Abridged spectral matrix inversion: parametric fitting of X-ray fluorescence spectra following integrative data reduction

Andrew M. Crawford,[a]*‡ Ben Huntsman,[a] Monica Y. Weng,[a] Olena Ponomarenko,[a] Cheyenne D. Kiani,[a] Simon J. George,[b] Graham N. George[a,c]* and Ingrid J. Pickering[a,c]*

[a]Molecular and Environmental Sciences Group, Department of Geological Sciences, University of Saskatchewan, 114 Science Place, Saskatoon, Saskatchewan S7N 5E2, Canada, [b]Simon Scientific, 2000 Allston Way, Unit 232, Berkeley, CA 94701, USA, and [c]Department of Chemistry, University of Saskatchewan, 110 Science Place, Saskatoon, Saskatchewan S7N 5C9, Canada. *Correspondence e-mail: andrew.crawford@northwestern.edu, g.george@usask.ca, ingrid.pickering@usask.ca

Recent improvements in both X-ray detectors and readout speeds have led to a substantial increase in the volume of X-ray fluorescence data being produced at synchrotron facilities. This in turn results in increased challenges associated with processing and fitting such data, both temporally and computationally. Herein an abridging approach is described that both reduces and partially integrates X-ray fluorescence (XRF) data sets to obtain a fivefold total improvement in processing time with negligible decrease in quality of fitting. The approach is demonstrated using linear least-squares matrix inversion on XRF data with strongly overlapping fluorescent peaks. This approach is applicable to any type of linear algebra based fitting algorithm to fit spectra containing overlapping signals wherein the spectra also contain unimportant (non-characteristic) regions which add little (or no) weight to fitted values, e.g. energy regions in XRF spectra that contain little or no peak information.

1. Background

X-ray fluorescence (XRF) analysis is a powerful tool for interrogating the elemental makeup in a wide variety of samples, including living tissues and cells. The method uses characteristic XRF emissions arising from relaxation of higher level electrons following core level electron photoexcitation by an X-ray beam. Since all atoms with absorption edges below the energy of the incident X-ray beam become excited and emit XRF, studies using XRF are capable of simultaneous quantification of all elements with absorption edges at or below the energy of the incident X-ray beam. In most cases, data are acquired using solid-state energy-dispersive detectors, the energy resolution of which, between 100 and 300 eV, is adequate to resolve emissions from most elements following spectral deconvolution. In recent years the method has been expanded, primarily using synchrotron radiation sources, to use micro- and nano-focused X-ray beams to image the elemental distribution in structured samples in a method known as X-ray fluorescence imaging (XFI) (Pushie et al., 2014; Carmona et al., 2019). Additionally, XRF has also been adapted to flow cytometry (Crawford et al., 2016; Crawford & Penner-Hahn, 2018).

Modern XRF data sets are often multidimensional, with individual XRF spectra recorded, on a spatial or time basis, from one or more solid-state detectors. Typically, individual
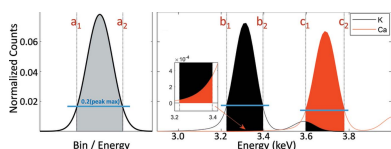
XRF spectra are digitized to 2^{11} or 2^{12} energy points. Two major methods have been used to store and access such multidimensional data sets. The first method is to store the data completely, with emission energy points for each detector channel at every sample location, or time, being stored, typically as 16-bit unsigned integers. A second method is to store the data sparsely by only recording photon event information. This method is used for large array detectors (Ryan *et al.*, 2010, 2014; Siddons *et al.*, 2014); every time a photon is registered by a detector the measured photon energy, the detector element number, and the pixel (or time for flow cytometry) coordinate are stored. Since XRF spectra are typically very sparsely populated, such sparse data storage makes for much more manageable storage requirements and, consequently, much faster data transfer during subsequent data reduction and analysis.

Typically, raw XRF spectra show primary fluorescence from both sample and background, elastic and inelastic scattering, and scatter-excited secondary fluorescence from such sources as the sample holder and the experimental hutch. When processing XRF data, the simplest analytical approach for XFI data is to sum all the counts in an energy region of interest (ROI) corresponding to a particular elemental emission, often referred to as binning. This method is commonly used for real-time data visualization during data acquisition; however, due to the limited energy resolution of the solid-state detectors, and the associated inability to resolve overlapping emissions (*e.g.* K $K\beta$ and Ca $K\alpha$), this ROI approach suffers from overlap of peaks for most samples, and the values obtained should be considered approximations (Pushie *et al.*, 2014; Penner-Hahn, 2014; Crawford *et al.*, 2018). Accurate quantitative analysis requires fitting of the full peak shape (Crawford *et al.*, 2018). A number of analytical software packages exist for the processing of XFI data: *MAPS* (Vogt, 2003), *PyMCA* (Solé *et al.*, 2007), *PyXRF* (Li *et al.*, 2017), *GeoPIXE* (Ryan & Jamieson, 1993; Ryan *et al.*, 1995), *AXIL* (Janssens *et al.*, 1996), *M-BLANK* (Crawford *et al.*, 2019) and *SMAK* (Webb, 2011), with *SMAK*, when fitting, utilizing an embedded version of *PyMCA* (Solé *et al.*, 2007).

For fitting, XRF spectra are typically treated as the sum of a background function together with one or more fluorescence peaks, typically modeled as modified Gaussians (Campbell & Maxwell, 1997; Van Grieken & Markowicz, 2002). The most common approach to account for a background is to initially remove a baseline (Janssens *et al.*, 1996; Van Grieken & Markowicz, 2002; Yi *et al.*, 2015; Ren *et al.*, 2014; Ryan *et al.*, 1988), in many cases, calculated on a per-pixel basis. However, it has recently been shown that per-pixel baselines can introduce significant, systematic errors in quantitation (Crawford *et al.*, 2019). Alternatively, one can usually take advantage of the fact that most XRF images contain regions of pixels which are off the sample of interest; these non-sample pixels can be used to define a background in the form of a blank (Crawford, 2020) (*e.g.* the X-ray fluorescence from the hutch, scatter and supporting template or substrate, but *not* the sample). If a blank is not possible or unavailable, a constant baseline (or no baseline) can be used to background correct the data.

A number of different mathematical treatments can be used for fitting. Currently, the most common approaches include non-negative matrix factorization, non-linear least-squares, and linear least-squares matrix inversion; of the three, the last is fastest. However, regardless of the specifics of fitting, once fitted, each pixel's fluorescence counts are converted to elemental mass per area (*e.g.* μg cm^{-2}) (an areal density) via simple comparison with the counts from standards of known composition collected under essentially identical conditions.

