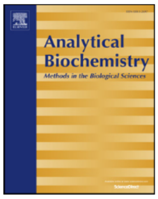
Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio



Analysis and uncertainty quantification of DNA fluorescence melt data: Applications of affine transformations

Paul N. Patrone*, Anthony J. Kearsley, Jacob M. Majikes, J. Alexander Liddle

National Institute of Standards and Technology, USA

ABSTRACT

Fluorescence-based measurements are a standard tool for characterizing the thermodynamic properties of DNA systems. Nonetheless, experimental melt data obtained from polymerase chain-reaction (PCR) machines (for example) often leads to signals that vary significantly between datasets. In many cases, this lack of reproducibility has led to difficulties in analyzing results and computing reasonable uncertainty estimates. To address this problem, we propose a data analysis procedure based on constrained, convex optimization of affine transformations, which can determine when and how melt curves collapse onto one another. A key aspect of this approach is its ability to provide a reproducible and more objective measure of whether a collection of datasets yields a consistent "universal" signal according to an appropriate model of the raw signals. Importantly, integrating this *validation* step into the analysis hardens the measurement protocol by allowing one to identify experimental conditions and/or modeling assumptions that may corrupt a measurement. Moreover, this robustness facilitates extraction of thermodynamic information at no additional cost in experimental time. We illustrate and test our approach on experiments of Förster resonance energy transfer (FRET) pairs used study the thermodynamics of DNA loops.

1. Introduction

Experimental studies of DNA systems must constantly ensure that measurements are reproducible. However, cost and complexity render this task challenging [1]. DNA preparation typically requires many expensive ingredients. Moreover, large samples are difficult to maintain at uniform temperatures, and, in the case of annealing studies, require infeasibly long heating cycles [1,2]. As a result, sample volumes are often limited to tens of microliters or less, so that even small pipetting and preparation errors yield signal-to-noise ratios that are smaller than one. In addition, these problems sometimes conflate with human error and sample contamination, making it difficult to identify the precise factors affecting data quality [1].

For fluorescence-based studies of DNA melting, these problems are so severe that scientists have tended to underutilize annealing data for quantitative measurements of thermodynamic parameters. Most analyses of melt curves only rely on identifying relative changes in their shape (e.g. to identify shifts in melt temperatures or detect for the existence of a mutation [3]) and do not leverage the absolute scale of data [4,5]. In some cases, data is discarded altogether [6]. Moreover, measurements based on the van't Hoff equation, which could be used to extract thermodynamic information about DNA, remain challenging at best and lack methods for estimating uncertainties [7]. Thus, the biology community would benefit from development of analyses that: (i) provide a greater understanding of the physical causes of variation in fluorescence data; (ii) separate measurement signals from background and noise; and (iii) enable uncertainty quantification (UQ) of experimentally determined quantities [8].

To address these problems, we propose a class of convex optimization techniques based on affine transformations that can account for and remove sources of variability in fluorescence data. The key idea behind this approach is to model observed measurement signals $\mathscr{S}$ as linear combinations of various background corrections $\mathscr{B}_n$ and a "universal" signal $\mathscr{U}$, which we wish to deduce. Importantly, the relative contribution of each source is encoded in a transformation parameter that can be used to express $\mathscr{U}$ in terms of $\mathscr{S}$. To determine these parameters, we minimize an objective function that compares independent realizations of $\mathscr{U}(\mathscr{S})$, which amounts to the requirement that $\mathscr{U}$ is independent of $\mathscr{S}$ (i.e. universal). Inequality and equality constraints are also imposed to test the feasibility (in a mathematical sense) of achieving the desired data collapse in a physically meaningful way. Applications to Förster resonance energy transfer (FRET) data confirm the validity of this approach and illustrate its benefits and limitations [9–11].[1]

A key insight motivating this work is the idea that the active physical processes are often the same in "identically" prepared systems, even if the raw data is not. Stated empirically, pipetting errors generally do not change the available mechanisms of fluorescence, and all else

* Corresponding author.
E-mail address: paul.patrone@nist.gov (P.N. Patrone).

[1] Sample data and scripts that execute our analyses are available upon request.

https://doi.org/10.1016/j.ab.2020.113773
Received 30 September 2019; Received in revised form 22 April 2020; Accepted 10 May 2020
Available online 08 June 2020
0003-2697/ Published by Elsevier Inc.

being equal, background signals alone can cause significant discrepancies between datasets due primarily to different amounts of impurities in well plates. These observations immediately suggest that *concentration-dependence* is frequently responsible for variation in realizations of $\mathscr{S}$ and motivates the generic model of the form

$$\mathscr{S}_i(T) = \hat{\tau}_{0,i}\mathscr{U}(T) + \sum_{n=1}^{N} \hat{\tau}_{n,i}\mathscr{B}_n(T) \tag{1}$$

where i indexes (experimental) datasets, T is the temperature and $\hat{\tau}_{n,i}$ are the unknown transformation coefficients associated with the ith dataset. Notably, these coefficients admit a simple interpretation as the concentration of the nth background source in sample i. We pursue further refinements of this model in later sections.

A key theme that permeates this work is the practice of integrating UQ into all steps of data analysis [12,13]. Here, we adopt a broad definition of UQ as comprising those tasks that assess the quality of raw data and characterize confidence in the predictions and information extracted from it [8]. As we show, this perspective is useful in the context of fluorescence experiments because it may be necessary to distinguish "corrupt datasets" (e.g. due to impurities) from those that are poorly scaled. Constrained optimization plays a critical role in this exercise as a tool to identify datasets that are inconsistent with data collapse, and thus candidates for rejection or further examination. Likewise, constraints can verify whether $\mathscr{U}$ has properties consistent with physical intuition. Such intermediate-stage assessments are important because they prevent us from continuing with downstream analyses and experiments when bad data might unknowingly distort results.[2] Moreover, we have found that this practice often leads to a more systematic accounting of experimental procedures, thereby creating opportunities to make them more robust. *As a result, thermodynamic properties of DNA can be extracted more reliably and with smaller uncertainties, often with little to no additional experimental overhead.*

The examples we consider also highlight a related aspect of our analysis: we often do not need to specify the functional forms of any of the terms in Eq. (1). Rather, we use this model to motivate a hierarchy of transformations applied directly to the experimental data. This fact is important, since it eliminates any issues associated with making poor choices about the actual form of $\mathscr{S}$, $\mathscr{U}$, and $\mathscr{B}_n$, aside from needing to satisfy Eq. (1). *In starker terms, knowing only that the data is described by an unspecified system of equations in the spirit of Eq. (1) may be sufficient to determine $\mathscr{U}$ and all of the $\mathscr{B}_n$ without further assumptions.* From the standpoint of uncertainty quantification (UQ), this surprising feature of the analysis is especially useful, since it reduces so-called *model-form* errors and simplifies downstream error estimates.

Despite these benefits, it is important to note that our analysis has subjective elements.[3] Equation (1) remains a model, and as such, it makes assumptions that may not always be valid, e.g. fluorescence is linear in fluorophore concentration. *Thus, any interpretation of the "true" signal $\mathscr{U}$ is exactly that, an interpretation that can only be understood in the context of these assumptions.* We highlight this point because in some cases, the remarkable degree of agreement between transformed datasets can mislead one into thinking that the corresponding $\mathscr{U}$ is "correct." Thus we cannot overemphasize: *data is not correct because it is visually pleasing.* In a related vein, we must guard against using constraints that force the data to look the way we want. While appropriate formulations of our analysis avoid this problem, their interpretations still require some consideration, especially relative to unconstrained optimization. Further discussion of these points is reserved for Sec. 5.

We also note that development of methods based on Eq. (1) is difficult to achieve *in vacuo*, since the goal is always to analyze experimental data about which we often have intuition or *a priori* information. Thus, our exposition follows the analysis of fluorescence data associated with FRET pairs that can detect the opening and closing of DNA loops [11]. We point out extensions and generalizations at appropriate places.

Finally we acknowledge that the main scientific content of this manuscript (i.e. mathematical theory and analysis) may at times be distinct from the core interests of our target audience (i.e. biologists and biophysicists). Thus, in an effort to make this work more accessible to readers, we have given each of the technical sections a theory-example structure. While tending towards more abstract, the former motivates many of the choices we make in the example sections. Taken collectively, the latter carry the reader through a full implementation of our transformation analysis applied to FRET experiments.

With this in mind, the rest of the manuscript is organized as follows. Section 2 discusses sources of fluorescence in DNA samples, with an eye towards motivating specific realizations of Eq. (1) for FRET data. Section 3 develops the key modeling equations that describe this data, including physics-based constraints. Section 4 presents the main optimization tools we use to compute the affine transformations on datasets $\mathscr{S}$ and explores the effects of using constraints. Section 5 discusses our approach in the context of other data analysis methods and presents our main conclusions. The Appendix provides an overview of the experimental procedure used to generate the data; see also Ref. [11].

2. Provenance and characteristics of experimental data

2.1. Typical experimental data collection

Polymerase chain-reaction (PCR) machines are a common tool for generating fluorescence data because this information can be used in Single Nucleotide Polymorphism (SNP) genotyping to identify single base mutations between alleles [14]. Of note, these techniques often use intercalating dyes rather than FRET pairs [15–17]. While the true universal signal should effectively be identical for either FRET or intercalating dye experiments, the sample preparation and data analysis procedures are likely to be distinct, owing to different baseline behavior in the fluorophores; see Sec. 5 for more discussion on these points. For concreteness, we restrict our attention to FRET-pair data.

A typical annealing protocol begins with sample preparation, which may be automated or done by hand [18,19]. The sample is then pipetted into wells on a plastic tray, which are typically arranged in a row-column format. Ideally, some subset of these wells contain N independent and identically prepared samples. This tray is loaded into a heating block molded to the shape of the former. A plastic lid covers this tray, and an additional heating block covers the top to prevent condensation. Importantly, the top heating block has a small aperture over each well so that optics can measure the fluorescence emitted from each well. Electronics and post-processing algorithms then output fluorescence counts per well per temperature increment into a data format accessible to a computer and/or end-user.

A typical dataset output by this process is composed of a vector of N_T temperatures and number of fluorescence counts per temperature. We denote this pair $(\mathbf{T}, \mathbf{S}_i)$, where $\mathbf{T} = (T_1, T_2, \ldots, T_{N_T})$ and $\mathbf{S}_i = (S_i(T_1), S_i(T_2), \ldots, S_i(T_{N_T}))$. For notational simplicity, we hereafter discontinue use of indices on T, recognizing that it is a discrete variable.

