JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

Received 6 February 2026
Accepted 21 June 2026

Edited by S. Bone, Forschungszentrum Jülich, Germany

Keywords: spectro-microscopy; experimental design; sparse acquisition; matrix completion.

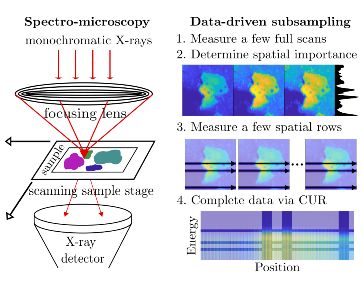
Supporting information: this article has supporting information at journals.iucr.org/s

Reducing acquisition time and radiation damage: data-driven subsampling for spectro-microscopy

Maike Meier,[a,b]* Lorenzo Lazzarino,[a,c] Boris Shustin,[a] Hussam Al Daas[a]* and Paul Quinn[d]

[a]Scientific Computing Department, STFC, Rutherford Appleton Laboratory, Harwell Campus, Didcot OX11 0QX, United Kingdom, [b]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, 9747 AG Groningen, The Netherlands, [c]Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom, and [d]Ada Lovelace Centre, STFC Scientific Computing, Rutherford Appleton Laboratory, Didcot OX11 0QX, United Kingdom. *Correspondence e-mail: m.meier@rug.nl, hussam.al-daas@stfc.ac.uk

Spectro-microscopy is an experimental technique which can be used to observe spatial variations in chemical state and changes in chemical state over time or under experimental conditions. As a result it has broad applications across areas such as energy materials, catalysis, environmental science and biological samples. However, the technique is often limited by factors such as long acquisition times and radiation damage. We present two measurement strategies that allow for significantly shorter experiment times and total doses applied. The strategies are based on taking only a small subset of all the measurements (*e.g.* sparse acquisition or subsampling), and then computationally reconstructing all unobserved measurements using mathematical techniques. The methods are data-driven, using spectral and spatial importance subsampling distributions to identify important measurements. As a result, taking as little as 4–6% of the measurements is sufficient to capture the same information as in a conventional scan.



1. Introduction

X-ray absorption spectroscopy (XAS) is a powerful technique for investigating the chemical state or short-range bonding within a material. Spectro-microscopy combines XAS and microscopy to also understand the spatial variations of these properties in materials. In a typical spectro-microscopy experiment, the sample is rastered through the beam and the absorption or X-ray fluorescence (XRF) across the raster grid to form an image. A schematic overview of a spectro-microscopy experiment is shown in Fig. 1(*a*). This process is repeated for various energies above and below the absorption edge of the element of interest, see Fig. 1(*b*). The resulting stack of absorption or XRF images versus energy typically corresponds to tens or hundreds of thousands of individual spectra. The application of the technique to large regions or *in situ* processes can be limited by long acquisition times and sample degradation due to high X-ray dosage.

Here we set out to mitigate these limitations by reducing the number of measurements taken. When performing a spectro-microscopy experiment, the fact that there are only a few chemical states allows us to greatly reduce the data using techniques like principal component analysis (PCA) (Lerotic *et al.*, 2004; Lerotic *et al.*, 2005). Datasets are reduced to a few core spectral components and a map describing the mixture of those components. Knowing that there are only a few spectral components and the data can be reduced or compressed, our acquisition process can be designed to sample the data in different ways to exploit this. In designing an efficient acquisition scheme, it is necessary to incorporate (mechanically) practical rastering or continuous scanning. In particular, we propose two novel sparse data acquisition schemes that are data-driven in nature. The schemes result in different strategies for selecting subsampled raster scans, see Fig. 1(*c*). Whereas every spatial row of pixels is measured in full raster scans, only a subset of the rows are measured in a subsampled raster scan.

Townsend *et al.* (2022) proposed a sampling scheme, *robust random raster sampling*, where the sampled rows are selected at random for each energy, while ensuring that every row of the specimen is sampled at least once. It is shown that with this approach one can achieve a subsampling ratio of 10–20% routinely, without noticeably sacrificing the quality of the dataset. The subsampling ratio refers to the percentage of measurements that are taken compared with the total number of measurements for a normally sampled scan. The datasets are of the same dimension, yet only 15–20% of the entries will be measured. The rest of the entries are reconstructed with a matrix completion algorithm. The aforementioned matrix completion approach was further generalized to tensor completion methods by Townsend *et al.* (2025), allowing further reduction in the subsampling ratio to 10% without significant degradation of the dataset.

The Robust random raster sampling approach is used as a baseline. We improve upon this by adopting a data-driven, adaptive approach to measurement selection, rather than relying on fully random sampling. This method intelligently chooses which spatial rows to sample based on the data itself, enabling a significant reduction in subsampling ratios—well beyond what has previously been achieved—without compromising the quality of the reconstructed dataset. Our results show that high-quality reconstructions are possible even at subsampling ratios as low as 4–6%. Across all tested ratios, the data-driven techniques consistently outperform uniform random sampling.

Reducing the measurement time, while still addressing the statistical significance and signal-to-noise required across the energy range in XAS, is critical to broadening the impact of XAS spectro-microscopy. While the increased flux of new sources helps to accelerate measurements, subsampling can further accelerate these approaches and help to manage dose.

1.1. Spectro-microscopy

In a spectro-microscopy experiment (Braun, 2005; Miller *et al.*, 2018; Hitchcock *et al.*, 2005), we have a sample that is exposed to X-rays of varying energy, measuring essentially spatially resolved XAS scans, as depicted in Figs. 1(*a*) and 1(*b*). We argue here that the data obtained during a spectro-microscopy experiment are fundamentally low-dimensional, which allows the subsampling of measurements.

In a spectro-microscopy experiment, one measures the absorption $I(x, y, E)$ of the sample at position (x, y) and at energy E, having fixed the incident flux $I_0(E)$. The optical density $D(x, y, E)$ is determined as $D(x, y, E) = -\ln[I(x, y, E)/I_0(E)]$ and is related to the absorption spectrum and thickness of the sample. Specifically, if the sample consists of one material with photon-energy-dependent linear absorption $\mu(E)$ at energy E, then the transmission image at (x, y) is $I(x, y, E) = I_0(E)\exp[-\mu(E)\,t(x, y)]$, where $t(x, y)$ is the thickness of the absorbing material. As a result, the ground truth optical density is linear in thickness $t(x, y)$ and absorption $\mu(E)$. In particular, for the exact ground truth optical density, $D_{\rm exact}(x, y, E) = \mu(E)\,t(x, y)$.

If there are S spectroscopically distinguishable components in the sample with absorption spectra $\mu_s(E)$ and thickness maps $t_s(x, y)$, $1 \le s \le S$, then we have

$$D_{\rm exact}(x, y, E) = \sum_{s=1}^{S} \mu_s(E)\, t_s(x, y). \qquad (1)$$

Mechanically, in a spectromicroscopic experiment, the incident beam is set to some energy and the sample is moved through the beam in a raster-like fashion, covering the sample either row-wise or column-wise (see Fig. 1). The spatial layout of the sample is divided into a grid of pixels, say $n_X \times n_Y$, for a total of $N = n_X n_Y$ pixels and the transmission image is measured for each of the pixels. This procedure is repeated for various discretized energies $E_1, \ldots, E_{n_E}$. Finally, after normalizing with respect to the incident flux and computing the logarithm, we have n_E datasets of size $n_X \times n_Y$ of the optical density. Throughout this work, we flatten these datasets to a single matrix $\mathbf{D}$ of size $n_E \times N$, as is typically done for analysis purposes (*e.g.* Lerotic *et al.*, 2004; Lerotic *et al.*, 2005), where each row represents a vectorized XRF map and each column corresponds to the measured absorption spectrum for a single pixel. Assuming noiseless data, equation (1) translates into a PCA-like factorization $\mathbf{D}_{\rm exact} = \mathbf{MT}$, where $\mathbf{M}$ is the discretized spectra and $\mathbf{T}$ is the thickness map. In practice the

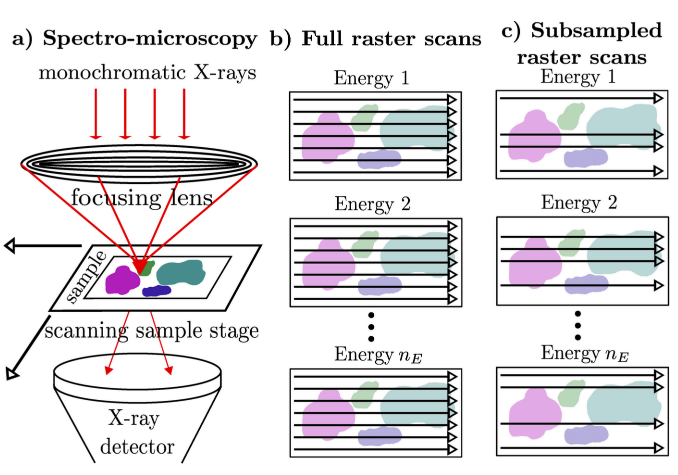


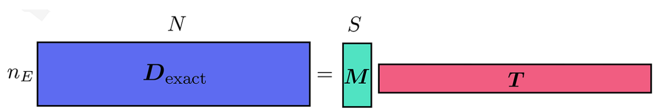
Figure 1
(*a*) A schematic overview of a spectro-microscopy beamline [inspired by Hitchcock *et al.* (2005)]. X-rays are focused onto a sample consisting of multiple chemical states with distinct absorption spectra. The transmitted signal is then measured at the detector. The sample is rastered through the focused beam, pixel-by-pixel, row-by-row. This process is repeated for different X-ray energies as shown schematically in (*b*). (*c*) An alternative measurement approach, where not every row of pixels is measured for each energy. This strategy is later referred to as *random raster sampling*.

data collected are noisy and we have $\mathbf{D} = \mathbf{MT} + \mathbf{G}$, with $\mathbf{G}$ representing the noise. The fundamental goal of a spectro-microscopic experiment is to approximate $\mathbf{M}$ and $\mathbf{T}$ (or only $\mathbf{M}$) using the noisy dataset $\mathbf{D}$.

1.2. Low-dimensionality of spectro-microscopy data and redundancy in measurements

The flattened dataset $\mathbf{D}$ of the sample's optical density has underlying low-dimensionality. The low-dimensional structure of the spectro-microscopy data is crucial to both subsampling techniques as well as the standard computational data analysis routine.

The ground-truth optical density is exactly low-rank; it can be represented as the product of the matrix of spectra and the matrix of (flattened) thickness maps:



This is a visual display of (1); the data in $\mathbf{D}_{\text{exact}}$ can be represented as the matrix-multiplication of $\mathbf{M}$ and $\mathbf{T}$, where $\mathbf{M}$ contains the absorption spectra of the different chemical states and $\mathbf{T}$ contains the thickness of the chemical states per pixel. Since there are much fewer chemical states (S) than energies (n_E) and pixels (N), the matrices $\mathbf{M}$ and $\mathbf{T}$ have one small dimension which results in the low-dimensional structure of $\mathbf{D}_{\text{exact}}$.

The measured optical density $\mathbf{D}$ is naturally noisy, so it will not have exact low-rank structure. However, the directions in the data corresponding to the S different materials remain dominant over the directions corresponding to noise—leading to approximate low-dimensionality.

