Isothermal Chemical Denaturation: Data Analysis, Error Detection, and Correction by PARAFAC2

Augustijn, Dillen; Kulakova, Alina; Mahapatra, Sujata; Harris, Pernille; Rinnan, Åsmund

Published in:
Analytical Chemistry

Link to article, DOI:
[10.1021/acs.analchem.9b05748](https://doi.org/10.1021/acs.analchem.9b05748)

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Augustijn, D., Kulakova, A., Mahapatra, S., Harris, P., & Rinnan, Å. (2020). Isothermal Chemical Denaturation: Data Analysis, Error Detection, and Correction by PARAFAC2. *Analytical Chemistry*, 92(10), 6958–6967. <https://doi.org/10.1021/acs.analchem.9b05748>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Isothermal Chemical Denaturation: Data Analysis, Error Detection, and Correction by PARAFAC2

Dillen Augustijn, Alina Kulakova, Sujata Mahapatra, Pernille Harris, and Åsmund Rinnan*

Cite This: <https://dx.doi.org/10.1021/acs.analchem.9b05748>

Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Characterization of a protein's conformational stability is a key step in the development of biotherapeutics, where protein unfolding leads to adverse properties, such as aggregation and loss of efficacy. Isothermal chemical denaturation (ICD) can be applied to determine chemical stability, aiming to identify the optimal solvent conditions, in terms of pH, salt concentration, and added excipients. For seven monoclonal antibodies, this study investigates the observed intrinsic protein fluorescence emission spectra as a function of denaturant concentration. Protein formulations are screened in two experimental series. We show how the peak shapes of folded and unfolded proteins are preserved under added salt (0–140 mM NaCl) and added excipients concentrations, as typically found in biotherapeutic formulations and that only minor effects in tryptophan fluorescence peak tailing are observed over a large pH range (5.5–9.0). The data of seven mAbs, where GuHCl was a suitable denaturant, are modeled using PARAFAC2. PARAFAC2, a linear decomposition method, is well suited for the data and yields robust, valid, and automated models that allow for the detection of erroneous measurements. Analysis of the errors show correlation with the well-based experimental setup, and differences in observed errors between the two experimental series. We additionally show a correction method for these outliers based on PARAFAC2 model scores, such that full transition curves can be retrieved, increasing the accuracy of any subsequent analysis.

In biotherapeutic development, the determination of the stability of prospective proteins plays a key role both in the selection of suitable candidates for further development and in the formulation development of lead candidates. Determination of the thermal stability of proteins is important in many fields of the bio- and life sciences. In the context of biopharmaceutical development, the conformational stability of a protein demonstrates its propensity to, partially, unfold; thus, assessing this parameter under laboratory-stressed conditions will yield partial insight into the long-term stability parameters.^{1–3} Protein unfolding is important to evaluate, as it can be induced by bioprocessing stresses, storage, and pharmaceutical delivery. Investigation of unfolding propensities of proteins and formulations thereof yield insights into aggregation, degradation, and loss of efficacy.^{1,4,5}

In conformational stability estimation, various techniques are employed, each with various drawbacks. These techniques are based on gradual protein unfolding, where a reversible thermodynamic path is favored, and rely on strictly controlled experimental parameters while incrementally altering few experimental parameters. All of these techniques aim at quantifying or ranking protein unfolding stability. Each of these techniques have merits and downsides, and isothermal chemical denaturation (ICD) is most commonly applied as a complimentary technique or when alternative, high-throughput, techniques fail.

ICD in Formulation Screening. ICD measures protein conformational stability by perturbation of the structure using chaotropic agents. Herein, we utilize the protein's intrinsic fluorescence, and study conformational stability based on the same fluorescence peak shift process as is employed in DSF measurements, and further deliberated on in the [Protein Fluorescence and Denaturation](#) section. Multidomain proteins will show a limited set of unfolding steps, dependent on the stability of the individual domains and only to a lesser extent on the chaotropic strength of the used denaturant. Under isothermal conditions, the transitions can be described in terms of the transition midpoint $c_{1/2}$ or c_m , indicating 50% unfolded protein in terms of denaturant concentration or by calculation of thermodynamics as Gibbs free energy of unfolding, ΔG .⁶

In ICD experiments, formulated protein samples are mixed with denaturant and incubated in order to reach equilibrium. Measuring equilibrated samples allows for better estimation of the parameters. A tradeoff with other techniques is the lower number of data points collected as each concentration has to

Received: December 20, 2019

Accepted: April 23, 2020

Published: April 23, 2020

be pipetted and incubated separately. A clear advantage of ICD in protein formulation development is the lack of temperature ramp, such that the physical properties of, for instance, excipients do not change. Furthermore, the assessment of fluorescent signals is not impeded by the temperature induced, and fluorophore dependent decrease of signal that otherwise coincides with unfolding. ICD has some drawbacks, making it less standard in a biopharmaceutical screening setting: a relatively high sample consumption and low throughput. However, it does offer complementary insights into protein stability, especially when the drawbacks of other techniques hinder high-throughput assessment.

Current ICD techniques employ highly automated systems, and the resulting data are regularly evaluated based on the emission peak maximum wavelength or the ratio of fluorescence signals. Additionally, often the ICD of a protein is not only measured one single time but aims to compare formulation conditions of proteins. By performing such a screening, a comprehensive data set over a diverse range of formulation conditions is obtained.⁷ Herein, multivariate data analysis (MDA), as commonly found in traditional analytical chemistry, allows for the development of new approaches for data analysis, preprocessing and outlier detection.