2.2. Example applied to FRET experiments

The FRET experiments that underpin our examples are designed to measure the fraction of DNA rings that are open or closed as a function of temperature. A system is composed of roughly 10^{12} DNA loops, each

[2] In the language of the broader UQ community, it is more conventional to say that our emphasis is on validation, i.e. the task of assessing the extent to which a model faithfully represents data [12]. The latter is always assumed to be "true." Given the technical challenges of generating high quality fluorescence data, however, we adopt a modified perspective acknowledging that the experiment may deviate from the intended "model" due to issues arising from contamination, spoiling of ingredients, etc.

[3] Our approach is not unique in this regard. Data analysis in general is a subjective endeavor.

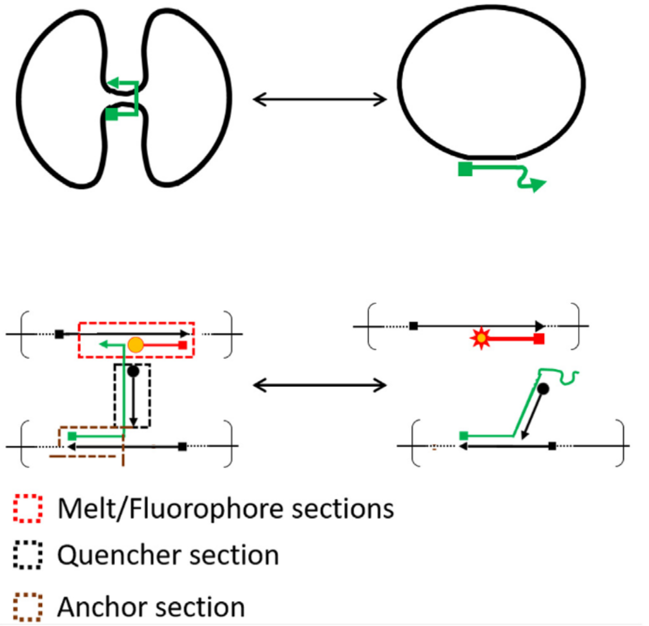


Fig. 1. Cartoon of the system used to generate the data in this manuscript. The system consists of the M13mp18 genome scaffold, which can be folded by the oligomer shown in green. The folding is reported by two strands labeled with a quencher/fluorophore pair. Top: the M13mp18 genome is folded into large loops of DNA with a "fold" oligomer (green) that binds to two locations: a high-melting-temperature "anchor" site and a low-melting-temperature "melt" site. The center section of this oligomer is hybridized with another oligomer (black) that carries a quencher. Hybridization of the melt site is energetically favored at low temperatures, whereas the unbound state is favored at high temperatures. Bottom: When the M13mp18 loop is folded, a fluorophore (yellow) attached to an oligomer (red) adjacent to the melt site is quenched by the quencher-functionalized oligomer (black). When the M13mp18 loop is open, the fluorophore is unquenched and can emit light (red star). Thus, the amount of emitted light should provide a quantitative measurement of the temperature response of ring opening for an ensemble of these systems. Different fold positions are achieved by varying the anchor sequence. The melt, fluorophore, and quencher oligomer sequences are conserved to ensure that the quencher/fluorophore pair experiences the same local nucleotide environment, and thus has the same functional form for its temperature-dependent fluorescence rate. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

having an attached oligomer that can close the ring by binding to an interaction site located at a predetermined second location. When this oligomer closes the ring, a FRET quencher/fluorophore pair is brought together so that a fluorescent marker at the interaction site is quenched (i.e. it stops emitting light). When the ring opens, the quencher is removed from the local environment of the fluorophore, which becomes active again. Thus, the total fluorescence F coming from the many copies should be proportional to the total number of open rings. See Fig. 1.

For these systems, it is well known that binding of the fold oligomer to the melt site on the DNA has a lower energy than the open state. Thus, at low enough temperatures (i.e. below 300 K), the rings should predominantly be folded, so that the fluorescence signal goes to zero. Conversely, at high enough temperatures, there is sufficient thermal energy in the system to open essentially all of the rings, so that $F(T)$ should achieve its maximum value. In both extremes, the slope of $F(T)$ should also vanish, with a monotone increasing function connecting the two temperature regimes. This function should attain its midpoint (corresponding to the melt temperature) between 310 K and 320 K. See the top plot of Fig. 2.

Actual measurements of the $F(T)$ relationship follows the procedure described in the previous section. These measurements are complicated by the fluorescence efficiency dependence of fluorophores on their local environment, including temperature, neighboring DNA bases, and pH [18]. As such, the fluorescence-temperature dependence was separately measured in samples containing all but the fold oligomer, which is used to bring the fluorophore and quencher in proximity to one another (cf. the Appendix). Because of the difficulty of measuring the temperature-response of a fully quenched (i.e. essentially dark) fluorophore, we assumed that unquenched temperature response was an appropriate approximation thereof. However, we note that neglecting the quenched temperature response may not be appropriate for all systems. In this case, Eq. (1) would require modification and corresponding experiments to account for such effects. See also Sec. 5.

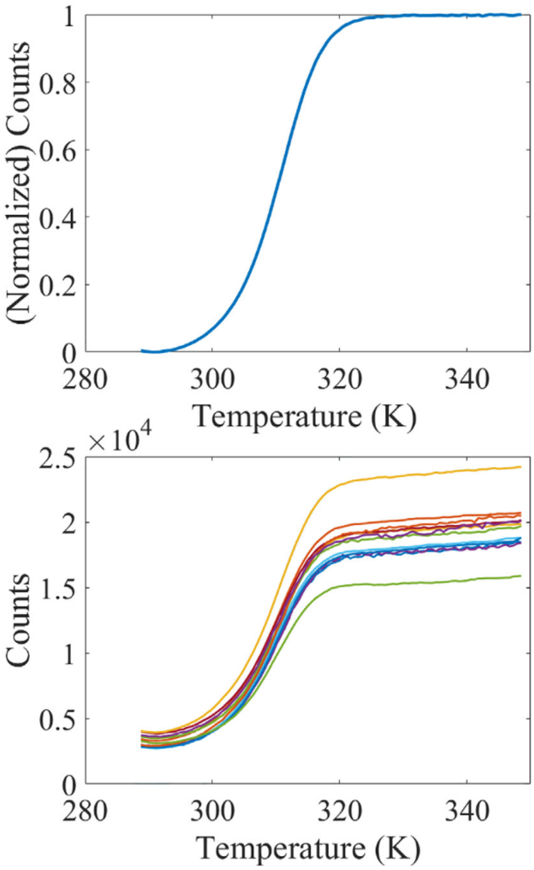


Fig. 2. Idealized and characteristic fluorescence data from FRET experiments (loop with 7450 bases and no anchor on the fold oligomer). *Top:* An idealized and normalized fluorescence curve. Note in particular the (near) zero slopes at low and high temperatures. *Bottom:* Raw data extracted from a PCR machine. Note that the experimental data exhibits several features not present in the idealized curve, such as non-zero slope at high temperature and large spread between datasets.

The bottom plot of Fig. 2 illustrates raw experimental data collected from a PCR machine. In contrast to the top plot, this data exhibits several features not present in the idealized fluorescence curve. In particular, the minimum fluorescence value is offset from zero and the slopes of each curve do not approach zero at high temperatures. Moveover, it is obvious that there is $O(1)$ variation between datasets. Our main goal in the rest of this manuscript is to reconcile the differences between these two plots.

2.3. *Sources of fluorescence in DNA samples*

With this overall picture in mind, there are several sources of fluorescence and related effects that inform a physical interpretation of the raw data.

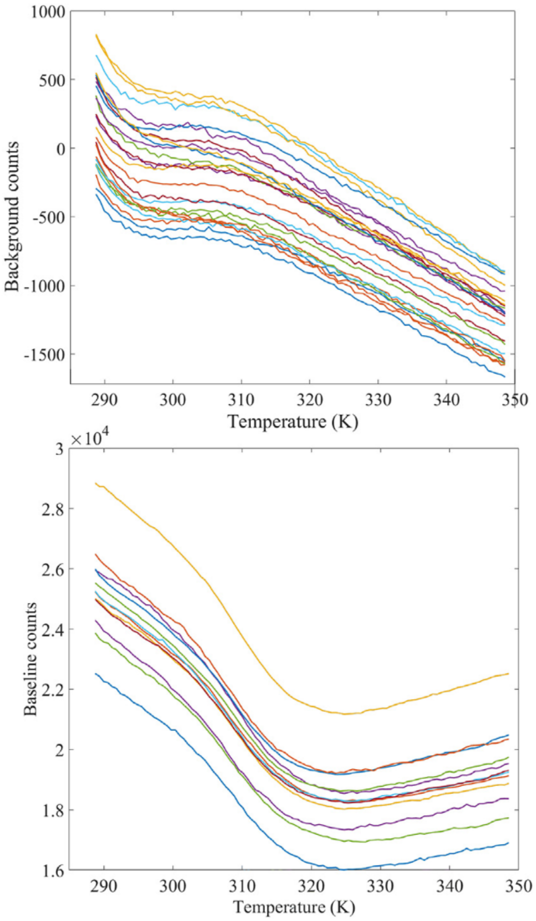


Fig. 3. Examples of background signals and temperature-dependence of fluorescence. *Top*: Fluorescence signals from empty wells in a PCR machine. The negative counts are physically meaningless and appear to arise from a vertical offset in the data. Note also that the variation in scale (roughly 2000 counts) is on the order of 10% of the full scale in Fig. 2. *Bottom*: Temperature dependence of the fluorescence rate for fluorophore-functionalized oligomers in the absence of quenchers. The top and bottom curves *cannot* be collapsed by simply translating the latter upwards, since they vary by 8000 and 6000 counts, respectively. Note also that the temperature dependence of the fluorescence rate could plausibly explain the non-zero slope of the raw data at high temperatures in the bottom plot of Fig. 2.

Background Signals: Measuring the fluorescence of empty trays often yields small but non-negligible and temperature-dependent signals; see Fig. 3. Physically, we attribute this background reading to several sources. In particular, it is well known that many polymers (including plastics in well-plate trays) exhibits some degree of autofluorescence [20]. Automated calibration routines on empty trays attempt to negate these effects, but such methods may not account for temperature-dependence of the background if only run at a single temperature. Moreover, calibration cannot account for natural variation in autofluorescence of different trays.

It is also important to consider the basic operating principles of the photodetectors that collect the fluorescent light. Invariably these devices convert optical power into either current or (analog) voltage. Amplifier circuits may be required to boost the raw signals to amplitudes that are readable by a computer. In each of these situations, semiconductor components exhibit dark currents or voltages that effectively *offset the raw signal by some constant but unknown amount, which may differ by dataset; see* Fig. 3. In certain cases, we have found that these and related offsets may be as much as 10% of the nominal fluorescence signal, thereby limiting reproducibility of measurements.

