some assays, invalidating the heuristic picture described below Eq. (16). Thus it is not clear that exact harmonization is possible. However, we can still recover the weaker notion of approximate harmonization given by Def. V. In particular, were it possible to determine the  $\Delta g_{r,n}$ , one could define a transformation

$$T(n, \hat{c}, r) = \hat{c} \exp(-\Delta g_{r,n}), \quad (21)$$

where we now reveal the explicit dependence of  $T$  on the reference material. Combined with Eq. (20), this would imply that

$$T(n, \hat{c}_{s,n,r}, r)(1 + \epsilon_{s,n}) = T(n', \hat{c}_{s,n',r}, r)(1 + \epsilon_{s,n'}) \\ = \chi_{s,r} = \chi_{s,r} y_r \quad (22)$$

$$\epsilon_{s,n} = \exp(\Delta g_{s,n}) - 1 \quad (23)$$

yields a consensus value.

A notable conclusion of Eq. (20) is that lack of harmonization has nothing to do with the choice of reference material. It is due solely sample-assay dependent effects  $\epsilon_{s,n}$ , since negating these implies that Eq. (21) is an exact harmonization rule according to Definition V. Physically, this conclusion arises from the simple fact that all samples are normalized by the same reference material, so they share a common bias associated with  $\Delta g_{r,n}$ . Our companion manuscript validates this result for a collection of several monoclonal antibodies [20]. The next section motivates an analysis to test the validity of Eq. (22).

<sup>10</sup> We do not explore the theoretical question of whether separability is *necessary* for harmonization.

## V. INDUCED PROBABILISTIC PERSPECTIVE

Equation (22) begs two questions: (i) how do we validate the underlying model; (ii) how do we use it to harmonize assays? Our main goal in Sec. V A is to show how Eq. (22) induces a probabilistic model that describes the serology measurements. In Sec. V B, we show how to perform a regression analysis on this model that answers questions (i) and (ii).

### A. The Forward Problem: Modeling Harmonized Measurements

Observe that the quantities  $\Delta g_{s,n}$ ,  $\Delta g_{r,n}$ , and thus  $y_{s,r}$  are in general unknown, since it is unreasonable to perform detailed measurements characterizing equilibrium constants for all samples and assays. For a fixed assay, however,  $\Delta g_{r,n}$  is common to all samples, whereas  $\Delta g_{s,n}$  is sample dependent. Rearranging Eq. (20) implies that

$$\ln(\hat{y}_{s,n,r}) - \Delta g_{r,n} + \Delta g_{s,n} = \ln(y_{s,r}) \quad (24)$$

which suggests interpreting  $\Delta g_{r,n}$  as a constant, reference-dependent *bias* and  $\Delta g_{s,n}$  as a sample-assay-dependent realization of a random variable.

Because the  $\Delta g_{s,n}$  correspond to Gibbs Free-Energy changes associated with sample antibodies, we treat this quantity as an  $s$ -dependent realization of a mean-zero normal random variable with variance  $\zeta_n^2$ . That is, we assume

$$\Delta g_{s,n} = \mathcal{N}_s(0, \zeta_n^2). \quad (25)$$

where  $\zeta_n^2$  is an unknown constant, and  $\mathcal{N}_s$  are independent and identically distributed normal random variables; i.e. the expectation  $\mathbf{E}[\Delta g_{s,n} \Delta g_{s',n'}] = 0$  if  $s \neq s'$  or  $n \neq n'$ . If we also treat  $\Delta g_{r,n}$  and  $y_{s,r}$  as unknown scalars, Eq. (24) yields a model of the normalized data  $\hat{y}_{s,n,r}$  that nominally solves the forward problem. Moreover, determining the unknowns in Eq. (25) provides all of the information needed to construct  $T$  and  $\epsilon$  in Eq. (22), thereby harmonizing the assays according to Def. VI.

In order for such an analysis to be meaningful, however, it is necessary to account for uncertainty inherent in the measurement process. The true  $\hat{y}_{s,n,r}$  are never known exactly due to effects such as pipetting error, instrument artifacts, etc. Thus, it is important to ensure that the associated measurement variability is not confused with the  $\Delta g_{r,n}$  and  $\Delta g_{s,n}$ . We therefore postulate that the quantity  $\bar{y}_{s,n,r}$  one measures is related to  $\hat{y}_{s,n,r}$  via the equation

$$\ln(\hat{y}_{s,n,r}) = \ln(\bar{y}_{s,n,r}) + \delta_{s,n,r} \quad (26)$$

where  $\delta_{s,n,r}$  is a “within-lab” uncertainty that models the effects described immediately above; see Ref. [35]. The dependence of  $\delta$  on  $s$ ,  $n$ , and  $r$  is largely incidental,

since this quantity should be estimated separately for each such triple. The  $\delta$  does not account for assay and reference-specific effects. Combining Eqs. (24) and (26) then yields

$$\ln(\bar{y}_{s,n,r}) - \Delta g_{r,n} + \Delta g_{s,n} + \delta_{s,n,r} = \ln(y_{s,r}). \quad (27)$$

In practice, the values of  $\bar{y}_{s,n,r}$  and  $\delta_{s,n,r}$  are given by direct measurement outputs, while the remaining quantities must be determined from the data via regression; such details are postponed until Sec. V B.

We end this section with two small technical issues.

First, Eq. (27) nominally complicates our ability to determine whether the  $\Delta g_{s,n}$  are reference independent, since it re-introduces reference-dependence into residuals of the data. However, we always find that  $\delta_{s,n,r}$  exhibits separability of the form

$$\delta_{s,n,r} = \delta_{s,n} + \delta_{r,n}. \quad (28)$$

Under this condition,  $\delta_{r,n}$  can be absorbed into  $\Delta g_{r,n}$ , and we again deduce that the residuals of bias-corrected antibody measurements (relative to the consensus value) should be reference independent. As a result, Eq. (28) restores our ability to validate the model.

In the context of single-point normalization having the form of Eq. (5a), it is straightforward to prove that Eq. (28) is exact. In particular, we can assume that  $F_{s,n}$  and  $F_{r,n}$  are measured up to some relative uncertainty, e.g.  $F_{s,n} = \bar{F}_{s,n}(1 + \varepsilon_{s,n})$ , where  $\bar{F}_{s,n}$  is a true or expected fluorescence, and  $\varepsilon_{s,n}$  is measurement uncertainty. Then ratios of the form  $\ln(F_{s,n}/F_{r,n})$  can be expressed directly in the form of Eq. (28). For the analysis discussed in Sec. IV A, it is more difficult to prove that Eq. (28) holds, but in practice, we find that it is valid. See Figs. 3 and 4, for example.

Second, we can validate the Eq. (27) by testing the degree to which the  $\Delta g_{s,n}$  empirically depend on the reference. Despite the assumption that these quantities are random, Eq. (24) indicates that remain they *point-wise constant* with changing  $r$ , provided the  $\Delta g_{r,n}$  are correctly determined. This restrictive criterion implies that the residuals between the bias-corrected and consensus antibody estimates should be invariant to the reference, which can be directly checked. We pursue this and related ideas in Sec. V B.

### B. The Reverse Problem: Harmonization via Regression

We now address the task of estimating the quantities  $\Delta g_{r,n}$ ,  $\Delta g_{s,n}$ , and  $\delta_{s,n,r}$ .

While there are a variety of methods for estimating  $\delta_{s,n,r}$  (e.g. [36, 37]), we consider the common situation wherein measurements are repeated. Extensions and alternatives are discussed in Sec. VI. Assume therefore that we are given  $S$  samples,  $A$  assays, and one reference material. Each sample and reference are measured

