Leveraging local continuity of connected measurements for dimensionality reduction and separation of backround, signal, and noise....

1. Introduction

Experiments undertaken at synchrotron and XFEL light sources can generate large volumes of data that are rich in information but plagued by noise and artifacts. There is often an intrinsic connectedness between successive measurements taken on a progression of different physical states: for example, a series of thermodynamic states in a phase diagram. There may also be further data redundancy, as many methods–such as powder diffraction and serial crystallography–require repeated measurements of the same system to accumulate sufficient statistics or orientational averaging. Consequently, large spectroscopic or diffraction datasets generated at third- or fourth-generation light sources are likely to contain sequences of measurements that combine smooth variation with respect to underlying physical parameters with stochastic variation due to counting statistics, detector characteristics, sample nonuniformity and other artefacts. These underlying physical parameters can be treated as ordinal variables to assemble a spectroscopic or (1D) diffraction dataset into a tensor of the form

$$X_{i_1 i_2 ... i_N, iq} \tag{1}$$

where the first $i1$ to i_N index the ordinal variables and iq indexes energy or momentum transfer in the case of XRD or spectroscopy, respectively.

The connectedness of X with respect to indices $i_1, i_2, ...i_N$ describes both the sample-derived signal and any smoothly-varying background, while other contributions to the measured X-ray intensity–such as shot noise, one-off artifacts, and stochastic variations across different portions of the sample–are independent or at least substantially stochastic. We label the signal and stochastic components of X as S and η, respectively: $X = S + \eta$.

elaborate that, if the dataset is underssampled. η is not entirely a result of statistical uncertainty; i.e. it also includes epistemic uncertainty

In the particular setting of powder diffraction analysis one can additionally exploit other prior information about the data: for example, the signal is concentrated in peaks that are narrow in q-space, which allows further separation of S into crystalline and diffuse diffraction components.

In this work we present an approach for separating connected measurements into their components S and η. We then apply this separation method to two germane x-ray diffraction analyses: (1) robust identification of powder diffraction peaks and (2) a feature extraction method for combinatorial materials discovery datasets that addresses the problem of peak-shifting, a long-standing impediment to the automated reconstruction of phase diagrams in the field of high-throughput materials discovery. To demonstrate the method's robustness and generalization we consider cases in which the space of measurement conditions are one- and two-dimensional: respectively, a temperature study of xxxx material and a combinatorial study of the CoNiTi ternary system.

These results suggest that handling connected datasets in the proposed fashion brings two significant benefits: (1) we get simple signal-to-uncertainty heuristics that allow more principled modeling and interpretation of measurements, and (2) prior physical information becomes captured in geometry, since Euclidean distance between sites in the data tensor X has a direct relation to distance in measurement condition-space.

there's probably a better way of saying this. and probably fits better in the abstract or conclusion. maybe elaborate on item (1), mentioning its its importance to (2) and any other downstream analysis where one would want to use signal-to-noise heuristics.

say why this is new and good, motivate and compare to prior efforts in the community

2. Methods

2.1. Prior approaches

Summarize approaches using simple similarity measures, such as cosine distance in the diffraction space.

2.2. Background subtraction

The first step to estimating the background is to identify the peak regions, which must be excluded before calculating the background. In the case of typical diffraction data a simple high-pass DFT filter does not cleanly extract the peaks when phase is retained in the inverse Fourier transform. This is because of the presence of ringing artifacts as well as the high concentration of signal in the diffraction peaks, which pile on top of the background in the low-frequency region of the power spectral density. On the other hand we can identify peak regions by padding the signal in the q dimension, taking a 1D DFT, and suppressing low-frequency components with a simple filter:

$$N(0,\sigma) \circledast |F^{-1}(HF(q))|, \tag{2}$$

where N is a unit Gaussian in q with standard deviation σ is chosen to match the diffraction peak width, $\circledast$ denotes convolution, and H is a Blackman window