In the last ten years, improvements to X-ray detectors, and associated electronics, have led to considerable improvements in readout speeds and scan times. This has resulted in significantly increased amounts of XFI data, both from increases to the total file size per image and to the total number of images produced (Siddons *et al.*, 2014). For example, Maia (Siddons *et al.*, 2014; Boesenberg *et al.*, 2018; De Andrade *et al.*, 2014; Guimarães *et al.*, 2018; Ryan *et al.*, 2014), a 384 pixel silicon monolith detector, routinely produces the equivalent of more than 3 terabytes (TB) of data for a single XFI scan of 1000 × 1000 pixels. Such high data acquisition rates mean that processing of experimental XRF data may be the rate-limiting step. This in turn results in increased challenges when processing and fitting such data, both in terms of computational limitations as well as the real time required to fit the data sets. The development of rapid fitting methods is therefore important. Given time constraints, image size, and a push for improved resolution, per-pixel dwell times are typically set as short as possible to maximize sample turnover; as a consequence, XRF data are typically collected at or near the limit of detection. As such, XRF spectra are frequently inherently sparsely populated, especially for biological samples where some elements exist at trace and ultra-trace levels; therefore, storing and accessing such data sparsely helps to decrease the overall storage requirements and processing time. Regardless, the need to process and fit data still presents a practical problem, and the need for enhanced or abridged approaches to data processing and reduction remains real. Herein we describe an abridging approach that both reduces and partially integrates XRF data sets to obtain a fivefold total decrease in processing time. We demonstrate our approach using linear least-squares matrix inversion, but note that it is applicable with any type of linear matrix factorization approach [*e.g.* non-negative matrix factorization and dynamic analysis (Ryan & Jamieson, 1993; Ryan *et al.*, 1988, 1995; Ryan, 2001)].

2. Methods

2.1. Sample preparation

2.1.1. Epidermal peel. Leaf tissues from mature *Astragalus bisulcatus* plants were harvested from wild plants growing within the University of Saskatchewan campus near the eastern banks of the South Saskatchewan River. Epidermal peels, with prominent trichomes (leaf hairs), were made from the dorsal leaf epidermis by grasping the epidermis at the leaf edge with fine forceps and carefully peeling the epidermis

away from the remaining leaf tissues. For mounting in the X-ray beam, samples were supported on 250 µm-thick Thermanox plastic coverslips (Nunc) and covered with 6 µm-thick polypropylene film (SPEX SamplePrep, Metuchen, NJ, USA) to prevent sample dehydration. The epidermal peel sample was maintained at room temperature throughout data acquisition.

2.1.2. Nephrogenic systemic fibrosis (NSF) autopsy tissue. The NSF tissue data used for comparison herein have been previously published (George *et al.*, 2010). For detailed specifics of sample preparation, please see George *et al.* (2010) and its associated supplemental information. Briefly, the tissue had been acquired from an autopsy of a 64-year-old patient diagnosed with end-stage renal failure at age 43, who had received a gadolinium-based contrast agent during an MRI and presented NSF symptoms within days. The tissue was formalin-fixed and embedded in paraffin. The section thickness was 20 µm.

2.2. X-ray fluorescence data

2.2.1. Epidermal peel. Using beamline 2-3 of the Stanford Synchrotron Radiation Lightsource (SSRL; Menlo Park, CA, USA), XFI data were collected in fly scan mode with an effective step size of 2 µm, an effective dwell time of 50 ms, and an incident energy of 13.45 keV using a Si(111) double-crystal monochromator. This excitation energy is just below the Br *K* edge and avoids exciting bromine contaminants in plastic components of the beamline. The sample and the detector were oriented at 45° and 90° to the incident X-ray beam in order to minimize the intensity of scattered X-rays. Micro-focusing of the incident X-ray beam to a full width at half-maximum of ∼2 µm was done using a pair of Kirkpatrick–Baez mirrors. X-ray fluorescence was monitored using a Vortex single-element detector running Xspress3 electronics by Quantum Detectors. The size of the collected image was 681 (1.362 mm) × 731 (1.462 mm) for a total of 497811 pixels.

2.2.2. NSF autopsy tissue. XRF microscopy on the 20 µm NSF tissue section was performed using the SSRL beamline 2-3 with a 2.8 × 2.0 µm beam. For further detailed specifics of XRF data collection, please see George *et al.* (2010) and its associated supplemental information.

2.3. Fitting

For the data herein, fitting was performed using the *M-BLANK* (Crawford *et al.*, 2019) software package (https:// drive.google.com/drive/folders/1QSAF_DkE38KEetFQxPEq 2Y9_J5T7Sa2W).

Specifics of data treatment, fitting and peak shape modeling definitions can be found in the supplemental information associated with the work of Crawford *et al.* (2019). Briefly, peak shapes were modeled similarly to the modified Gaussians outlined in the *Handbook of X-ray Spectrometry* (Van Grieken & Markowicz, 2002), and emission energies and branching ratios were referenced using the *xraylib* program from ESRF (Schoonjans *et al.*, 2011) and the *CXRO X-ray Data Booklet* (Thompson *et al.*, 2009), with branching ratios

empirically optimized during parameter creation. Once calculated, peak shapes were held constant and only the amplitudes were floated. This is reflected in the use of equation (2) below. Self-absorption was assumed negligible and was neglected in quantification of the elements in both abridged and full-spectrum fitting.

XFI data were fitted according to the following relations:

$$A_{(C \times E)} \, k_{(E \times P)} = F_{(C \times P)} \tag{1}$$

$$k_{(E \times P)} = \left[A_{(E \times C)}^{\mathrm{T}} A_{(C \times E)} \right]^{-1} A_{(E \times C)}^{\mathrm{T}} F_{(C \times P)} \tag{2}$$

where A is a $C \times E$ (number of digitized *channels* × number of spectral *equations*) matrix of parameterized equations representing the constituent spectral components (*e.g.* fluorescence emissions, elastic and Compton scatter); k is an $E \times P$ (number of spectral *equations* × number of *pixels*) matrix of coefficients; and F is the $C \times P$ (number of *channels* × number of *pixels*) raw data matrix of photon counts. The coefficients of k are solved for in a very straightforward manner using equation (2) to invert the matrix by multiplying both sides of equation (1) by $(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}$.

If the size of F and/or A can be reduced, then the computational burden of matrix inversion can also be reduced, thereby increasing the rate at which the XRF data can be fitted and processed. Additionally, a reduction in the size of F will result in decreased overhead associated with retrieval of the raw data from disk and the subsequent creation of the appropriately sized matrix, F, for algebraic matrix inversion.

This is possible, and the computational requirements associated with fitting an XRF data set can be decreased by reducing one of three things: the number of equations, the number of pixels, or the number of channels (energy points) in the XRF spectra. Since the equations are what allow one to solve for the various spectral components (*e.g.* Ca and K *K*-line emissions) discarding these will result in an under-parameterized equation set (matrix A) and hence errors arising from artifactual changes in the amplitudes of the solved coefficients caused by least-squares minimization of the residual with the under-parameterized set. Similarly, discarding or averaging pixels degrades the spatial resolution of an image, which is also undesirable. A more viable option for reduction of the data size is to decrease the number of channels (energy points) used in fitting. Classically, this has been done by narrowing the fitting frame to the spectral region containing the emission peaks of interest. Here, we take this approach to an extreme by decreasing the size of F to an abridged, discontinuous, set of channels (energy points) constituting the characteristic regions containing the constituent spectral components. This, in turn, also reduces the size of A.