Temperature-dependent fluorescence rates: In addition to probing DNA topology (for example), fluorophores also exhibit their own systematic responses to generic thermodynamic variables such as temperature; see Sec. 2.2. Generally speaking, the latter behavior is problematic because its effects must be decoupled from the measurement signal before meaningful information can be extracted. Fig. 3 also shows an example of the baseline temperature response of fluorophore functionalized oligomers without quenchers. Notably, the signal is nonlinear and changes by 25% or more. The raw datasets exhibit different vertical offsets (similar to the background data), suggesting that it may difficult to accurately identify the relative change in fluorescence due to temperature effects alone.

Non-identical Samples: Robust experimental protocols often call for repeating experiments on independent and identically prepared samples. In practice, however, it is difficult to ensure that they are all identical to within detection thresholds, especially for typical DNA experiments. Even when samples are pulled from the same feedstock, pipetting errors can lead to variations in the volume of liquid in each well. Thus, the characteristic scale of the data, which depends on the total number of fluorophores in a sample, will fluctuate accordingly.

3. Key modeling equations

3.1. General case

In measurements designed to extract thermodynamic information from DNA, the previously discussed sources of variation never appear alone. Moreover, they may interact in such a way that they need to be removed simultaneously. As a first step to achieving this, we invert Eq. (1) for $\mathscr{U}$, which yields

$$\mathscr{U}(T) = \tau_{0,i}\mathscr{S}_i(T) - \sum_{n=1}^{N} \tau_{n,i}\mathscr{B}_n(T) \tag{2}$$

where $\tau_{0,i} = 1/\hat{\tau}_{0,i}$ and $\tau_{n,i} = \hat{\tau}_{n,i}/\hat{\tau}_{0,i}$. Even if the $\tau_{n,i}$ were known, Eq. (2) only provides a single relationship between $\mathscr{S}_i$ and the unknown quantities $\mathscr{U}$ and $\mathscr{B}_n$. However, it is often possible (and in fact, necessary) to design experiments that determine various background signals. Thus, we temporarily assume that the $\mathscr{B}_n(T)$ are known on some grid of temperatures $\mathbf{T}$. In the following example, we show how extensions of Eq. (2) can be used to deduce the various $\mathscr{B}_n(T)$ for FRET data. For later convenience, we also define the vector of transformations parameters as $\boldsymbol{\tau}_i = (\tau_{0,i}, \tau_{1,i}, ..., \tau_{N,i})$ and $\mathbf{B} = (\mathscr{B}_1, ..., \mathscr{B}_n)$ as the vector of backgrounds.

Equation (2) merits several comments. In particular, the left-hand side (LHS) is a realization independent quantity; that is, it does not depend on i. This implies that the realization-dependence of the right-hand side (RHS) has been eliminated. If we interpret $\mathscr{U}(T)$ as the "true" or universal fluorescence signal associated with an idealized experiment, Eq. (2) is a prescription for subtracting all sources of noise and variation in the experimental data. We further recognize that the RHS of Eq. (2) can be viewed as a mapping $M(\boldsymbol{\tau}_i)$ acting on $(\mathscr{S}_i, \mathbf{B})$ with the property that $M(\boldsymbol{\tau}_i; S_i, \mathbf{B}) = \mathscr{U}$ for the correct transformation parameters. *It is also important to note that for arbitrary* $\boldsymbol{\tau}$ *and constants* x, w, *the mapping M has a bilinearity property that*

$$\begin{aligned} M(\boldsymbol{\tau}; x\mathscr{S}_i + w\mathscr{S}_j, x\mathbf{B}_i + w\mathbf{B}_j) \\ = xM(\boldsymbol{\tau}; \mathscr{S}_i, \mathbf{B}_i) + wM(\boldsymbol{\tau}; \mathscr{S}_j, \mathbf{B}_j), \end{aligned} \tag{3}$$

$$M(x\boldsymbol{\tau} + w\boldsymbol{\tau}'; \mathscr{S}, \mathbf{B}) = xM(\boldsymbol{\tau}; \mathscr{S}, \mathbf{B}) + wM(\boldsymbol{\tau}'; \mathscr{S}, \mathbf{B}). \tag{4}$$

This observation plays a critical role in our constrained optimization formulation and interpretation of subsequent results.

3.2. Example applied to FRET data

To make the interpretation of Eq. (2) more concrete, we postulate a corresponding model for FRET data informed by the discussion in Sec. 2. In particular, we assume

$$f_i(T) = \hat{a}_i F(T)R(T) + \hat{b}_i B(T) + \hat{c}_i, \tag{5}$$

where f_i is the fluorescence signal associated with the ith identically prepared sample, $F(T)$ is the "true" fluorescence signal associated with the intended number of fluorophores (i.e. sample volume) assuming no pipetting errors, $R(T)$ is the temperature-dependence of the fluorescence rate, and $B(T)$ is the temperature-dependent background signal. We may substitute $\mathscr{U}(T) = F(T)R(T)$ in Eq. (5), rendering it consistent with Eq. (2). Note also that F, R, and B are to be determined. The sample-dependent, unknown coefficients $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ respectively characterize: (i) the degree to which the number of fluorophores deviates from its nominal value; (ii) the level of background as determined by the concentration of autofluorescing compounds in the ith well; and (iii) any unknown constant offset associated with effects such as bias voltages in the photodetector.

Provided $R(T) \neq 0$ we may invert Eq. (5) to find

$$F(T) = a_i \frac{f_i(T)}{R(T)} + b_i \frac{B(T)}{R(T)} + \frac{c_i}{R(T)}, \tag{6}$$

where a_i, b_i, and c_i are functions of their counterparts in Eq. (5). While not strictly necessary to characterize the sources of fluorescence variation, we supplement Eq. (6) with additional information about the anticipated behavior of $F(T)$. Such considerations are system specific and play an important role in the optimization and UQ steps that follow. In particular, we recall that at low and high temperatures, FRET pairs should be fully quenched and unbound, respectively. Thus, it is reasonable to assume that

$$\lim_{T\to 0} \frac{\mathrm{d}F}{\mathrm{d}T} = \lim_{T\to\infty} \frac{\mathrm{d}F}{\mathrm{d}T} = 0. \tag{7}$$

Moreover, for the purposes of extracting thermodynamic information from van't Hoff plots, it will eventually be necessary to scale the data to [0,1], which we encode via

$$\lim_{T\to 0} F(T) = 0 \tag{8a}$$

$$\lim_{T\to\infty} F(T) = 1. \tag{8b}$$

We consider further refinements of Eqs. (8a) and (8b) in Sec. 4 in the context of constrained optimization.

By itself, Eq. (6) does not offer a route for simultaneously determining F, R, and B, since R is coupled to F and the coefficients (a_i, b_i, c_i) are all unknown. To address this problem, we assume it is possible to measure the fluorescence of fluorophores without the quenchers,[4] which we model via the equation

$$\begin{aligned} r_i(T) &= \hat{\alpha}_i R(T) + \hat{\beta}_i B(T) + \hat{\kappa}_i \\ \Rightarrow R(T) &= \alpha_i r_i(T) + \beta_i B(T) + \kappa_i. \end{aligned} \tag{9}$$

See Ref. 11 and the Appendix for more details of these experiments. See also the bottom plot of Fig. 3. By analogy with Eq. (6), $r_i(T)$ is a specific realization of the fluorescence signal, $R(T)$ is a universal fluorescence signal associated with the intended number of fluorophores in each well, and the triple $(\alpha_i, \beta_i, \kappa_i)$ are unknown coefficients associated with deviations from the nominal sample size, magnitude of background effects, and constant offsets in measurements of functionalized oligomers without quenchers. The remaining unknown function $B(T)$ can likewise be determined by measuring the fluorescence of empty wells,[5] which we model with

$$\begin{aligned} \mathfrak{b}_i(T) &= \hat{\mathfrak{a}}_i B(T) + \hat{\mathfrak{c}}_i \\ \Rightarrow B(T) &= \mathfrak{a}_i \mathfrak{b}_i(T) + \mathfrak{c}_i. \end{aligned} \tag{10}$$

Again, $\mathfrak{b}_i(T)$ is the ith realization of a background signal and $B(T)$ is a universal background associated with a fixed concentration of autofluorescing compounds in the sample wells. See the top plot of Fig. 3.

4. Optimization

4.1. Formulating the objective function

4.1.1. General case

If we assume N_D total datasets indexed by i, then our method for determining the set of unknown coefficients $\{\tau\} = \{\tau_1, \ldots, \tau_D\}$ is motivated by the observation that $\mathscr{U}(T)$ is realization independent. That is, Eq. (2) implies

$$\mathscr{U}(T) = \mathscr{U}_i = M(\tau_i; \mathscr{S}_i, \mathbf{B}) = M(\tau_j; \mathscr{S}_j, \mathbf{B}) = \mathscr{U}_j, \tag{11}$$

for all pairs (i, j), where $\mathscr{U}_i$ is the ith realization of the universal signal. This obviously implies that $\mathscr{U}_i(T) - \mathscr{U}_j(T) = 0$, and moreover, for N_D datasets, the sum of the $\binom{N_D}{2}$ differences squared is zero. In other words, the non-negative objective function

$$\mathscr{L}_{\mathscr{U}}(\{\tilde{\tau}\}) = \sum_{i,j,T} [M(\tilde{\tau}_i; \mathscr{S}_i, \mathbf{B}) - M(\tilde{\tau}_j; \mathscr{S}_j, \mathbf{B})]^2 \geq 0 \tag{12}$$

should activate the inequality (i.e. $\mathscr{L}_{\mathscr{U}} = 0$) for the correct set of transformation parameters $\{\tau\}$. For real data, transformed signals may still contain small amounts of noise that do not entirely cancel, ensuring an inactive inequality $\mathscr{L}_{\mathscr{U}} > 0$. However, it is nonetheless reasonable to estimate the coefficients τ_i by minimizing $\mathscr{L}$. That is,

$$\{\tau\} = \underset{\{\tilde{\tau}\}}{\operatorname{argmin}}\ \mathscr{L}_{\mathscr{U}}(\{\tilde{\tau}\}) \tag{13}$$

are the affine parameters that yield $\mathscr{U}$ from the realizations of $\mathscr{S}$. Physically, Eqs. (12) and (13) define the optimal transformation coefficients as those that yield the smallest differences squared when summed over all pairs of datasets at all temperatures.[6]

4.1.2. Example applied to FRET data

It is useful to see realizations of Eq. (12) applied to the model equations for FRET data. If we denote $\mathscr{L}_\star$ as the objective corresponding to the universal functions $\star = B$, R, or F, we find

$$\mathscr{L}_B(\{\mathfrak{p}\}) = \sum_{i,j,T} [\mathfrak{a}_i \mathfrak{b}_i(T) + \mathfrak{c}_i - \mathfrak{a}_j \mathfrak{b}_j(T) - \mathfrak{c}_j]^2 \tag{14}$$

$$\begin{aligned} \mathscr{L}_R(\{\pi\}) = \textstyle\sum_{i,j,T} \Big[& \alpha_i r_i(T) + \beta_i B(T) + \kappa_i \\ & - \alpha_j r_j(T) - \beta_j B(T) - \kappa_j]^2 \end{aligned} \tag{15}$$

$$\begin{aligned} \mathscr{L}_F(\{p\}) = \textstyle\sum_{i,j,T} R(T)^{-2} \Big[& a_i f_i(T) + b_i B(T) + c_i \\ & - a_j f_j(T) + b_j B(T) + c_j]^2, \end{aligned} \tag{16}$$

[4] The corresponding measurements are performed on samples having all of the reactants except for the fold oligomer, which contains the quencher; see Fig. 1. Thus, the characterization of $R(T)$ aims to measure the temperature-dependence of the fluorophores as controlled by their local, unquenched environment.