Scope. In this work, we investigate the use of multivariate data analysis (MDA) on a screening study,⁷ where ICD has been measured as a function of proteins, salt concentrations, pH, and excipient. We apply MDA to fluorescence emission spectra for seven mAbs, formulated in two screening series with 27 and 16 formulations each over 48 denaturant (GuHCl) concentrations. The first series investigates the effect of pH and ionic strength. The second screening series investigates the effect of three common biopharmaceutical excipients⁸ (proline, arginine, and sucrose) over pH and ionic strength. Where the common approach is to analyze each set of ICD data separately, we show how the data can be analyzed together. This efficiently simplifies the comparison between different protein formulations. This also gives the unique opportunity to assess and understand which measurements are subject to errors originating in the measurement system and which measurements follow the protein stability overall trends. We will show that we, through the use of the advanced MDA method PARAFAC2^{9,10} (PARAllel FACTor analysis 2), are able to detect outliers and subsequently correct these measurement points. The observed outliers are related back to the instrumental setup, where error frequencies appear systematically, with minor differences within and between the two screenings. As stated, all of the proteins investigated are mAbs, and as such they are all multistate proteins. The goal of this manuscript is to correct the data, preparing it for further analysis (such as ΔG and c_m) and not to subsequently fit the corrected data to any multistate model and compare our approach to the current approach. We show how PARAFAC2 readily can be used to improve ICD measurement systems as diagnostic, as well as give more robust protein unfolding curves.

EXPERIMENTAL SECTION

Materials and Experimental Screening. Two series of experimental screenings were set up, as schematically given in Figure 1. The first screening, focusing on pH and NaCl effects, tests 27 formulations conditions without excipients, as represented by blue markers in Figure 1. The studied proteins are listed in Table 1. Formulations of protein PPI-17 at pH 8.0,

Figure 1. Experimental design of the formulations. In blue the first series, without excipients, for formulations with all combinations of nine pHs (in steps of 0.5, pH 5.0–7.5 buffered by 10 mM histidine, pH 8–9 buffered by 10 mM Tris) and three salt concentrations (0, 70, and 140 mM NaCl). The second series is represented by the four columns, in purple, yellow, orange, and blue, indicating combinations of two pH values, 5.0 and 6.5, two levels of NaCl, 0 and 140 mM, and three excipients. In the second series, the corresponding formulations were also produced without excipient (pH 5.0 and 6.5 at 0 and 140 mM NaCl), corresponding to replicates of the first series.

Table 1. Presence of Tyrosine and Tryptophan Residues in the Studied Monoclonal Antibodies^a

protein	M_w (kDa)	no. AA	tyrosine residues	tryptophan residues
PPI-01	144.8	1334	64 (4.8%)	24 (1.8%)
PPI-02	148.2	1324	52 (3.9%)	24 (1.8%)
PPI-03	144.7	1330	64 (4.8%)	20 (1.5%)
PPI-08	204.4			
PPI-10	144.2	1326	58 (4.4%)	24 (1.8%)
PPI-13	146.3	1338	58 (4.3%)	26 (1.9%)
PPI-17	145.1	1322	56 (4.2%)	20 (1.5%)

^aThe amino acid sequence of PPI-08 is not given. PPI-08 is a bispecific monoclonal antibody.

8.5, and 9.0 were not collected due to strong protein aggregation when dialyzed at pH 8.5. The second screening series tested 16 formulation combinations. Protein samples were formulated with one of three excipients (280 mM proline, 140 mM arginine, or 280 mM sucrose) or without an added excipient, in combination with two pH levels (5.0 and 6.5) and two salt concentration levels (0 and 140 mM). The samples without excipient, four per protein, are thus replicates of the four of the formulations from the first screen. The second series was carried out ten months after the first.

The studied proteins are IgG1-type PPI-01, PPI-02, PPI-03, PPI-10, and PPI-13; bispecific IgG1 PPI-08; and IgG2-type PPI-17. Chemicals and materials used in sample preparation were: L-histidine (J.T.Baker), L-histidine monohydrochloride (J.T.Baker), Trizma base (Sigma-Aldrich), Trizma hydrochloride (Sigma-Aldrich), and sodium chloride (Sigma-Aldrich). Dialyses were performed using Slide-A-Lyzer dialysis cassettes (Thermo Fisher Scientific).

ICD studies were performed with an Unchained Laboratories HUNK system - AVIA ICD 2304 (Unchained Laboratories, Pleasanton, USA). This fully automated system required protein stock solutions, formulation buffers and denaturation buffers. Protein stock solutions were prepared by dialysis of the antibodies in 10 mM histidine at pH 5.5 and pH 7.0, and in 10 mM tris at pH 8.5. Each dialyzed protein stock was diluted to 1 g/L and used for ICD experiments performed within this pH $\pm$ 0.5. For example, the protein stocks of 10 mM histidine pH 5.5 were used for ICD experiments performed at pH 5.0, 5.5, and 6.0 in the presence of 0, 70,

and 140 mM NaCl, covering nine different formulation conditions in total. For the second series, all studied antibodies were dialyzed in 10 mM histidine at pH 5.0 and 6.5. For the pH and NaCl effect, buffer solutions were prepared at different pH (from 5 to 9) in combination with 76 and 152 mM of NaCl. The protein stock did not contain NaCl, and adding it to the formulation buffer decreased the final concentration of NaCl to, respectively, 70 and 140 mM. For the second series, formulation buffers contained excipients: 328 mM sucrose, 328 mM proline, 164 mM arginine, or without excipient, which resulted in 280, 280, 140, and 0 mM reference mixtures, respectively. Finally, denaturation buffers were prepared as formulation buffers with addition of 6 M GuHCl, resulting in an experimental chaotropic range of 0–5.5 M GuHCl, over 48 steps of 0.12 M GuHCl increments. Final conditions were robotically mixed, incubated, and measured. Incubation was set for 1134 min (~19 h). For each protein, 1 g/L protein stock was used to set the spectrometer gain. The spectroscopic measurements were performed with an excitation wavelength of 285 nm, and the emission spectra were recorded from 300 to 500 nm, in steps of 2 nm.