2.4. Abridged spectral matrix inversion via data reduction and partial integration

The abridging approach demonstrated herein consists of two methods. The first method is a data reduction step wherein the total data to be fitted is reduced in size by focusing only on spectrally characteristic regions of the data. The second

Table 1
Comparison of total fitting speeds. The relative speeds of fitting are compared using the data reduction approach and the partial integration approach, both individually and in tandem. Comparison of times via application of the abridging threshold applied at data loading, sparse array creation, and fitting are shown. The fitting times were determined by taking the mean measured time from ten iterations. Threshold in the table is expressed as a fraction of the peak amplitude.

Peel – comparing total fitting speeds.

		Full data fitting #1	Data reduction (at fitting) #2	Data reduction (at restructuring) #3	Data reduction (at loading) #4	Partial integration #5	Data reduction (at loading) and partial integration #6	#7
Threshold		–	0.2	0.2	0.2	–	0.2	0.4
					Time (s)			
Indexing	a)	–	0.00032	4.04	0.002	1.37	0.83	0.60
Loading	b)	1.38	1.37	1.45	2.01	1.42	2.05	1.49
Restructuring	c)	20	21	11.9	11.8	15.9	8.63	7.37
Fitting	d)	33.3	8.49	7.52	8.05	1.07	0.96	1.13
				Total time (s) (summation of rows a through d)				
	e)	54.7	30.9	24.9	21.9	19.8	12.5	10.6
				Percentage (rows a through b divided by row e)				
Indexing	f)	–	0	16.2	0	6.9	6.7	5.7
Loading	g)	2.5	4.4	5.8	9.2	7.2	16.4	14.1
Restructuring	h)	36.6	68.0	47.8	54.0	80.5	69.2	69.6
Fitting	i)	60.9	27.6	30.2	36.8	5.4	7.7	10.6
			Relative times (rows b through d of columns #2 through 7 divided by column #1)					
Loading	j)	–	0.99	1.05	1.46	1.03	1.49	1.08
Restructuring	k)	–	1.05	0.6	0.59	0.8	0.43	0.37
Fitting	l)	–	0.25	0.23	0.24	0.03	0.03	0.03
			Relative fitting speeds (row e column #1 divided by columns #2 through #7)					
	m)	–	1.8	2.2	2.5	2.8	4.4	5.2

method further reduces the total data for fitting by partially integrating the data to be fitted to groups of single channels. The two methods can be performed independently (*i.e.* data reduction or partial integration; columns 2–5, Table 1) or tandemly (*i.e.* data reduction and then partial integration; columns 6–7, Table 1).

2.4.1. Identifying characteristically important spectral regions. The first step in abridging the data is to identify the spectroscopically important regions. These regions are defined based on the user-designated emissions for fitting. The most basic way of identifying which emissions are present in a given data set is to use the integrated spectrum sum. Identification of the elemental emissions present in a given sample is most easily done during optimization of peak fitting parameters. In *M-BLANK*, peak fitting parameters are calculated using the integrated spectra sum (from a sample) corrected by a background function [typically a blank (Crawford, 2020), although baseline correction is feasible]. During fitting parameter optimization, the identified peaks are used to calculate, iterate, and optimize the associated variables (*e.g.* detector energy calibration and response function) governing the peak shapes used for fitting. For a thorough explanation of the equations used for peak fitting, please refer to Crawford *et al.* (2019) and the associated supplemental information. Additionally, during the treatment of individual data sets, it may be necessary to include or remove certain elemental emissions based on

changes in samples (*e.g.* switching from mouse model brain tissue to Cd-doped yeast cells imaged during the same data collection). This does not pose a problem as the incorporation of the new emission lines updates the spectral regions to be retained during fitting. Furthermore, the identification of missing spectral lines is straightforward from an inspection of the residual from fitting.

2.4.2. Data reduction to spectroscopically characteristic regions. XRF data sets contain a series of peaks, and the areas under those peaks correspond to elemental emissions from the sample, and therefore the information one seeks when fitting XRF. As such, regions of the energy spectrum that contain no peaks are less important and can be neglected. These can be eliminated by simply discarding the unimportant energy points (channels). To identify 'unimportant' spectral channels, matrix A [equations (1) and (2)] was used, together with a user-set global threshold, T. This threshold is applied to each column of matrix A, such that all indices with a value less than T multiplied by the maximum value for that column are designated as 'unimportant' (for that column). Then, if a specific channel is designated as 'unimportant' for all the columns, it is then rejected (or not included) during processing/fitting. Fig. 1(A) shows a calculated Gaussian, representative of an arbitrary X-ray emission peak, with a_1 and a_2 indicating the points between which the values of that Gaussian are greater than the threshold being applied (0.2, the

maximum peak amplitude in the figure). Likewise, Fig. 1(B) shows the computed overlapping emissions for K and Ca *K* emission lines. As can be seen, when the threshold for K is applied, all the channels that are identified using that threshold result in those channels being retained for both K and Ca [Fig. 1(B), inset of zoomed region], and vice versa. As can be seen in Fig. 1(C), 0.02 (2%) threshold has been applied to the equations for the various spectral components (*e.g.* Si *K* and K *K*). The resulting discarded spectral regions from each component have been shaded and the spectra have been
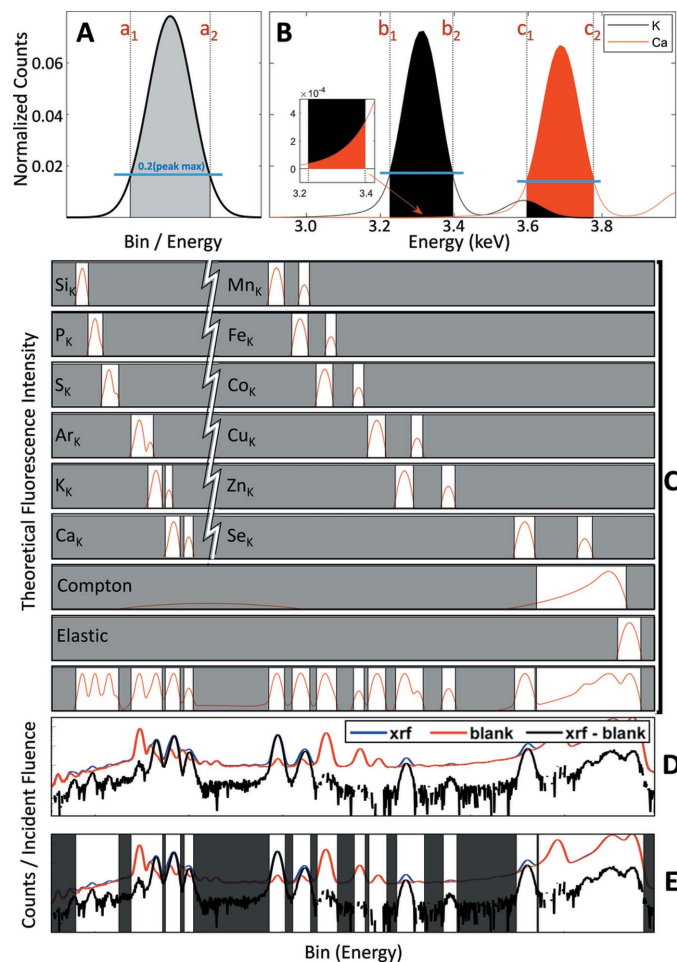