This low-dimensionality means there is a large redundancy in the measurements. We demonstrate this using a low-rank decomposition technique called a *CUR decomposition*. Assume for simplicity that we have a dataset $\mathbf{D}_{\text{exact}}$ corresponding to a sample with $S = 2$ spectroscopically different components. Select two energies and measure full raster scans for these. These data correspond to two rows of $\mathbf{D}_{\text{exact}}$, which we will call $\mathbf{R} \in \mathbb{R}^{2\times N}$. Similarly, imagine we could measure the full absorption spectra for just two pixels, *i.e.* two columns of $\mathbf{D}_{\text{exact}}$, and collect these data in $\mathbf{C} \in \mathbb{R}^{n_E\times 2}$. We can reconstruct all of $\mathbf{D}_{\text{exact}}$ using just $\mathbf{C}$ and $\mathbf{R}$.

In particular, under only the assumption that the rows in $\mathbf{R}$ and the columns in $\mathbf{C}$ are linearly independent, we can form the CUR decomposition. Let $\mathbf{U} \in \mathbb{R}^{2\times 2}$ denote the intersection of $\mathbf{C}$ and $\mathbf{R}$, then $\mathbf{D}_{\text{exact}} = \mathbf{CU}^{-1}\mathbf{R}$. This is graphically displayed in Fig. 2.

If we were able to take noiseless measurements, this means only $2(n_E + N)$ measurements rather than $n_E N$ measurements are required to map the chemical state of a sample. This represents an enormous speed-up in the experiment without losing any information!

Our measurements are, however, naturally noisy, which means the dataset $\mathbf{D}$ is not exactly low-rank, but only

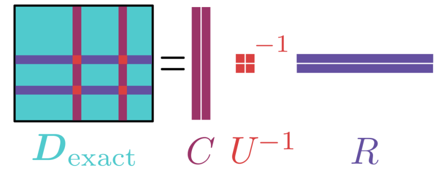


Figure 2
A matrix of exact rank S (in this case $S = 2$) can be represented using just S of its columns and S of its rows using the CUR decomposition. The only requirement is that the rows and columns are linearly independent. Here, the construction of $\mathbf{C}$, $\mathbf{U}$ and $\mathbf{R}$ from $\mathbf{D}_{\text{exact}}$ is shown with color: $\mathbf{C}$ consists of the burgundy columns of $\mathbf{D}_{\text{exact}}$, $\mathbf{R}$ of the two purple row, and $\mathbf{U}$ of the two-by-two matrix where the rows and columns overlap. Experimentally, $\mathbf{C}$ represents two pixels measured for all energies and $\mathbf{R}$ represents two energies for which all pixels are measured.

approximately low-rank. As a result, any low-rank decomposition like the CUR decomposition will only approximate $\mathbf{D}$. Yet, it will still be true that $\mathbf{D} \simeq \mathbf{CU}^{-1}\mathbf{R}$ due to the low-dimensional nature. Practically, we will have to *oversample* slightly: measuring a few more than two columns and rows to obtain an accurate approximation. We elaborate on this in Section 2.1.

The CUR decomposition is just one approach to illustrate the redundancy present in $\mathbf{D}$. Townsend *et al.* (2022) exploit the fact that there exists an approximate low-rank decomposition $\mathbf{D} \simeq \mathbf{XY}$, where $\mathbf{X} \in \mathbb{R}^{n_E\times r}$ and $\mathbf{Y} \in \mathbb{R}^{r\times N}$ have one small dimension $r \ll n_E, N$. They suggest using subsampled raster scans chosen at random to reduce the acquisition time, see Fig. 1(*c*), combined with a matrix completion algorithm, a looped alternating gradient-descent algorithm (LoopedASD), that tries to find such $\mathbf{X}$ and $\mathbf{Y}$. Generally speaking, the problem of matrix completion is to find the unknown entries of a matrix that is partially observed but has some underlying structure, such as low-dimensionality.

The process is illustrated in Fig. 3. By employing the subsampled raster scans, we observe just bits of the matrix $\mathbf{D}$. Let $\mathcal{P}(\mathbf{D})$ denote the matrix $\mathbf{D}$ with each unobserved entry set to zero. We then look for $\mathbf{X}$ and $\mathbf{Y}$ such that $\mathcal{P}(\mathbf{D}) \simeq \mathcal{P}(\mathbf{XY})$; that is, $\mathbf{X}$ and $\mathbf{Y}$ are fitted only to the known entries. Although we have then only observed a subset of the measurements in $\mathbf{D}$, we hope that if $\mathcal{P}(\mathbf{D}) \simeq \mathcal{P}(\mathbf{XY})$ it will be true that $\mathbf{D} \simeq \mathbf{XY}$. Townsend *et al.* (2022) use an alternating gradient descent-type algorithm, the so-called LoopedASD, to compute these matrices $\mathbf{X}$ and $\mathbf{Y}$. This method is described in detail by Townsend *et al.* (2022).

Fundamentally, low-dimensionality not only allows us to compress data or analyze it with a PCA-like technique but also to measure just a fraction of it and predict the rest reliably.

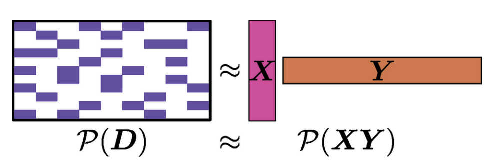


Figure 3
A graphical representation of random raster subsampling combined with the LoopedASD algorithm. Random raster sampling results in a dataset where the measurements correspond to blocks of entries for each row (left side). The LoopedASD algorithm aims to find two factors $\mathbf{X}$ and $\mathbf{Y}$ so that $\mathbf{XY}$ fits the observed measurements: $\mathcal{P}(\mathbf{D}) \simeq \mathcal{P}(\mathbf{XY})$.

1.3. The analysis of spectro-microscopic data

In a perhaps more familar context, the low-dimensional nature is already exploited in the analysis of $\mathbf{D}$. Specifically, the analysis exploits the fact that the number of actual chemical states within the sample under study will be small and significantly less than the number of sampled positions (Lerotic *et al.*, 2004). This allows the data to be described, to a high degree of accuracy, by a mixture of a small set of components.

In particular, one first computes the eigenimages and eigenspectra corresponding to the data. Let $\mathbf{D} = \mathbf{Q}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$ be a singular value decomposition of $\mathbf{D}$, where $\mathbf{Q} \in \mathbb{R}^{n_E \times n_E}$ and $\mathbf{V} \in \mathbb{R}^{N \times n_E}$ are orthonormal matrices and $\boldsymbol{\Sigma}$ is diagonal. The leading eigenimages are the first few rows of $\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$, and the corresponding eigenspectra are the leading columns of $\mathbf{Q}$.

Cluster analysis is performed on the eigenimages to cluster pixels that have similar eigenspectra. Finally, the average absorption spectra for a cluster can easily be found using the mean of the measured data for each cluster. We denote these spectra by $\hat{\mathbf{M}}_{\mathrm{cluster}}$. In the ideal case—if the clusters correspond to a singular component—these cluster spectra resemble the true spectra $\mathbf{M}$; if the clusters contain different components the cluster spectra will be linear combinations of the true spectra.

The cluster-analysis approach from Lerotic *et al.* (2004) is used to analyze spectro-microscopic data on the I14 beamline at Diamond Light Source, which is why we use it as a tool for comparison. However, we note that any analysis method can be used in combination with our proposed algorithms since our approaches reconstruct full datasets.

1.4. Subsampling spectro-microscopy measurements and data-driven approaches

We have argued that it is possible to reduce the amount of measurements, while still recovering a good approximation to the full dataset $\mathbf{D}$. There are two main questions that now arise: (1) what measurements should we take? (2) how do we then reconstruct the full data? The answer to the first question is restricted by the mechanical apparatus. Traditional fly scanning approaches necessitate the scanning of full spatial sample rows (or columns), rather than individual pixels. As a result, the prior question boils down to a technique to select which spatial rows to sample.

In previous a work (Townsend *et al.*, 2022), for each energy, sample rows were selected uniformly at random. The completion of the missing data was then done using LoopedASD. Another sampling approach, which is more closely related to our approaches, is presented by Quinn *et al.* (2024). In the aforementioned work, energy samples are selected via a model order reduction technique, using a low-rank approximation of reference spectra or spectra measured in small parts of the specimen. Then, the full spectrum is recovered via the previously formed low-rank approximation.

Recently, another line of works for subsampling the spectra uses methods of Baysian optimization. An AI-base approach of selecting energies to sample was presented by Du *et al.* (2025), where Bayesian optimization and Gaussian Process are the underlying techniques. In Ito *et al.* (2025), Bayesian experimental design with Gaussian Process Regression is the underlying technique.

We propose two new strategies for selecting which measurements to take—both are *data-driven*. One of these strategies is based on the CUR decomposition (Drineas *et al.*, 2008) and naturally allows for an optimization-free completion. The second strategy employs the same LoopedASD completion, although we note that any other matrix completion algorithm could also be used, as the sampling is independent of the matrix completion algorithm.

We motivate our data-driven approaches over (uniform) random subsampling approaches in Fig. 4 by comparing three different subsampling strategies. *Random raster sampling* refers to the method of Townsend *et al.* (2022), where for each energy a number of spatial rows is measured according to a uniform random distribution, and the matrix completion is done with LoopedASD. *Data-driven raster sampling* is similar, but the spatial rows are selected based on an importance distribution on the sample (details on the importance distribution are given in Section 2.3). *Data-driven CUR sampling* refers to taking measurements that correspond to full rows and columns in $\mathbf{D}$. That is, full XRF scans for some energies (rows) and measurements of some spatial rows for every energy (column blocks). As previously, the spatial rows are selected based on an importance distribution on the sample.

Fig. 4 shows that we are able to retrieve a very large portion of the information accurately with the two new data-driven methods for very small subsampling ratios. The data-driven nature allows us to concentrate measurements where they are most useful, achieving subsampling ratios as small as 4% while accurately retrieving 90% of the information. Additionally,

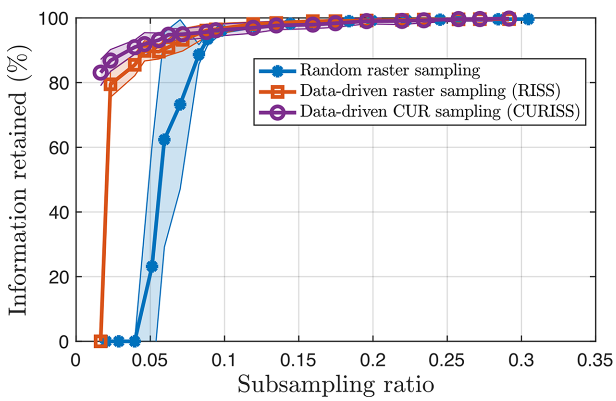


Figure 4
The amount of information retained with different subsampling strategies. The information measure is $(1 - \|\mathbf{D} - \mathbf{D}_{\mathrm{completed}}\|_{\mathrm{F}}) / (1 - \|\mathbf{D} - \mathbf{D}_{\mathrm{optimal}}\|_{\mathrm{F}}) \times 100\%$, where $\mathbf{D}_{\mathrm{completed}}$ is the completed dataset (with filled in entries) and $\mathbf{D}_{\mathrm{optimal}}$ is the best rank-10 approximation to $\mathbf{D}$ according to the Eckart–Young–Mirsky theorem (Eckart & Young, 1936), *i.e.* truncated rank-10 singular value decomposition (SVD). Solid lines show the mean across trials, while shaded bands represent ± 1 standard deviation. The dataset is DS1 from Townsend *et al.* (2022). For the raster sampling approaches, completion is done with LoopedASD, and the data shown are the average of ten iterations. For the CUR sampling approaches, the CUR decomposition is the completion and the data shown are the average of 100 iterations.

the figure shows that CUR completion is much more reliable than LoopedASD completion for very small subsampling ratios.