[5] It may be more appropriate to determine $B(T)$ from a mixture of all reagents, omitting only the fluorophore. In our specific examples, however, the

(footnote continued)

restrictive set of admissible affine transformations and agreement between datasets after subtracting off background effects suggests that $B(T)$ was sufficiently well measured. Note that Eqs. (9) and (10) likely ignore the effects of impurities on $B(T)$ and $R(T)$, which might contribute additional multiplicative terms.

[6] Other norms may provide equally valid transformation coefficients. For example, replacing $[M(\tilde{\tau}_i; \mathscr{S}_i, \mathbf{B}) - M(\tilde{\tau}_j; \mathscr{S}_j, \mathbf{B})]^2$ with $|M(\tilde{\tau}_i; \mathscr{S}_i, \mathbf{B}) - M(\tilde{\tau}_j; \mathscr{S}_j, \mathbf{B})|^p$ (i.e. the L^p norm) would likely yield reasonable (but slightly different) transformations. We find L^2 to be a convenient choice since it is well known and seems suited for our examples. See also Sec.V.

where we denote the corresponding transformation parameters $\mathfrak{p}_i = (\mathfrak{a}_i, \mathfrak{b}_i)$, $\pi_i = (\alpha_i, \beta_i, \kappa_i)$, and $p_i = (a_i, b_i, c_i)$.

Note that Eqs. (14)–(16) suggests a hierarchical framework within which we can determine $F(T)$. Specifically, we first minimize $\mathscr{L}_B$, which yields an estimate of $B(T)$ at the vector of temperatures $\mathbf{T}$ in terms of the mean values B_i; viz

$$B(T) \approx \frac{1}{N_D}\sum_{i=1}^{N_D} B_i(T) = \frac{1}{N_D}\sum_i \mathfrak{a}_i b_i(T) + \mathfrak{c}_i \tag{17}$$

for the optimal $\{\mathfrak{p}\}$. We may then use this function to minimize $\mathscr{L}_R$ and thereby estimate $R(T)$ in terms of the mean of $R_i(T)$. With these quantities in hand, $\mathscr{L}_F$ can then be minimized to estimate $F(T)$. An important benefit of this approach is that we need not identify a functional form for the intermediate quantities $B(T)$ and $R(T)$, since they can be determined point-wise from the data directly.

4.2. Regularization and sufficient constraints

4.2.1. General case

Equation (13) does not lead to a unique set of coefficients. For example, the objective $\mathscr{L}_B$ given in Eq. (14) is equal to zero if we set $\mathfrak{a}_i = 0$ for all i and let $\mathfrak{c}_i = \mathfrak{c}_j$ for all pairs (i, j). That is, transforming all of the data to an arbitrary constant yields a meaningless, degenerate solution for $B(T)$. Equally problematic, for non-zero $\mathfrak{a}_i$, minimization of $\mathscr{L}_B$ only fixes the relative differences $\mathfrak{c}_i - \mathfrak{c}_j$, so individual $\mathfrak{c}_i$ are only determined up to an additive constant. Conceptually, this issue of uniqueness conflates two problems that must be understood (if not addressed) separately.

First, Eqs. (12) and (13) only define the universal signals on a relative scale. The coefficients $\{\tau\}$ remain unchanged if we apply the same affine transformation to all mappings $M(\tau_i; \mathscr{S}_i, \mathbf{B})$, i.e. if we substitute

$$aM(\tau_i; \mathscr{S}_i, \mathbf{B}) + ab = M(\tau_i; a\mathscr{S}_i, a\tilde{\mathbf{B}}) \tag{18}$$

into $\mathscr{L}_{\mathscr{U}}$, where a and b are arbitrary constants and $\tilde{\mathbf{B}}$ is the same as $\mathbf{B}$ with an added constant offset source of noise proportional to b. In this case, one finds that $\mathscr{L}_{\mathscr{U}} \to a\mathscr{L}_{\mathscr{U}}$, with the constant b canceling. As the location of a minimum of a function is insensitive to the overall scale of the function, $\{\tau\}$ is therefore unchanged. This yields the somewhat ironic conclusion that optimization of affine-parameters can only be accomplished up to an affine transformation of the form given by Eq. (18). To address this problem, we can add to the objective function a regularization term

$$\varepsilon\mathscr{L}_\tau = \varepsilon\sum_{i,n}\tau_{n,i}^2 \tag{19}$$

where ε is a small parameter. Although discussed later in more detail, we typically take $\varepsilon = 10^{-3}\delta$, where δ is a characteristic scale of the data. Equation (19) has the effect of penalizing large values of the transformation coefficients, thereby forcing the optimization algorithm to converge to the smallest admissible transformation coefficients that yield agreement between signals.

A second problem arises from the fact that Eqs. (6), (9) and (10) assume contributions from one or more realization independent sources, i.e. a background signal and/or constant offset. Thus, one always minimizes $\mathscr{L}_{\mathscr{S}}$ by employing transformations that eliminate any contribution from S_i (i.e. by setting $\tau_{0,i} = 0$) and project the signal entirely onto background. The addition of a regularization term without additional constraints will yield the unique but trivial solution in which all transformation parameters are zero. Invariably, we recognize this as a failure to impose a requirement that the signal $\mathscr{S}_i$ have a characteristic scale that is $\mathcal{O}(1)$ relative to $\mathscr{U}$; i.e., the former must be informative of the latter if the data analysis is to be meaningful. To achieve this, we impose the constraint

$$\tau_{0,i} \geq \tau_{\min}, \tag{20}$$

where $\tau_{0,i}$ is always the coefficient multiplying $\mathscr{S}_i(T)$ in Eq. (2) and $\tau_{\min} > 0$ is a user-defined, positive constant.

Several comments are in order. Firstly, the combination of regularization and a scale constraint are sufficient conditions to yield a unique and non-trivial set of affine parameters. *Thus, Eqs. (13), (19), and inequality (20) comprise the "simplest" formulation of the optimization problem that remains well-posed. However, we are not guaranteed that the corresponding transformation parameters are physically meaningful!* Moreover, Eqs. (19) and (20) are not necessary conditions insofar as other constraints can yield unique (but possibly different) transformation parameters. Thus, our simple formulation of the optimization amounts to a modeling choice that may require scrutiny in special circumstances. We consider such issues in more detail in Sec. 5.

4.2.2. Example applied to FRET data

Here we illustrate the outcomes of applying our simple formulation to the hierarchy of Eqs. (14)–(16) for FRET data. We take $N_D = 24$ datasets for background and $N_d = 12$ for all other experiments, which yields 276 and 66 distinct terms in the corresponding objective functions (excluding regularization). An important preliminary step is to normalize the datasets r_i, b_i, and f_i according to

$$\mathscr{S}_i \to \frac{\mathscr{S}_i}{\max_{j,T}[|\mathscr{S}_j(T)|]}, \tag{21}$$

where $\mathscr{S}$ stands for r, b, or f. This has the benefit of ensuring that all signals (background, rate-dependence, FRET fluorescence) are normalized to be $\mathcal{O}(1)$, so that the ratios b_i/a_i and c_i/a_i are rough estimates of the noise-to-signal ratios. We can then directly compare transformation parameters as a means for deciding whether the optimization has over-corrected for noise. We also set $\tau_{\min} = 1$ for all optimizations, since this requires that the universal signals be $\mathcal{O}(1)$. Moreover, having normalized all of the data implies that the characteristic scale of the data is unity, so that we may pick a small regularization parameter $\varepsilon \ll 1$. For the analyses in this section, we choose $\varepsilon = 10^{-4}$.

Fig. 4 shows the results of applying this analysis to the 24 background datasets in Fig. 3 by solving the optimization problem

$$\{\mathfrak{p}\} = \underset{\{\tilde{\mathfrak{p}}\}}{\operatorname{argmin}}\left[\mathscr{L}_B(\{\tilde{\mathfrak{p}}\}) + \varepsilon\sum_i(\mathfrak{a}_i^2 + \mathfrak{c}_i^2)\right] \tag{22a}$$

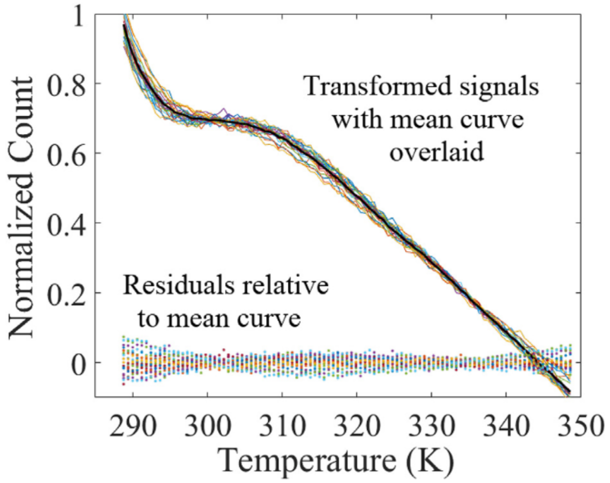


Fig. 4. Transformed background signals (color) with mean curve (black) overlaid on top. The dots around zero are residuals of each curve relative to the mean. Note that for the raw data in Fig. 3, the background signal is roughly 10% of $F(T)$, while the residuals are roughly 10% of the background signal. Thus, the uncertainty induced in using the mean background to correct $f_i(T)$ should be 1% or less. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

$$\mathfrak{a}_i \geq 1. \tag{22b}$$

We find that the simple bound on $\mathfrak{a}_i$ is active for three of the transformations parameters, i.e. $\mathfrak{a}_i = 1$ for only three indices i, with all other $\mathfrak{a}_j > 1$. Having estimated the $\mathfrak{p}_i$, we can now use Eqs. (10) and (17) to estimate the function $B(T)$ pointwise in temperature for use in computing $R(T)$ and $F(T)$.[7] Figs. 5 and 6 show results of solving the corresponding problems for $R(T)$ and $F(T)$. As before, the average of the realizations $R_i(T)$ are used in the model for F.