Figure 1
Application of 0.2 abridging threshold: (A) representative Gaussian function demonstrating the use of a 0.2 (20%) threshold (blue line) resulting in all channels not between a_1 and a_2 being discarded. Similarly, (B) shows the retention of values for both equations for the overlapping modified Gaussians for the *K* lines of K and Ca. For both equations, all channels between b_1 and b_2 and c_1 and c_2 are retained; the portion of the K *Kβ* overlapping with the channels identified by the Ca *Kα* peak is retained; (inset) likewise, the low-energy Ca tail which overlaps with the channels identified by the K *Kα* peak is also retained. The *y*-axis units for (A) and (B) are counts normalized to a unit area of 1. (C) shows application of a 0.02 (2%) threshold to the parameterized equations for *K* emissions of Si, P, S, Ar, K, Ca, Mn, Fe, Co, Cu, Zn, and Se, as well as Compton (inelastic) and elastic scatter. The summation of all the parameterized equations is at the bottom. (D) The integrated spectra from the *Astragalus* peel (Fig. 4) along with the blank and corrected spectrum. (E) The same spectra as (D) with the discarded or rejected regions shaded in gray.

plotted with a log scale for *y*. The summation of all the spectral components has been plotted in the bottom row of Fig. 1(C) to illustrate that the important spectral regions are retained in the full spectra. As an example application to actual spectra, Fig. 1(D) shows the integrated spectrum from the *Astragalus* peel used herein, with Fig. 1(E) masking the regions identified in Fig. 1(C) that were determined to be unimportant using the 0.02 threshold.

2.4.3. Partial integration. Once non-characteristic channels have been eliminated (or not), the data can be further reduced in size by partially integrating (*i.e.* summing) spectral regions corresponding to peaks (consisting of multiple channels) to groups of single channel values. Fig. 2 schematically illustrates how each peak is separated and integrated into groups of single-value channels. Data reduction to spectroscopically characteristic regions was not performed for this illustration. To identify the key spectral regions, the equations used in fitting, matrix *A* [equation (2)], are summed horizontally and the second derivative is used to identify inflection points [red dashed lines in Fig. 2(A)]. This operation typically reduces each spectral constituent (*e.g.* Fe *Kα* and Zn *Kα*) to four values representing, respectively, four distinct regions: a low-energy tail, low-energy peak shoulder, high-energy peak shoulder and high-energy tail. An example is demarcated by letters a, b, c, and d, respectively, for the Zn *Kβ* peak in Fig. 2(A). Similarly, letters e, f, g, and h refer to the same respective regions for the Se *Kα* peak. Fig. 2(B) shows the application of the regions to the integrated spectrum of the *Astragalus* peel which reduces 2048 channels to 55. The blue lines with the black circular markers indicate the single values of each region, obtained by integrating the counts between the two corresponding vertical dashed red lines for the blank-corrected spectrum. For example, integration of the regions for Zn *Kβ* and Se *Kα* [*i.e.* letters a–h in Fig. 2(A)] are demarcated in Fig. 2(B) as a′, b′, c′, de′, f′, g′, and h′. Note that there is no separation between the low-energy tail of Se *Kα* [letter e, Fig. 2(A)] and the high-energy tail for Zn *Kβ* [letter d, Fig. 2(A)]. Because of this the high-energy tail of Zn *Kβ* and the low-energy tail of Se *Kα* are integrated into a single representative value [letter de′, Fig. 2(B)].

3. Results

3.1. Effects on fitting speeds

There are three main steps associated with the fitting of XRF data: step (i) the raw data are loaded, *i.e.* transferred from disk to RAM (Table 1, rows b, g, and j); step (ii) once in RAM, the data are restructured into the raw data matrix, *F*, from equation (1) (Table 1, rows c, h, and k); and step (iii) the data are then fitted (Table 1, rows d, i, and l) using the parameterized equations [matrix *A*, equation (2)]. For the abridged methods illustrated herein, identification of the indices for the spectroscopically relevant channels (identified as shown in Fig. 1) introduces a fourth, independent step (Table 1, rows a and f). These indices can be applied to abridge the data at any of the steps (i)–(iii). Application of the energy
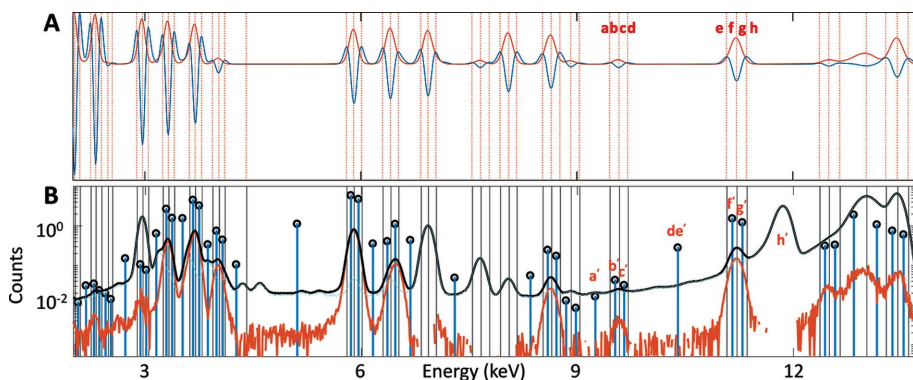
Figure 2
Integration of individually characteristic spectral regions: (A) the summation of the unit area normalized spectral functions used for linear least-squares fitting are summed into a single vector (orange) and the second derivative (blue) is taken to identify the inflection points (vertical dashed red lines). (B) The raw fluorescence from the mean spectrum from the peel (black) has been corrected by the blank spectrum (cyan) to yield the blank-corrected spectrum (red). All channels between each inflection point (vertical black lines) have been integrated into a single value (blue bars with the black circles), from the blank-corrected spectrum, to reduce the data set from 2048 channels to 55.

indices used for fitting as applied during steps (i), (ii), or (iii) is schematically illustrated in Fig. 3, steps a, b, and c, respectively.

The XFI data sets of the epidermal peel used in the fitting comparisons are shown in Fig. 4, including elemental maps [Figs. 4(A) and 4(C)] comprising 24 trichomes, as well as the mean blank-corrected XRF traces for the trichomes [Fig. 4(B)]. To compare the impact of the two abridging methods (data reduction and partial integration), alone or in tandem, the required time for fitting was recorded in Table 1. Additionally, Table 1 also reports the required time for each of the steps (i)–(iii), along with the total time required for processing/fitting, for applying the abridging indices at each of the three steps. The effect of partial integration is also tabulated.