The methods we suggest are based on recent theory in randomized numerical linear algebra, where the CUR decomposition is a well established tool. To approximate a general (approximately low-rank) matrix with a few number of its rows and columns, one needs to carefully select particular rows and columns that contain the most necessary information. The measures for the 'information-richness' of rows or columns are called *leverage scores* (Mahoney, 2011). Our data-driven methods are based on leverage-score-type schemes to ensure we are optimally exploiting every measurement. In Section 2 we explain the importance distributions, and in Section 3 we suggest schemes for reducing spectro-microscopy acquisition times.

1.5. Overview of proposed methods

In this subsection, we summarize our main contributions. Firstly, we propose two data-driven sampling approaches for spectro-microscopy, aimed at reducing experiment time and dose applied on the specimens. Both strategies incorporate the use of statistical leverage scores to determine the importance distributions, both of the energy levels (when a spectral dictionary is available) and the spatial rows of the sample (Section 3). The first sampling strategy allows the inpainting of missing entries via any matrix completion algorithm such as LoopedASD (Subsection 3.1). We call this strategy RISS. The second sampling strategy is designed specifically to allow the use of the CUR matrix factorization for approximating the missing entries (Subsection 3.2). We call this strategy CURISS.

Secondly, we propose and explore adaptive approaches and stopping criteria for sampling (Subsection 3.2.1). We call this strategy ACURISS.

The RISS measurement strategy consists of the following steps:

Step 1. Determine a *spectral importance subsampling distribution* (see Fig. 5) on the different energies using an (approximate) dictionary of chemical states. If an energy is assigned a larger value in the subsampling distribution, that particular energy should contain more information. If a dictionary is not available, a uniform distribution can be used.

Step 2. Sample a small number of energies from the probability distribution (that is, more important energies are sampled with higher probability). For the sampled energies, scan all the pixels as in a standard spectro-microscopy experiment [see Figs. 1(*a*) and 1(*b*)].

Step 3. Use the small number of full scans to determine a *spatial importance subsampling distribution* (see Fig. 6) on the spatial rows of the sample. Again, a larger value represents more information being present in the particular spatial row of the sample.

Step 4. For all the energies that were not yet measured, we sample a small number of spatial rows from the probability distribution. A new set of rows is sampled for each energy. These rows are then measured.

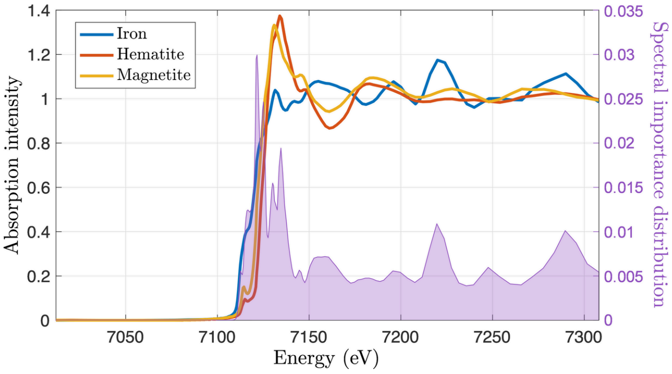


Figure 5
The spectral importance subsampling distribution (purple) for prior knowledge on the absorption spectra of three (potentially) present iron/iron-oxides. The incident energies (on the *x*-axis, in eV) where the most variance is explained are subsampled with higher probabilities. Visually, it can be noted that the highest probabilities occur near the edges of the spectra. Scans at these energies will likely contain the most discriminating information about the sample.

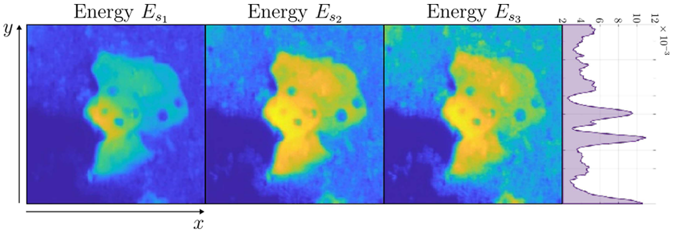


Figure 6
The spatial importance subsampling distribution (purple) on the spatial rows. Three ($s_E = 3$) full raster scans at energies E_{s_1}, E_{s_2} and E_{s_3} are used to determine rows with high variance that are assigned a greater subsampling probability. Visually, one can see that the peaks of the distribution correspond to spatial rows where these are many different levels of absorption. One image has dimensions 50 µm × 50 µm, and the step of the scan in the *x*- and *y*-directions was 250 nm.

Step 5. Complete the missing dataset with a matrix-completion algorithm as done by Townsend *et al.* (2022) (see Fig. 3).

The CURISS measurement strategy consists of the same first three steps. However, the fourth and fifth step are different:

Step 4. Draw one set of rows from the spatial importance probability distribution. For all the energies that were not yet measured, measure these rows.

Step 5. Complete the missing dataset using the *CUR algorithm* (see Fig. 2).

1.6. Outline

We first discuss how to design data-driven importance subsampling distributions for the energy levels and spatial rows in Section 2. The resulting subsampling and completion algorithms are presented in Section 3. In Section 4 the algorithms are compared with the existing approach of random sampling.

2. Spectral and spatial importance subsampling

We describe how spectral and spatial subsampling importance distributions can be formed. To this end we first introduce the general CUR decomposition.

2.1. The CUR decomposition and columns/row subset selection

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a matrix with a small number of large singular values, and many small trailing singular values corresponding to noise. The rank-k CUR decomposition (Goreinov *et al.*, 1997; Drineas *et al.*, 2006) without oversampling of this matrix is defined by two index sets: $\mathcal{I}$ for the rows and $\mathcal{J}$ for the columns, both of which contain k indices. We define the factors $\mathbf{C} = \mathbf{A}(:, \mathcal{J})$, $\mathbf{R} = \mathbf{A}(\mathcal{I}, :)$, and $\mathbf{U} = \mathbf{A}(\mathcal{I}, \mathcal{J})$, which are respectively the columns of $\mathbf{A}$ corresponding to $\mathcal{J}$, the rows of $\mathbf{A}$ corresponding to $\mathcal{I}$, and the intersection of these entries. The CUR approximation is given by $\mathbf{A}_{\rm CUR} = \mathbf{C}\mathbf{U}^{-1}\mathbf{R} \simeq \mathbf{A}$. The CUR decomposition (without oversampling) exactly interpolates the measured entries, *i.e.* $\mathbf{A}(\mathcal{I}, :) = \mathbf{A}_{\rm CUR}(\mathcal{I}, :)$ and likewise for the columns. All other rows and columns are approximated by linear combinations of the measured rows and columns, respectively.

We can also define a CUR decomposition where one of the index sets is oversampled (Drineas *et al.*, 2008; Mahoney & Drineas, 2009). Say $|\mathcal{I}| = k$ but $|\mathcal{J}| = k + p$ for some oversampling parameter p. We can then similarly define the rank-k approximation $\mathbf{A}_{\rm CUR} = \mathbf{C}\mathbf{U}^{\dagger}\mathbf{R} \simeq \mathbf{A}$, where $\mathbf{U}^{\dagger}$ indicates the pseudo-inverse of $\mathbf{U}$. This approximation will still interpolate the rows of $\mathbf{A}$ [*i.e.* $\mathbf{A}(\mathcal{I}, :) = \mathbf{A}_{\rm CUR}(\mathcal{I}, :)$], but this will not generally hold for the columns. We are naturally restricted to the CUR decomposition with oversampling because in raster sampling whole blocks of columns are measured, rather than individual columns (pixels).

The quality of a CUR decomposition depends on the columns and rows that are chosen. That is, $\mathcal{I}$ and $\mathcal{J}$ determine the magnitude of the error $\|\mathbf{A} - \mathbf{A}_{\rm CUR}\|$. One particularly effective way of selecting good rows and columns is to use the associated leverage scores. The scores are a measure for the relative importance of a row or column. Focus on rows for simplicity, and let l_i^2 denote the leverage score of row i. The leverage scores are determined by the row-norms squared of an orthogonal basis for the dominant column space. That is, if $\mathbf{A} = \mathbf{U}_1\boldsymbol{\Sigma}_1\mathbf{V}_1^{\rm T} + \mathbf{U}_2\boldsymbol{\Sigma}_2\mathbf{V}_2^{\rm T}$ is a singular value decomposition (SVD) of $\mathbf{A}$, where $\mathbf{U}_1 \in \mathbb{R}^{m\times r}$ is the dominant column space, then the leverage scores for the rows of $\mathbf{D}$ are $l_i^2 = \|\mathbf{U}_1(i, :)\|_2^2$.

Now, the row leverage scores determine an excellent sampling distribution for the row index set $\mathcal{I}$. If we sample rows according to a probability distribution determined by l_i^2 [and use a similar process for the columns of $\mathbf{D}(\mathcal{I}, :)$], then we have good theoretical guarantees on the quality of the CUR approximation. That is, an excellent strategy is to sample row i with probability l_i^2/n_E. Restricting the matrix $\mathbf{D}$ to the set of selected rows before determining the column indices is highly recommended for the CUR approximation (see Park & Nakatsukasa, 2025).

Exact leverage scores require access to the full matrix; however, in the case of a subsampled matrix as in this paper, estimations of the leverage scores are required. Various adaptive leverage score sampling strategies have been proposed in recent years. These strategies do not compute the leverage scores once and then sample from the resulting distribution—instead they update leverage scores based on what rows have already been sampled. One such algorithm is Adaptive Randomized Pivoting (ARP) (see Cortinovis & Kressner, 2024). Given an orthonormal basis of dimension k for the columns space of $\mathbf{A}$, *i.e.* $\mathbf{U}_1$, the algorithm finds k row indices that will result in small approximation errors for the CUR decomposition.

Of course, when it comes to our dataset $\mathbf{D}$, we do not have access to the leverage scores or a basis for the dominant row or columns space, because we have not yet observed the data. The main contribution of this work is a data-driven strategy that allows us to find good measurements to take without observing the dataset.

2.2. Determining a data-driven spectral importance distribution

We use the existing theory on the CUR decomposition and row subset selection problem to determine a *spectral importance distribution*. This distribution is defined on the n_E available energy levels, and should giver higher sampling probabilities to energies where more variance is explained. The leverage scores of $\mathbf{D}$ would be a great choice, but unfortunately we do not have access to those. However, an approximation is readily available if we have access to a dictionary of possible absorption spectra in the sample.

Considering the exactly low-rank (noiseless) matrix $\mathbf{D}_{\rm exact}$, which is close to $\mathbf{D}$. Since $\mathbf{D}_{\rm exact} = \mathbf{MT}$, we can show that the leverage scores of $\mathbf{D}_{\rm exact}$ are the same as the leverage scores of $\mathbf{M}$, see Appendix B.1. If we have an approximation to $\mathbf{M}$, this means we can approximate the leverage scores of $\mathbf{D}_{\rm exact}$, and by that approximate the leverage scores of $\mathbf{D}$. Let $\mathbf{M}_{\rm dictionary}$ denote a dictionary of $\hat{S}$ spectra. We note that $\mathbf{M}_{\rm dictionary}$ need not be exactly equal to $\mathbf{M}$ or even the same dimension as $\mathbf{M}$; the number of spectra in the dictionary $\hat{S}$ can be smaller or larger than the true number of chemical states S.