The bottom-left plot of Fig. 7 shows the results of minimizing the hierarchy of equations for the datasets shown in the top-left plot of the figure (minimization for $B(T)$ and $R(T)$ is not shown but are qualitatively the same as Figs. 4 and 5). In this example, optimization yields a unique, non-trivial solution, but it is clear from variation in the individual realizations $F_i(T)$ that there is no apparent way we can identify a universal signal. The corresponding plots on the right show a different case in which the data collapse is reasonable, but the universal signal we find differs significantly from what we expect to find. In particular, the high-temperature behavior exhibits significant negative slope, inconsistent with the behavior of DNA rings that remain open. In this situation, it is plausible that too much background has been subtracted from the raw signals. In the next section, we discuss the origins of these problems and point to methods for addressing them.

4.3. Problem-specific constraints

While Eqs. (12), (19) and (20) are sufficient to guarantee a mathematically well-posed analysis, they do not necessarily yield physically meaningful transformation parameters. As noted above, Fig. 7 illustrates a case in which transformed signals are obviously meaningless, as well as one in which it is difficult to objectively discern if the analysis was successful. Conceptually, there are two interrelated issues at stake.

First, we almost always have qualitative (or even quantitative) information about how data should behave, e.g. as expressed in Eq. (7) – (8b). However, our simple analysis does not take this behavior into account, only seeking to minimize $\mathscr{L}$ subject to the scaling constraint. This suggests that when possible, it is useful to leverage additional information about $\mathscr{U}$ as part of the optimization procedure to restrict the admissible transformations to a set that yields physically reasonable results. This can be achieved by imposing additional constraints, which act as new modeling assumptions on the data.

Second, we must always allow for *and test* the possibility that there is no meaningful set of transformations to $\mathscr{U}$, either because the raw data is somehow too corrupted or $\mathscr{U}$ is modeled incorrectly. Here again, constraints play a critical role. We can impose, for example, the restriction that all transformed signals have small deviations from their average. The ability to simultaneously satisfy this constraint along with all of the other modeling assumptions leads to an "unbiased" criterion for deciding that $\mathscr{U}$ exists and can be approximated. In the language of constrained optimization, the mathematical *feasibility* of the problem provides this information. Infeasibility can be interpreted as an indication that no meaningful transformations exists or at least motivate further examination of the data.

Invariably, the first of these issues is problem specific, whereas the second is equally applicable to any data. With this overall picture With this overall picture With this overall picture With this overall picture With this overall picture With this overall picture We thus explore the latter in the context of our general formulation of the optimization framework in the next section, and reserve discussion of problem-specific considerations to Sec. 4.3.2.

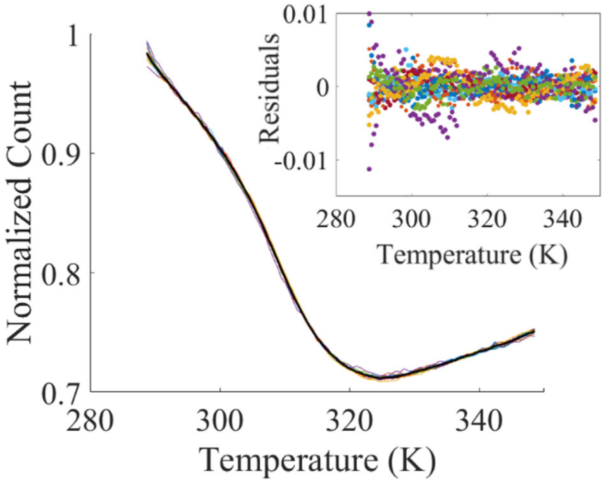


Fig. 5. Transformation of data appearing in the bottom plot of Fig. 3. As before, transformed datasets are in color, with the mean overlaid in black. Note that the residuals are all less than 10^{-2}, which is within 1% of the absolute scale of the data. This is consistent with our observation on the residuals of the background data in Fig. 4. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

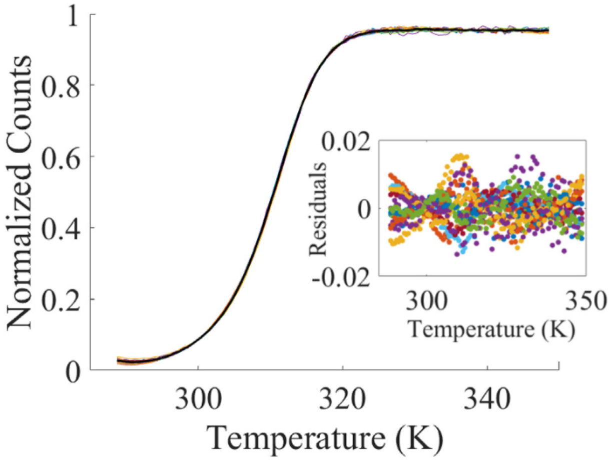


Fig. 6. Transformed signals (color) and mean curve (black) corresponding to the data in Fig. 2. As before, note that the residuals are roughly 1% of the mean curve. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4.3.1. General formulation of error constraints

If we assume N_D total datasets, an estimate of the mean universal signal is given by

$$\begin{aligned}\overline{\mathscr{U}}(T) &= \frac{1}{N_D}\sum_{i=1}^{N_D}\mathscr{U}_i(T)\\ &= \frac{1}{N_D}\sum_i \tau_{0,i}\mathscr{S}_i(T) - \sum_n \overline{\tau}_n\mathscr{B}_n(T)\end{aligned} \tag{23}$$

where $\overline{\tau}_n = N_D^{-1}\sum_i \tau_{n,i}$ is the sample mean value of the nth noise coefficient. An obvious formulation of an error constraint is

$$|\overline{\mathscr{U}} - \mathscr{U}_i| \leq \sigma \tag{24}$$

for some threshold σ. In the context of constrained optimization, it is advantageous to reformulate Eq. (24) as a linear inequality constraint, since this allows us to recast the optimization in terms of *quadratic programming*, for which there are a variety of numerical algorithms to minimize objective functions [21]. Thus, we re-express Eq. (24) in terms of a pair of equivalent constraints

[7] Moreover, the independent realizations $B_i(T)$ can be used individually to propagate uncertainties associated with our inability to exactly determine the universal signal, although we do not pursue this task here.

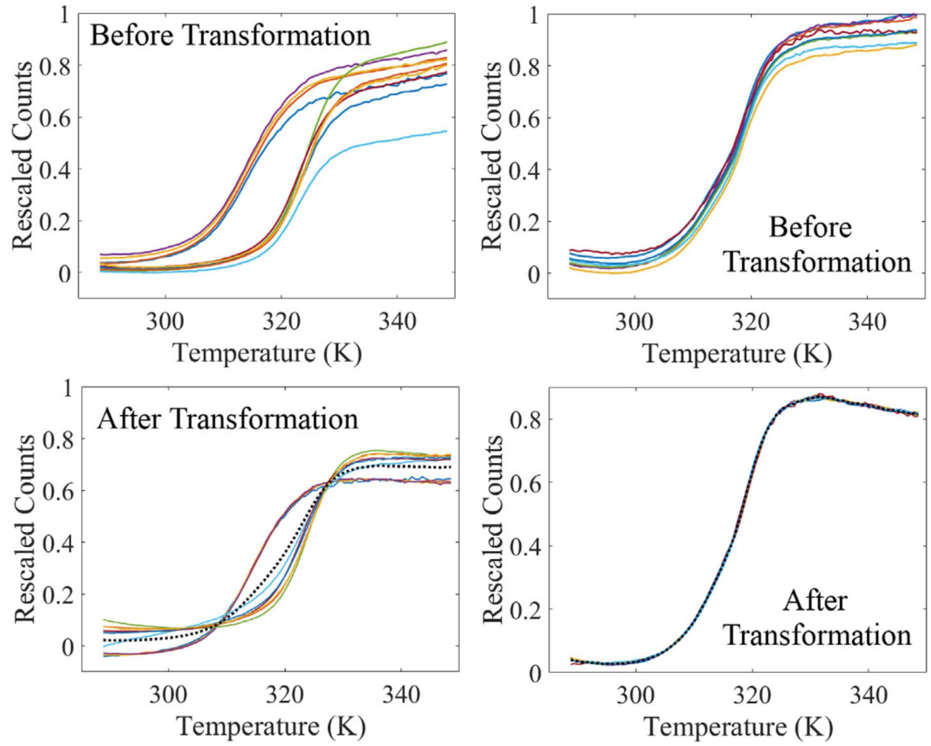


Fig. 7. Two examples of transformations computed according to the minimally well-posed formulation. Note that the ability to identify unique, non-trivial transformation parameters is not sufficient to guarantee that the resulting melt curves are physically meaningful (Fig. 4 and 5 are representative of the corresponding $B(T)$ and $R(T)$.). *Left*: Data before (top) and after (bottom) affine transformations. While the bottom curves yield the minimal sum of errors squared between all datasets, they nonetheless fail to collapse onto a single curve. The mean is shown in dotted black and does not by itself indicate a problem without reference to the individual curves in color. *Right*: Data before (top) and after (bottom) affine transformations. Note that while the transformed curves agree nicely with the mean (dotted black), they have a roughly 10% decrease in value at high temperatures. This behavior is inconsistent with our physical expectations for this dataset and may indicate over-subtraction of noise. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

$$\overline{\mathscr{U}} - \mathscr{U}_i \le \sigma \tag{25a}$$

$$\mathscr{U}_i - \overline{\mathscr{U}} \le \sigma. \tag{25b}$$

It is important to note that the value of σ is a free parameter, which allows modelers a degree of flexibility when constraining estimates of $\mathscr{U}$. As discussed in the previous section, good practice entails scaling the raw data to an interval whose range is $\mathscr{O}(1)$. In this case, the numerical value of σ is comparable to $\delta = \max_T[\overline{\mathscr{U}}(T)] - \min_T[\overline{\mathscr{U}}(T)]$. In particular, we may interpret the ratio σ/δ as the characteristic maximum allowable noise-to-signal ratio for the realizations $\mathscr{U}_i(T)$.

A key consequence of introducing the error-bound constraint is that the expanded optimization problem encoded in Eqs. (13), (19), (20) and (25) may be *infeasible*, meaning that the constraints cannot be simultaneously satisfied. In other words, there are no transformation parameters that collapse the data to within the desired threshold. The benefit to allowing such infeasible problems is that the analysis automatically encodes a fixed criterion by which we can judge the analysis to be possible or not. While this criterion is partially subjective, it does provide a mathematical and reproducible test that is not subject to the opinions of a modeler.