During step (i) (data transfer from disk to RAM), the data can either be fully loaded in one sweep through the disk, or an abridged set can be loaded by reading specific segments and
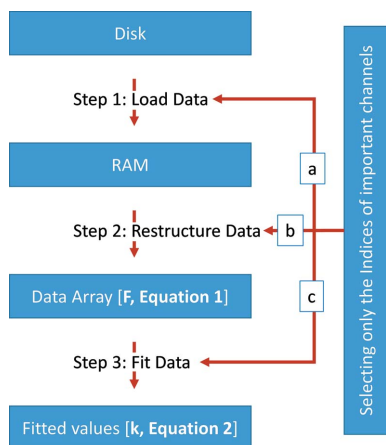

Figure 3
Schematic illustration of indices application: the indices used during data reduction can be applied during one of three different steps, a, b or c, in the above flowchart.

skipping data chunks between them (Table 1, column #4). The latter approach, which applies the indices directly in step (i), resulted in step (i) taking longer (45% more time relative to full data fitting), but resulted in significantly improved times for steps (ii) and (iii) (31% and 76% less time, respectively, relative to full data fitting) and led to a 2.5-fold improvement in the relative total fitting time (Table 1, column #4, row m). This can be seen by looking at Table 1, rows b–d, and comparing columns #4 and #1. Applying the abridging indices in steps (ii) and (iii) yielded improvements in relative fitting speed of 2.2 and 1.8, respectively.

For the application of the abridging method, data reduction by discarding or ignoring non-essential channels to reduce the total amount of data had the greatest overall impact on fitting speed when implemented at the stage of transferring data from disk to RAM and resulted in a total fitting time enhancement of 2.5-fold (Table 1, column #4, row m). Conversely, partial integration alone resulted in an enhancement of 2.8-fold (Table 1, column #5, row m). Using tandem data reduction and partial integration with a 0.2 threshold led to a relative total fitting time that was 4.4-fold faster (Table 1, column #6, row m). Moreover, based on fit quality comparisons (see next section), a 0.4 threshold was more than feasible, yielding further improvements in fitting times. Tandem application of that threshold with partial integration of the remaining spectrally characteristic regions led to an overall fitting time 5.2-fold faster (Table 1, row m, columns #6 and #7, respectively).

3.2. Effects on fitting quality

To investigate the quality of the fits obtained from abridged data, R^2 values [Fig. 5(A)] and per-pixel percent root-mean-squared deviation [Fig. 5(B)] were calculated between the fitted counts for each element from fitting the full data set and those obtained from fitting the abridged data sets created using thresholds from 0% to 99% for Se, Zn, Fe, Mn, Ca, and K.

To investigate the effect, if any, of the abridging methods on absolute quantitation, the 24 trichomes observed in the epidermal peel [Fig. 4(A)] were used to calculate ratios of elemental content derived from fitting the abridged spectra to that from fitting the full spectra. These ratios were then plotted as a function of the maximum elemental value from full-spectral fitting [Fig. 5(C)]. This demonstrates that a 0.4 threshold was more than feasible, enabling the additional improvements in fitting times described above.

3.3. Effects on overlapping emissions

To broaden our study to include more complex spectra, we tested our approach on spectra obtained from NSF autopsy
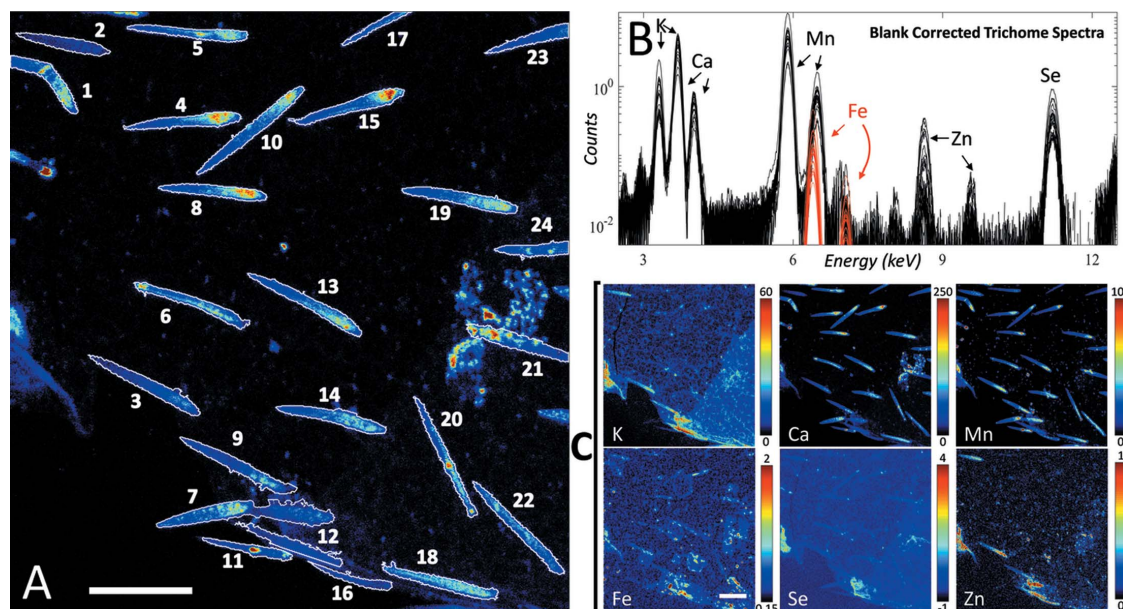
Figure 4
XFI data of the epidermal peel: (A) the Ca image from the fitted XRF from the epidermal peel showing partitioning into 24 individual trichomes. (Note: abridged-integrated fitting and full-spectrum fitting yielded identical images.) (B) Mean blank-corrected XRF traces for the trichomes identified in (A); the overlay of the fitted Fe signal is shown in red. (C) Elemental maps for K, Ca [same as (A)], Mn, Fe, Se, and Zn. Length bars are 500 μm. Intensity scale bar units are μg cm^{-2}.

tissue to check the ability to reliably deconvolute competing L and K emissions. Since NSF tissue contains Gd $L3$, $L2$, and $L1$ emissions which overlap with Mn, Fe, and Zn K emissions, fitting such spectra can be problematic, even when full-spectral fitting is used. Fig. 6 compares the fitted images for full-spectral fitting [Figs. 6(A) and 6(C)] and abridged spectral fitting using a 70% threshold [Figs. 6(B) and 6(D)], with Fig. 6(D) demonstrating how much data were ignored. Fig. 7(A) shows the difference maps calculated by subtracting the map from full-spectral fitting from the maps created using the 70% threshold (obtained in Fig. 6). These maps were then multiplied by 100 divided by the maximum pixel value for that image obtained from full-spectral fitting to give absolute percentage error. The values from each difference map were then histogrammed [Fig. 7(B)] and the corresponding histograms were fitted to a Gaussian [Fig. 7(C)] to calculate the percent standard error associated with abridged spectral fitting of the NSF autopsy tissue using a 70% threshold. As can be seen, all emissions have fractional percentage standard errors, except for Gd $L1$ which has an associated standard error of 1.7%.

4. Discussion

Based on Table 1 results, it is clear that the rate-determining step in processing data when using abridged spectral data sets becomes the rate at which data can be restructured in RAM to the appropriately sized matrices for fitting. This step still takes the greatest amount of time (70% using a threshold of 0.4 with data rejection and subsequent partial integration) and a reduction of the time at this step could significantly enhance the rate further.