We show an example distribution determined by leverage scores based on the dictionary spectra of iron, hematite and magnetite in Fig. 5. One can interpret leverage scores as indicators for the the amount of variation in a dataset that is explained by a particular row.

If a dictionary of true spectra $\mathbf{M}_{\rm dictionary}$ is not known, one may resort to random sampling a few energies. In Section 4.2.2 we show how this adaptation affects the quality of the technique.

2.3. Determining a data-driven spatial importance distribution

Selecting good columns of $\mathbf{D}$ to sample is a similar problem to energy selection. However, as we do not have access to a dictionary for the thickness maps, we need a different strategy.

An additional mechanical limitation is that we can only measure entire spatial rows, not individual pixels (columns). We thus introduce the concept of a *spatial importance subsampling distribution*. This probability distribution on the spatial rows of the sample represents which rows contain the most information. These rows are then sampled in our algorithms with a higher probability. In this section, we describe how to compute such a distribution.

A spatial row of the sample corresponds to a block of columns in $\mathbf{D}$. To mitigate this we use a different unfolding of the data so that a row of the new dataset physically corresponds to a spatial row of the sample. This is then used to determine a good index set for the columns of $\mathbf{D}$ using an *adaptive* leverage score sampling strategy.

In the previous section we discussed how to determine an importance distribution on the energies. Assume we have selected a small number of energies, and that we have taken full raster scans at these energies. This will correspond to a few full rows of $\mathbf{D}$. These XRF scans provides us with an initial idea of the spatial layout of the sample, which will help us determine an approximate importance subsampling distribution. Denote the XRF scans with $\mathbf{F}_{s_1}, \ldots, \mathbf{F}_{s_E} \in \mathbb{R}^{n_Y \times n_X}$, measured at respective energies $E_{s_1}, \ldots, E_{s_E}$.

Next consider a different unfolding of the data. Define $\mathbf{F} \in \mathbb{R}^{n_Y \times s_E n_X}$ as the matrix with individual XRF images stacked horizontally (see Fig. 6 for an example). That is, $\mathbf{F} = \mathbf{F}_{s_1}, \ldots, \mathbf{F}_{s_E}$. A row of $\mathbf{F}$ now corresponds to a spatial row of the sample. As a result, we can look at row selection for $\mathbf{F}$ to directly determine good sample positions.

If we wish to sample s_R spatial rows, we first compute the rank-s_R truncated SVD of $\mathbf{F} = \mathbf{U}_1 \boldsymbol{\Sigma}_1 \mathbf{V}_1$. The dominant column space $\mathbf{U}_1 \in \mathbb{R}^{n_Y \times s_R}$ is used to approximate the leverage scores associated with the spatial rows. We then use the adaptive leverage score sampling algorithm ARP to select s_R spatial rows to sample from. The adaptive strategy allows us to assign smaller sampling probabilities to the areas of the sample that we have already sampled from, so increasing the probability to observe all important spatial areas.

Fig. 6 shows an example distribution on the spatial rows using three XRF scans. The peaks of the distribution correspond to rows with high levels of variation in spectral makeup.

We have described methods to determine importance sampling distributions on the energies and spatial columns. In the next section we propose two different data-driven subsampling strategies for spectro-microscopy based on these ideas.

3. Experimental approaches

Armed with strategies for determining importance distributions on the energies and spatial rows, we propose two data-driven subsampling schemes. The first is a raster subsampling scheme similar to the scheme proposed by Townsend *et al.* (2022). The difference is that we use the importance distributions to sample from, rather than using uniform distributions. We call this approach Raster Importance Sampling for Spectro-microscopy (RISS). The second scheme is CUR-based, where the completion method is naturally the CUR approximation. We call this approach CUR Importance Sampling for spectro-microscopy (CURISS).

Both schemes start in the same way: (1) A spectral dictionary $\mathbf{M}_{\rm dictionary}$, if available, is used to determine a spectral importance distribution (as in Fig. 5), (2) sample a few energies from the spectral importance distribution and take full XRF scans at those energies, and (3) use these scans to determine a spatial importance distribution (as in Fig. 6). The spectral dictionary can be formed from XAS standards, from fully sampled scans taken previously during the experiment or a fusion of both. Otherwise, we can resort back to uniformly random sampling for the energies. We display the process in pseudocode in Appendix *A*. In this algorithm we take the subsampling ratio as input, and then choose the number of sampled energies s_E and the number of sampled spatial rows s_R as approximately equal.

Note that small changes of temperature or mechanical movement of the equipment in the experiment can lead to spatial drifts in the specimen during the experiment. When a full scan is available, this can be compensated for during the stacking process via an alignment procedure using cross-correlation (Lerotic *et al.*, 2014). In general, for subsampled scans, spatial registration to correct drifts cannot be performed in the same way. However, in our method, full raster scans at selected energies are available, thus various alignment methods can be still used to deal with drift. Although we do not implement it here, one can easily incorporate an alignment procedure into our method. Moreover, even if there is a small drift in the spatial rows, it is reasonable to assume that close by entries have similar importance, neglecting the effect of small drifts in our approach.

3.1. RISS

RISS is a data-driven variant of the raster subsampling method of Townsend *et al.* (2022). In this method, for each energy a small number of spatial rows is measured. These spatial rows are selected using a uniform random distribution at each energy. The adaptation we propose is to replace the uniform distribution with the spatial importance distribution.

The measured dataset will have a small number of rows measured fully; these correspond to the full raster scans at the sampled energies. All other measurements will be column-wise blocks as in Fig. 3. These blocks will be concentrated around spatial rows with high sampling probabilities. The structure of the missing entries forces us to use a matrix completion algorithm. We choose the LoopedASD algorithm introduced by Townsend *et al.* (2022). See pseudocode in Appendix *A*.

3.2. CURISS

CURISS is a CUR-based approach that makes use of the importance sampling strategies previously described and the CUR approximation as the completion method. The main steps are collected as in the beginning of this section.

Firstly, given a subsample ratio we calculate the number of full XRF scans and spatial rows that is possible to measure. Then, when possible, we determine the importance distribution of the XRF scans using the leverage scores associated with a dictionary of absorption spectra $\mathbf{M}_{\rm dictionary}$, as described in Section 2.2. Alternatively, the spectral importance distribution is defined uniformly. We use these scores to select s_E XRF scans to measure. These scans correspond to s_E rows of $\mathbf{D}$. As explained in Section 2.3, in order to determine spatial importance distribution we reshape the measured data into an $n_Y \times s_e n_X$ matrix $\mathbf{F}$. We can now obtain importance scores for the rows of $\mathbf{F}$, corresponding to spatial rows of the sample using the ARP algorithm.

The difference between RISS and CURISS is that we sample from this distribution just once, and measure these spatial rows for all remaining energies. As a result, the observed entries of the dataset $\mathbf{D}$ are full rows and blocks of columns. This allows us to employ the CUR decomposition as a tool for completion, rather than a gradient descent or other matrix completion algorithm. See the pseudocode in Appendix *A*.

3.2.1. Adaptive approaches and stopping criteria

In this subsection, we suggest a few approaches for adaptiveness and stopping criteria, which we explore empirically in Section 4.2.1. In Section 3, we describe our proposed algorithms, where we assume a subsampling ratio is given *a priori*, *i.e.* sampling budget, and it determines the stopping criteria for the number of full XRF scans and number of spatial rows to be sampled. Furthermore, the spectral importance distribution and the spatial importance distribution are set in advance as described in Section 3, and do not change as in the previous subsections.

However, the approaches above lead to the following two questions: how many measurements are needed to obtain a good enough completion, *i.e.* can the subsampling ratio depend on some stopping criteria, apart from the given sampling budget; how can we refine our sampling set and select full XRF scans and spatial rows in an adaptive manner. To address these questions, we propose the following Adaptive CURISS (ACURISS) approach: estimate the accuracy of a small CUR approximation, and, if needed, refine it. This process has two main components: the refinement steps, and the stopping criteria. The algorithm is detailed and explained in the following paragraphs.

First, we provide further details on adding new scans. Suppose that we have performed the CURISS method for an initial subsampling ratio, p_0, to obtain $\hat{\mathbf{D}}_0$. Then, we refine the approximation by including more XRF scans. To avoid a significant increase in the dead time needed to mechanically change energy, we propose to only add XRF scans while updating the CUR approximation after each new sample. That is, we only use the spatial rows already included in $\hat{\mathbf{D}}_0$. At each step, for a new XRF scan, we can use the available leverage scores of $\mathbf{M}_{\rm dictionary}$ and set to zero the scores of the already selected XRF scans, thus ensuring the selection of new energies. Afterwards, we update the CUR approximation with the newly scanned XRF to obtain $\hat{\mathbf{D}}_1$, and compute the corresponding subsampling ratio p_1. We repeat the process until some stopping criteria are satisfied. Finally, $\hat{\mathbf{D}}$ and p are obtained.

In addition, we propose two options for stopping criteria. After each refinement of the CUR approximation is done via the additional sample, we suggest using the change between refinements, inspired by the error metrics used for the experiments [equations (4)) and (5)], and using only the sampled entries. Note that, in practice, we cannot compute equations (4) and (5), as they require having a full scan available. In addition, while the change in the completion error can be measured in the sampled entries, the spectral error requires access to the full matrix. Thus, we resort to comparing the completed matrix in the current refinement step, $\hat{\mathbf{D}}_i$, with the one obtained in the previous refinement step $\hat{\mathbf{D}}_{i-1}$. We define *completion variation* as a stopping criteria relating to the completion error. For $i \geq 1$, stop when successive updates are sufficiently close,

$$\text{Completion variation} = \|\boldsymbol{\Delta} D_i\| := \|\hat{\mathbf{D}}_i - \hat{\mathbf{D}}_{i-1}\|_{\rm F} \leq \eta_{\mathbf{D}}, \quad (2)$$

where $\|\ldots\|_{\rm F}$ denotes the Frobenius norm, and $\eta_{\mathbf{D}} \geq 0$ is a user-prescribed tolerance parameter.

After each refinement step, we can perform a clustering analysis as described in Section 1.3 and obtain approximations to the spectra, $\mathbf{M}_{\rm cluster}(\hat{\mathbf{D}}_i)$. We compare these with those from the previous refinement step, $\mathbf{M}_{\rm cluster}(\hat{\mathbf{D}}_{i-1})$, and consider the magnitude of the update. The *spectral variation* stopping criteria for $i \geq 1$ is met when

$$\text{Spectral variation} = \|\boldsymbol{\Delta} M_i\| := \|\mathbf{M}_{\rm cluster}(\hat{\mathbf{D}}_i) - \mathbf{M}_{\rm cluster}(\hat{\mathbf{D}}_{i-1})\|_{\rm F} \leq \eta_{\mathbf{M}}, \quad (3)$$

where $\eta_{\mathbf{M}} \geq 0$ is again a tolerance parameter.