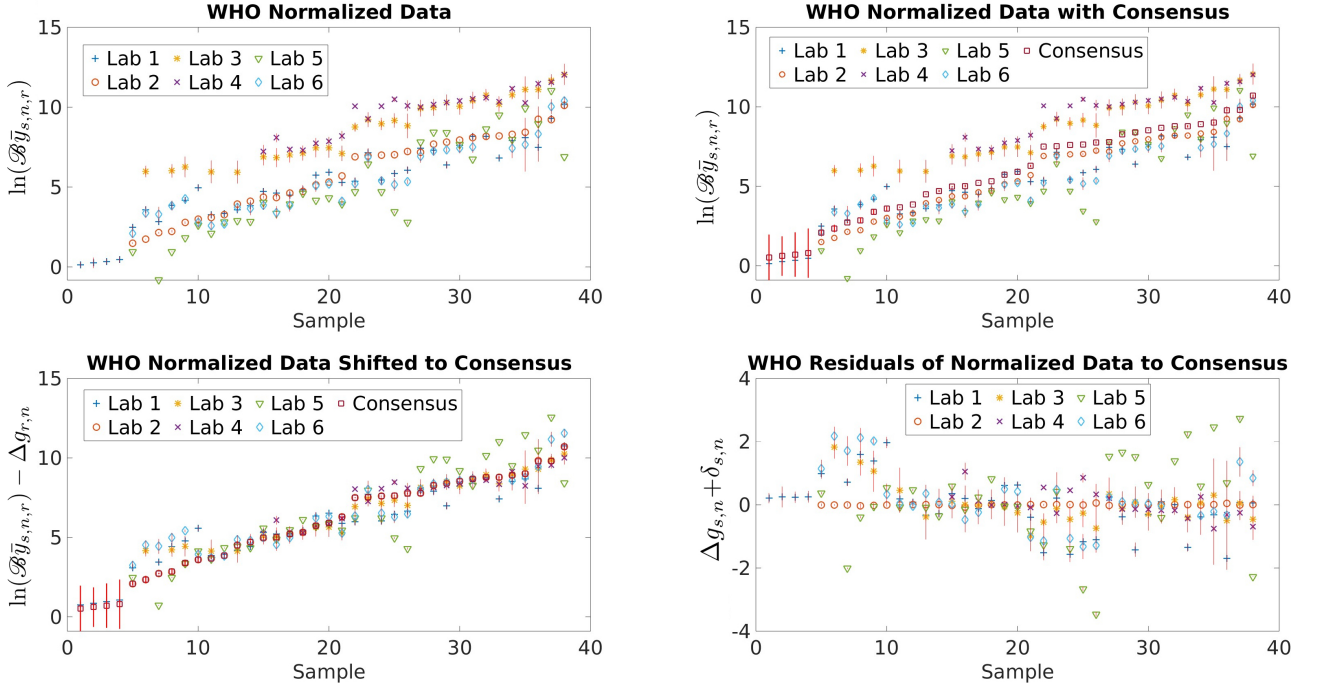


FIG. 3. Example of the analysis leading to the solution given by Eq. (34). Symbols have the same meanings in all figures and are defined in the top plots. *Top left:* Normalized data for 38 positive samples measured via 6 different assays. *Top right:* Normalized data with consensus estimates given by Eq. (34). *Bottom left:* Harmonized data, i.e. normalized data corrected for the assay-dependent biases  $\Delta g_{r,n}$  so that all samples are distributed about the consensus values. *Bottom right:* Difference between the harmonized data and consensus values. In all figures, the vertical red bars centered at each datapoint are one-sigma confidence intervals. For the lab-specific datapoints, these confidence intervals are given by  $\delta_{s,n,r}$ . For the consensus values, the confidence intervals are computed from the distribution of values arising from the jackknife analysis, as described in the main text.

$\tau$  times for each assay, where  $t \in \{1, 2, \dots, \tau\}$  indexes these time-points. We use the analysis of Appendix. B to normalize all antibody measurements relative to the reference dilution curve measured at the same time. Denote the corresponding antibody levels  $\tilde{y}_{s,n,r,t}$ . Assuming that these measurements are independent in the  $t$  index, we construct the arithmetic mean estimator

$$\ln(\bar{y}_{s,n,r}) = \frac{1}{\tau} \sum_{t=1}^{\tau} \ln(\tilde{y}_{s,n,r,t}) \quad (29)$$

and sample standard uncertainty [38]

$$\sigma_{s,n,r}^2 = \frac{1}{\tau(\tau-1)} \sum_{t=1}^{\tau} [\ln(\bar{y}_{s,n,r}) - \ln(\tilde{y}_{s,n,r,t})]^2, \quad (30)$$

where we approximate the variance  $\text{Var}[\delta_{s,n,r}] = \sigma_{s,n,r}^2$ . Note that the estimate for  $\bar{y}_{s,n,r}$  corresponds to a geometric mean of antibody concentrations. See Eqs. (26) and (27).<sup>11</sup> Given a few replicates  $\tau$  per sample, we make the

additional minimal assumption that  $\delta_{s,n,r}$  is a mean-zero normal random variable with variance  $\sigma_{s,n,r}^2$ .

Returning to Eq. (27),

$$\ln(\bar{y}_{s,n,r}) - \Delta g_{r,n} + \Delta g_{s,n} + \delta_{s,n,r} = \ln(\mathbf{y}_{s,r}),$$

assume that  $r$  is fixed and observe that the scaled and consensus values  $\bar{y}_{s,n,r}$  and  $\mathbf{y}_{s,r}$  depend on the reference material  $r$ . Provisionally assume that the assay-sample dependent effects  $\Delta g_{s,n}$  are independent of  $r$ , which we check after-the-fact by varying the reference.

In light of Eq. (27), the  $\ln(\bar{y}_{s,n,r})$  distributed as

$$\ln(\bar{y}_{s,n,r}) = \mathcal{N}(\ln(\mathbf{y}_{s,r}) - \Delta g_{r,n} - \ln(\bar{y}_{s,n,r}), \zeta_n^2 + \sigma_{s,n,r}^2),$$

which has the corresponding probability density

$$P(\ln(\bar{y}_{s,n,r}) | \Delta g_{r,n}, \mathbf{y}_{s,r}, \zeta_n^2) = \frac{e^{-\frac{(\ln(\mathbf{y}_{s,r}) - \Delta g_{r,n} - \ln(\bar{y}_{s,n,r}))^2}{2(\zeta_n^2 + \sigma_{s,n,r}^2)}}}{\sqrt{2\pi(\zeta_n^2 + \sigma_{s,n,r}^2)}}. \quad (31)$$

<sup>11</sup> Recall that the logarithm of concentration is linear in the Gibbs free energy. Thus Eq. (29) can also be viewed as an estimate of the average  $\Delta G$ .

In the event that  $\bar{y}_{s,n,r}$  falls below a detection threshold  $y_\theta$  (corresponding to dilution curves for which no measurement is above the signal-to-noise of 1/10 in the ex-

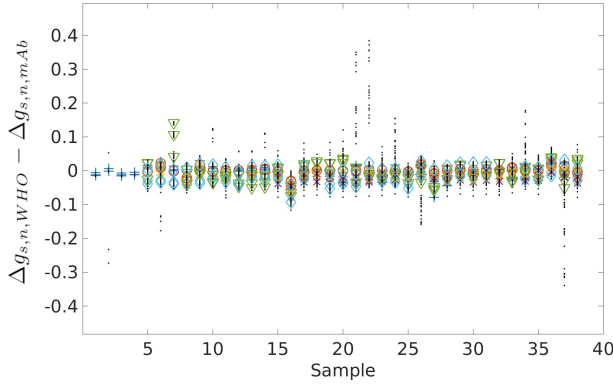


FIG. 4. The difference of residuals  $\Delta g_{s,n,WHO} - \Delta g_{s,n,mAb}$  computed from the WHO standard and the 3 other mAbs used in the interlab study associated with Ref. [20]. Symbols and colors have the same meaning as in Fig. 3. Each point with a fixed color and symbol is associated with a different mAb. The difference in residuals is typically less than 0.05, which corresponds to a relative variation in antibody number of 5%. This demonstrates that the quantity  $\Delta g_{s,n}$ , which is associated with the sample-assay dependent randomness about the consensus value, does not depend on the reference material. The black dots are the distributions of values obtained from the jackknife analysis as described in the main text. Note that only 0.43% of the values in the plot have a magnitude greater than 0.1. Also, for the three samples with the lowest consensus values, only one lab was able to yield uncensored concentrations. As a result, the jackknife analysis yields a total of eight datapoints that fall outside the scale of the plot. These datapoints have characteristic values of 2 in the units of these axes.

amples above), we can at most define the censored probability

$$\mathcal{P}(\bar{y}_{s,n,r} \leq y_\theta) = \int_{-\infty}^{\ln(y_\theta)} P(\bar{z} | \Delta g_{r,n}, \mathbf{y}_{s,r}, \zeta_n^2) d\bar{z} \quad (32)$$

in terms of Eq. (31), where we take  $y_\theta$  to be the smallest measured value of  $\ln(\bar{y}_{s,n,r}) > -\infty$ , and  $\bar{z}$  is a dummy integration variable representing  $\ln(\bar{y}_{s,n,r})$ ; see also Refs. [39, 40]. [For simplicity of notation, we have suppressed the dependence of  $\mathcal{P}(\bar{y}_{s,n,r} \leq y_\theta)$  on  $(\Delta g_{r,n}, \mathbf{y}_{s,r}, \zeta_n^2)$ .] Given these quantities, we define a regularized, negative log-likelihood objective function to be

$$\begin{aligned} \mathcal{L}_l = & - \sum_{\substack{s,n: \\ \bar{y}_{s,n,r} > y_\theta}} \ln [P(\ln(\bar{y}_{s,n,r}) | \Delta g_{r,n}, \mathbf{y}_{s,r}, \zeta_n^2)] \\ & + \tilde{\epsilon}_3 \left[ \sum_n \Delta g_{r,n} \right]^2 - \sum_{\substack{s,n: \\ \bar{y}_{s,n,r} \leq y_\theta}} \ln [\mathcal{P}(\bar{y}_{s,n,r} \leq y_\theta)] \quad (33) \end{aligned}$$

by summing Eq. (31) over all samples and assays. Observe that  $\mathcal{L}_l$  is a function of the consensus values  $\mathbf{y}_{s,r}$ , reference-dependent free energy bias  $\Delta g_{r,n}$ , and assay-dependent free-energy variances  $\zeta_n^2$ . The value of  $\tilde{\epsilon}_3$  can be set to any positive value (we set  $\tilde{\epsilon}_3 = 1$ ), as the reg-

ularization only serves to remove a connected set associated with ambiguity in the value of  $\Delta g_{r,n}$ . (That is, we can only determine the  $\Delta g_{r,n}$  up to an additive constant, necessitating the regularization.) Minimizing Eq. (33) yields estimates of these parameters via

$$\{\Delta g_{r,n}^*, \{\mathbf{y}_{s,r}^*\}, \{\zeta_n^{2,*}\}\} = \underset{\{\Delta g_{r,n}, \{\mathbf{y}_{s,r}\}, \{\zeta_n^2\}\}}{\operatorname{argmin}} \mathcal{L}_l. \quad (34)$$