4.3.2. *Example of problem-specific constraints for FRET data*

Equations (7) and (8) motivate constraints that can be used to test the feasibility of data collapse conforming to physical expectation. Consider first a reinterpretation of Eq. (7) in terms of linear inequality constraints. Specifically, the condition that $\frac{\mathrm{d}F}{\mathrm{d}T} \to 0$ for sufficiently large and small T implies that a least-squares fit of a line to the corresponding transformed data should have a slope close to zero. It is straightforward to show that for a vector of N_T temperatures $\mathbf{T}$ and corresponding realizations of the universal signal $\mathbf{F}$, the least-squares slope can be expressed as

$$m(\mathbf{T}, \mathbf{F}) = \frac{\sum_i F(T_i)[T_i - \overline{T}]}{\sum_i T_i[T_i - \overline{T}]}, \tag{26}$$

where $\overline{T} = N_T^{-1} \sum_i T_i$ is the sample mean temperature. It is also important to note that m is a linear operator in $\mathbf{F}$, so that we may impose the inequality constraints

$$-m_\ell \le m(\mathbf{T}, \overline{F}(\mathbf{T})) \le m_h \tag{27}$$

where m_ℓ and m_h are lower and upper bounds on the slope and $\overline{F}(\mathbf{T})$ is the mean universal signal [computed according to Eq. (23)] evaluated at the vector of temperatures $\mathbf{T}$ at which we expect the slope to vanish. As before, scaling F to be $\mathscr{O}(1)$ allows us to interpret m_ℓ and m_h in terms of fractional change per degree Kelvin. In the examples that follow, we set $m = 2.5 \times 10^{-4}$ K^{-1}. Note that this limits change in the absolute value of the universal signal to less than a 0.25% over a 10 K interval.

Fig. 8 shows the result of this analysis applied to the raw data in the top-right plot of Fig. 7. Here we include error constraint according to inequalities (24) with $\sigma = 0.03$ (roughly 3%) relative error, but this is not active. Amazingly, the slope constraint yields transformed data that is visually indistinguishable from horizontal. The inset shows a closer inspection of this data. As expected the average curve (dotted black) has a total change in height that deviates by less than 1% of the scale.

Fig. 9 illustrates how sensitive inequality (27) is to data quality by showing an example for which the constrained optimization is infeasible. In this case, it is impossible to satisfy the inequalities (24) and (27) with $\sigma = 0.03$ and $m_\ell = m_h = 2.5 \times 10^{-4}$ K^{-1}. The inset in the middle plot shows, for example, how the optimization algorithm was unable to satisfy the error constraint. Closer examination of the top plot shows that the high-temperature behavior of the blue curve is partly to blame, as it displays an upward trend inconsistent with the other data. The bottom plot shows the effects of removing this curve, which yields

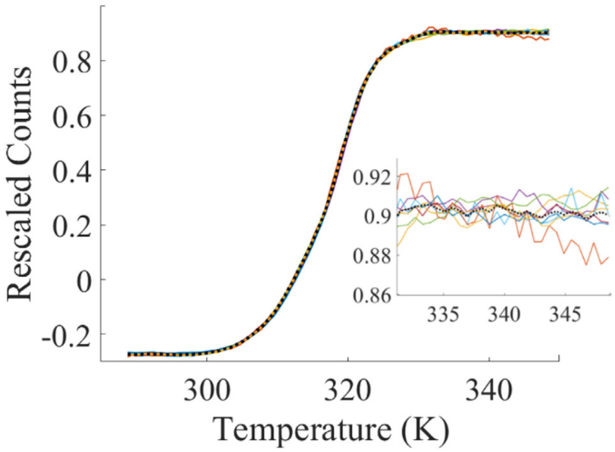


Fig. 8. Transformations of the data in the top-right of Fig. 7 with the constraints given by inequalities (24) and (27). As before, the mean curve is in dotted black. The inset shows the high temperature behavior of the data. Note that the variation in average slope is less than 1% of the range of the data.

a feasible problem. Thus, the infeasibility of the optimization provides a motivation for either holding out this dataset from the analysis or at least considering its behavior in more detail.

5. Discussion, limitations, and conclusions

5.1. Objectivity of the method

As a general rule, data analysis is an inherently subjective endeavor. This arises from the fact that no one has “direct access” to the underlying processes that create data.[8] Fluorescence-based measurements of DNA, for example, rely on a chain of instruments to amplify microscopic signals to make them accessible to an observer. Thus, the measurements are at best indirect. Even the processes of signal amplification demand an element of trust on the part of the instrument user because they involve microscopic phenomena that are equally difficult to audit. Data must therefore always be interpreted in the context that it was collected. *This defines the role of modeling* [e.g. *Eq. (1)*] *as the task of providing that context.* We have repeatedly emphasized this point because modeling is the point in analysis where subjectivity is first introduced by way of choices about how to interpret data. As these choices can have unexpected and/or unintended consequences on the meaningfulness of an analysis, it is important to be open-eyed about both their limitations and potential for misuse. These considerations play a critical role in all extensions of the work we have presented and therefore require further scrutiny.

We begin by addressing the question of how constrained optimization provides context for interpreting fluorescence data. In the previous sections we have demonstrated that physically motivated constraints can lead to transformations of raw data that yield well behaved and reasonable universal signals. However, comparison of Figs. 7 and 8 could lead one to conclude that the method forcibly transforms data to behave as we wish, thereby providing meaningless and/or even misleading context. The argument might proceed along the lines that, issues of well-posedness aside, an analysis should allow data to “speak for itself” more in the spirit of our minimal formulation of the optimization.

However, a more in-depth look at the latter provides contrast for interpreting fully constrained optimization. In particular, the minimal formulation determines the universal signal by solving Eq. (13) subject only to the requirement that the data provide some meaningful

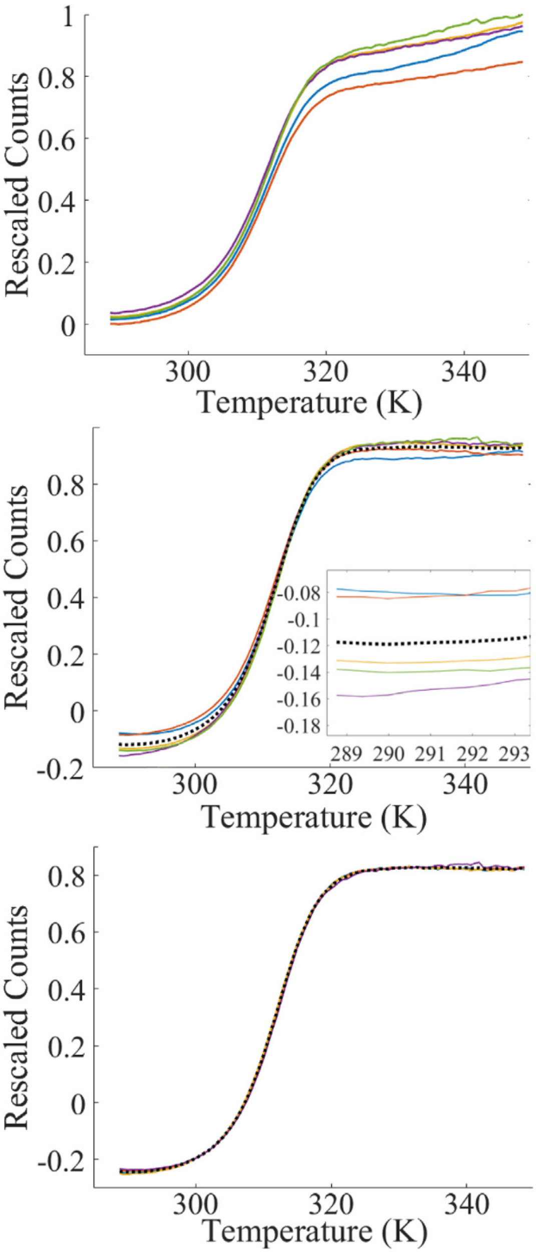


Fig. 9. Example of dataset for which the affine transformations cannot simultaneously satisfy the error and slope constraints. *Top*: Original datasets. *Middle*: The optimizer's best attempt at satisfying the inequality constraints and minimizing the objective. Note that while the slope of the average curve is close to zero at low temperatures, the individual curves deviate by more than $\sigma = 0.03$ from the mean. *Bottom*: The transformations are successful after removal of the blue curve in the middle plot. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

contribution.[9] The interpretation of this approach is that the “true” transformation parameters are those that minimize the sum of differences-squared, all else being irrelevant. By analogy, this is like hiring the smartest person one can find without regard to their qualifications for the job. Constraints are therefore useful to ensure that we only consider those candidates with the right qualifications, of whom there

[8] We mean this statement largely as it pertains to experimental data, but even some simulation data arises from stochastic or chaotic behavior.

[9] As an aside, it is worth noting that an equivalent approach to our minimal formulation can be achieved by setting $\tau_{0,1} = 1$ and solving Eq. (12) without additional constraints. This method scales all datasets to the first.

may be none (infeasibility of the problem). Viewed in this light, *constrained optimization provides a practical context for analyzing data by only allowing those interpretations that are deemed to be physically meaningful.*

Despite the usefulness of this approach, constraints introduce new choices into the modeling process, and *one must take care not to abuse this freedom.* This is more likely to occur through omission rather than inclusion (too many constraints often leads to infeasibility; see Sec. 5.3). For example, the error constraint expressed via inequalities (24) plays a critical role in ensuring that the slope constraints are applied uniformly to each of our datasets. Without the former, it is almost guaranteed that one could transform arbitrary data in such a way that the mean universal signal satisfies inequalities (27), irrespective of any wild behavior in the individual sets; see, e.g. Fig. 9. Thus, checks for internal consistency between datasets must not be omitted in favor of physically justified constraints.

Similar considerations apply to model construction in support of either constrained or unconstrained optimization. In particular, it is possible to postulate enough noise sources $\mathscr{B}_n(T)$ that any collection of random datasets could be brought into agreement consistent with arbitrary (or no) constraints. This could be achieved, for example, by assuming the $\mathscr{B}_n(T)$ are Fourier modes on the domain of temperatures, in which case the affine transformations would amount to a “filtering” exercise that would always seem to magically collapse data. In this case, the transformations are obviously meaningless, with the collapse being a consequence of the mathematical fact that (essentially) any function can decomposed into a collection of Fourier modes. More generally, however, it is not necessary that the $\mathscr{B}_n(T)$ have such a well defined structure. Even postulating too many noise sources of arbitrary form can span a large enough domain of functions, so that collapse is almost guaranteed. *Thus, it is critical that any realization of Eq. (1) have terms that are rooted in rigorous modeling and/or experimentally validated phenomena to avoid overfitting the data.*

The examples provided in this manuscript suggest one route to avoid such problems: construct noise terms directly from experimental data. In particular, our background and baseline function $\mathscr{B}_n(T)$ and $R(T)$ were built from measurements of various control experiments, which therefore provided justification for using them in Eq. (5). Moreover, we did not need to assume a functional form for these effects, which removed our freedom to give them unnecessary structure. From our broader perspective of UQ, this aspect of our approach is beneficial because it reduces model-form errors and instills confidence that the analysis assumes the minimum necessary to arrive at meaningful conclusions.