We remark that for both equations (2) and (3) the average variation of several refinement steps can be used as alternative stopping criteria, rather than using a variation between only two refinment steps. As we demonstrate in Section 4.2.1, equation (3) gives us an indicator of the accuracy of the completion process. See the pseudocode in Appendix *A*.

4. Experiments

4.1. Experimental method

The measurements were conducted on I14, the Hard X-ray Nanoprobe beamline at Diamond Light Source. The I14 endstation, high stability scanning stages, room and beam stability control measures are described elsewhere (Quinn *et al.*, 2021*a*; Kelly *et al.*, 2022; Cacho-Nerin *et al.*, 2020; Quinn *et al.*, 2021*b*), but, in brief, the experiment was conducted in a temperature controlled hutch using a nano-focusing Kirkpatrick–Baez (KB) mirror, with the sample rastered through the beam, in atmospheric conditions, using a voice coil actuated stage with interferometric control of the stage position, but no relative or differential control between the stage and KB. The X-ray fluorescence signal (XRF) was collected using

Table 1
The dimensions of each of the datasets used in Section 4.

Datasets DS1–DS5 are identical to DS1–DS5 from Townsend *et al.* (2022).

Name	$n_X \times n_Y$	n_E	Name	$n_X \times n_Y$	n_E
DS1	101 × 101	149	DS5	80 × 80	152
DS2	92 × 79	150	DS6	76 × 79	152
DS3	101 × 101	149	DS7	86 × 73	151
DS4	40 × 40	152	DS8	85 × 84	250

a four-element silicon drift detector and energy changes and sparse scan motions were implemented using Diamond's acquisition software (Basham *et al.*, 2018).

The sample consisted of Fe_2O_3 (hematite) and Fe_3O_4 (magnetite) powders, ground, mixed and drop cast on a silicon nitride membrane to provide different mixtures and a well known reference. While simulated spectra could have been used, experimental data capture realistic sample drift, beam artifacts such as top-up and other artifacts which allow us to test the approaches against real conditions. The XRF maps at each energy were collected in fly scanning mode. The dwell time per point over the absorption edge was constant (15 ms).

Sample areas scanned ranged from 3 µm × 3 µm to 10 µm × 10 µm using 50, 75 or 100 nm steps. The step size and area were chosen to keep the scanning time to a reasonable level during the time available for the overall experiment, rather than being a limit in resolution or size.

We summarize the dimension of the datasets in Table 1. Each of the datasets are obtained at I14 as described above.

4.2. Results

We use two different measures to determine the quality of the sampling strategies and the completion. The first is the distance between the completed dataset, say $\hat{\mathbf{D}}$, and the full dataset $\mathbf{D}$. We define

$$\text{Completion error} = \frac{\|\mathbf{D} - \hat{\mathbf{D}}\|_F}{\|\mathbf{D}\|_F}. \tag{4}$$

It is important to note that the size of this error is lower bounded by the error with respect to the optimal low-rank approximation (Eckart & Young, 1936). Fundamentally, when approximating a matrix of full rank by a matrix that has significantly smaller rank, the error is controlled by the trailing singular values. As a result this error will likely not be less than, say, 0.01, because it will now correspond to 1% noise present in the data.

Secondly, we define an error related to the further analysis. As described in Section 1.3, the most common way to analyze spectro-microscopic data is to perform cluster analysis on the eigenimages and use average absorption spectra per cluster as estimates for the true absorption spectra $\mathbf{M}$. Denote the cluster spectra resulting from this analysis on the full dataset by $\mathbf{M}_{\text{cluster}}(\mathbf{D})$ and on the completed dataset by $\mathbf{M}_{\text{cluster}}(\hat{\mathbf{D}})$. We define

$$\text{Spectral error} = \frac{\|\mathbf{M}_{\text{cluster}}(\mathbf{D}) - \mathbf{M}_{\text{cluster}}(\hat{\mathbf{D}})\|_F}{\|\mathbf{M}_{\text{cluster}}(\mathbf{D})\|_F}. \tag{5}$$

The completion errors for eight different datasets are shown in Fig. 7 (note the log-scale). We observe that RISS and CURISS generally perform well for subsampling ratios as small as 5%, where random raster sampling is generally unable to provide a satisfactory completion at this subsampling ratio. In most cases, CURISS outperforms RISS for subsampling ratios smaller than approximately 8%, after which RISS is the most successful technique. We note that the subsampling ratio is the percentage of total measurements that are measured; for CURISS that includes the necessary full XRF scans.

The spectral errors for the same datasets are displayed in Fig. 8, with a close-up of small subsampling ratios provided in

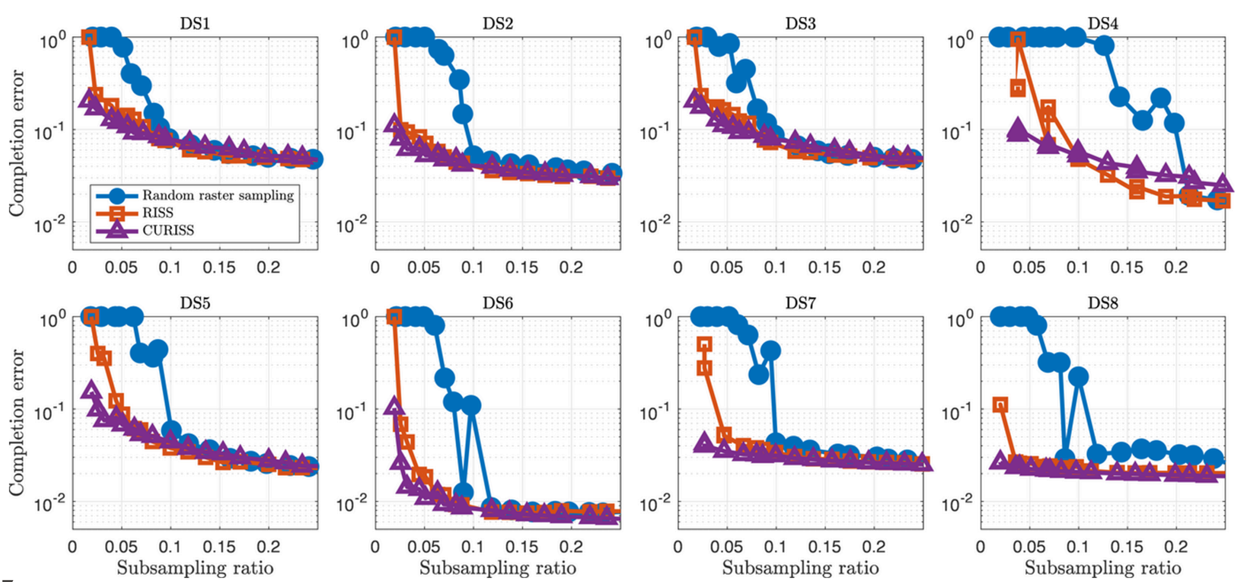


Figure 7
The completion error [defined in equation (4)] associated with each of the discussed algorithms: random raster sampling, data-driven raster sampling (RISS) and data-driven CUR raster sampling (CURISS). The data shown for random raster sampling and RISS are the average of ten trials, whereas the data shown for CURISS are the average of 100 trials. LoopedASD is set to maximum rank 5 and tolerance 10^{-5}. If LoopedASD fails, the error is set to equal 1.

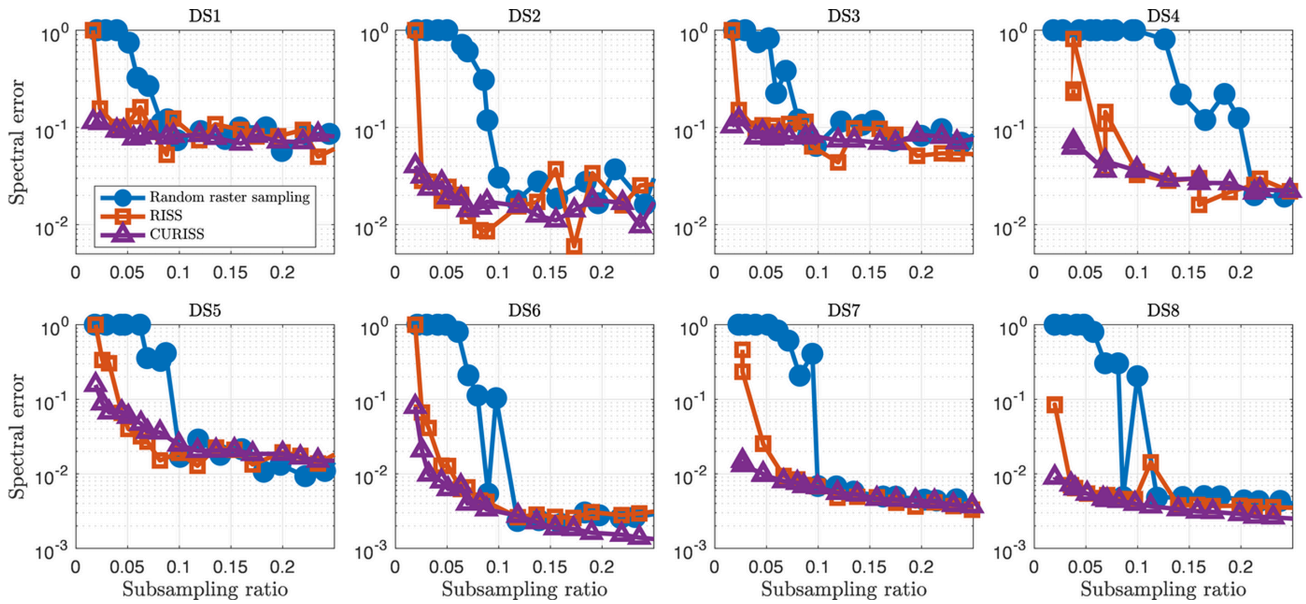


Figure 8
The spectral error [defined in equation (5)] associated with each of the discussed algorithms: random raster sampling, data-driven raster sampling (RISS) and data-driven CUR raster sampling (CURISS). The experimental set-up is the same as in Fig. 7. We provide a close-up of the results for subsampling ratios less than 10% in Fig. 9.

Fig. 9. It is even more evident in these figures that CURISS leads to good results for ratios as small as 3%. The CUR completion apparently captures the leading singular vectors and values (which are used for the cluster analysis) very well. RISS outperforms random raster sampling by at least an order of magnitude for these small subsampling ratios.

Because the methods, and consequently the experiments, are random in nature, the reported completion and spectral error curves represent averages over multiple trials. As a result, they are not strictly monotonic with increasing subsampling ratio, and the small oscillations observed in some of the plots are a natural consequence of this randomness. To better illustrate the variability of the discussed methods, in Fig. 10 we show the average error in log scale across trials together with shaded regions representing ± 1 standard deviation across runs for the dataset DS2.

We note furthermore that either of the methods that employ LoopedASD (random raster sampling and RISS) take considerably more computational effort, due to the optimization steps that they require.