While Eq. (34) is ultimately a choice of how to define consensus values, it admits an interpretation that is consistent with physical intuition. In particular, recognize that  $\Delta g_{s,n}$  (i.e. the sample-assay component of free energy) is treated as random because in general, we do not know *a priori* how a sample will interact with an assay. In minimizing  $\mathcal{L}_l$ , we therefore seek the consensus values (up to a constant offset fixed by the regularization) and assay-dependent variability that maximizes the probability of the measured data. Thus, we interpret Eq. (34) as the most likely probabilistic representation of the data [41]. Moreover, including  $\sigma_{s,n,r}^2$  in the definition of  $\mathcal{L}_l$  accounts for the fact that measurements with high uncertainty contribute less to our knowledge of the corresponding consensus values and assay-dependent uncertainties. See Sec. VI for further analysis of this model.

To characterize uncertainties associated with the parameters  $\Delta g_{r,n}^*$ ,  $\mathbf{y}_{s,r}^*$ , and  $\zeta_n^{2,*}$ , one can perform, for example, a jackknife style analysis [42, 43]. In the examples that follow, we perform this analysis for fixed reference  $r$  by omitting from the optimization each  $\bar{y}_{s,n,r}$  one time [44].<sup>12</sup> We then take the standard deviations from the distributions of parameter estimates as the corresponding confidence intervals.

Figure 3 shows the results of this analysis applied to mAb 1 considered in Ref. [20]. The top plots show the normalized log-concentrations  $\ln(\mathcal{B}\bar{y}_{s,n,r})$  with and without the consensus values for six different labs and 38 positive samples, where  $\mathcal{B} = 1.2 \times 10^6$  is the number of ng/mL of antibody in the undiluted reference. The bottom plots show bias-corrected (i.e. harmonized) estimates  $\ln(\mathcal{B}\bar{y}_{s,n,r}) - \Delta g_{r,n}$ . The remaining variation is quantified by  $\Delta g_{s,n} + \delta_{s,n}$  [recall Eq. (28)]. Figure 4 shows the differences in residuals between harmonized concentrations and consensus values according to Eq. (34) using

<sup>12</sup> In principle, jackknife methods can be used to quantify and remove a bias associated with an estimator. In the experiments accompanying this manuscript, however, the measurements for some samples fell below the detection threshold for all but one lab. Leaving out the single uncensored concentration for that sample therefore couples the corresponding consensus value to only the terms  $\Delta g_{r,n}$  and  $\zeta_n^2$ . In practice, we find that this yields jackknife bias correction terms that are unphysically large, so that we do not consider them further. Moreover, censoring is known to cause problems for resampling-based techniques [44]. As a result, we only consider the standard deviations from the jackknife distributions, which we treat as reasonable proxies for the true parameter uncertainties. A further analysis of this situation is beyond the scope of the present manuscript.

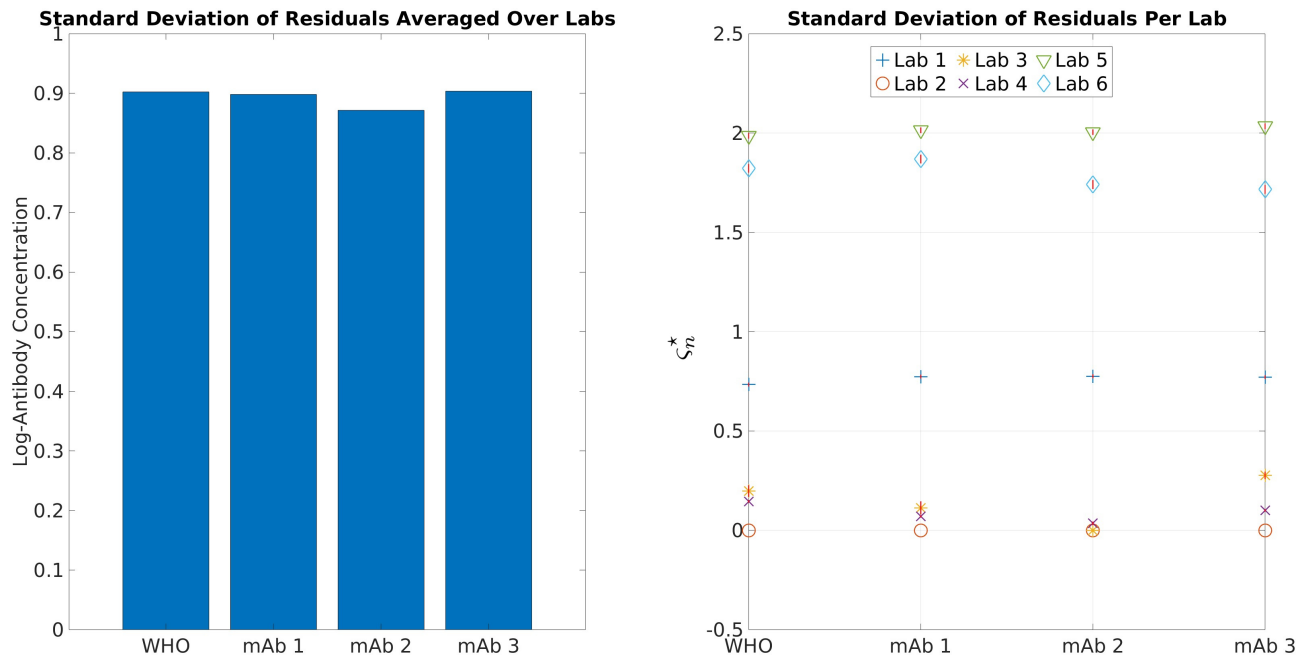


FIG. 5. Further confirmation that approximate harmonization via Eq. (34) is reference independent. *Left*: Square root of the average of  $(\zeta_n^*)^2$  over all  $n$  assays. Note that the standard deviation is approximately constant across all standards. *Right*: Maximum likelihood estimate of  $\zeta_n^*$  as a function of assay. On the right plot, the vertical red bars for each datapoint are one standard deviation confidence intervals associated with the jackknife analysis described in the main text. These confidence intervals are generally the same size as or smaller than the corresponding symbols and in some cases, are not visible.

the WHO standard and three different mAbs described in Ref. [20]. The bottom plot in particular shows that the residuals change by less than roughly 0.05 on a log scale when switching reference material, which corresponds to roughly 5% relative variation in antibody concentration. This validates that to within good approximation, the residuals can be expressed only in terms of  $s$  and  $n$  via  $\Delta g_{s,n} + \delta_{s,n}$ , as predicted by our thermodynamic model. In other words, the lack of exact harmonization is due almost entirely to coordinated assay-sample effects.

To further illustrate this last point independently from the noise terms  $\delta_{s,n}$  and  $\delta_{s,n,r}$ , recognize that the variance  $(\zeta_n^*)^2$  is an estimate of the statistical properties of the  $\Delta g_{s,n}$  via Eq. (25). Figure 5 shows the both the average of these variances over all  $n$ , as well as their assay-specific values. All estimates are independent of the reference material used.

## VI. DISCUSSION: BROADER IMPLICATIONS FOR SEROLOGY

### A. Deeper Comparison with Past Works

Several studies have considered the impacts of normalization and harmonization, both from the standpoint of establishing reference materials [14, 15, 24] and deploying them in real-world settings [3–6]. However, none of these works formally defined the relationship between normal-

ization and harmonization, implicitly taking these tasks to be identical. More specifically, the authors tended to use manufacturer-specified scaling values (relative to a pooled standard) to harmonize measurements without the bias-correction term corresponding to  $\Delta g_{r,n}$ . In terms of our analysis, this amounts to the assumption that  $T(n, \hat{c}) = \hat{c}$ ; i.e. the equilibrium constants are separable.

This lack of distinction between the concepts of harmonization and normalization may therefore be responsible for significant confusion within the serology community. For example, a universal conclusion of Refs. [3–6] has been that “harmonized” (i.e. normalized) measurements are not interchangeable, even when using human-derived standards. Interestingly, the authors also observed at least two common trends: (i) normalized antibody measurements are correlated between different assays; and (ii) for a fixed sample, variability of normalized concentrations between assays increases with increasing antibody titers.