For the in-depth analysis of the *Astragalus* peel presented herein, the trichome contents of Se, Zn, Fe, Mn, Ca, and K were compared. For reference, the associated images are presented in Fig. 4(C). As can be seen in Fig. 5(C), Mn, Ca, and K vary negligibly up to a threshold of 0.99; and Se is shown to vary negligibly for thresholds ≤ 0.95. However, Zn and Fe vary negligibly for thresholds ≤ 0.4. The poor performance with Zn and Fe (relative to Se, Mn, Ca, and K) is caused by the extremely low signals of these two elements in trichomes and reflects an introduction of noise [Figs. 5(D) and 5(E) inset table]. The mean counts per pixel for Zn and Fe were $1.8(\pm 2.0)$ and $2.3(\pm 1.5)$, respectively [Fig. 5(E)]. This also can be seen by looking at the spectra in Fig. 4(B) as well as the images in Fig. 4(C). Both Fe and Zn have relatively similar signal intensity; however, Fe has a much worse signal-to-background ratio due to overlap of the Fe $K\alpha$ emission with the intense Mn $K\beta$ peak. Overall, the decreased precision in Zn and Fe is more indicative of the absence of these two elements in the epidermal peel [Figs. 4(C) and 5(D)], rather than being a contraindication of our abridging methods.

Taking the fit of the full data set as ground truth, Fig. 5 explored fidelity of elemental specification and quantification. The reason the trends in Fig. 5(A) fall off at high threshold is a reflection of increased noise at the per-pixel level. This is reflected in Fig. 5(B) which shows the corresponding standard error associated with the comparisons in Fig. 5(A). However, by reflecting on Fig. 5(C), it can be seen that the total content of each of the trichomes does not change, and that such deviations have a negligible impact on the quantitative interpretations of the data.

Ultimately, the applicability of the abridging methods herein depends on the stability of the parameterized equations

used for peak fitting, and is therefore likely sensitive to parameter correlation. As with any fitting methods, the absence (or avoidance) of parameter correlation is very important. For example, one of the major ways that parameter correlation is avoided with many XRF fitting programs is through the use of non-negativity constraints. This is because a large physically impossible negative value for one parameter can offset a large positive value for another parameter during least-squares minimization. In all cases, this type of correlation is caused by inaccurate peak shape parameters. However, with the blank-corrected XRF spectra herein, it is necessary to avoid the use of non-negativity constraints because doing so can lead to false positives (Crawford *et al.*, 2019). As such, the increased scrutiny performed during peak shape parameterization necessary for blank correction fitting of XRF, in many ways, ensured the high degree of fidelity for elemental

specificity and quantification demonstrated for fitting the data herein using our abridging methods [Figs. 5(A), 5(B), and 5(C); and Figs. 6 and 7].

In the epidermal peel, taken from a specimen of the selenium hyperaccumulating plant *A. bisulcatus*, the trichomes were shown to contain low levels of selenium and high levels of calcium. A more complete evaluation of these and other XFI data from *A. bisulcatus* will be provided in a forthcoming paper.

Importantly, our abridged fitting procedure is able to handle the typical challenges of XRF fitting such as pile-up, peak deconvolution, and escape peaks. For the comparisons and trials presented herein, we limited the scope of the study to data that can be fitted using linear least-squares. Issues such as broadening due to count rates that may differ from pixel to pixel, or background that strongly varies from pixel to pixel, must be dealt with using non-linear fitting methods because the shape of the associated equations changes, not just the amplitudes. In this case, the data reduction step, of the abridged approaches presented herein, should still be applicable and yield some improvement in the speed of data processing since the total computational overhead is reduced. This remains to be tested in future studies.

That being said, many issues with XRF data are inherent functions of either the choice of detector (*e.g.* escape peaks) or of the readout electronics employed, and are therefore independent of the fitting method employed. For example, pile-up and changes in the peak shape as a function of count rate both depend heavily upon the readout electronics employed. Thus, analog readout electronics frequently incorporate measures to discriminate pile-up pulses which are then simply discarded, and the approaches employed by the various types of digital signal processing electronics range from ignoring pile-up altogether to employing sophisticated model-based pulse processing algorithms to attempt pile-up recovery. While in these cases we anticipate little difference between full-spectrum and abridged fitting, in the case of some other artifacts, such as distortions of spectral peak shapes at high count rates, post-processing methods such as conventional full-spectrum fitting may be better able to cope.
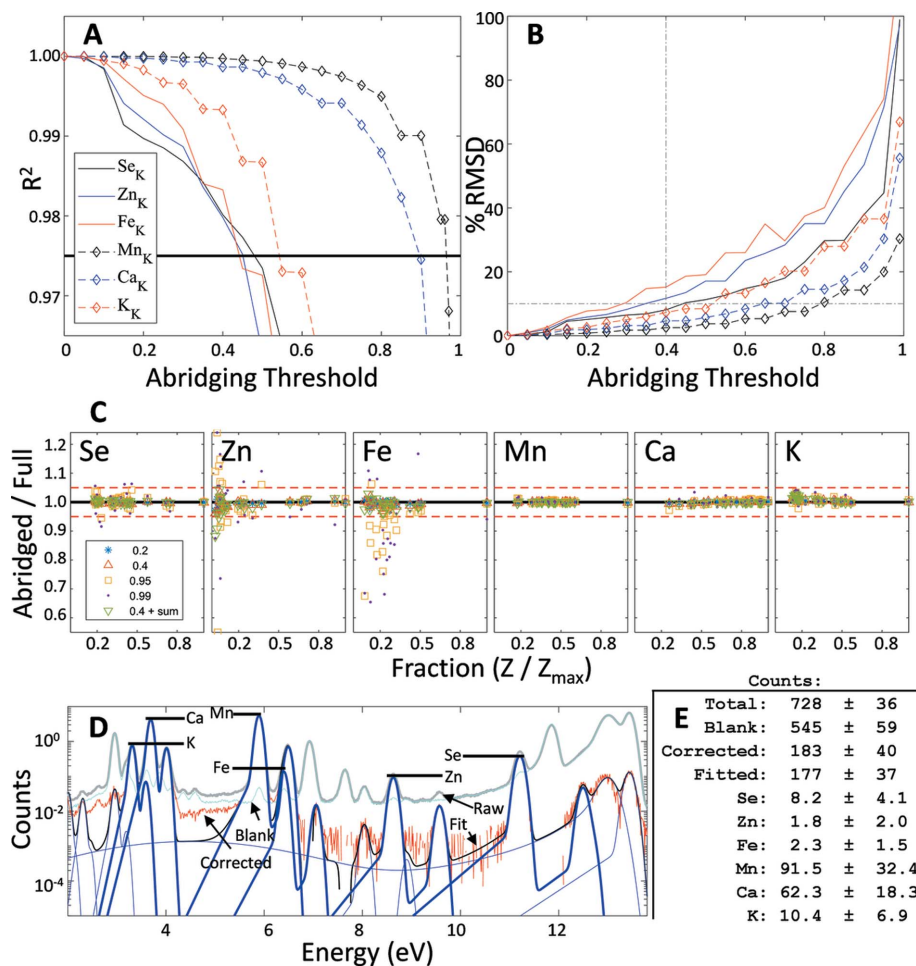