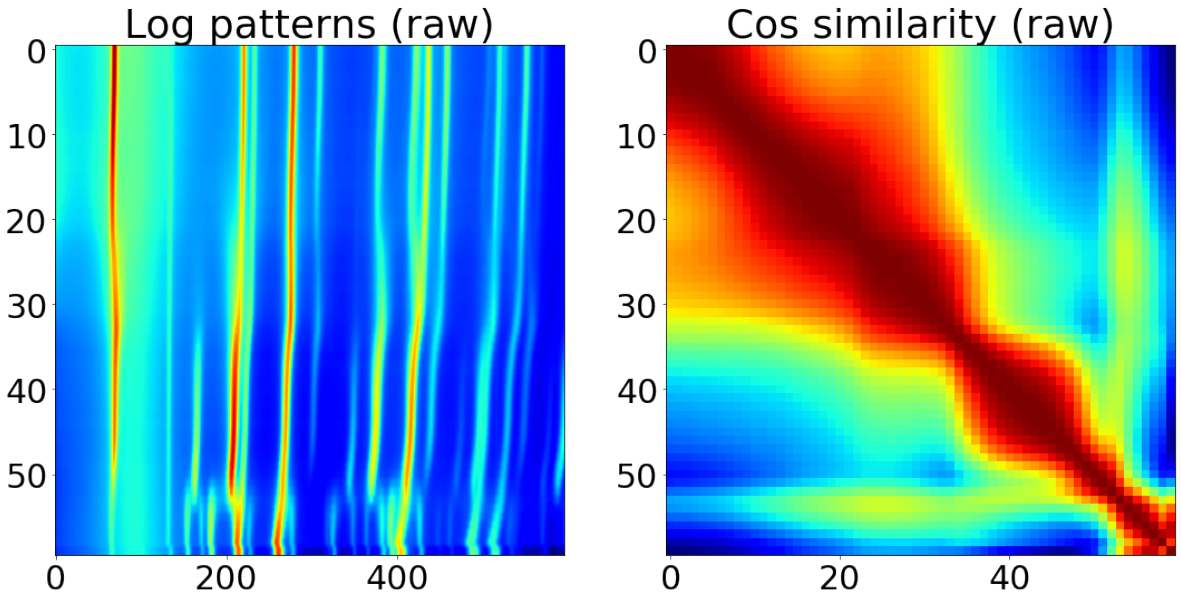


Figure 1.

this part will be fleshed out. some steps are still omitted

The background is then estimated by an interpolation using data from non-peak regions to fill in background intensities within the peak regions (Fig. 1(c)). Finally, it is subtracted from the denoised data (Fig. 1(d)).

To make for a simple extension to datasets of arbitrary dimensions we estimate the background by linear interpolation in the q dimension alone, together with an N-dimensional nearest-neighbor assignment to fill in points out of range for interpolation.

2.3. Noise estimation

While in the case of intensity variation in the q dimension we assume that high-frequency features belong to the informative (i.e. physical) component of the signal, in the case of the ordering dimensions we make the opposite assumption that the physically relevant part of the signal varies smoothly from one XRD pattern to any adjacent one, while any high-frequency, uncorrelated deviations from this progression are due to either noise (from e.g. detector characteristics or Poisson statistics) or artifacts (such as insufficient orientational sampling of the diffracting crystallites). Under this assumption the signal and noise lack substantial overlap in the N-1-dimensional Fourier transform along ordering dimensions; thus we can use a simple DFT filter to separate them (Fig. 2)

The above estimations of background and noise components separate the signal into estimates of the signal component S (which in turn separates into background + diffuse scattering and diffraction) and uncertainty values that are interpreted as single samples of an unknown underlying noise distributions. Despite its poorly-characterized nature, the latter of these can be used to propagate uncertainties to any downstream