Protein Fluorescence and Denaturation. In this study we focus on the intrinsic fluorescence emission of the protein. Folded protein fluorescence emission spectra are a combination of tryptophan and tyrosine contributions, especially as the excitation spectra of both these amino acids are highly similar and proteins are typically excited at wavelengths above 280 nm. [Free form tryptophan has an excitation maximum at 285 nm, and free form tyrosine has an excitation maximum at 280 nm.] To reduce the influence of tyrosine, an excitation wavelength between 285 and 300 nm is commonly used,⁴ ensuring that the behavior of tryptophan becomes the predominant factor in studying the proteins.^{11,12}

For protein-bound tryptophan, the protein emission peak is shifted to lower wavelength, based on its electrostatic interactions with sterically neighboring amino acids. The underlying mechanisms have been well described by Vivian and Callis.¹³ In terms of protein unfolding, this means that a protein's tryptophan residues fluorescence show red-shifted emission as the polarity of their surroundings increases, i.e. on moving from a hydrophobic environment to hydrophilic environment.^{14,15} Tryptophan residues in native proteins are typically buried and during unfolding they become solvent exposed, moving into a polar environment. The proteins unfolding can be characterized by measuring the ratio between solvent exposed and buried tryptophan fluorescence signal peak shapes, as is used in DSC, intrinsic fluorescence (IF) DSF, nDSF, and IF ICD.

The intensity of the tryptophan fluorescence signal is subject to changes based on the electrostatic environment of the fluorophores, and some tryptophan residues can be completely quenched in the native state.^{16,17} Typically, in denaturation experiments and under isothermal conditions, the signal intensity increases during protein unfolding. Various other factors can affect the observed signal intensity, as relevant for the data collected for this study: for sufficiently dilute samples, the signal increases linearly with protein concentration. Protein aggregation into larger aggregates is known to decrease the signal. Also, the solvent conditions can affect the intensity, most prominently in the denatured state, by changing the electrostatic environment of exposed tryptophan. Within a typical biopharmaceutical pH range, the pH can have a small, nonlinear effect on the fluorescence intensity. For the purpose

of this study and the applied MDA technique, the fluorescence intensity is less important than the peak shape. The effect of solvent conditions (pH and ionic strength) including excipients, as typical for biopharmaceutical formulations, on the fluorescence peak shape are shown in the results.

The fluorescence peak shift is monitored to study protein thermal unfolding. This can be done by observing the signal at two different wavelengths, e.g., as used in intrinsic fluorescence differential scanning fluorimetry, or by taking a full emission spectrum. For ICD the incubation period is long; a single measurement is taken at each distinct denaturant concentration, and a full fluorescence emission spectrum is obtained. Access to the full emission spectra gives us the opportunity to apply advanced data analysis methods to (a) further study the fluorescence peak shift, (b) better characterize the shift and unfolding ratio, and (c) better distinguish between different unfolding transitions.

For the studied set of monoclonal antibodies, the number of tyrosine and tryptophan residues is given in Table 1. It is not known to which extent each of the residues contribute to the observed fluorescence signal.

Monoclonal antibodies have multiple thermal unfolding steps,^{4,11,18,19} and the antibodies in this study have shown multiple transitions in corresponding DSF studies across the formulation conditions.^{1,7,20} The methodology relies on the presence of tryptophan residues that have nonzero quantum yield, are not solvent exposed in the native state, and are solvent exposed in a denatured or intermediate state. Furthermore, for multidomain proteins with multiple unfolding stages - such as mAbs¹⁹ - this implies that each structural change must include at least one tryptophan residue in order to be characterized separately. Not all domains will unfold under the same denaturant range consistently. For mAbs the common unfolding domains are, in order of energy, the C_H2, Fab and then C_H3 domain. The Fab region is found to be highly variable in ΔG_{Fab} due to changes in the Fv domain, and this transition can coincide with the other domains depending on protein and formulation.^{20–22} It therefore becomes important to have error free data in order to be able to identify these transitions.

Data Collection and Treatment. Each protein formulation was measured as a function of 48 denaturant concentrations, 0–5.5 M GuHCl, in steps of 0.12 M, measuring intrinsic fluorescence emission between 300 and 500 nm, in steps of 2 nm. The first screen contains 180 samples, while screen 2 consists of 112 samples. Therefore, the total number of emission spectra for the two screens is 8640 and 5376, respectively.

As an example of the data input, we show a single formulation experiment, over 48 denaturant concentrations (see Figure 2). From Figure 2, we can appreciate the red-shift in the maximum emission peak, as well as the increase of the signal as the protein is unfolded.

Data were exported into MATLAB (MATLAB 2018b, The MathWorks Inc., Natick) and analyzed therein. The used MATLAB code related to the PARAFAC2 decomposition is based on open source libraries, *The N-way toolbox*²³ and *parafac2.m*,²⁴ both available on www.models.life.ku.dk/algorithms.

MULTIVARIATE ANALYSIS

Our approach to the data analysis of the ICD data is to keep the natural dimensionality of the data as formulations (and

Figure 2. Protein fluorescence emission spectra of PPI-03, formulated at pH 7.5 and 70 mM NaCl, as a function of denaturant concentration.

proteins) by emission spectra by denaturation concentration, and analyze this data array by a multiway method, named PARAFAC2. The advantage of applying PARAFAC2 rather than, e.g., unfolding the data and employing singular value decomposition (SVD) is that PARAFAC2 does not suffer from rotational ambiguity as SVD does. Furthermore, the outcome of PARAFAC2 will be chemical meaningful, indicating that the components in the different modes of the data are directly linked to chemical signals of the measured sample. This is opposed to SVD that “only” will maximize the explained variation in the data, seldom related to specific chemical signals in the recorded data.

PARAFAC2. The applied advanced data analysis method, PARAFAC2, creates a trilinear decomposition of the data. The algorithm is described by Kiers et al.,¹⁰ and we focus on the results of the decomposition and its implications pertaining to ICD data. The notation follows that used by Kamstrup-Nielsen et al.,²⁵ and the significance of the notation, regarding the decomposition of ICD data, are given in Table 2.

To apply PARAFAC2, the data is arranged in a three-dimensional tensor, by stacking the denaturation profiles along the sample axis. We create the tensor $\underline{\mathbf{X}}$, of size $[K \times J \times I]$. In its first dimension there are K samples, as protein and formulation combinations, the second dimension there are J variables, as denaturant concentrations, and the third dimension there are I variables, over the emission wavelengths. For the first experiments series, this results in a tensor $\underline{\mathbf{X}}$ of 180 samples by 48 denaturant concentrations by 97 emission wavelengths. Strictly, the number of denaturant conditions for each sample is variable and erroneous measurement at a specific denaturant concentration can be excluded. The method does not require other preprocessing of the data.