Figure 5
Quantitative comparison: (A) R^2 values calculated between the fitted counts from the full data set and those obtained from abridged data using thresholds from 0% to 99% for Se, Zn, Fe, Mn, Ca, and K. (B) The associated % root mean-squared deviation as a function of abridging threshold applied. (C) The ratios of the fitted elemental values from abridged fits to the same values obtained from fitting the full spectra as a function of the maximum fitted value from the full unabridged fits for the 24 trichomes from Fig. 4. The x axis (Z/Z_{max}) is the ratio of the fitted elemental values from full-spectrum fitting to the corresponding maximum fitted value. The values in the legend are the threshold values applied during data reduction; '+ sum' indicates partial integration was also used. (D) The fitted integrated spectrum for the XRF data for the peel presented herein. (E) The tabulated mean fitted counts per trichome for the 24 trichomes with the uncertainty due solely to background noise (*i.e.* not the uncertainty arising from counting statistics).
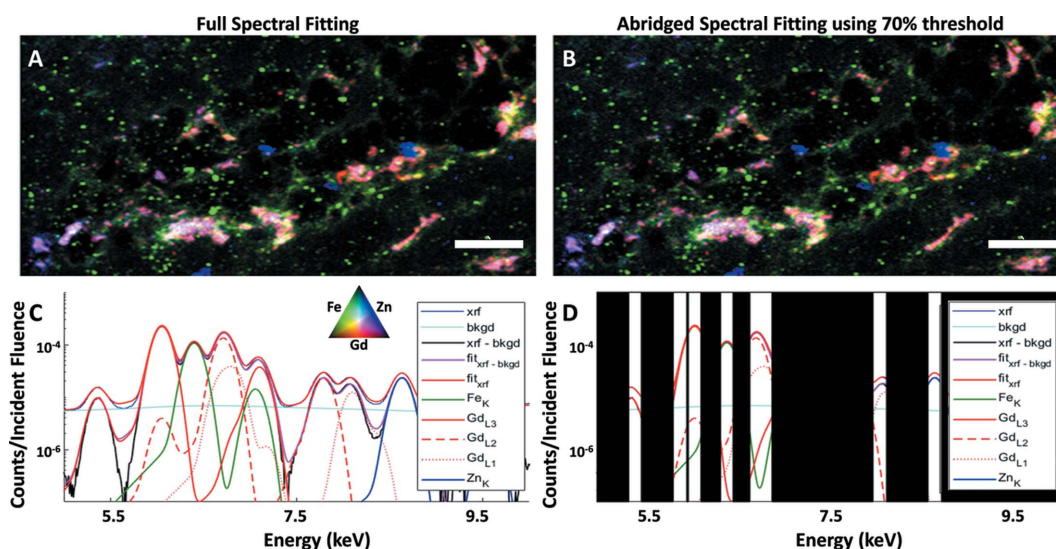
Figure 6
NSF autopsy tissue: (A) the tricolor RGB map (Gd, Fe and Zn, respectively) obtained from full-spectrum fitting. (B) The tricolor RGB map, identical in construction to (A) except the image was obtained using abridged spectral fitting with a 70% threshold. (C) shows the spectral region from 4 to 10 keV (containing the overlapping fluorescence lines for the K emissions from Mn, Fe and Zn, with the $L1$, $L2$ and $L3$ emissions for Gd), for the summation of the fitted spectra from full-spectral fitting (A). (D) shows the same spectral region as (C) for the summation of fitted spectra using abridged spectral fitting with a 70% threshold. Scale bar measures 100 µm horizontally and 10 µm vertically.
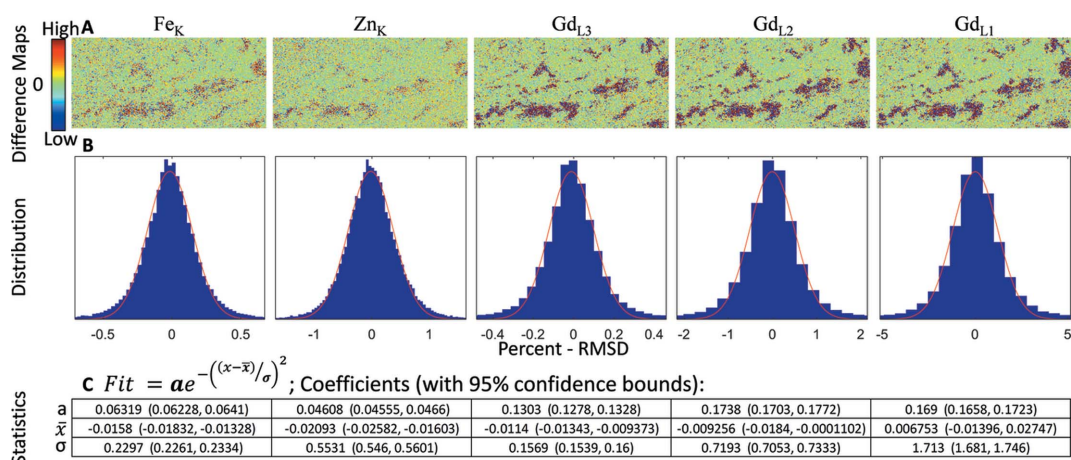

	Fe$_K$	Zn$_K$	Gd$_{L3}$	Gd$_{L2}$	Gd$_{L1}$
a	0.06319 (0.06228, 0.0641)	0.04608 (0.04555, 0.0466)	0.1303 (0.1278, 0.1328)	0.1738 (0.1703, 0.1772)	0.169 (0.1658, 0.1723)
$\bar{x}$	-0.0158 (-0.01832, -0.01328)	-0.02093 (-0.02582, -0.01603)	-0.0114 (-0.01343, -0.009373)	-0.009256 (-0.0184, -0.0001102)	0.006753 (-0.01396, 0.02747)
σ	0.2297 (0.2261, 0.2334)	0.5531 (0.546, 0.5601)	0.1569 (0.1539, 0.16)	0.7193 (0.7053, 0.7333)	1.713 (1.681, 1.746)

C $Fit = ae^{-\left((x-\bar{x})/\sigma\right)^2}$; Coefficients (with 95% confidence bounds):

Figure 7
Error calculations: (A) the percentage difference maps for Fe K, Zn K, Gd $L3$, Gd $L2$, and Gd $L1$, calculated by subtracting the map from full-spectral fitting from the maps created using the 70% threshold (Fig. 6), and dividing by the maximum pixel value for that image. (B) The percentage residuals from (A) were histogrammed and the resulting histograms were fitted to a Gaussian function to calculate the % standard error between full-spectral fitting and abridged spectral fitting. (C) The fitted value for the Gaussian functions calculated in (B) with 95% confidence bounds.