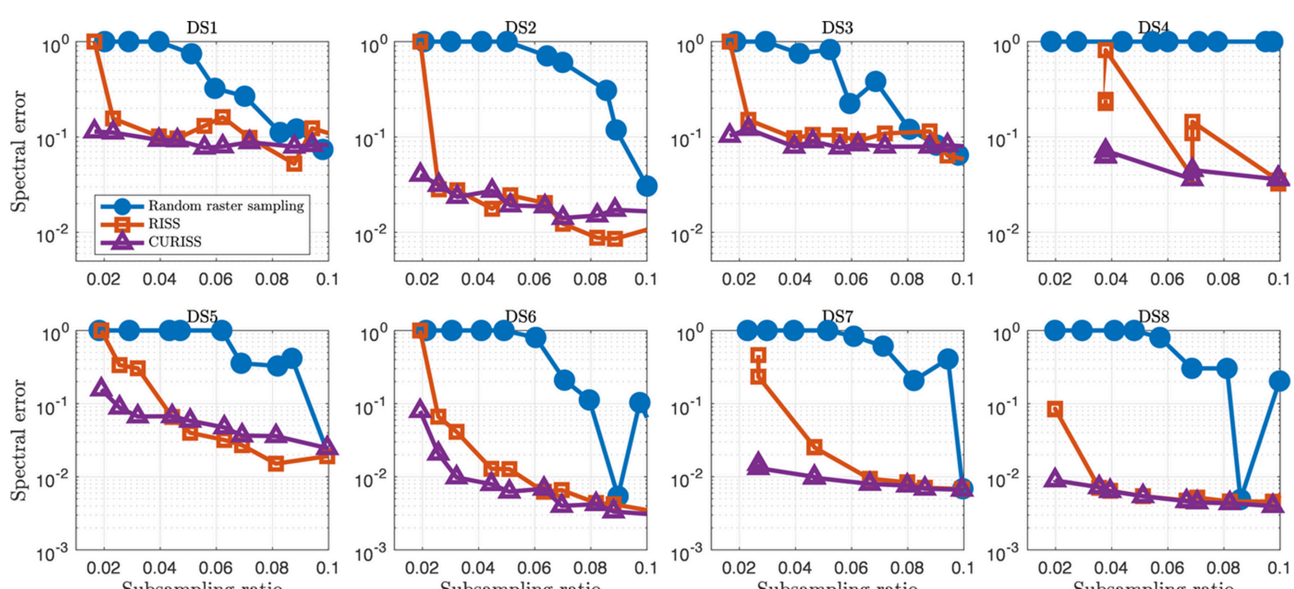


Figure 9
A close-up of Fig. 8 to observe subsampling ratios smaller than 10%. We see that, for many of the datasets, CURISS performs exceptionally well even for very small subsampling ratios. The best possible accuracy is approximately attained at ratios as small as 2%, and consistently attained for 6% subsampling. RISS is comparable in performance, except for the smallest subsampling ratios. Both methods outperform random raster sampling considerably.

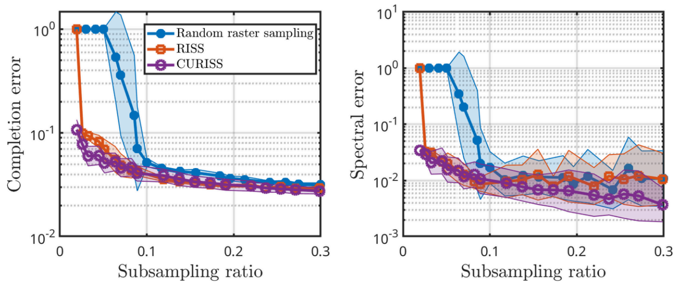


Figure 10
The completion and spectral errors for DS2 associated with each of the discussed algorithms: random raster sampling, data-driven raster sampling (RISS), and data-driven CUR raster sampling (CURISS). Solid lines show the mean across 100 trials for CURISS and 10 trials for random raster sampling and RISS. Shaded bands represent ± 1 standard deviation in the log-scale. The standard deviations show that the variance is low and the probability that RISS or CURISS fails is very small.

Selecting an appropriate subsampling ratio remains a challenge. However, based on the previous experiments, a subsampling ratio of 6% appears to be a reliable rule of thumb when using the proposed methods in practical settings, consistently yielding satisfactory accuracy.

4.2.1. Adaptive approaches

As described in Section 3.2.1, the CURISS method allows for an adaptive counterpart that, starting with a small subsampling ratio, refines the CUR approximation by including new XRF scans as described in Section 3.2.1. We perform experiments on different datasets to compare the accuracy—measured in completion error—of CURISS versus ACURISS. We begin by running ACURISS for a fixed number of iterations. At each step, we store the used subsampling budget, p_i, and use it to compute the corresponding CURISS completed matrix. The choice of the initial subsampling ratio p_0 is crucial for the success of the adaptive strategy. We have observed that for most of the analyzed datasets $p_0 = 0.02$ is good enough. However, for other datasets a larger p_0 is needed to obtain completion errors competitive with the one obtained by CURISS. Specifically, DS3 needs $p_0 = 0.04$, while $p_0 = 0.1$ is necessary for DS8. Fig. 11 shows, for some datasets, that, with suitable p_0, the completion errors of the two strategies are comparable. This motivates further investigation of this technique.

As mentioned, an adaptive strategy requires a stopping criteria. Thus, we analyze how the two proposed techniques perform on different datasets. Since the considered variations (2) and (3) are absolute measures, we compare them with the respective absolute errors. Specifically, we compute refined approximations via the ACURISS and measure their completion absolute error. We then compute the completion variation [equation (2)] and the spectral variation [equation (3)]. For analysis purposes, we also compute the spectral approximations attainable with the full matrix $\mathbf{D}$. In Fig. 12, we present the main scenarios that may occur. In the best cases, the spectral variation is really close to the spectral error, *e.g.* for DS2 and DS3. Otherwise, the variations exhibit similar behavior to the actual errors but at different magnitudes. In

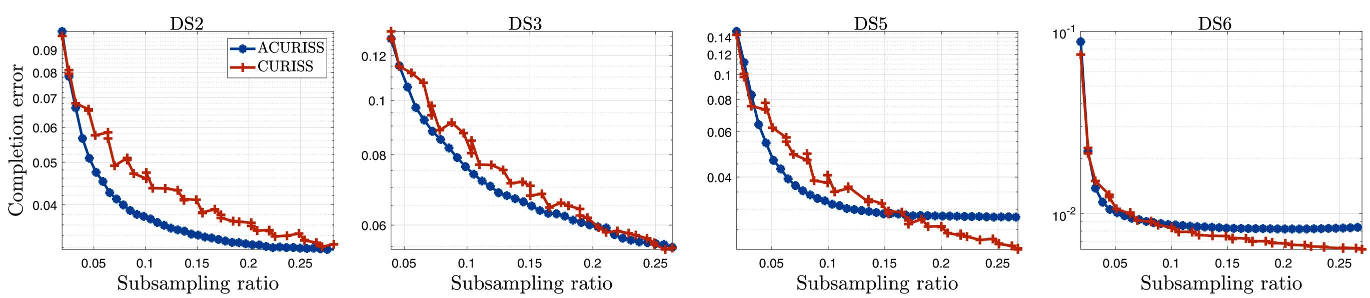


Figure 11
Comparison of completion errors for CURISS (orange) and ACURISS (blue). Displayed is the the mean error of 100 trials. The analyzed dataset shows that ACURISS's adaptivity can yield better approximation quality at very low subsampling ratios. However, beyond a certain point its performance plateaus, whereas futher improvements would be achievable using CURISS.

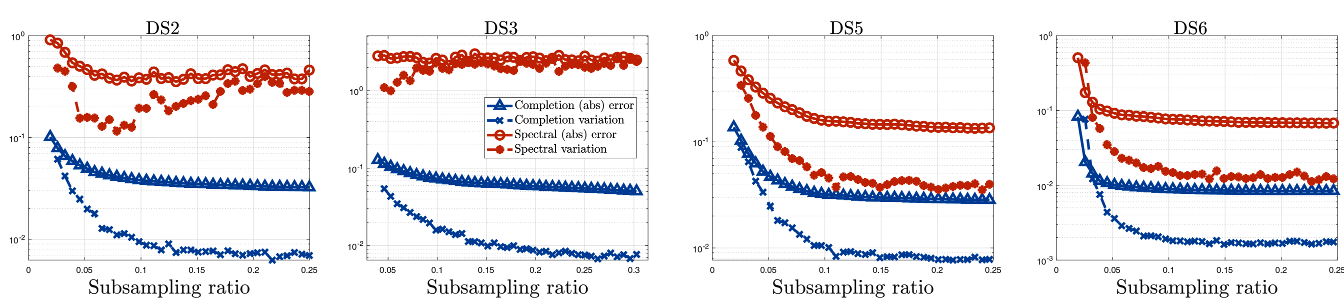


Figure 12
Study of the proposed stopping criteria for the ACURISS strategy. After 100 trials, the means of the completion absolute error for ACURISS (blue triangles), the completion stopping criteria (blue crosses), the spectra approximation absolute error computed from the full matrix (orange open circles), and the spectral stopping criteria (orange asterisks) and are presented. We observe that the behavior of the stopping criteria can mirror that of the actual errors, but with potentially different magnitudes, suggesting that their stagnation may serve as a useful indicator that little further improvement in ACURISS is achievable.

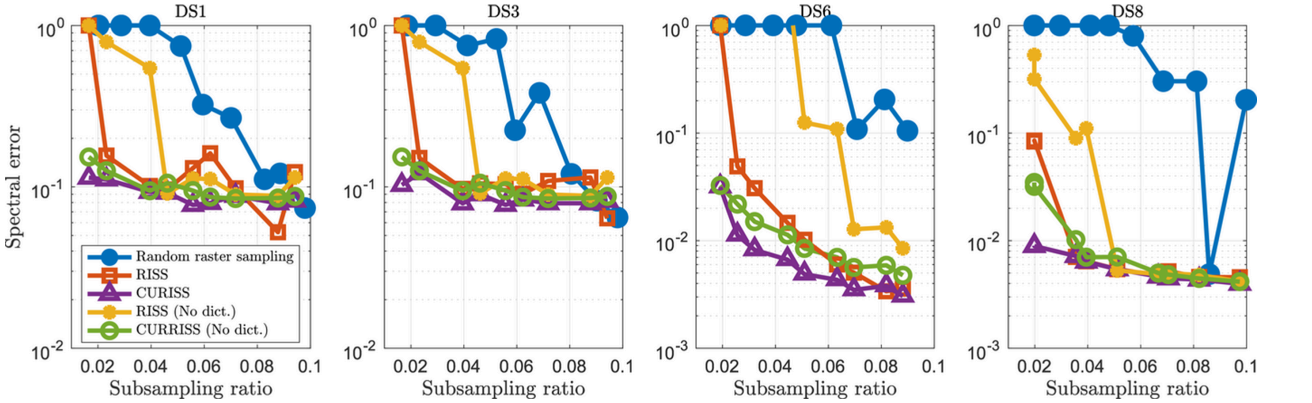


Figure 13
The spectral error [defined in equation (5)] associated with each of the discussed algorithms, as well as RISS and CURISS without the use of a spectral dictionary $\mathbf{M}_{\text{dictionary}}$ (No dict.).