The trilinear decomposition splits the tensor $\underline{\mathbf{X}}$ into three modes, $\mathbf{D}$, $\mathbf{B}$, and $\mathbf{A}$, and an error tensor $\mathbf{E}$ containing the residuals. The number of components (or factors), F , is the tunable parameter in PARAFAC2. Each component contains a separate piece of chemical information of the data, and determination of the appropriate number of components is discussed in the Model validation section. For the ICD data, with two components we can model both the native ($f = 1$) state and the denatured ($f = 2$) state. Other variations in the data, that are not modeled by the components, end up in the residuals or error, $\mathbf{E}$. PARAFAC2 is an unsupervised method, meaning that it finds the components autonomously by finding the patterns in the data that allow for the minimization of the

Table 2. Applied Mathematical Notations and Terms, with Their Relation to the Experimental Data

notation	
a	lowercase or uppercase and italic indicates a scalar or index
$\mathbf{a}$	lowercase and bold indicates a vector
$\mathbf{A}$	uppercase and bold indicates a two-dimensional matrix
$\underline{\mathbf{A}}$	uppercase, underlined, and bold indicates a three-dimensional tensor
terms	
$k \cdots K$	sample index k , a specific combination of a protein and formulation, of all combinations K
$i \cdots I$	wavelength i , for all wavelengths I
$j \cdots J$	denaturant concentration j , between $j = 1$ at 0 M GuHCl and $j = J$ at 5.5 M GuHCl
$\underline{\mathbf{X}}$	three-way data tensor, of size $K \times J \times I$, to be modeled by PARAFAC2
$\mathbf{D}$	modeled fluorescence intensity for each sample k , as a value for each component
$\mathbf{A}$	modeled fluorescence spectrum or spectrum shift, for each emission wavelength i , as a value for each component
$\mathbf{B}$	modeled mixing of components as a function of denaturant concentration, determined for each sample k and over the denaturation concentrations $j \cdots J$, denoted as $\mathbf{B}_k$
$\mathbf{H}$	common part of mode $\mathbf{B}$
$\mathbf{P}_k$	formulation and protein specific part of mode $\mathbf{B}$
$\mathbf{E}$	residuals or error tensor, representing unmodelled variation
$f \cdots F$	PARAFAC2 component f , each representing spectroscopically relevant states

squared residuals, and not by prescribed patterns or fluorescence spectra.

Each mode has F number of components and lies along the original directions of the tensor. The first mode, mode $\mathbf{D}$, lies along the sample dimension K and represents the relative fluorescence intensity of each sample for each model component. The second mode, mode $\mathbf{B}_k$, lies along the denaturation concentrations and gives the mixing of the components for each sample k at different denaturation concentrations ($j \cdots J$). The third mode, mode $\mathbf{A}$, lies along the spectroscopic dimension I and represents the fluorescence emission spectrum. In PARAFAC2, mode $\mathbf{B}_k$ is split into a common part $\mathbf{H}$, and a sample specific part $\mathbf{P}_k$. The term $\mathbf{P}_k$ is a vector that describes the denaturation process for a specific formulation. For two components ($F = 2$) the overall PARAFAC2 model is given graphically, in Figure 3. Modes $\mathbf{D}$ and $\mathbf{A}$ consist of F vectors each as $\mathbf{d}_f$ and $\mathbf{a}_f$ of length K and I respectively. Mode $\mathbf{B}$, split into $\mathbf{H}$ and the denaturant dependent term $\mathbf{p}_{f,k}$, and it contains $K \times J \times F$ values.

The applicability of PARAFAC2 relies on the trilinearity of the structure underlying the data.¹⁰ In chemical terms, this means that each spectrum can be split into independent

Figure 3. Two component PARAFAC2 model, the 3-way tensor $\underline{\mathbf{X}}$ is split into modes $\mathbf{D}$, $\mathbf{B}_k$, and $\mathbf{A}$, and a residual tensor $\mathbf{E}$. For ICD data, the denaturation curves are obtained in the second mode $\mathbf{B}_k$, commonly written as $(\mathbf{P}_k \cdot \mathbf{H})$, with common component $\mathbf{H}$ and formulation dependent $\mathbf{P}_k$ for each sample k . The three modes for each component $f \cdots F$ are vectors, denoted $\mathbf{d}_f$, $\mathbf{p}_{f,k}$, $\mathbf{H}$, and $\mathbf{a}_f$. Mode $\mathbf{D}$ represents the relative concentration of the sample for each component. Mode $\mathbf{A}$ represents the fluorescence profile.

contributions from distinguishable compounds and that the resulting modes cannot have an interaction term. For the ICD data, this lack of interaction term forces us to combine the formulation and proteins into the same dimension, since both the protein and the formulation would influence the overall fluorescence spectrum intensity and shape. Vice versa, in order to create a PARAFAC2 model only the denaturant concentration can influence the peak shift. Therefore, if the formulation has an interaction effect on the protein peak shift in combination with the denaturant, PARAFAC2 cannot be used. This recognizes the importance of the effect of formulation on the fluorescence emission peak shape, and should be investigated before application of PARAFAC2.

The number of PARAFAC2 components is chosen based on the expected number of spectroscopically relevant contributions. Increasing the number of components will always decrease the residuals, thus improving the model fit. However, if too many components are used, some will be chemical meaningless, and/or some components are split into multiple, semiequivalent components. The use of too many factors in a model is termed overfitting, and should be avoided. If too few components are chosen, the model is underfitted and valuable information is left in the residuals.^{10,26} Multiple validation strategies for deciding on the model complexity are available (see the [Model Validation](#) section).