5.2. Generalizations of the analysis

5.2.1. Generic considerations

The analysis presented herein allows for many generalizations. Roughly speaking, modification may take place at one of three stages: model construction, choice of objective, and formulation of constraints.

While the first and third tasks are problem specific, we can make general observations about the impact of their structure on the choice of objective. In particular, the formulation of Eqs. (6), (9) and (10) has the important property of being a *linear* function of the unknown coefficients $\mathfrak{p}$, π, and p, despite the sources of fluorescence being nonlinearly coupled. In such cases, an objective function can always be expressed as a quadratic function [i.e. sum of differences squared, in the spirit of Eq. (12)], which facilitates quadratic programming if the constraints are linear. Nonlinear models in the unknown coefficients can also be employed, but in such cases the objective function may be more complicated and/or not amenable to well-established optimization algorithms. We leave such tasks for future work.

Regarding the objective, our choice of the L^2 norm [i.e. the sum of differences squared between signals in Eq. (12)] seems well suited for the task at hand. However, equally useful transformations would likely arise from norms of the form

$$\mathscr{L}_{\mathscr{U}}(\{\tilde{\tau}\}) = \sum_{i,j,T} \left| M(\tilde{\tau}_i;\, \mathscr{S}_i,\, \mathbf{B}) - M(\tilde{\tau}_j;\, \mathscr{S}_j,\, \mathbf{B}) \right|^p \tag{28}$$

for $p > 0$, or even more generic measures of “distance” in the spirit of the Kullback-Leibler divergence [22]. In a related vein, one could imagine weighting temperatures in the objective differently so as to enforce agreement more strictly in certain regions.[10] Finally, we have assumed no uncertainty in the temperature variable, which may not be realistic for all PCR protocols. Thus, “orthogonal” objectives that minimize perpendicular distances (instead of vertical distances) between curves may be appropriate in some cases. In all of these situations, however, increased complexity of the objective will induce additional computational costs.

5.2.2. Comparison of FRET pairs, intercalating dyes, and related systems

In the context of affine transformations, FRET pairs and intercalating dyes should yield identical universal signals $F(T)$. However, the corresponding melt curves differ in several fundamental ways, so that experimental design and analysis considerations must be tailored to each system. We discuss such issues now.

In quantitative data analysis, both FRET pairs and intercalating dyes exhibit baseline physical behavior that may contribute a temperature-dependent response to the measured fluorescence signal independent of their interactions with the DNA; see Eqs. (6) and (9). As mentioned in Sec. 2.2, this baseline fluorescence may change with the local environment (i.e. quenched versus unquenched states), which should be characterized experimentally if deemed to be relevant. (For our purposes we ignored differences between these two states.) Moreover, both FRET pairs and intercalating dyes can modify the thermodynamics of hybridization, although it is far easier to do so accidentally by adding the latter in too high of a relative concentration. Such considerations are largely system specific and could depend, for example, on other temperature-dependent changes to the secondary structure immediately adjacent to a fluorophore.

Because FRET pairs are typically bonded to 5' or 3' positions of DNA, and energy transfer only occurs when their separation is between 1 and 10 nm, the local environmental states are easier to characterize. By contrast, fluorescent dyes are relatively promiscuous molecules which will intercalate between base pairs or bind to the backbone, resulting in a several order of magnitude increase in fluorescence [23]. Owing to their non-specific nature, dyes have access to many more “local environment” states. These include: (i) free dye; (ii) ssDNA bound dye; (iii) dsDNA bound dye; and (iv) and potentially AT/GC sequence dependent variations for both ssDNA and dsDNA. Given that each of these states may have a different temperature-dependent baseline fluorescence behavior, a model of such systems might take the form

$$\begin{aligned} f_i(T) = {} & \hat{a}_{i,1}F(T)R_1(T) + \hat{a}_{i,2}F(T)R_2(T) + \ldots \\ & + \hat{a}_{i,n}F(T)R_n(T) + \hat{b}_i B(T) + \hat{c}_i, \end{aligned} \tag{29}$$

where (as before), the index $\boldsymbol{i}$ corresponds to the experimental realization and $\hat{a}_{i,j}$ corresponds to the relative fraction of molecules in local environment state $\boldsymbol{j}$ having baseline rate $R_j(T)$, assuming $\boldsymbol{n}$ total such states.

We also note that our methods are likely to be useful for analyzing experiments in which other mechanisms (e.g. static quenching) control the fluorescence; see, e.g. Ref. [24]. In such cases, it may be possible to carry over our analysis with few (if any) modifications, provided the low and high-temperature asymptotic behaviors of an idealized signal are constant. However, it is important to recognize that the affine transformations do not set an absolute scale for the universal signals, which requires outside information about the system. In the FRET

[10] In more mathematical language, we could make our objective function non-convex.

experiments, this information was extracted by the observation that the fluorophores are all off or on at low and high temperatures. As this will not in general be true, it may be necessary to impose additional constraints that set an appropriate scale of the data. Such issues are the topic of a manuscript in preparation.

In a related vein, we anticipate that our approach can be applied to non-equilibrium PCR curves associated with fast heating rates. In such cases, the degree to which any given sample is out of equilibrium will likely correlate strongly with the sample size, since larger volumes respond more slowly to temperature changes. Thus, the relative error constraint may play an important role in assessing the extent to which the samples are uniformly out of equilibrium.

Ultimately, however, models of the fluorescence efficiency, its temperature dependence, and the various local environmental factors will depend on the system at hand for FRET pairs, intercalating dyes, and other potential systems of interest. In all such cases it is critical that one evaluate the assumptions as to which effects may be ignored, preferably at an early stage so as to facilitate later data analysis and uncertainty quantification.

5.2.3. Comparison with methods based on finite-differences

A large class of high-resolution melt (HRM) PCR measurements attempt eliminate background effects by computing finite-differences of a melt curve and defining T_m as the location of the corresponding global maximum. In the present work, we refrain from such techniques due to considerations arising from UQ. While a detailed discussion of such issues is best reserved for a separate manuscript, we highlight several key points.

For one, it is well known in the mathematics and physics community that finite differences amplify noise. This observation stems from the recognition that noise η in experimental systems if often not differentiable or even continuous. Thus, finite differences of the form

$$\left|\frac{\eta(t+\Delta t)-\eta(t)}{\Delta t}\right| = \mathcal{O}(\sigma/\Delta t), \tag{30}$$

[where t is a continuous parameter such as temperature, and $\sigma > 0$ is the standard deviation of η] tend to blow up as Δt becomes small.[11] Such phenomena complicate measurement and uncertainty quantification because data analysis routines must contend with increased variation in the data. Moreover, smoothing techniques do not necessarily simplify the problem, since these can add layers of subjectivity whose impacts on measurements can be difficult to quantify.

In addition to this, Eq. (5) reveals that *finite differences do not eliminate background effects and temperature-dependence of fluorophores*. Explicitly computing the maximum in terms of the derivative yields

$$\hat{a}_i[F'(T)R(T) + F(T)R'(T)] + \hat{b}_i B'(T) = 0. \tag{31}$$

In general, the temperature for which this equation is true does not correspond to $F'(T) = 0$, owing to non-trivial dependence of $R(T)$ and $B(T)$ on T. Given that these quantities can lead to variations of 10% or more in the total fluorescence, it is likely that they may severely bias estimates of T_m if not properly addressed as part of the data analysis.

We also note that the definition of T_m for the DNA systems considered here is ambiguous, which affects comparison with finite-difference approaches. Fundamentally this issue arises from the fact that molecular disassociation of DNA occurs over a range of temperatures (as opposed to a single melt temperature) and is a manifestation of the Boltzmann statistics for this process. The melt curve can be likened to a cumulative distribution function, while its derivative is the corresponding probability density function (PDF). Ambiguity therefore arises from the way in which we extract a single number, i.e. T_m, from a probability distribution. Computing this quantity from the maximum of the PDF defines the melt temperature in terms of a mode, whereas the definition that we use is a median. In general, these quantities will not be the same, and care must be taken when such estimates are to be compared with the results of other measurement techniques.

5.3. Constraints that may be problematic

The behavior implied by Eq. (7) is local in the sense that it characterizes the behavior of a function at a point. With noisy data sampled at discrete intervals (e.g. temperatures), it can be difficult to deduce this type of behavior with significant accuracy. Indeed, this phenomenon is well understood in terms of the maxim that "derivatives amplify noise." In the context of the examples discussed herein, we could have replaced the least-squares estimate of slope [Eq. (26)] with a finite-difference approximation (and in fact we tried this). However, the resulting increased noise in this estimate would necessitate such large bounds m_ℓ and m_h in Eq. (27) that it would become meaningless. The use of Eq. (26) was therefore justified on the grounds that a non-local characterization of the behavior of $F(T)$ yields more stable estimates of slope. The more general takeaway is the observation that local or point-wise estimates *of change* should be avoided when formulating constraints. It is important to distinguish these from point-wise estimates of the function value itself. For example, inequality (24) can be computed at each temperature because the aim is to constrain fluctuations about a mean at a common value of T.

It is also important to keep in mind that constraints always decrease (or more accurately, never increase) the domain of possible solutions to Eq. (13). Thus, adding too many increases the chances that a formulation of the optimization will be infeasible. One should take care not to overconstrain the problem, lest physically meaningful solutions be unnecessarily excluded from consideration.

5.4. Extensions of UQ

While a primary goal of this work is to formulate the analysis in a way that facilitates UQ, our main focus is *validation*, i.e. tasks that assess the extent to which a model accurately describes the data. *Uncertainty propagation, the process of estimating uncertainty in a prediction based on variability in inputs* (e.g. *raw data*) *remains largely outside of our scope*. In general, this latter task is best reserved for downstream or decision-making analyses that leverage information about fluorescence data, since each one will have its own requirements that inform uncertainty budgets, methods of estimating uncertainties, etc.

We emphasize, however, that the constrained optimization method we propose is amenable to a variety of numerical propagation techniques, in large part because it is inexpensive to run. Among the simplest are Monte Carlo style approaches in which an analysis is repeatedly applied to different realizations of a dataset to generate a distribution of predictions. In the context of our hierarchical modeling Eqs. (6), (9) and (10), this could be done as follows. First, compute transformed realizations $B_i(T)$ of the background signal through the procedure described in Sec. 4. Next, use each of these realizations to estimate independent realizations of $R(T)$. Given, for example, 12 experimental datasets for $B_i(T)$ and 12 for $R(T)$, this would yield 144 total realizations $R_{i,j}$. We could then combine these estimates with additional datasets for $F(T)$. Twelve such experiments would yield 1728 distinct realizations $F_{i,j,k}(T)$, which could then be propagated through some downstream analysis to quantity $X_{i,j,k}$. While these datasets would not necessarily be statistically independent, their range of corresponding estimates would nonetheless provide an empirical and reasonable confidence interval for the mean value of X, for example.