Our analysis provides a likely explanation for both observed results. Considering (i), the implicit choice  $T(n, \hat{c}) = \hat{c}$  ignores the reference-dependent bias term  $\exp(-\Delta g_{r,n})$  appearing in Eqs. (21) and (22). Thus, it is not surprising that the harmonized values are proportional but not identical on average. We predict this missing constant of proportionality should be  $\exp(\Delta g_{r,n} - \Delta g_{r,n'})$  for two different assays  $n$  and  $n'$ . Concerning (ii), the increasing uncertainty with antibody concentrations

is a direct manifestation of the uncertainty  $\epsilon_{s,n}$  appearing in Eq. (22), which we have already shown yields the *relative* uncertainty  $T(n, \hat{c}_{s,n,r}, r)\epsilon_{s,n}$  about consensus values. We anticipate that a post-hoc analysis of the results in Refs. [3–6] would lead to consistent predictions of the aforementioned physical quantities. A potentially insurmountable, and thus disconcerting corollary is that the sample-assay dependent effects may be so large in real-world settings as to induce significant uncertainty, even using the harmonization techniques that we propose.

Another notable point of comparison is the calibration study performed in Ref. [24]. The authors determined that in establishing the reference material, normalization via a human-derived, pooled standard harmonized samples to within a factor of two or better on average.<sup>13</sup> This contrasts with Ref. [20], wherein we found that *after* correcting for reference-induced biases (i.e.  $\Delta g_{r,n}$ ), harmonization could only be achieved to within a factor of approximately 2.5 ( $\exp(0.9)$ ) on average; see Fig. 5. Without the bias correction, harmonization was only achieved to within a factor of roughly 12 on average [20]. The result of Ref. [24] is especially surprising because the study therein included IgG measurements from SARS-CoV-2 spike, receptor binding domain (RBD), and nucleocapsid (N) assays, whereas we only considered the first two. In the former study, one might expect that the corresponding set of  $\Delta g_{r,n}$  and  $\Delta g_{s,n}$  would be larger, considering that more types of assays were used.

A deeper analysis of the experimental design in Ref. [24] suggests a resolution to this disparity. It is noteworthy that the corresponding validation study only attempted to harmonize 5 samples (compared to our 38), four of which were also used to develop the standard. In the context of Eq. (24), it is likely that this choice of validation samples causes cancellation between the terms  $\Delta g_{r,n}$  and  $\Delta g_{s,n}$ . Ultimately, this would lead one to underestimate the true uncertainties associated with harmonization via normalization alone (i.e. without using our bias-correction). Thus, we predict that a different validation study *not using samples also found in the standard* would yield results comparable to ours. This prediction is consistent with Refs. [3–6].

## B. On the Mathematical Interpretation of a Consensus Value and Antibody Standard

It is notable that in Figs. 3 and 4, the measurements from Lab 2 are nearly identical to the consensus values.

<sup>13</sup> In contrast, using different study designs and analyses, Refs. [3–6] found that normalization still yielded up to 20-fold systematic discrepancies between assays. Comparing these results to Ref. [24] is challenging due to differences in the way results were reported. Reference [24] provided aggregate coefficients of variation across assays, whereas the other studies explicitly correlated normalized results between assays head-to-head.

This begs the obvious question of why the analysis yields such a result, and more generally, what is the interpretation of our consensus value?

In this instance, we note that the inter-day variation  $\delta_{s,n,r}$  for Lab 2 is nearly a decade smaller than the corresponding uncertainties for the next closest lab. Thus, it stands to reason that estimation of the consensus value is dominated by those measurements having the smallest uncertainties. Ostensibly this is problematic: a precise measurement is not necessarily accurate. However, we recall that antibody concentrations can only be estimated up to an unknown multiplicative factor, since the equilibrium constants are rarely, if ever determined; see Secs. IV A and IV B.

These observations suggest that a reasonable definition of consensus is one that minimizes disagreement between the results of different labs. The maximum likelihood estimate given by Eq. (34) interprets disagreement as corresponding to low probability of the observed measurements under the assumption that the joint sample-assay component  $\Delta g_{s,n}$  of the free energy is random. [This modeling choice is justified by the fact that  $\Delta g_{s,n}$  characterizes the immune response of an individual who is arbitrarily selected from a large population.] In this context it is reasonable that a measurement with high uncertainty contributes less to the consensus estimate; its underlying true value may be closer to the remaining measurements than is reflected by its nominal value.

The subjectivity of this decision highlights the fact that from a purely performance standpoint, there may not exist a universal, best reference material for antibodies without further knowledge of equilibrium constants. This stands in contrast to hardened metrological standards based on fundamental physical constants, e.g. as established by the connection between mass and Planck’s constant [45, 46]. Absent such relationships, we are forced to *choose* definitions for “harmonization” and “consensus antibody value,” and these necessarily control what we mean by a best standard. In our analysis, the concept of “best” is associated with the reference that induces the least additional uncertainty (again a choice) into harmonization, which is in turn fixed by Def. V, Eq. (34), and the supporting modeling assumptions. However, our choices are not unique, and others may lead to distinct notions of a best standard.

A key challenge for the serology community is therefore to agree on a harmonization convention, which is necessary before standards can even be fully established. The difficulty of addressing this problem is seen in Defs. IV and V. They provide a generic structure of what harmonization entails, but critically, do not propose a functional form of the mapping  $T(c, n)$  or noise  $\epsilon_{s,n}$ . This latter task is complicated by virtue of being a task in mathematical modeling, although historically it has not been viewed as such.

Here we propose that the modeling paradigm should be dictated by its usefulness. In this respect, our approach has certain benefits grounded in its connection to physics.

Because the underlying statistical model is induced by a thermodynamic description of antibody kinetics, it provides an intuitive justification for various choices. For example, Eq. (29), which is a geometric mean over antibody number, is revealed to be an arithmetic mean over Gibbs free-energies, a quantity for which this type of averaging may be appropriate. Perhaps more importantly, the thermodynamics reveals how the reference-dependent bias can be removed as a source of uncertainty. Critically, the resulting equivalence of all standards for purposes of harmonization enables one to consider a broader definition of fitness of purpose. Issues such as development times, manufacturing and distribution constraints, and traceability can become deciding factors in what constitutes a best serology standard.

Ultimately the generality of Def. V permits multiple interpretations of harmonization, and it is plausible that other approaches may further reduce uncertainty. While we believe that our underlying approach is useful, our primary goal in this section and the previous is to highlight the *importance of rigorously defining and distinguishing the concepts of harmonization and normalization*. These definitions and their realizations (e.g. via mathematical models) play a fundamental metrological role in ensuring reproducibility of measurements and developing reference materials.

### C. On the Physical Interpretations of Gibbs Free-Energies and Consensus Values

Antibodies in serology samples are typically polyclonal, as evidenced by the fact that distinct SARS-CoV-2 antigens (for example) can be detected in the same blood [20, 24]. Thus, the reaction process described by Eq. (11) is a simplification of the true chemistry underlying serology assays. A more accurate representation would consider a collection of simultaneous reactions for each type of antibody with associated reaction kinetics. From a mathematical standpoint, however, this is problematic since one does not know *a priori* how many reactions to model. Moreover, it is reasonable to assume that one type of antibody (or perhaps a small subset thereof) dominates the chemistry of a single assay, which justifies Eq. (11).

This suggests a need to re-interpret  $\Delta g_{s,n}$ . In Sec. IV, this quantity represents the sample-assay specific contributions to the free-energy under the assumption that a single antibody interacts with the assay. For a human-derived sample, we must add at a minimum that the specific type of antibody interacting with the assay can vary with the latter. While this seems obvious – e.g. some SARS-CoV-2 assays distinguish anti-nucleocapsid from anti-spike antibodies – it weakens the concept of a consensus value. That is,  $\chi_{s,r}$  is not the total or even average concentration of antibodies of a specific type (e.g. anti-spike) in a sample. At best, we can say that it is a characteristic concentration conditioned on the number

and types of assays used to construct it. In the context of Fig. 3, for example, we might say that the consensus is the typical concentration of anti-SARS-CoV-2 antibodies across all types considered in Ref. [20].

It is important to note that these observations do not change the underlying structure of our analysis, only its interpretation. Likewise, these conclusions do not meaningfully change if we take the reference material to be a human-derived, pooled standard. In such cases, we must re-interpret  $\Delta g_{r,n}$  in the same way as we have done for  $\Delta g_{s,n}$ . For human-derived standards, it is also reasonable that the raw, normalized antibody concentrations might exhibit less variation relative to the consensus as compared to mAbs, since the reference may have a collection of antibodies that will respond to each assay as might a test sample. Indeed, this effect is evident in Ref. [20]. *However*, as Figs. 4 and 5 illustrate, this decrease in variance of the raw data does not impact the final uncertainty estimates of  $\Delta g_{s,n}$  or reference-induced uncertainty, which we find to be constant. Nor does it in any way change our harmonization method.