Model Validation. PARAFAC2 models should be checked and validated, which is done in a combination of methods. Multiple PARAFAC2 models are made, using random initialization to avoid local minima during optimization, and the reinitializations of models over an increasing number of components can be compared. Validation and comparison of the models check (a) for uniformity in solutions and local minima, (b) that modes make “sense” from a chemical and spectroscopic perspective,^{27,28} (c) the size of the fit, as this is important to ensure the model is based on a significant portion of the data, (d) the residual matrix for unmodelled structure, indicating a underfitted model, and (e) model complexity by checking the core consistency. Core consistency is a common tool for the estimation of the appropriate number of PARAFAC2 components.²⁵

Outlier Detection. We aim to detect outliers, measurements that lie outside the overall pattern of the samples. The outliers originate from measurement errors, mixing errors, or buffer preparation errors. These outliers are characterized by the mode of the data it manifests and, depending on the phase where the error occurred, how widespread the error is. Assuming correct buffer preparation, we focus on outliers caused by instrumental errors in either sample and buffer mixing or in the spectroscopic measurement. Such outliers manifest along the spectroscopic dimension, i.e., for a single combination of protein, formulation, and denaturant concentration. To detect these outliers, we look at the PARAFAC2 residuals, $\underline{E}$, for each unique combination of sample and denaturant concentration.

We propose the following workflow for analysis of ICD data, as given in [Figure 4](#). In the scheme the data is collected and input into a first PARAFAC2 model step, where the number of components is estimated and the resulting model is validated. To assess each measurement, we compare errors within each denaturant experiment. We sum over the spectral dimension i , to obtain the error at denaturant concentration j . We use a one-sided cutoff value of three standard deviations to indicate significant high errors. Once all outliers are marked, their data

Figure 4. Data analysis scheme. The data input is modeled in its entirety by PARAFAC2, this model is optimized for number of components and validated. Outliers are detected using a cutoff based on their contribution to the residual tensor $\underline{E}$. The outliers are removed, and another PARAFAC2 model is obtained and validated. Outliers are again searched for, and the model is remade. Outliers can then be predicted in the denaturant dependent mode $\underline{B}$, based on surrounding points.

is removed from data input, and a second PARAFAC2 model is made and validated. For our case, no more outliers were found after the first round, and the second PARAFAC2 model is thus also the final model. From this model we obtain (1) a sample and denaturant concentration independent *spectral mode* $\underline{A}$, (2) an emission wavelength and denaturant concentration independent sample mode $\underline{D}$, and (3) an emission wavelength independent denaturant response mode $\underline{B}$. We use mode $\underline{B}$ to correct the observed outlier found earlier, for which a method is described in the [Outlier Correction](#) section. For further analysis, it is possible to investigate mode $\underline{B}_k$ (as $\underline{P}_k \cdot \underline{H}$), or reconstitute the original data as $\hat{\underline{X}}_k$ with or without the corrected outliers, for each sample k . The error term is removed, and $\underline{P}_k$ is of varying length or contains outlier corrected values.

Outlier Correction. We used an iterative smoother over the PARAFAC2 $\underline{B}_k$ term to fill missing data. The method uses a second order Savitzky–Golay filter,²⁹ where the marked outlier itself and neighboring points marked as outliers are not used in the polynomial fits. For the 48 points in $\underline{B}_k$ over a full denaturant series a width of seven datapoints is used to interpolate the outliers. This is done in an iterative procedure, where the fit is repeated for all outliers in $\underline{B}_k$, ranked according to residual between the smoothed $\underline{B}_k$ and the estimated $\underline{B}_k$. After the first fit, $\underline{B}_k$ is updated and the Savitzky–Golay filtering is repeated using the updated points. This procedure is repeated until all outlying points have been estimated. This results in iterative interpolation and possible extrapolation, depending on the position of the corrected data point.

RESULTS

Fluorescence Peak Characterization. In order to assess if the PARAFAC2 model is appropriate for the ICD data, it is necessary to assess the peak shape deformation by formulation factors. We therefore investigate the emission spectra of the native state, at low denaturant concentration, and of the denatured state, at high denaturant concentration. For the investigated proteins, the first data series was used to assess the effect of salt and pH. The second data series, formulated at four combinations of salt and pH, with and without excipients, shows the effect of these excipients on the peak shape. Outliers are visible and are identifiable by either a low signal, a peak deformation, or interference patterns in the background.

To illustrate the peak deformations, we normalize the signal intensity by the protein peak maximum of the fluorescent emission above 306 nm. The acquired emission wavelengths, 300–500 nm, covers also part of the first order Rayleigh scatter peak, originating from excitation at 285 nm. Results shown herein are for monoclonal antibody PPI-03 and PPI-13, at selected formulations. Graphs of other proteins and formulation measurements are found in the [Supporting Information](#); note that these graphs include the curves of outliers, which may appear as inflated backgrounds or low intensity spectra.

The effect of the salt concentration, at 0, 70, and 140 mM NaCl, for PPI-03 at pH 6.5 with 5.38 M GuHCl without added excipient is shown in [Figure 5A](#). It is evident that the emission spectra are highly similar, and by investigating similar plots for all other mAbs and GuHCl concentrations, we conclude that the addition of NaCl does not significantly affect fluorescence emission peak shape in any of the investigated mAbs. Changes in pH mainly affect the fluorescence signal intensity, but a small effect on peak shape is observed. By investigating all of the mAbs, the data show a small peak tailing effect, as is most pronounced for the native state, e.g., see [Figure 5B](#), where PPI-13 is shown at 0 mM NaCl and 0.12 M GuHCl. The peak deformation is minor when compared to the peak shift between the native and denatured states, and peak maxima remain intact. From the second experimental series, the data show that effects of excipients on the native and denatured protein fluorescence emissions spectrum are minimal. In [Figure 5C](#) we show the peak shape of PPI-03 at 5.38 M GuHCl, 6.5 pH and 0 mM NaCl, colored according to excipient. It is clear that the peak shape remains constant.