Fig. 10 provides a small illustration of this exercise using 12 experimental datasets for $F(T)$ with $R(T)$ and $B(T)$ taken as their mean values from 12 to 24 experimental realizations of these quantities, respectively. Here we have performed a van't Hoff analysis on the data, fitting the function

[11] Technically speaking, the equality in Eq. (30) should be understood as being true in an average sense.

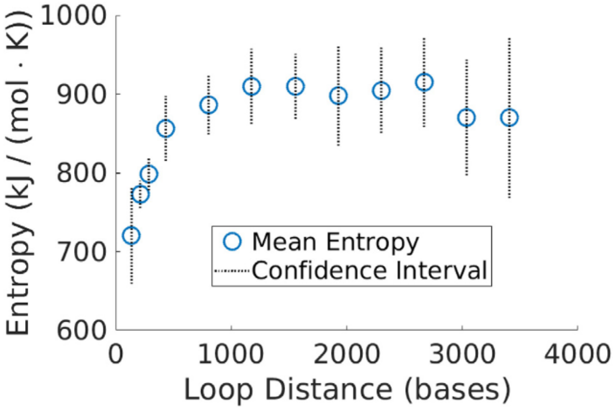


Fig. 10. Van't Hoff analysis of data according to Eq. (33). The qualitative agreement with theory suggests that the affine transformations have produced meaningful interpretations of the raw data.

$$K = \frac{F(T)}{1 - F(T)} \tag{32}$$

(normalized to [0,1]) to the function

$$\log(K) = \frac{\Delta H_0}{T} + \Delta S_0 \tag{33}$$

where ΔH_0 is the enthalpy change upon DNA melting, and ΔS_0 is the entropy change upon melting (i.e. $F(T) = 1/2$). [Full details of the van't Hoff analysis and experimental procedure are provided in Ref. 11 and will not be repeated here.] By propagating 12 datasets through this analysis, we recover uncertainty estimates that are small enough to reveal the anticipated behavior of this data. See Ref. 11 for more details.

5.5. *Final thoughts: overarching perspectives on data analysis*

Data analysis is so fundamental to science that it is sometimes seen as routine, if not mundane. Indeed, the availability of software packages that facilitate the *implementation* of various fitting algorithms can lead to the misconception that data analysis *itself* is easy.

A primary objective of our manuscript has been to counter this perspective by resurfacing many of the challenges of data analysis as they pertain to fluorescence characterization of DNA binding. Our approach has been to recast this task as an exercise in mathematical modeling, which reveals its most difficult elements: model formulation and validation of underlying assumptions. Invariably, the difficulty in these elements arises from their subjectivity. Good models may be informed by generations of empirical knowledge, but they remain (educated) guesses nonetheless. Validation thus becomes an exercise in assuring oneself that a guess is still a useful tool for extracting information about a system and making predictions. This begs the question: when has a model been sufficiently validated?

While we cannot answer this question because it is situation dependent, we have provided computational tools that can help scientists more efficiently address it for themselves. Specifically, we have demonstrated how modeling and validation can be integrated into a single task by way of constrained optimization. In the past these have often been treated separately, leading to the possibility that costly experiments and modeling would be carried out, only to be invalidated at the very end. In other fields (e.g. materials modeling), we are even aware of instances in which lack of internal validation in the data analysis is a primary contributor to unacceptably large uncertainties and/or misleading predictions [25]. Thus, another goal of our work has been to promote a new perspective, namely that UQ (in its broader sense) can and should be more tightly integrated into both modeling and data analysis. Ultimately we believe that such practices will lead to more robust measurement protocols and facilitate reproducibility in the biology community.

Author contribution

Paul Patrone: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation, Visualization, Investigation, Software, Validation, Writing- Reviewing and Editing.

Anthony Kearsley: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation, Visualization, Investigation, Software, Validation, Writing- Reviewing and Editing.

Jacob Majikes: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation, Visualization, Investigation, Software, Validation, Writing- Reviewing and Editing.

J. Alexander Liddle: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation, Visualization, Investigation, Software, Validation, Writing- Reviewing and Editing.

Acknowledgements

This work is a contribution of the National Institute of Standards and Technology and is not subject to copyright in the United States.

Appendix. Experimental details

We provide a brief overview of the experimental details; see also Table 1. See also Ref. [11] for more details.

We considered all combinations of DNA systems based on the M13mp18 genome or “scaffold”. Thirteen different fold geometries were produced by varying the anchor sequence (Fig. 1). Five different persistence lengths were generated by converting different proportions of the ssDNA scaffold to dsDNA by using sets of complementary 32 and/or 37 base oligomers, at locations distant from the fold and fluorophore-labeled oligomers. For each pair of design parameters, we generated 12 replicates of the system in order to quantify measurement uncertainties. (However, infeasibility of the optimization leads us to exclude some datasets.) To ensure a consistent sampling of pipetting error simultaneously with manageable sample preparation, 12 replicates of all buffer and oligomers except the fold oligomer were independently prepared. These master replicates were used as a base stock for the 12 replicates of each fold distance by multichannel pipettor through the two plates for each scaffold persistence length. As such, pipetting variability within any fold distance is uncorrelated. However, uncertainties arising from pipetting are correlated between replicates of folding distances.

The scaffold used was M13 MP18 ssDNA acquired from Tillibit.[12] DNA oligomers were acquired from Integrated DNA Technologies. Modified oligos were HPLC purified and unmodified oligos were not purified at all before use. Cacodylate buffer, at 50 mM, was used for its ability to hold pH constant with temperature and was acquired in 8x stock from Electron Microscopy Sciences. The buffer was supplemented with Magnesium Acetate at 12.5 mM. To minimize extraneous signal, all other oligomers were added in excess of the fluorophore oligomer, as shown in the Table 1.

[12] Certain commercial products are identified in this work in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

To generate fluorescence curves, we used a StepOnePlus Real-Time PCR, with a melt protocol consisting of an initial denaturing step (80C for 1 min), an annealing sequence (75C–15C, 0.61C steps, 3.5 min hold) and finally melting (15C–75C, 0.61C steps, 3.5 min hold). The raw fluorescence intensity in the green channel was used in place of the (internally computed) multicomponent intensity, since uncertainties associated with the latter were unknown.

In addition to these experiments, we also repeated fluorescence measurements on 12 replicates of a baseline control containing no fold oligomer, as well as 24 empty wells.

Table 1
Concentrations of oligomers in each sample.

	Absolute	Excess
	Concentration	(Relative to scaffold)
Scaffold	45 nM	–
Fold Strand	225 nM	5×
Quencher	450 nM	10×
Fluorophore	30 nM	0.67×

References

[1] J.A. O'Rawe, S. Ferson, G.J. Lyon, Trends Genet. 31 (2015) 61.

[2] T.C. Lorenz, JoVE (2012) e3998.

[3] T. Ishikawa, Y. Kamei, S. Otozai, J. Kim, A. Sato, Y. Kuwahara, M. Tanaka, T. Deguchi, H. Inohara, T. Tsujimura, T. Todo, BMC Mol. Biol. 11 (2010) 70.

[4] R.H. Lipsky, C.M. Mazzanti, J.G. Rudolph, K. Xu, G. Vyas, D. Bozak, M.Q. Radel, D. Goldman, Clin. Chem. 47 (2001) 635.

[5] M. Li, R. Palais, L. Zhou, C. Wittwer, Anal. Biochem. 539 (2017) 90.

[6] D. Jost, R. Everaers, Biophys. J. 96 (2009) 1056.

[7] P. Atkins, J. de Paula, Atkins' Physical Chemistry, OUP Oxford, 2010.

[8] A.L. Plant, L.E. Locascio, W.E. May, P.D. Gallagher, Nat. Methods 11 (2014) 895 EP.

[9] T. Förster, Ann. Phys. 437, 55..

[10] C. dos Remedios, P. Moens, J. Struct. Biol. 115 (1995) 175.

[11] J.M. Majikes, P.N. Patrone, D. Schiffels, M. Zwolak, A.J. Kearsley, S.P. Forry, J.A. Liddle, Nucl. Acids Res. 48 (10) (04 June 2020) 5268–5280, https://doi.org/10.1093/nar/gkaa283.

[12] IEEE Std 1012-2012 (Revision of IEEE Std 1012-2004) - Redline, (2012), p. 1.

[13] R. Smith, Uncertainty Quantification: Theory, Implementation, and Applications, Computational Science and Engineering, SIAM, 2013.

[14] R. Rapley, S. Harbron, Molecular Analysis and Genome Discovery, Wiley, 2011.

[15] D.N. Birdsell, T. Pearson, E.P. Price, H.M. Hornstra, R.D. Nera, N. Stone, J. Gruendike, E.L. Kaufman, A.H. Pettus, A.N. Hurbon, J.L. Buchhagen, N.J. Harms, G. Chanturia, M. Gyuranecz, D.M. Wagner, P.S. Keim, PLoS One 7 (2012) 1.

[16] C.T. Wittwer, G.H. Reed, C.N. Gundry, J.G. Vandersteen, R.J. Pryor, Clin. Chem. 49 (2003) 853.

[17] G.H. Reed, C.T. Wittwer, Clin. Chem. 50 (2004) 1748.

[18] N. Kretschy, M. Sack, M.M. Somoza, Bioconjugate Chem. 27 (2016) 840.

[19] A. Vologodskii, M.D. Frank-Kamenetskii, Phys. Life Rev. 25 (2018) 1.

[20] M. Reiter, M. Pfaffl, Biotechnol. Biotechnol. Equip. 22 (2008) 824.

[21] A.J. Kearsley, J. Res. Natl. Inst. Stand. Technol. 111 (2006) 121.

[22] S. Kullback, R.A. Leibler, Ann. Math. Stat. 22 (1951) 79.

[23] F.G. Loontiens, P. Regenfuss, A. Zechel, L. Dumortier, R.M. Clegg, Biochemistry 29 (1990) 9029.

[24] M.K. Johansson, H. Fidder, D. Dick, R.M. Cook, J. Am. Chem. Soc. 124 (2002) 6950.

[25] P. Patrone, A. Kearsley, and A. Dienstfrey, “The role of data analysis in uncertainty quantification: case studies for materials modeling,” in 2018 AIAA Non-deterministic Approaches Conference.