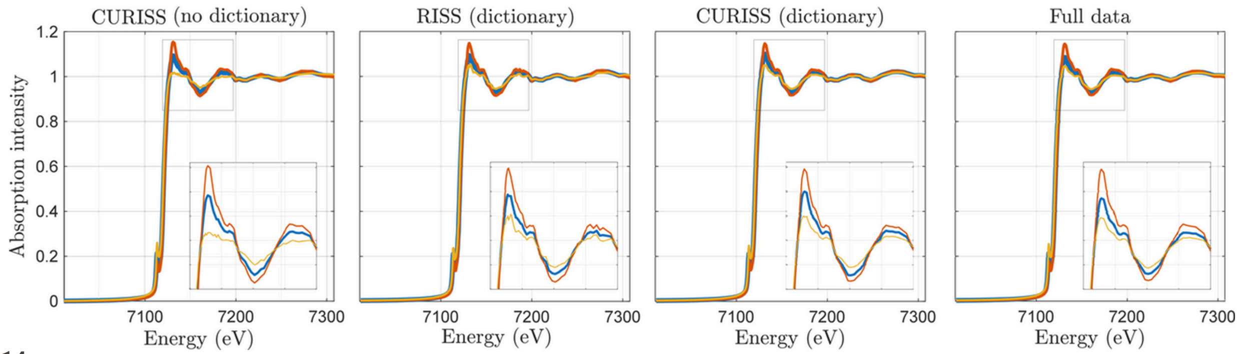


Figure 14
Spectra obtained from DS6 at 6% subsampling for different sampling strategies. The right figure shows the spectra obtained from the full data. The spectral errors are, from left to right, 0.006, 0.006 and 0.004. The spectra from RISS without a dictionary were significantly more noisy with spectral error 0.043.

both cases, it is clear that both stopping criteria can be used as indicators of the correspondent true errors. We note that the computational effort involved with updating the CUR decomposition and computing the stopping criteria is negligible. A similar scheme using LoopedASD would be more challenging, because the necessary time is much greater.

4.2.2. No *a priori* spectral knowledge

One might be concerned that the quality of RISS and CURISS depends on the availability of a good dictionary of the spectra $\mathbf{M}_{\text{dictionary}}$. We show here that we obtain comparable results if we do not have access to a spectral dictionary.

For each dataset, the quality of the completion for the algorithms with or without a dictionary is indistinguishable for subsampling ratios greater than 10%. The results for ratios smaller than 10% are shown in Fig. 13. For subsampling ratios smaller than 6%, RISS without a dictionary suffers from considerably larger errors, while CURISS is more robust to this deficiency.

We also show a visualization of the spectral errors in Fig. 13 on obtained spectra. Fig. 14 shows the recovered spectra for DS6 at 6% subsampling for CURISS without a dictionary (spectral error 0.006), RISS with a dictionary (spectral error 0.006), and CURISS with a dictionary (spectral error 0.004), and shows the spectra obtained from the full data. The errors and the visualization of the reconstructed spectra show that CURISS without a dictionary performs similar to RISS with a dictionary and slightly worse than CURISS with a dictionary. Fig. 14 shows that CURISS without a dictionary cannot fully reconstruct the yellow spectrum. As Fig. 13 also displays, without a dictionary, slightly larger subsampling ratios are required for similar results; say 9% instead of 6%. Nonetheless, Figs. 13 and 14 demonstrate a robustness of our method with respect to incomplete or absent spectral dictionaries.

5. Conclusions

We have proposed several data-driven approaches to subsampling during a spectro-microscopy experiment. The goal of these experiments is to determine bond length variations in materials with nanoscale spatial resolution. These experiments are generally limited by long acquisition times and radiation damage, as they require a raster XRF scan for various energy levels. Subsampling approaches to spectro-microscopy aim to mitigate this by omitting various measurements in a way that fits the raster acquisition of measurements, without changing the (spatial or spectral) resolution. Missing measurements are in-painted with low-rank matrix approximation methods using the available observations, *e.g.* matrix completion with LoopedASD, and matrix factorization with CUR. This is possible because of the low-dimensional nature of a spectro-microscopic dataset; the underlying dimension is equal to the number of spectroscopically different materials up to measurement noise.

The question that might arise when considering subsampling approaches is to what degree can spatial subsampling be applied and what are the limits of the approach. Fundamentally, the approaches depend on the information content in the measurements being representative of the full measurement A, *i.e.* to recover A, the chemical states in A must be represented in the sample set, otherwise it would not be recovered. Methods based on uniform randomly (raster) subsampling measurement, such as LoopedASD (Townsend *et al.*, 2022), suffer from the aforementioned drawback. We built on Townsend *et al.* (2022) by making the random sampling data-driven. In the data-driven scheme, entries that contain more information are measured with higher probabilities. This data-driven approach allows us to measure very small amounts of data while still retaining much of the relevant information. In particular, we find that, for the available datasets, data-driven approaches can reach subsampling levels of 2–5% whereas uniform random sampling would need 10–20% subsampling ratios.

We note especially the efficacy of the CUR-based approach. The interpretability of this approach, combined with the simplicity and speed of its corresponding completion, suggests potential for its application in different experiments with low-dimensionality. As far as the authors are aware, this is the first time that the CUR decomposition has been used as a tool for experiment design. Especially in the context of costly experiments, we believe this approach has great potential to speed up acquisition times and even improve resolution beyond current mechanical limits.

From the numerical experiments presented in the previous sections, it is possible to draw practical recommendations for the use of these approaches. In particular, a subsampling ratio of 6% emerges as a reliable guideline, consistently yielding accurate results across the datasets considered.

Future work includes further investigation of adaptive approaches and stopping criteria, as the adaptive strategy has shown promise as a viable alternative to the 6% guideline, but still requires further development to fully establish its reliability and effectiveness. Once an initial completion was performed for an initial subsampling ratio, adaptively adding a new XRF scan can be done via other criteria, such as restarting the leverage scores from each new completion. As for other stopping criteria, apart from testing the change between iterations via the error metrics, other approaches can be leveraged from numerical linear algebra, *e.g.* testing the singular values and numerical rank differences between iterations. In addition, our work can be generalized to the tensor framework, as done by Townsend *et al.* (2025), by using tensor generalizations for the notions of leverage scores and CUR decomposition. We leave it for future work as well.

APPENDIX *A*
Algorithms

In this section, we provide details and write in pseudocode form our proposed algorithms, which we explain in Section 3.

We begin with our algorithm to determine spectral and spatial importance distributions (Algorithm 1). The spectral importance distribution is based on the leverage scores of a dictionary of ground-truth spectra. In practice, this is achieved by performing a QR factorization of $\mathbf{M}_{\rm dictionary} = \mathbf{QR}$, followed by computing the row norms of $\mathbf{Q}$. Using these leverage scores as sampling weights, we then select s_E out of the n_E XRF scans to be measured. For the spatial importance distribution, we choose to apply Adaptive Randomized Pivoting (ARP) to the horizontally stacked XRF maps, $\mathbf{F} = \mathbf{F}_1, \ldots, \mathbf{F}_{s_E}$. ARP is an algorithm that adapts leverage scores while selecting rows to sample. The process begins similarly to standard leverage score sampling by computing the matrix $\mathbf{U}_1$ of dominant left singular vectors of $\mathbf{F}$ and using the row norms of $\mathbf{U}_1$ as initial sampling probabilities. The first index i_1 is randomly selected according to these normalized leverage scores. Next, $\mathbf{U}_1$ is updated to remove the selected column $\mathbf{U}_1(:, i_1)$, resulting in a new matrix $\bar{\mathbf{U}}_1$. This is achieved via orthogonal projection implemented by Householder reflectors. The second index is then sampled based on the updated leverage scores of $\bar{\mathbf{U}}_1$. This process repeats until all indices are selected and $\mathbf{U}_1$ has been fully reduced. Note that this process is able to pick at most r indices when applied to an $n \times r$ matrix. Thus, we cannot use it for the previous step, where from the $n_E \times \hat{S}$ dictionary $\mathbf{M}_{\rm dictionary}$, we want to obtain importance scores to select $s_E \geq \hat{S}$ XRF scans.

Algorithm 1 Importance Sampling for Spectro-microscopy: determining spectral and spatial importance distribtions.
Input: subsampling ratio p, number of spatial rows n_y and columns n_X of the sample, number of energies n_E, and a dictionary of absorption spectra $\boldsymbol{M}_{\rm dictionary}$ (optional).
Output: a small number of full XRF scans, a spectral importance distribution, and a spatial importance distribution

Set $s_E = \left\lceil \frac{p n_Y n_E}{n_Y + n_E} \right\rceil$ and $s_R = \left\lfloor \frac{n_Y (p n_E - s_E)}{n_E - s_E} \right\rfloor$.
Use the dictionary $\boldsymbol{M}_{\rm dictionary}$ to estimate the leverage scores on the energies and define the corresponding spectral importance distribution. If a dictionary is not available, define the spectral importance distribution as uniform.
Sample s_E energies $E_{s_1}, \ldots, E_{s_{s_E}}$ from the spectral importance distribution.
Measure full raster scans at energies $E_{s_1}, \ldots, E_{s_{s_E}}$. Name the corresponding XRF maps $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_{s_E}$.
Horizontally stack the XRF maps as $\boldsymbol{F} = \begin{bmatrix} \boldsymbol{F}_1 & \ldots & \boldsymbol{F}_{S_E} \end{bmatrix}$ and use $\boldsymbol{F}$ to determine a spatial importance distribution.

Next, we present pseudocode for the presented strategies. That is, RISS (Section 3.1) in Algorithm 2, CURISS (Section 3.2) in Algorithm 3, and ACURISS (Section 3.2.1) in Algorithm 4.

Algorithm 2 RISS: Raster Importance Sampling for Spectro-microscopy.
Input: subsampling ratio p, number of spatial rows n_y and columns n_X of the sample, number of energies n_E, and a dictionary of absorption spectra $\boldsymbol{M}_{\rm dictionary}$ (optional).
Output: an approximation to the dataset $\boldsymbol{D}$ using subsampling ratio p.

Use Algorithm 1 to measure a small number of full XRF scans at energies $E_{s_1}, \ldots, E_{s_{s_E}}$ and determine a spatial importance distribution.
for all energies E in $\{E_1, \ldots, E_{n_E}\} \backslash \{E_{s_1}, \ldots, E_{s_{s_E}}\}$ **do**
    Set the beam energy to E.
    Sample s_R spatial rows from the spatial importance distribution: $r_{s_1}^{(E)}, \ldots, r_{s_R}^{(E)}$
    Measure the sampled spatial rows: $r_{s_1}^{(E)}, \ldots, r_{s_R}^{(E)}$
end for
Complete the measured dataset using LoopedASD.

Algorithm 3 CURISS: CUR Importance Sampling for spectro-microscopy
Input: subsampling ratio p, number of spatial rows n_Y and columns n_X of the sample, number of energies n_E, and a dictionary of absorption spectra $\boldsymbol{M}_{\text{dictionary}}$ (optional).
Output: an approximation $\hat{\boldsymbol{D}}$ to the dataset $\boldsymbol{D}$ using subsampling ratio p.

Use Algorithm 1 to measure a small number of full XRF scans at energies $E_{s_1}, \ldots, E_{s_{s_E}}$ and determine a spatial importance distribution.
Sample s_R spatial rows from the spatial importance distribution.
for all energies E in $\{E_1, \ldots, E_{n_E}\} \backslash \{E_{s_1}, \ldots, E_{s_{s_E}}\}$ **do**
    Set the beam energy to E.
    Measure the sampled spatial rows
end for
Complete the measured dataset using CUR.

Algorithm 4 ACURISS: Adaptive CURISS
Input: initial subsampling ratio p_0, $\eta_{\boldsymbol{D}} \geq 0$ and $\eta_{\boldsymbol{M}} \geq 0$ are tolerance parameters for the stopping criteria, number of spatial rows n_Y and columns n_X of the sample, number of energies n_E, and a dictionary of absorption spectra $\boldsymbol{M}_{\text{dictionary}}$ (optional).
Output: an approximation $\hat{\boldsymbol{D}}$ to the dataset $\boldsymbol{D}$, and its corresponding subsampling ratio p.