Between studied proteins, the emission peak shape and intensities can vary. The spectroscopic detector gain was set per protein experimentally, and only the peak position λ_{max} can be compared between proteins. [Figure 6](#) shows the variation in emission peak maxima, in shift and intensity, over the studied proteins. Although measurements were taken per 2 nm in emission wavelength, a second order polynomial fit of 9 data points around λ_{max} was employed to obtain a more refined estimation of the λ_{max} . Within each protein formulation series, we can compare both the intensity maxima and shift between signals from folded and unfolded protein. This further illustrates the input data and sets expectations for PARAFAC2 model results and validation. Trends in fluorescence intensity as a function of pH and salt concentration hold only per protein and, not in general, and have shown both similar and opposing trends between native and denatured states; that is, the intensity of the denatured state's peak of PPI-03 increases with pH and the peak of PPI-13 decreases with pH. However, for both these proteins the native state peaks of formulations at pH 6.5 or 7.0 have the highest intensity. Native proteins show λ_{max} at emission wavelengths of 336–348 nm. For the denatured proteins the emission maximum is shifted to higher wavelengths, 353–358 nm. Denatured monoclonal antibodies can show an increase in fluorescence signal relative to the native state, which, depending on protein and formulation, are observed to increase the intensity by factors between approximately 1.1 and 6.9. [Samples are mixed and measured in different wells in the instrument but originate from one prepared formulation and buffer solution, instrumental pipetting and spectroscopic error margins are not included in the reported factors.] This increase is attributed to increased quantum yield of solvent exposed tryptophan.

Figure 5. Effect of formulation factors on mAb fluorescence emission peak shape. (A) The effect of salt concentration. Plotted PPI-03 at pH 6.5 with 5.5 M GuHCl without added excipient, Colored according to salt concentration. (B) The effect of pH. PPI-13 at 0 mM NaCl, 0.12 M GuHCl without added excipient, colored according to pH. (C) The effect of added excipients, from the second experimental series data of PPI-03 at pH 6.5 and 0 mM NaCl and 5.38 M GuHCl, colored according to excipient (no = no excipient, suc = sucrose, arg = arginine, and pro = proline).

As shown, pH, excipient, and NaCl concentration do not strongly influence peak shape. Of these, pH has the strongest influence. Proteins have varying peak maxima both in intensity and λ_{max} ; however, overall peak shape is preserved. We can make use of the structure in the data to model the chemical denaturation of proteins at varying formulation. Herein, we assume that, at various denaturant concentration, samples are characterizable as mixtures of the native and denatured state.

Protein Model. We have made two types of models: local models (one model per protein) and a global common model

Figure 6. Summary of protein maxima across formulations. Plotted are estimated protein peak maxima by their respective wavelength and intensity. Native proteins (0 M GuHCl) show $\lambda_{\max}$ at emission wavelengths between approximately 336 and 348 nm, whereas denatured proteins (5.5 M GuHCl) show intensities at a higher wavelength, over a small range, between 353 and 358 nm. Please be aware that the data contains outliers.

for all seven proteins. The local model for one single protein negates the differences in folded protein spectra between proteins, while the global model estimates the emission profile as mixtures of the same two underlying phenomena: the native and the denatured state. Thus, we do not need to estimate the number of transitions the system goes through (often a complex estimation) but rather how many underlying phenomena there exists (e.g., solved by investigating the core consistency). As we do not directly model the transition, but rather how the signal develops, even a protein undergoing several transitions will readily be modeled by this methodology. It was important to remove all emission wavelengths below 306 nm due to the Rayleigh scatter. The results from the local models are not shown, as they gave very similar overall results compared to the global model. The main difference is that the local models were better at estimating the emission spectra of the native and denatured state (results not shown).

A PARAFAC2 model is made on the data from all proteins and their formulations for each experimental series. These models are validated, a two-component model is selected, and outliers are identified. The model is recalculated without outliers and validated. Although the native states of the proteins are distinctly different, a global model is able to capture the peak shift by linear combination of the

components in the spectral mode. Figure 7 shows the model of the first experimental series. The spectrum mode A (Figure 7A) shows the two components, where a linear combination of the two components can be used to model the transition induced shift from native (336 and 348 nm) to denatured (353–358 nm). The protein and formulation mode D (Figure 7B) shows the 180 samples' overall intensity as a mixing of the two spectral shift components; therefore, the mode scores represent a mix of peak intensity and shift and can only be compared within a single protein's formulation series.

Figure 8 shows the model residual and the model scores of PPI-13 formulations, colored by formulation pH. At low denaturant concentration the positive component 1 scores in combination with low component 2 scores correspond to a low $\lambda_{\max}$ peak. Across increasing denaturant concentration, this peak shifts to a higher $\lambda_{\max}$ with negative component 1 mode scores. Outliers (seen as spikes in Figure 8A) are visible as sharp deviations from the trends in Figure 8B. Two transitions are visible.

Outlier Trends. From the combined PARAFAC2 model the distribution of errors within the instrument were analyzed and compared between the two experimental series. Outliers showed an uneven distribution across denaturant concentrations, observable as a disproportional number of high mean residuals at specific denaturant concentrations. The denaturation concentration, and the number of outliers, can be related back to the experimental setup. In the experimental setup, protein solution, blank buffer solution and denaturant buffer solutions were pipetted into four rows (A - D) of 12 wells robotically, with increasing denaturant concentration along the rows, i.e., A1 (0.00 M GuHCl), A2 (0.12 M GuHCl), ..., D12 (5.50 M GuHCl). The fraction of errors in both experimental screenings is shown in Figure 9. The first experimental screening (top) showed 41% outliers from measurements of well A1 and an overall high number of outliers in wells neighboring A1. The second experimental screen showed 37% outliers from well A1, and a much reduced but still elevated number of errors in wells directly neighboring well A1.

Outlier Correction. We apply the proposed outlier correction method to the outliers found from the PARAFAC2 model based on all emission spectra of the first experimental data series. In Figure 10A we show how, for PPI-01, pH 7.5, and 0 mM NaCl, outliers have been corrected in an iterative manner. This sample represents a denaturation series with the highest number of outliers (14 out of 48), in combination with neighboring outliers and outliers bordering the low denaturant concentration edge. To find and correct the outliers, first the

Figure 7. PARAFAC2 model scores on the spectral mode, mode A (A), and sample mode (B) for the global model.