### D. Probabilistic Connection to Neutralization

Neutralizing assays (cf. Appendix A) are often seen as a better (yet still imperfect) measurement for assessing immunity against a pathogen. Such assays are also significantly more expensive than the binding assays considered in this work. This begs the question of the extent to which we can use binding assays as a proxy for neutralization assays.

To better understand this issue, observe that Eq. (27) yields a probability density for a consensus value given a sample measurement from a specific assay normalized to reference  $r$ . We can denote this function by  $P(\underline{y}_{s,r}|\bar{y}_{s,n,r}, r, n)$ , meaning that the probability of a specific  $\underline{y}_{s,r}$  is conditioned on the triple  $(\bar{y}_{s,n,r}, r, n)$ . If we likewise construct a probability density  $P(\nu|\underline{y})$  of a neutralizing value  $\nu$  (e.g. expressed as a 50 % neutralizing titer or NT50) conditioned on the consensus value, then the probability density of a neutralizing value given a scaled binding value is [41]

$$P(\nu|\bar{y}_{s,n,r}, r, n) = \int d\underline{y} P(\nu|\underline{y})P(\underline{y}|\bar{y}_{s,n,r}, r, n). \quad (35)$$

Equation (35) provides actionable information. Defining  $\nu_t$  as a lower neutralizing threshold that guarantees a degree of immunity, one can quantify the probability  $\mathcal{P}(\nu \geq \nu_t)$  that a person is protected by computing the integral

$$\mathcal{P}(\nu \geq \nu_t|\bar{y}_{s,n,r}, r, n) = \int_{\nu_t}^{\infty} d\nu P(\nu|\bar{y}_{s,n,r}, r, n) \quad (36)$$

One can then find the minimum measured binding level  $\bar{y}_{\min}$  that guarantees the corresponding  $\nu$  is above  $\nu_t$  with confidence  $\mathcal{P}(\nu \geq \nu_t) > 95\%$ , for example. In this case,

$\bar{y}_{\min}$  could be interpreted as the 95% “correlate level” for  $\nu_t$ .

This suggests that the following may be a useful definition for correlates of protection. Let  $X$  and  $Z$  denote the outputs of two distinct types of measurements (e.g. a binding and neutralizing assay), and let  $p$  be a percentage satisfying  $0 < p \leq 1$ . We say that  $X^*$  is the  $p$ -**correlate level** for  $Z^*$  if it is the smallest value such that  $X \geq X^*$  implies that the probability of  $Z \geq Z^*$  is greater than  $p$ . Mathematically,

$$X^* = \min_{\mathcal{X}} \{X : X \geq \mathcal{X} \implies \mathcal{P}(Z \geq Z^*) \geq p\}, \quad (37)$$

where  $\mathcal{P}(Z \geq Z^*)$  is the probability that  $Z \geq Z^*$ . While seemingly abstract, this definition enables statements of the form, “A binding level of at least  $X^*$  implies that a neutralization level is at least  $Z^*$  with a probability of 95% or greater.” Note that  $X^*$  is a function of  $Z^*$  and the probability  $p$ .

To realize this type of analysis in practice, however, it is advisable to quantify uncertainties associated with the parameters appearing in Eq. (27). In addition to the variance associated with  $\delta_{s,n,r}$ , which can be estimated from repeat measurements, one should also estimate the uncertainty in the  $\Delta g_{r,n}$  and  $\zeta_n^2$ . This latter task can be accomplished, e.g. via a jackknife type estimator as we have done.

### E. Additional Limitations and Extensions

The primary limitation of this work is the thermodynamic model inducing the probabilistic analysis. Where the underlying physical assumptions are violated, our analysis may not be valid. Examples are discussed in the previous section. Other assays that may invalidate our analysis are those in which antibody avidity plays an important role, since the chemical reactions may be dominated by more complicated binding interactions not described by the Gibbs free-energy of Eq. (13) [31, 47]. However, the invariance of the residuals  $\Delta g_{s,n} + \delta_{s,n}$  with respect to the reference provides a powerful tool that can be used to check the appropriateness of the assumptions, and thus our analysis.

Despite this limitation, our analyses provides several routes for generalization and/or incorporation of new physical information. For example, the normalization procedure discussed in Sec. IV A can be augmented with constraints to test for the degree of collapse among dilution curves or different assumptions about the structure of the underlying curve. In the event that the residuals are too large, for example, one could hold out such samples for further investigation. See Ref. [48] for related methods. The probabilistic modeling of the within-lab uncertainties  $\delta_s, n, r$  can also be estimated via more sophisticated techniques, e.g. bootstrap-type methods [36, 37]

Finally, we observe that probabilistic modeling of the individual terms  $\delta_{s,n}$  and  $\delta_{r,n}$  appearing in Eq. (28) may

be useful in applications that seek to estimate correlates of protection in terms of  $P(\underline{y}_{s,r} | \bar{y}_{s,n,r}, \Delta g_{r,n}^*, (\zeta_n^*)^2)$ . While not needed to validate the analysis herein, the separability of  $\delta_{s,n,r}$  implies that the  $\Delta g_{r,n}$  have additional uncertainty associated the contribution from  $\delta_{r,n}$ . Thus, it may be desirable to treat  $\Delta g_{r,n}^*$  as a random variable whose mean is given by our MLE analysis and distribution by the  $\delta_{r,n}$ . Propagating the latter uncertainty into  $P(\underline{y}_{s,r} | \bar{y}_{s,n,r}, \Delta g_{r,n}^*, (\zeta_n^*)^2)$  would then provide more realistic estimates of this distribution.

### F. Concluding Thoughts

The main objective of this work is to provide a theoretical foundation for tasks such as antibody normalization, harmonization, and estimating correlates of protection. As exercises in metrology, however, these tasks are challenging because their uncertainties are dominated by significant epistemic effects, e.g. lack of knowledge as to which antibodies are being detected, how they interact with the measurement system, and even what we mean by an antibody concentration. This is not a criticism so much as an observation: serology testing and immunity are difficult to understand due to complicated thermodynamic effects and inherent multiscale phenomena. Our approach has therefore been to identify those aspects that cannot be made precise and leverage UQ as a means to quantify our lack of understanding. This approach is enticing because it allows one to make informed decisions based on imperfect knowledge. It also suggests routes for optimizing – both informally and more mathematically – aspects of serology testing, as well as diagnostics in general. Looking forward then, our hope is that this work motivates a wider adoption of UQ within the biomedical community as a route to establishing rigorous principles of biometrology.

*Acknowledgements:* The authors thank Drs. Ronald Boisvert, Charles Romine, and especially Barry I. Schneider for helpful discussion during the preparation of this manuscript. This manuscript is an official contribution of the National Institute of Standards and Technology and is not subject to copyright in the United States.

*The NIST Research Protections Office has approved the use of data described herein.*

### Appendix A: Introduction to Serology

To understand our motivations for studying serology, it is useful to have basic knowledge of the properties of antibodies, the processes that create them, and the methods by which they are measured.

Antibodies are a key part of the *humoral immune system*, which is mediated by macromolecules in extracellular fluid (i.e. “humors” or body-fluids). Humoral immunity is itself part of a larger *adaptive immune response* associated with those biological phenomena that change

in response to pathogens in order to better fight infections. A key process of this adaptive immunity is so-called “affinity maturation” wherein certain T-cells force B-cells to undergo a process of hyper-evolution. As a result, the latter create antibodies that strongly bind to a specific chemical target (epitope), e.g. part of a virus. Physically, this process tailors the Gibbs-free energy of reaction between the antibody and target such that binding of the two is heavily favored over the reverse reaction. The B-cells themselves act as a sort of “memory” of the disease, and antibodies act as long-term (months or longer) defenses that disable their target antigen upon re-exposure. In this way, the body can more quickly respond to re-infection, and hopefully one experiences less symptoms, if any at all.

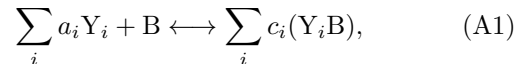
This overall picture explains why antibodies are objects of interest in diagnostic and public-health settings: they indicate when someone has been infected by a specific pathogen, especially after the disease has run its course. To a lesser extent, detecting antibodies also provides a degree of confidence that an individual has developed adaptive immunity to a disease, although such inferences are nuanced.

In fact, our description of humoral immunity is as oversimplified as it is intuitive. To highlight just a few relevant issues: (i) the human body produces not one, but at least five major types of antibodies, each with different structures and purposes; and (ii) any given antibody type will have a panoply of subtypes, each of which is specific to (i.e. binds with high affinity to) a different epitope, sometimes from the same pathogen. Antibody types are often called *immunoglobulins*, and convention dictates that they are denoted by the acronym “IgX” followed by the target epitope, where “X” stands for G, M, A, D, and E. IgG corresponds to the type best described by our picture above: they tend to have high-specificity and be long-lasting. IgM antibodies often appear within days to weeks of infection, are less specific, and do not last as long. IgA antibodies are often found in external fluid secretions and are often a relevant measurand when doing saliva testing. In the context of serology testing, the other two types of antibodies do not concern us.