Use CURISS (Algorithm 3) to obtain $\hat{\boldsymbol{D}}_0$.
$i = 0$
while $\|\boldsymbol{\Delta D}_i\| > \eta_{\boldsymbol{D}}$ or $\|\boldsymbol{\Delta M}_i\| > \eta_{\boldsymbol{M}}$ **do**
    Set to 0 the leverage scores of the previously selected energies.
    Sample one additional energy using the refined spectral importance distribution.
    Set $i = i + 1$.
    Update the subsampling ratio p_i.
    Update CUR with the additional XRF scan to obtain $\hat{\boldsymbol{D}}_i$.
    Set $s_E = s_E + 1$.
end while
return $\hat{\boldsymbol{D}} = \hat{\boldsymbol{D}}_i$ and $p = p_i$.

APPENDIX *B*
Mathematical foundations

In the following, we provide more details, mathematical background and implementation specifics about observations and statements in the main text.

B1. Further motivations on the use of a dictionary of true spectra

In Section 2.2 we have introduced the idea of using a dictionary of true spectra $\mathbf{M}_{\text{dictionary}}$ to obtain approximate leverage scores of the matrix $\mathbf{D}$. This idea is based on the observation that, since $\mathbf{D}_{\text{exact}} = \mathbf{MT}$, the leverage scores of $\mathbf{D}_{\text{exact}}$ are the same as the leverage scores of $\mathbf{M}$. We elaborate here.

Let $\mathbf{D}_{\text{exact}} = \mathbf{U}_1 \boldsymbol{\Sigma}_1 \mathbf{V}_1^{\text{T}}$ be the (reduced) singular value decomposition of $\mathbf{D}_{\text{exact}}$. That is, $\mathbf{U}_1 \in \mathbb{R}^{n_E \times S}$ and $\mathbf{V}_1 \in \mathbb{R}^{n_X n_Y \times S}$ are matrices with orthonormal columns corresponding to the leading singular vectors of $\mathbf{D}_{\text{exact}}$, and $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{S \times S}$ is a diagonal matrix containing the first S singular values of $\mathbf{D}_{\text{exact}}$. Then, by definition, the leverage scores of $\mathbf{D}_{\text{exact}}$ are

$$l_i^2 = \|\mathbf{U}_1(i,:)\|_2^2.$$

Now, let $\mathbf{M} = \mathbf{QR}$, with $\mathbf{Q} \in \mathbb{R}^{n_E \times S}$, $\mathbf{R} \in \mathbb{R}^{S \times S}$, and $\mathbf{T}^{\text{T}} = \tilde{\mathbf{Q}}\tilde{\mathbf{R}}$, with $\tilde{\mathbf{Q}} \in \mathbb{R}^{n_X n_Y \times S}$, $\tilde{\mathbf{R}} \in \mathbb{R}^{S \times S}$, be the QR-factorizations of $\mathbf{M}$ and $\mathbf{T}^{\text{T}}$. Then, we have

$$\mathbf{D}_{\text{exact}} = \mathbf{MT} = \mathbf{QR}\tilde{\mathbf{R}}^{\text{T}}\tilde{\mathbf{Q}}^{\text{T}}. \tag{6}$$

Considering the (full) singular value decomposition $\mathbf{R}\tilde{\mathbf{R}}^{\text{T}} = \hat{\mathbf{U}}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}^{\text{T}}$ and using it in (6), we obtain

$$\mathbf{D}_{\text{exact}} = \mathbf{Q}\hat{\mathbf{U}}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}^{\text{T}}\tilde{\mathbf{Q}}^{\text{T}} = \mathbf{U}_1\boldsymbol{\Sigma}_1\mathbf{V}_1^{\text{T}},$$

and, thus, $\mathbf{U}_1 = \mathbf{Q}\hat{\mathbf{U}}$. Finally, since $\hat{\mathbf{U}}$ is a square orthogonal matrix, and the Euclidean norm is invariant under orthogonal transformations, we have

$$l_i^2 = \|\mathbf{U}_1(i,:)\|^2 = \|\mathbf{Q}(i,:)\hat{\mathbf{U}}\|^2 = \|\mathbf{Q}(i,:)\|^2,$$

i.e. the leverage scores of $\mathbf{D}_{\text{exact}}$ are the same as the leverage scores of $\mathbf{M}$.

When we turn our attention to $\mathbf{D}$, the situation is slightly different. Since $\mathbf{D}$ is only approximately low-rank, its leverage scores will no longer coincide with those of $\mathbf{M}$. However, because we have $\mathbf{D} = \mathbf{MT} + \mathbf{G}$, where $\mathbf{G}$ represents relatively small noise (*i.e.* $\|\mathbf{G}\|$ is small), we can reasonably expect the leverage scores of $\mathbf{D}$ to remain close to those of $\mathbf{M}$.

B2. Theoretical guarantees on the quality of the CUR decomposition

In Section 2.1 we provided a technical introduction to CUR decomposition. In this subsection, we provide some theoretical guarantees from the vast literature on CUR. In Goreinov *et al.* (1997) and Boutsidis & Woodruff (2017), it is stated that if $\mathbf{A}$ is exactly rank-k, then there exists a CUR decomposition that exactly recovers $\mathbf{A}$. Drineas *et al.* (2008) show that there exists a CUR of rank k that provides, with high probability $(1 - \delta)$, the following,

$$\|\mathbf{A} - \mathbf{C}\mathbf{U}^{\dagger}\mathbf{R}\|_{\text{F}} \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\text{F}}, \tag{7}$$

where $\mathbf{A}_k$ is an optimal rank-k approximation of $\mathbf{A}$, via the truncated SVD, and $\epsilon > 0$ is an error parameter. The number of columns, c, that are selected from $\mathbf{A}$ using statistical leverage scores is $c = O[k^2 \log(1/\delta)/\epsilon^2]$ exactly or $c = O[k \log(k) \log(1/\delta)/\epsilon^2]$ in expectation, and the number of rows, r, that are selected from $\mathbf{A}$ using statistical leverage scores is $r = O[c^2 \log(1/\delta)/\epsilon^2]$ exactly or $r = O[c \log(c) \log(1/\delta)/\epsilon^2]$, respectively. Boutsidis & Woodruff (2017) provide a deterministic (not probabilistic) bound similar to equation (7) with the number of selected rows and columns reduced to $O(k/\epsilon)$.

Acknowledgements

We thank Taejun Park, Nathaniel Pritchard, Yuji Nakatsukasa, Johnathan M. Bulled, Julia Parker, and Miguel Gomez Gonzalez for useful discussions.

Conflict of interest

The authors declare no conflicts of interest.

Data availability

Data used to produce the figures and tables in this manuscript are available for downloading at the following open-access Zenodo repository: 10.5281/zenodo.18470992 (Meier *et al.*, 2026).

Funding information

The following funding is acknowledged: Ada Lovelace Centre, Scientific Computing Department, STFC.

References

Basham, M., Filik, J., Cobb, T., Mudd, J. J., Mosselmans, J. F. W., Dudin, P., Parsons, A. D., Quinn, P. D. & Dent, A. J. (2018). *Synchrotron Radiat. News* **31**(5), 21–26.

Boutsidis, C. & Woodruff, D. P. (2017). *SIAM J. Comput.* **46**, 543–589.

Braun, A. (2005). *J. Environ. Monit.* **7**, 1059–1065.

Cacho-Nerin, F., Parker, J. E. & Quinn, P. D. (2020). *J. Synchrotron Rad.* **27**, 912–922.

Cortinovis, A. & Kressner, D. (2024). *arXiv*:2412.13992.

Drineas, P., Kannan, R. & Mahoney, M. W. (2006). *SIAM J. Comput.* **36**, 158–183.

Drineas, P., Mahoney, M. W. & Muthukrishnan, S. (2008). *SIAM J. Matrix Anal. Applic.* **30**, 844–881.

Du, M., Wolfman, M., Sun, C., Kelly, S. D. & Cherukara, M. J. (2025). *arXiv*:2504.17124.

Eckart, C. & Young, G. (1936). *Psychometrika* **1**, 211–218.

Goreinov, S. A., Tyrtyshnikov, E. E. & Zamarashkin, N. L. (1997). *Linear Algebra Applic.* **261**, 1–21.

Hitchcock, A. P., Morin, C., Zhang, X., Araki, T., Dynes, J., StöVer, H., Brash, J., Lawrence, J. R. & Leppard, G. G. (2005). *J. Electron Spectrosc. Relat. Phenom.* **144–147**, 259–269.

Ito, Y., Takeichi, Y., Hino, H. & Ono, K. (2025). *Mach. Learn.: Sci. Technol.* **6**, 025037.

Kelly, J., Male, A., Rubies, N., Mahoney, D., Walker, J. M., Gomez-Gonzalez, M. A., Wilkin, G., Parker, J. E. & Quinn, P. D. (2022). *Rev. Sci. Instrum.* **93**, 043712.

Lerotic, M., Jacobsen, C., Gillow, J., Francis, A., Wirick, S., Vogt, S. & Maser, J. (2005). *J. Electron Spectrosc. Relat. Phenom.* **144–147**, 1137–1143.

Lerotic, M., Jacobsen, C., Schäfer, T. & Vogt, S. (2004). *Ultramicroscopy* **100**, 35–57.

Lerotic, M., Mak, R., Wirick, S., Meirer, F. & Jacobsen, C. (2014). *J. Synchrotron Rad.* **21**, 1206–1212.

Mahoney, M. W. (2011). *Found. Trends Mach. Learn.* **3**, 123–224.

Mahoney, M. W. & Drineas, P. (2009). *Proc. Natl Acad. Sci. USA* **106**, 697–702.

Meier, M., Lazzarino, L., Shustin, B., Al Daas, H. & Quinn, P. (2026). *Data supporting the research article entitled: Reducing acquisition time and radiation damage: data-driven subsampling for spectromicroscopy* https://doi.org/10.5281/zenodo.18470992.

Miller, E. C., Kasse, R. M., Heath, K. N., Perdue, B. R. & Toney, M. F. (2018). *J. Electrochem. Soc.* **165**, A6043–A6050.

Park, T. & Nakatsukasa, Y. (2025). *SIAM J. Matrix Anal. Applic.* **46**, 780–810.

Quinn, P. D., Alianelli, L., Gomez-Gonzalez, M., Mahoney, D., Cacho-Nerin, F., Peach, A. & Parker, J. E. (2021*a*). *J. Synchrotron Rad.* **28**, 1006–1013.

Quinn, P. D., Gomez-Gonzalez, M., Cacho-Nerin, F. & Parker, J. E. (2021*b*). *J. Synchrotron Rad.* **28**, 1528–1534.

Quinn, P. D., Sabaté Landman, M., Davis, T., Freitag, M., Gazzola, S. & Dolgov, S. (2024). *Chem. Biomed. Imaging* **2**, 283–292.

Townsend, O., Dolgov, S., Gazzola, S. & Kilmer, M. (2025). *arXiv*:2506.10661.

Townsend, O., Gazzola, S., Dolgov, S. & Quinn, P. (2022). *Opt. Express* **30**, 43237–43254.