Figure 8. PARAFAC2 model of PPI-13, shown per formulation over the denaturation concentrations and colored by formulation pH. (A) The residuals sorted according to increasing denaturant concentration $[GuHCl]$ (mol/L). (B) The chemically induced unfolding curves B_k , representing the mixing of the two components.

Figure 9. Fraction of detected outliers per well for (A) first and (B) second experimental screening.

outliers (marked blue) are identified based on the residuals. The model B_k mode scores (in black) are corrected to obtained corrected scores (in green).

The corrected outlier mode scores are combined with nonoutlier scores to obtain the reconstituted data, $\hat{X}$. Figure 10B compares the corrected denaturant intensity ratio I_{350}/I_{330} , to show improvement in extracted unfolding pattern of PPI-01, pH 7.5 at 0 mM NaCl. By comparing the black and the green line of Figure 10B it becomes evident that any extracted parameter (such as ΔG or c_m) from this ICD experiment will be more correct in the case of the green line than the corresponding line using the black line (with, or without the outliers marked with a red circle).

It should be noted that automatic outlier detection and correction is filled with pitfalls. The proposed method only detects large outliers, and is thus rather conservative in the detection. The correction uses the fact that most nonlinear behavior can be described by a second order polynomial fitting, if the range is short enough. We therefore deem that our suggested method is a gentle and conservative approach to outlier detection and outlier correction.

DISCUSSION

The limitations and extensions of the proposed method and the application thereof lie with the experimental series, and the

Figure 10. (A) Corrected B_k mode score of PPI-01, pH 7.5 and 0 mM NaCl. From the model scores (blue), the outliers (marked in red) are detected based on the residuals (Supporting Information). After iterative correction, the corrected denaturant scores are obtained (green). (B) Effect of reconstitution and correction on the ratio I_{350}/I_{330} , derived from the combined protein model of the first experimental series. The most compromised sample, PPI-01, formulated at pH 7.5 at 0 mM NaCl over denaturant concentration is shown. The ratio I_{350}/I_{330} of the input data (black) and the reconstituted data without correction (blue) and with correction (green) are given. Detected outliers are marked red.

data, with the errors and their structure and with the applicability of PARAFAC2 to the data and prospective data.

By choosing full factorial experimental designs in the two experimental series, the obtained information has multiple query points on few formulation parameters, such as that patterns therein can be discerned. With regards to the modifier of ionic strength, we used 0, 70, and 140 mM NaCl. It is important to note that the ionic strength arising from the use of 10 mM tris and histidine buffers is relatively small, i.e., 18 mM in its extreme at pH 5.0, but may have to be taken into account especially at 0 mM NaCl and 0 M GuHCl. For comparison of the denatured states, the added GuHCl strongly changes the ionic strength of the solution as well.

Over both experimental series the number of detected outliers averaged 8.4%. A structural distribution of the outliers with regards to the instrumentation may indicate issues with the instrument. For the current instrument, we therefore recommend randomization of the denaturant concentrations used across the wells to avoid these systematic offsets.

The combined experimental series resulted in 14 016 emission spectra. For a better overview of the data recorded, we would like to refer the reader to the [Supporting Information](#), and this data set will be available in the future. PARAFAC2 has earlier been applied to multiway chromatography data,³⁰ and here we show the application of PARAFAC2 in a novel application. In this model, we model two shifts, in the spectral mode and in the denaturant mode, simultaneously. It is therefore important to scrutinize created models with regard to rank deficiency in the number of model components. Modeling by PARAFAC2 assumes trilinearity of the data, and validated models show a high fit, lowest at 99.2% for the combined protein model. The high fit, in combination with a large number of outliers, is due to the overall low fluorescence intensity of the outliers. The fit is a valuable statistic in evaluating the validity of the model for nonoutlier data but not in estimation of the number of outliers or the outliers' influence on the model.

CONCLUSION

From investigation of the data of two large experimental screening of formulations in combination with seven mAbs by isothermal chemical denaturation, we conclude the following. (1) The NaCl concentration between 0 and 140 mM does not impact the peak shape of fluorescence emission spectra of the native and denature states. (2) Added excipients, 280 mM proline, 140 mM arginine, or 280 mM sucrose, do not significantly impact the peak shape of fluorescence emission spectra of the native and denature states. (3) The formulation pH has a small influence on the peak shape fluorescence emission spectra of the native and denature states. (4) Effects of NaCl and pH on fluorescence peak intensity do not show trends between mAbs.

The data can be fit using a PARAFAC2 model. Herein, the data is decomposed into a spectral mode, a sample mode, with a sample as a combination of protein and formulation, and a separate denaturant mode for each sample. The denaturant mode treats each denaturation series as a shifted mixing of the native and denatured state spectra. We show how two-component models are capable of capturing the denaturant transition peak shift.

It becomes possible to directly combine and simultaneously analyze the data from all seven mAbs using a single two-component PARAFAC2 model. This model uses the 8640

fluorescence emission spectra, obtaining a model fit of 99.2%, to model the emission shift. For both models we can use the model residuals to identify outliers and show how trends in the outliers are related to the experimental setup. The identified outliers can be corrected, using PARAFAC2 denaturant mode scores, and the original data can be reconstituted with corrected outliers and with removal of the residual variation. Further analysis of the cleaned-up data can be performed using traditional analysis methods, with more valid and robust results than current state-of-the-art methodology.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.9b05748>.

Additional results and divided into three sections: (1) raw data and effect of formulation on tryptophan peak shape, (2) PARAFAC2 models and core consistency, and (3) correction of modeled transition curves (PDF) Animated figure of mode score vs denaturant concentration (j) (MP4)

AUTHOR INFORMATION

Corresponding Author

Åsmund Rinnan – Department of Food Science, Faculty of Life Sciences, University of Copenhagen, DK-1958 Frederiksberg C, Denmark; orcid.org/0000-0002-7754-7720; Email: aar@food.ku.dk

Authors

Dillen Augustijn – Department of Food Science, Faculty of Life Sciences, University of Copenhagen, DK-1958 Frederiksberg C, Denmark; orcid.org/0000-0003-0843-5423

Alina Kulakova – Department of Chemistry, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; orcid.org/0000-0002-5994-1759

Sujata Mahapatra – Department of Chemistry, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; Novozymes A/S, 2800 Kongens Lyngby, Denmark

Pernille Harris – Department of Chemistry, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; orcid.org/0000-0002-6806-4903

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.analchem.9b05748>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was funded as part of the EU Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 675074.