Issues (i) and (ii) both lead to serious challenges when considering serology testing from a metrology standpoint, but to understand these, one must know how antibody measurements are performed. In an ideal biological setting, binding is described by a chemical reaction associated with Eq. (11). The goal of measuring antibody titers or “levels” is to estimate the concentration of bound complexes under limiting cases, i.e. when antibodies are the limiting reagent. Thus, the measurement process mirrors biology. In a *binding assay*, a blood or saliva sample is exposed to a *substrate*, i.e. a material containing copies of the target epitope of interest. After antibodies are bound, they are labeled with fluorescent tags, and the total fluorescence is measured as a proxy for number of bound complexes. *Neutralizing assays* are more complicated and involve a series of steps to determine what

concentration of a sample is sufficient to inhibit growth of a target pathogen. This is a more direct (albeit incomplete) measure of “immunity.”

In this context, the presence of many different antibody types complicates process of measuring the reaction associated with Eq. (11), especially when they target the same or similar antigens. In a binding assay, this leads to multiple reactions competing for the same epitope. Experimentalists have means of distinguishing signals from different antibody types, but the thermodynamics *is not* described by a simple reaction. One might hazard that the true chemical equation looks something like



where  $Y_i$  is the  $i$ th antibody type binding to substrate B,  $(Y_i B)$  is the corresponding bound complex, and the  $a_i$  and  $c_i$  are unknown coefficients describing the detailed balance associated with this system. The thermodynamics of this situation are full of unknowns, so that it is not entirely clear what one means by the concentration of a bound complex. We take a somewhat vague definition: the measurand is any bound complex reacting like a given antibody type (e.g. IgG) in a competitive binding environment. This may also be the most biologically relevant definition, since it corresponds to competitive reactions that happen *in vivo*. But then we cannot say that we are quantifying concentration of IgG antibodies, but rather only IgG-like molecules.

This problem is compounded by two other conventions in serology. First, it was normal during the SARS-CoV-2 pandemic to develop standards based on individual or pooled blood samples (i.e. drawn from multiple individuals and mixed together) having large but finite volumes. Second, common practice dictated defining concentration of antibodies for IgG, IgM, and IgA assays on the same scale relative to such standards. From the above discussion, it should be clear that this further confuses what we mean by concentration of antibodies. Pooled standards increase competitive binding, and combining multiple Ig types onto one scale means that we are abstracting to some generic notion of “concentration count.” On the surface this would seem to simplify the situation, since *number* can be understood easily in metrological terms. But as the main manuscript shows, this notion of “counting” is neither absolute nor even unambiguously relative. The issue comes back to the concept of Gibbs free energies, and the fact that only energy differences are meaningful.

For the purposes of this overview, our main takeaway is this: concentration measurements in serology use the language of counting in metrology, but fundamentally they are context-dependent. One does not measure the number of bound antibodies in isolation, but rather as impacted by effects such as competitive binding of other antibodies. A main goal of this manuscript is to show that the *assay* being used to perform the measurements is also a key part of that context.

To directly *compare* concentration measurements on a scale that permits quantitative comparison, it is necessary to account for this context and the uncertainty it induces. This leads to the concept of *harmonization*, i.e. the process of making measurements from different instruments interchangeable. In serology, this is often confused with *normalization*, i.e. the process of putting a measurement from a given instrument on a scale. This alone may be insufficient for harmonization if the context-dependent scales associated with each instrument are themselves different. In practice, this is in fact the case, since each assay contributes differently to a free-energy of reaction.

A proposed solution to some of these problems has been to use *monoclonal antibodies* (or mAbs) as reference materials. Monoclonals are synthetic (laboratory-made) antibodies whose properties can be tailored for different applications. A key feature of mAbs is the fact that they can be manufactured to have little, if any, variability between realizations of the same molecule. From a measurement standpoint this is desirable, since it eliminates the competitive binding problem for the reference material. Monoclonal antibodies also clarify interpretation of the measurements: harmonized results can be interpreted as quantifying the effective number of mAbs of a given type in a sample.

## Appendix B: Affine Transformations for Dilutions Series: Global Method for Antibody Normalization

In this section, we address the numerical issues associated with estimating the scaled antibody level  $\hat{y}_{s,n,r}$  in terms of the invariant quantity  $\alpha_s$  defined in Sec. IV A. The analysis herein generalizes the methods of Refs. [18, 29].

Without loss of generality, we assume that the measurement value  $F$  corresponds to MFI. Let  $x = cd$  denote the concentration of bound antibodies in a sample with concentration  $c$  and dilution factor  $d$ , and assume that there exists a range  $[x_{\min}, x_{\max}]$  over which  $F(x)$  is strictly monotone increasing function of  $x$ . The physical interpretation is straightforward: more antibodies yield more fluorescence. We supplement this with this assumption that  $F(x)$  approaches lower and upper limits  $F_{\min}$  and  $F_{\max}$  (which we can always take to be positive) as  $x \rightarrow 0$  or  $x \rightarrow \infty$ . Physically  $F_{\min}$  and  $F_{\max}$  can be interpreted as a noise-floor and detector saturation threshold. It is important to distinguish these sources of nonlinearity, which are instrument artifacts, from effects associated with nonlinear dependence of  $c_{s,n}$  on  $y_s$  as expressed by Eq. (12). We always assume that  $c$  (and thus  $x$ ) is linear in  $y$ , even when fluorescence  $F$  is nonlinear in  $c$ .

In practice, estimating the  $\alpha_s$  via data collapse is complicated by three issues.

First, measurements are often taken at a few serial dilutions  $d_i$  ( $i = 1, 2, \dots, D$ ) spanning several decades.

Thus  $F(x)$  tends to be given on a sparse grid whose characteristic spacing grows exponentially. To make the spacing more uniform, we take a logarithm of  $F$  and express the resulting function in terms of  $x = \ln(x)$ , which transforms the measurement domain to  $-\infty < x < \infty$ . This transformation also preserves strict monotonicity of  $f(x) = \ln(F(e^x)/U_F)$ . As an added benefit, we find that  $f(x)$  is typically sigmoidal. That is, for some inflection point  $x_I$ ,  $f(x)$  is convex (concave) when  $x \leq x_I$  ( $x \geq x_I$ ). While not strictly necessary, this assumption is so convenient that we leverage it throughout. Section VI proposes generalizations and limitations of this choice.

Second, the sparsity of the  $d_i$  means that  $f(x)$  is only known at a few points, which makes it challenging to determine whether two dilution series coincide. We address this problem by determining multiple  $\alpha_s$  simultaneously by requiring that they all fall on the same curve. Here the sigmoid structure of  $f(x)$  plays an important role by ensuring that this curve has a physically reasonable structure.

Third, we anticipate that the  $\alpha_s$  are to be determined by some numerical method that iteratively varies these parameters to find their optimal values. However, doing so makes the grid of  $x$  values dependent on the  $\alpha_s$ . This motivates us to treat the fluorescence values of each measurement as the independent variables, since these always define a fixed grid. By the strict monotonicity of  $f(x)$ , we may then write  $x = x(f)$  as a function of the fluorescence values, which effectively “flips” our perspective about the line  $x = f$ . Recalling that  $x = \ln(x) = \ln(cd)$ , we now see that the equality

$$\begin{aligned} x &= \ln(\hat{c}_{s,n,r}d) = \ln(c_{r,n}d/\alpha_s) \\ &= \ln(c_{r,n}d) - \ln(\alpha_s) \end{aligned} \quad (\text{B1})$$

reinterprets  $\ln(\alpha_s)$  as a constant vertical offset accounting for the difference between a reference and sample dilution series.

To realize these ideas mathematically, let  $f_{i,r}$  be the fluorescence measurements associated with the reference material at dilution  $d_i$ . Assume  $S$  samples indexed by  $s$  having unknown normalized concentrations  $\hat{c}_{s,n,r}$ . For each of these samples, we assume corresponding measurements  $f_{i,s}$  for dilutions  $d_i$ . In practice, the dilutions can be different for each sample, although for simplicity we assume the same set for each sample. Generalizations are trivial and left for the reader.