5. Conclusions

We have described an abridging method for data reduction of XFI data sets that provides a 5.2-fold reduction in data processing time relative to the conventional approach of full-spectrum fitting. Our method is generally applicable, functions well even when challenged by overlapping emission lines, and gives negligible changes in quantitation when compared with full-spectrum fitting. As the size of XFI data sets continues to increase due to developments in beamline technology, the implementation of fourth-generation synchrotron sources, and improvements in experimental methodology, speed of data reduction is increasingly important. Moreover, faster XFI data analysis at the beamline might enable real-time visualization of details which would otherwise need to wait for post-beamtime processing. Such capabilities could inform potential new directions for measurements leading to more efficient use of beamtime and providing new opportunities for researchers.

References

Boesenberg, U., Ryan, C. G., Kirkham, R., Jahn, A., Madsen, A., Moorhead, G., Falkenberg, G. & Garrevoet, J. (2018). *J. Synchrotron Rad.* **25**, 892–898.

Campbell, J. L. & Maxwell, J. A. (1997). *Nucl. Instrum. Methods Phys. Res. B*, **129**, 297–299.

Carmona, A., Zogzas, C. E., Roudeau, S., Porcaro, F., Garrevoet, J., Spiers, K. M., Salomé, M., Cloetens, P., Mukhopadhyay, S. & Ortega, R. (2019). *ACS Chem. Neurosci.* **10**, 599–609.

Crawford, A. M. (2020). *Adv. X-ray Anal.* **63**, 180–193.

Crawford, A. M., Deb, A. & Penner-Hahn, J. E. (2019). *J. Synchrotron Rad.* **26**, 497–503.

Crawford, A. M., Kurecka, P., Yim, T. K., Kozemchak, C., Deb, A., Dostál, L., Sun, C.-J., Brewe, D. L., Barrea, R. & Penner-Hahn, J. E. (2016). *J. Synchrotron Rad.* **23**, 901–908.

Crawford, A. M. & Penner-Hahn, J. E. (2018). *Cellular Heterogeneity: Methods and Protocols*, 1st ed., edited by N. S. Barteneva & I. A. Vorobjev, pp. 97–112. Humana Press.

Crawford, A. M., Sylvain, N. J., Hou, H., Hackett, M. J., Pushie, M. J., Pickering, I. J., George, G. N. & Kelly, M. E. (2018). *J. Synchrotron Rad.* **25**, 1780–1789.

De Andrade, V., Ganne, J., Dubacq, B., Ryan, C. G., Bourdelle, F., Plunder, A., Falkenberg, G. & Thieme, J. (2014). *22nd International Congress on X-ray Optics and Microanalysis.* Bristol: IOP Publishing Ltd.

George, S. J., Webb, S. M., Abraham, J. L. & Cramer, S. P. (2010). *Br. J. Dermatol.* **163**, 1077–1081.

Guimarães, D., Roberts, A. A., Tehrani, M. W., Huang, R., Smieska, L., Woll, A. R., Lin, S. & Parsons, P. J. (2018). *J. Anal. At. Spectrom.* **33**, 1616–1630.

Janssens, K., Vekemans, B., Adams, F., van Espen, P. & Mutsaers, P. (1996). *Nucl. Instrum. Methods Phys. Res. B*, **109–110**, 179–185.

Li, L., Yan, H., Xu, W., Yu, D., Heroux, A., Lee, W., -k, Campbell, S. I. & Chu, Y. S. (2017). *Proc. SPIE*, **10389**, 103890U.

Penner-Hahn, J. E. (2014). *Metallomics and the Cell*, edited by L. Banci. Dordrecht: Springer.

Pushie, M. J., Pickering, I. J., Korbas, M., Hackett, M. J. & George, G. N. (2014). *Chem. Rev.* **114**, 8499–8541.

Ren, L. Q., Wu, D., Li, Y. H., Chen, W. R. & Liu, H. (2014). *Proc. SPIE*, **8944**, 89440H.

Ryan, C. G. (2001). *Nucl. Instrum. Methods Phys. Res. B*, **181**, 170–179.

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H. & Cousens, D. R. (1988). *Nucl. Instrum. Methods Phys. Res. B*, **34**, 396–402.

Ryan, C. G. & Jamieson, D. N. (1993). *Nucl. Instrum. Methods Phys. Res. B*, **77**, 203–214.

Ryan, C. G., Jamieson, D. N., Churms, C. L. & Pilcher, J. V. (1995). *Nucl. Instrum. Methods Phys. Res. B*, **104**, 157–165.

Ryan, C. G., Kirkham, R., Hough, R. M., Moorhead, G., Siddons, D. P., de Jonge, M. D., Paterson, D. J., De Geronimo, G., Howard, D. L. & Cleverley, J. S. (2010). *Nucl. Instrum. Methods Phys. Res. A*, **619**, 37–43.

Ryan, C. G., Siddons, D. P., Kirkham, R., Li, Z. Y., de Jonge, M. D., Paterson, D. J., Kuczewski, A., Howard, D. L., Dunn, P. A., Falkenberg, G., Boesenberg, U., De Geronimo, G., Fisher, L. A., Halfpenny, A., Lintern, M. J., Lombi, E., Dyl, K. A., Jensen, M., Moorhead, G. F., Cleverley, J. S., Hough, R. M., Godel, B., Barnes, S. J., James, S. A., Spiers, K. M., Alfeld, M., Wellenreuther, G., Vukmanovic, Z. & Borg, S. (2014). *22nd International Congress on X-ray Optics and Microanalysis.* Bristol: IOP Publishing.

Schoonjans, T., Brunetti, A., Golosio, B., Sanchez del Rio, M., Solé, V. A., Ferrero, C. & Vincze, L. (2011). *At. Spectrosc.* **66**, 776–784.

Siddons, D. P., Kirkham, R., Ryan, C. G., De Geronimo, G., Dragone, A., Kuczewski, A. J., Li, Z. Y., Carini, G. A., Pinelli, D., Beuttenmuller, R., Elliott, D., Pfeffer, M., Tyson, T. A., Moorhead, G. F. & Dunn, P. A. (2014). *J. Phys. Conf. Ser.* **499**, 012001.

Solé, V. A., Papillon, E., Cotte, M., Walter, P. & Susini, J. (2007). *At. Spectrosc.* **62**, 63–68.

Thompson, A., Atwood, D., Gullikson, E., Howells, M., Kim, K.-j., Kirz, J., Kortright, J., Lindau, I., Liu, Y., Pianetta, P., Robinson, A., Scofield, J., Underwood, J., Williams, G. & Winick, H. (2009). *X-ray Data Booklet*, 3rd ed. Lawrence Berkeley National Laboratory, University of California, USA.

Van Grieken, R. E. & Markowicz, A. A. (2002). *Handbook of X-ray Spectrometry*, 2nd ed. Marcel Dekker.

Vogt, S. (2003). *J. Phys. IV Fr.* **104**, 635–638.

Webb, S. M. (2011). *10th International Conference on X-ray Microscopy*, pp. 196–199. AIP Publishing.

Yi, L. T., Liu, Z. G., Wang, K., Chen, M., Peng, S. Q., Zhao, W. G., He, J. L. & Zhao, G. C. (2015). *Nucl. Instrum. Methods Phys. Res. A*, **775**, 12–14.