REFERENCES

- (1) Svilenov, H.; Markoja, U.; Winter, G. *Eur. J. Pharm. Biopharm.* **2018**, *125*, 106–113.
- (2) Frokjaer, S.; Otzen, D. E. *Nat. Rev. Drug Discovery* **2005**, *4* (4), 298–306.
- (3) Mahler, H.-C.; Friess, W.; Grauschopf, U.; Kiese, S. *J. Pharm. Sci.* **2009**, *98* (9), 2909–2934.
- (4) Temel, D. B.; Landsman, P.; Brader, M. L. Orthogonal methods for characterizing the unfolding of therapeutic monoclonal antibodies: differential scanning calorimetry, isothermal chemical denaturation,

and intrinsic fluorescence with concomitant static light scattering. In *Methods in Enzymology*; Feig, A. L., Ed.; Elsevier: Amsterdam, 2016; Vol. 567, pp 359–389.

(5) Freire, E.; Schön, A.; Hutchins, B. M.; Brown, R. K. *Drug Discovery Today* **2013**, 18 (19–20), 1007–1013.

(6) Clarkson, B. R.; Schön, A.; Freire, E. *Drug Discovery Today* **2016**, 21 (2), 342–347.

(7) Gentiluomo, L.; Svilenov, H.; Augustijn, D.; El Bialy, I.; Greco, M. L.; Kulakova, A.; Indrakumar, S.; Mahapatra, S.; Morales, M.; Pohl, C.; Roche, A.; Tosstorff, A.; Curtis, R.; Derrick, J.; Noergaard, A.; Khan, T.; Peters, G.; Pluen, A.; Rinnan, Å.; Streicher, W.; van der Walle, C.; Uddin, S.; Winter, G.; Roessner, D.; Harris, P.; Friess, W. *Mol. Pharmaceutics* **2020**, 17 (2), 424–440.

(8) Kamerzell, T. J.; Esfandiary, R.; Joshi, S. B.; Middaugh, C. R.; Volkin, D. B. *Adv. Drug Delivery Rev.* **2011**, 63 (13), 1118–59.

(9) Kolda, T. G.; Bader, B. W. *SIAM Rev.* **2009**, 51 (3), 455–500.

(10) Kiers, H. A.; Ten Berge, J. M.; Bro, R. J. *Chemom.* **1999**, 13 (3–4), 275–294.

(11) Lakowicz, J. R. *Principles of Fluorescence Spectroscopy*, 3rd ed.; Springer Science & Business Media: Berlin, 2006.

(12) Garidel, P.; Hegyi, M.; Bassarab, S.; Weichel, M. *Biotechnol. J.* **2008**, 3 (9–10), 1201–11.

(13) Vivian, J. T.; Callis, P. R. *Biophys. J.* **2001**, 80 (5), 2093–2109.

(14) Eftink, M. R. *Methods Biochem. Anal. Analysis* **2006**, 35, 127–205.

(15) Eftink, M. R. *Biophys. J.* **1994**, 66 (2), 482–501.

(16) Alston, R. W.; Urbanikova, L.; Sevcik, J.; Lasagna, M.; Reinhart, G. D.; Scholtz, J. M.; Pace, C. N. *Biophys. J.* **2004**, 87 (6), 4036–4047.

(17) Najor, M. S.; Olsen, K. W.; Graham, D. J.; Mota de Freitas, D. *Protein Sci.* **2014**, 23 (10), 1392–1402.

(18) Brader, M. L.; Estey, T.; Bai, S.; Alston, R. W.; Lucas, K. K.; Lantz, S.; Landsman, P.; Maloney, K. M. *Mol. Pharmaceutics* **2015**, 12 (4), 1005–17.

(19) Lazar, K. L.; Patapoff, T. W.; Sharma, V. K. *mAbs* **2010**, 2 (1), 42–52.

(20) Augustijn, D.; Mahapatra, S.; Streicher, W.; Svilenov, H.; Kulakova, A.; Pohl, C.; Rinnan, Å. *Eur. J. Pharm. Biopharm.* **2019**, 142, 506–517.

(21) Garber, E.; Demarest, S. J. *Biochem. Biophys. Res. Commun.* **2007**, 355 (3), 751–757.

(22) Ionescu, R. M.; Vlasak, J.; Price, C.; Kirchmeier, M. J. *Pharm. Sci.* **2008**, 97 (4), 1414–26.

(23) Andersson, C. A.; Bro, R. *Chemom. Intell. Lab. Syst.* **2000**, 52 (1), 1–4.

(24) Bro, R.; Andersson, C. A.; Kiers, H. A. J. *Chemom.* **1999**, 13 (3–4), 295–309.

(25) Kamstrup-Nielsen, M. H.; Johnsen, L. G.; Bro, R. J. *Chemom.* **2013**, 27 (5), 99–105.

(26) Bro, R. Multi-way analysis in the food industry-models, algorithms, and applications. University of Amsterdam: Amsterdam, 1998.

(27) Bro, R.; Vidal, M. *Chemom. Intell. Lab. Syst.* **2011**, 106 (1), 86–92.

(28) Thygesen, L. G.; Rinnan, Å.; Barsberg, S.; Møller, J. K. *Chemom. Intell. Lab. Syst.* **2004**, 71 (2), 97–106.

(29) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, 36, 1627–1639.

(30) Bevilacqua, M.; Bro, R.; Marini, F.; Rinnan, Å.; Rasmussen, M. A.; Skov, T. *TrAC, Trends Anal. Chem.* **2017**, 96, 42–51.