From this data, we create a single vector of elements  $f_j$  comprised of the  $f_{i,s}$  and  $f_{i,r}$  in ascending order and without regard to their sample number or status as a reference. It is not problematic if values of  $f_j$  are repeated. Also let  $\alpha_{s_j}$  and  $d_{i_j}$  denote the corresponding value of  $\alpha$  and  $d$  for the  $j$ th fluorescence value, where  $s_j$  can be a specific value of  $s$  or denote the reference  $r$ . To find the  $\alpha_{s_j}$ , we postulate the existence of true log-antibody numbers  $\hat{x}_j$ , which should be sufficiently close to the values predicted by the  $c_{r,n}d_{i_j}/\alpha_{s_j}$ . In fact, under noiseless

conditions, one expects

$$\hat{x}_j - \ln(c_{r,n}d_{i_j}) + \ln(\alpha_{s_j}) = 0. \quad (\text{B2})$$

In practice, there will be noise, which suggests the objective

$$\hat{\mathcal{L}} = \sum_j [\hat{x}_j - \ln(c_{r,n}d_{i_j}) + \ln(\alpha_{s_j})]^2. \quad (\text{B3})$$

Assuming a value for  $c_{r,n}$ , which we can take to be 1 for convenience, we minimize Eq. (B3) as a function of the  $\hat{x}_j$  and the  $\alpha_{s_j}$ , subject to the constraints that

$$\hat{x}_j = \hat{x}_{j'} \quad \text{if} \quad f_j = f_{j'} \quad (\text{B4})$$

$$\alpha_{s_j} = 1 \quad \text{if} \quad s_j = r. \quad (\text{B5})$$

Note that setting  $c_{r,n} = 1$  amounts to an arbitrary rescaling of the reference concentration, which can be undone by multiplying all concentrations by the appropriate units and scale factor at the end of all calculations; see Ref. [26].

By itself, Eq. (B3) does not define a well-posed optimization problem. For example, if the  $f_j$  all correspond to distinct samples, then any values of the  $\alpha_{s_j}$  will yield  $\hat{x}_j$  that yield  $\mathcal{L} = 0$ . *In such cases, we require that our analysis reduce to single-point normalizations based on Eq. (5a), which enforces theoretical consistency with past work, e.g. Refs. [18, 19].* We therefore add two constraints and a small regularization. First, letting  $j_k$  denote indices associated with unique values of  $f_j$  in ascending order, we require that

$$\hat{x}_{j_{k+1}} \geq \hat{x}_{j_k}. \quad (\text{B6})$$

That is, the antibody levels must be increasing with increasing fluorescence. Second, we require that  $x(f)$  be concave for  $f_{j_k}$  up to some inflection point  $f_I$  and convex for all  $f_{j_k} \geq f_I$ . This is equivalent to the sigmoid assumption transformed about the line  $x = f$ . To enforce this constraint, we construct a second-order, finite difference matrix  $A_{m,j_k}(p)$  in terms of an undetermined inflection index  $p$  using the procedure in the appendix of Ref. [49], where

$$\sum_{j_k} A_{m,j_k} x_{j_k} \leq 0 \quad m \leq p \quad (\text{B7a})$$

$$\sum_{j_k} A_{m,j_k} x_{j_k} \geq 0 \quad m > p. \quad (\text{B7b})$$

Third, we modify the objective function to be

$$\hat{\mathcal{L}} \rightarrow \mathcal{L} = \hat{\mathcal{L}} + \tilde{\epsilon}_1 \sum_{k=k_{\text{low}}}^{k_{\text{high}}} \left( \frac{x_{j_{k+1}} - x_{j_{k-1}}}{f_{j_{k+1}} - f_{j_{k-1}}} - 1 \right)^2 + \tilde{\epsilon}_2 \sum_{m=1}^{N_{j_k}} \left( \sum_{j_k} A_{m,j_k} x_{j_k} \right)^2 \quad (\text{B8})$$

where  $\tilde{\epsilon}_1$  and  $\tilde{\epsilon}_2$  are small regularization parameters, and  $k_{\text{low}}$ ,  $k_{\text{high}}$  are user-defined lower and upper limits between which we expect the fluorescence signals to be approximately linear with antibody number. Thus, the regularization term associated with  $\tilde{\epsilon}_1$  ensures that the reconstructed dilution curve has a linear region when there is only one measurement per sample. The regularization associated with  $\tilde{\epsilon}_2$  penalizes excessive curvature. These parameters are chosen to have values that are roughly three decades smaller than the characteristic value of  $\mathcal{L}$  near its minimum, or, if  $\mathcal{L} = 0$  is in the feasible set, we define  $\tilde{\epsilon}_1 = \tilde{\epsilon}_2 = 10^{-3}$ .

To determine the remaining parameters, we minimize Eq. (B8) with respect to the  $\hat{x}_j$ ,  $\alpha_{s_j}$ , and  $p$ , subject to Eqs. (B4)–(B5) and the inequality constraints (B6)–(B7b). It is straightforward to show that when the data is noiseless and the dilution curve is linear, the minimum of Eq. (B8) is unique and yields the true values of  $\alpha_s$  and  $x$ . Thus, our normalization procedure generalizes the techniques in Refs. [18, 19] and reduces to these approaches when only analyzing a single dilution associated with each sample and reference.

Figure 2 illustrates the results of this analysis applied to a collection of 38 SARS-CoV-2 positives samples and a mAb reference material, all measured using a ligand binding assay. For this analysis we set  $\tilde{\epsilon}_1 = \tilde{\epsilon}_2 = 10^{-3}$  and  $k_{\text{low}} = k_{\text{high}}$  to be the index associated with the measurement closest to the median fluorescence. By eye, it is clear that the raw dilution curves all have the same approximate shape (top subplot). After collapse, we find that with the exception of a few low-fluorescence data points, the characteristic deviation from the estimated dilution curve  $x$  is less than 5 %, which is well within characteristic uncertainties associated with pipetting and sample preparation. We speculate that the few data points showing significant deviation are exhibiting noise associated with being near the instrument noise floor. Note also that we do not need to specify either a linear range or functional form of the dilution curve. See Ref. [20] for more examples of this analysis applied to an interlab study with multiple distinct reference materials and assays.

As a cautionary remark, the low-fluorescence data of Fig. 2 reveals the challenges of dealing with data fully at the noise-floor or upper saturation threshold. In such cases, the amount of relevant physical information is dwarfed by instrument artifacts, which violates the assumptions underlying the optimization. Thus, while our analysis does not need a linear fluorescence region *per se*, it does require the signal-to-noise ratio of the data to be suitably large. To ensure this is the case, we remove from our analysis all data-points for which the signal-to-noise is less than roughly 1/10. While this choice is subjective, we find for the examples herein that it yields reasonable results.

- [1] J. Abbasi, *JAMA* **326**, 1781 (2021).
- [2] R. M. West, A. Kobokovich, N. Connell, and G. K. Gronvall, *mSphere* **6**, 10.1128/msphere.00201 (2021), <https://journals.asm.org/doi/pdf/10.1128/msphere.00201-21>.
- [3] L. Muller, J. Kannenberg, R. Biemann, M. Hönemann, G. Ackermann, and C. Jassoy, *Journal of Clinical Virology* **155**, 105269 (2022).
- [4] T. Perkmann, N. Perkmann-Nagele, T. Koller, P. Mucher, A. Radakovics, R. Marculescu, M. Wolzt, O. F. Wagner, C. J. Binder, and H. Haslacher, *Microbiology Spectrum* **9**, e00247 (2021).
- [5] D. Giavarina and M. Carta, *Diagnosis* **9**, 274 (2022).
- [6] M. Infantino, M. Pieri, M. Nuccetelli, V. Grossi, B. Lari, F. Tomassetti, G. Calugi, S. Pancani, M. Benucci, P. Casprini, M. Manfredi, and S. Bernardini, *International Immunopharmacology* **100**, 108095 (2021).
- [7] J. Prechl, *Biologia Futura* **72**, 37 (2021).
- [8] “Serological sciences network,” <https://www.cancer.gov/research/key-initiatives/covid-19/coronavirus-research-initiatives/serological-sciences-network>, accessed: 2022-10-11.
- [9] F. Krammer, *The Lancet* **397**, 1421 (2021).
- [10] S. Feng, D. J. Phillips, T. White, H. Sayal, P. K. Aley, S. Bibi, C. Dold, M. Fuskova, S. C. Gilbert, I. Hirsch, H. E. Humphries, B. Jepson, E. J. Kelly, E. Plested, K. Shoemaker, K. M. Thomas, J. Vekemans, T. L. Villafana, T. Lambe, A. J. Pollard, M. Voysey, S. Adlou, L. Allen, B. Angus, R. Anslow, M.-C. Asselin, N. Baker, P. Baker, T. Barlow, A. Beveridge, K. R. Bewley, P. Brown, E. Brunt, K. R. Buttigieg, S. Camara, S. Charlton, E. Chiplin, P. Cicconi, E. A. Clutterbuck, A. M. Collins, N. S. Coombes, S. A. C. Clemens, M. Davison, T. Demissie, T. Dinesh, A. D. Douglas, C. J. A. Duncan, K. R. W. Emary, K. J. Ewer, S. Felle, D. M. Ferreira, A. Finn, P. M. Folegatti, R. Fothergill, S. Fraser, H. Garland, L. Gatcombe, K. J. Godwin, A. L. Goodman, C. A. Green, B. Hallis, T. C. Hart, P. T. Heath, H. Hill, A. V. S. Hill, D. Jenkin, M. Kasanyinga, S. Kerridge, C. Knight, S. Leung, V. Libri, P. J. Lillie, S. Marinou, J. McGlashan, A. C. McGregor, L. McInroy, A. M. Minassian, Y. F. Mujadidi, E. J. Penn, C. J. Petropoulos, K. M. Pollock, P. C. Proud, S. Provstgaard-Morys, D. Rajapaska, M. N. Ramasamy, K. Sanders, I. Shaik, N. Singh, A. Smith, M. D. Snape, R. Song, S. Shrestha, R. K. Sutherland, E. C. Thomson, D. P. J. Turner, A. Webb-Bridges, T. Wrin, C. J. Williams, and the Oxford COVID Vaccine Trial Group, *Nature Medicine* **27**, 2032 (2021).
- [11] J. Perry, S. Osman, J. Wright, M. Richard-Greenblatt, S. A. Buchan, M. Sadarangani, and S. Bolotin, *PLOS ONE* **17**, 1 (2022).
- [12] J. Abbasi, *JAMA* **327**, 115 (2022).
- [13] A. B. Karger, J. D. Brien, J. M. Christen, S. Dhakal, T. J. Kemp, S. L. Klein, L. A. Pinto, L. Premkumar, J. D. Roback, R. A. Binder, K. W. Boehme, S. Boppana, C. Cordon-Cardo, J. M. Crawford, J. L. Daiss, A. P. Dupuis, A. M. Espino, A. Firpo-Betancourt, C. Forconi, J. C. Forrest, R. C. Girardin, D. A. Granger, S. W. Granger, N. S. Haddad, C. D. Heaney, D. T. Hunt, J. L. Kennedy, C. L. King, F. Krammer, K. Kruczyński, J. LaBaer, F. E.-H. Lee, W. T. Lee, S.-L. Liu, G. Lozanski, T. Lucas, D. R. Mendu, A. M. Moormann, V. Murugan, N. C. Okoye, P. Pantoja, A. F. Payne, J. Park, S. Pinninti, A. K. Pinto, N. Pisanic, J. Qiu, C. A. Sariol, V. Simon, L. Song, T. L. Steffen, E. T. Stone, L. M. Styer, M. S. Suthar, S. N. Thomas, B. Thyagarajan, A. Wajenberg, J. L. Yates, and K. Sobhani, *medRxiv* (2022), 10.1101/2022.02.27.22271399.
- [14] P. A. Kristiansen, M. Page, V. Bernasconi, G. Mattiuzzo, P. Dull, K. Makar, S. Plotkin, and I. Knezevic, *The Lancet* **397**, 1347 (2021).
- [15] I. Knezevic, G. Mattiuzzo, M. Page, P. Minor, E. Griffiths, M. Nuebling, and V. Moorthy, *The Lancet Microbe* **3**, e235 (2022).
- [16] L. Niu, K. N. Wittrock, G. C. Clabaugh, V. Srivastava, and M. W. Cho, *Frontiers in Immunology* **12** (2021), 10.3389/fimmu.2021.647934.
- [17] B. C. Braden and R. J. Poljak, in *Idiotypes in Medicine: Autoimmunity, Infection and Cancer*, edited by Y. Shoenfeld, R. C. Kennedy, and S. Ferrone (Elsevier Science B.V., Amsterdam, 1997) pp. 37–50.
- [18] A. Frey, J. Di Canzio, and D. Zurakowski, *Journal of Immunological Methods* **221**, 35 (1998).
- [19] R. W. Barrette, J. Urbonas, and L. K. Silbart, *Clinical and Vaccine Immunology* **13**, 802 (2006).
- [20] L. Wang, P. N. Patrone, A. J. Kearsley, and S. Lin-Gibson, Submitted (2023).
- [21] R. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*, Computational Science and Engineering (Society for Industrial and Applied Mathematics, 2013).
- [22] J. R. Tate, R. Johnson, and M. Legg, *Clin Biochem Rev* **33**, 81 (2012).
- [23] R. W. McLawhon, *Clinical Chemistry* **57**, 936 (2011), <https://academic.oup.com/clinchem/article-pdf/57/7/936/32655704/clinchem0936.pdf>.
- [24] T. J. Kemp, J. T. Quesinberry, J. Cherry, D. R. Lowy, and L. A. Pinto, *Journal of Clinical Microbiology* **60**, e00995 (2022).
- [25] E. M. Bentley, E. Atkinson, P. Rigsby, W. Elsley, V. Bernasconi, P. Kristiansen, H. Harvala, L. C. Turtle, S. Dobson, S. Wendel, R. Anderson, S. Kempster, J. Duran, D. Padley, N. Almond, N. J. Rose, M. Page, and G. Mattiuzzo, *Expert Committee on Biological Standardization*, 1 (2022).
- [26] L. Wang, P. N. Patrone, A. J. Kearsley, J. R. Izac, A. K. Gaigalas, J. C. Prostko, H. J. Kwon, W. Tang, M. Kosikova, H. Xie, L. Tian, E. B. Elsheikh, E. J. Kwee, T. Kemp, S. Jochum, N. Thornburg, L. C. McDonald, A. V. Gundlapalli, and S. Lin-Gibson, *International Journal of Molecular Sciences* **24** (2023), 10.3390/ijms242115705.
- [27] K. Nixon, S. Jindal, F. Parker, N. G. Reich, K. Ghobadi, E. C. Lee, S. Truelove, and L. Gardner, *The Lancet Digital Health* **4**, e738 (2022).
- [28] L. Dron, V. Kalatharan, A. Gupta, J. Haggstrom, N. Zariffa, A. D. Morris, P. Arora, and J. Park, *The Lancet Digital Health* **4**, e748 (2022).
- [29] U. Andreasson, A. Perret-Liaudet, L. J. C. van Waalwijk van Doorn, K. Blennow, D. Chiasserini, S. Engelborghs, T. Fladby, S. Genc, N. Kruse, H. B.

- Kuiperij, L. Kulic, P. Lewczuk, B. Mollenhauer, B. Mroczko, L. Parnetti, E. Vanmechelen, M. M. Verbeek, B. Winblad, H. Zetterberg, M. Koel-Simmelink, and C. E. Teunissen, *Frontiers in Neurology* **6** (2015), 10.3389/fneur.2015.00179.
- [30] T. Holland and H. Holland, *Journal of Microscopy* **214**, 1 (2004).
- [31] B. Božič, S. Čučnik, T. Kveder, and B. Rozman, in *Autoantibodies (Third Edition)*, edited by Y. Shoenfeld, P. L. Meroni, and M. E. Gershwin (Elsevier, San Diego, 2014) third edition ed., pp. 43–49.
- [32] R. Pathria, *Statistical Mechanics* (Elsevier Science, 2016).
- [33] H. P. Erickson and L. Corbin Goodman, *Biochemistry* (2022), 10.1021/acs.biochem.2c00291.
- [34] D. Lay, S. Lay, and J. McDonald, *Linear Algebra and Its Applications* (Pearson, 2016).
- [35] A. L. Rukhin, *Metrologia* **46**, 323 (2009).
- [36] R. STINE, *Sociological Methods & Research* **18**, 243 (1989).
- [37] M. R. Chernick, W. González-Manteiga, R. M. Crujeiras, and E. B. Barrios, “Bootstrap methods,” in *International Encyclopedia of Statistical Science*, edited by M. Lovric (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011) pp. 169–174.
- [38] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML, “Evaluation of measurement data — Guide to the expression of uncertainty in measurement,” Joint Committee for Guides in Metrology, JCGM 100:2008.
- [39] A. Clifford Cohen, *Truncated and censored samples* (CRC Press, 2016).
- [40] N. H. K. Tony, “Censoring methodology,” in *International Encyclopedia of Statistical Science*, edited by M. Lovric (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011) pp. 221–224.
- [41] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2005).
- [42] The Annals of Mathematical Statistics **29**, 614 (1958).
- [43] M. H. QUENOUILLE, *Biometrika* **43**, 353 (1956).
- [44] S. Portnoy, *Computational Statistics and Data Analysis* **72**, 273 (2014).
- [45] D. Haddad, F. Seifert, L. S. Chao, A. Possolo, D. B. Newell, J. R. Pratt, C. J. Williams, and S. Schlamming, *Metrologia* **54**, 633 (2017).
- [46] H. Fang, F. Bielsa, S. Li, A. Kiss, and M. Stock, *Metrologia* **57**, 045009 (2020).
- [47] B. Alberts, A. Johnson, J. Lewis, P. Walter, M. Raff, and K. Roberts, *Molecular Biology of the Cell 4th Edition: International Student Edition* (Routledge, 2002).
- [48] P. N. Patrone, E. L. Romsos, M. H. Cleveland, P. M. Vallone, and A. J. Kearsley, *Analytical and Bioanalytical Chemistry* **412**, 7977 (2020).
- [49] P. N. Patrone, G. Cooksey, and A. Kearsley, *Phys. Rev. Applied* **11**, 034025 (2019).