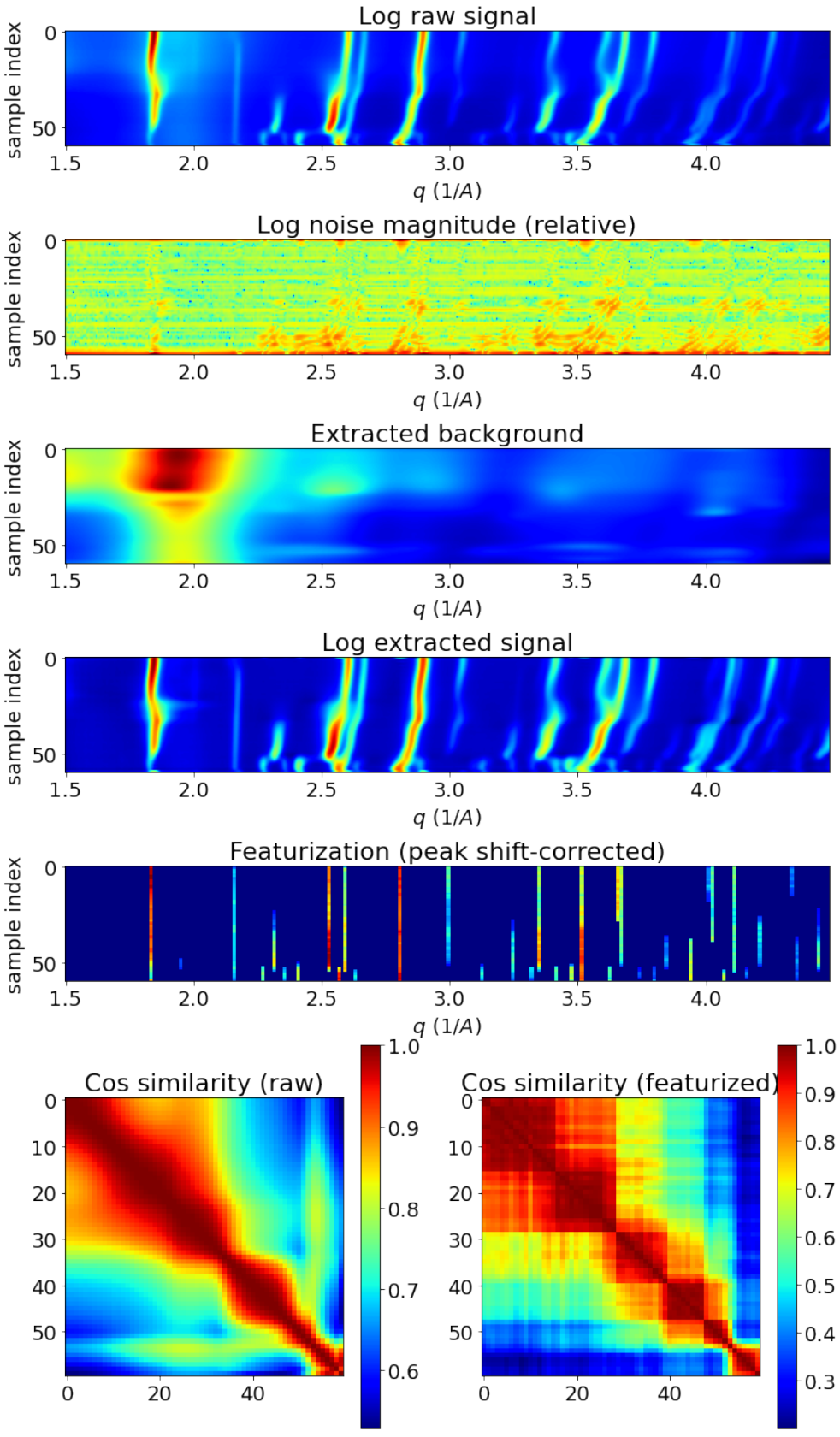


Figure 2. (a) Heatmap of XRD dataset X_{iq} corresponding to a temperature scan of XRD patterns for xxxxx system. The vertical index i and horizontal index q index temperature and momentum transfer, respectively.

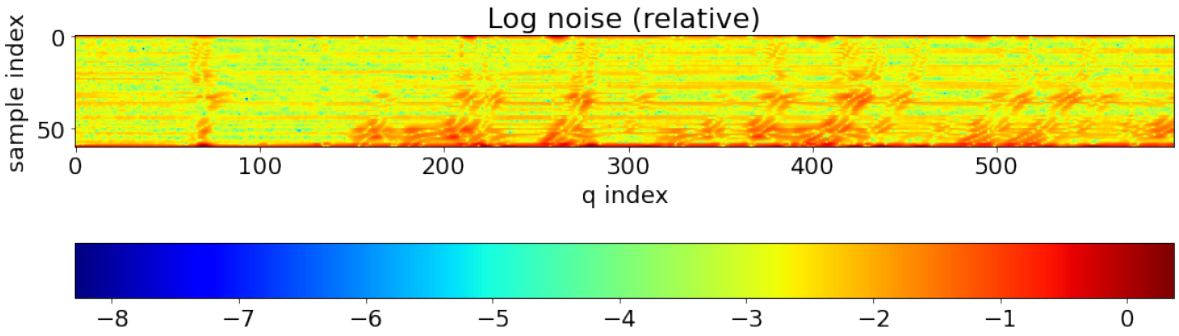


Figure 3.

analyses.

We label the signal and background $\hat{S}$ and B, respectively, so that $X = \hat{S} + B + \eta$.

some discussion is needed on the difference in interpretation of the noise estimate in undersampled vs. sufficiently sampled datasets

3. Methods and Discussion

3.1. Application: feature extraction

The above analysis addresses the data itself, with no scientific interpretation aside from the separation between crystalline and diffuse contributions to the scattering signal. However, we find that the same ordering and smoothness properties are useful for reducing the data into salient *information*, where the goal is to find physically-meaningful boundaries in the ordering dimensions corresponding to, e.g., the surface separating a single-phase region from surrounding multi-phase regions in a combinatorial diffraction dataset. Specifically, we identify diffraction peaks in every XRD pattern, independently, and then link each peak in a pattern to any q-adjacent peaks in neighboring patterns (i.e. patterns that are adjacent in $i_1, i_2 ... i_N$).

Each contiguous set of peaks linked in this way is denoted as a feature, and the ensemble of features defines a new representation of the dataset $Y_{i_1,i_2...i_N,j}$, where j indexes the peak features. The entries of Y are 0 where the corresponding pattern lacks feature j, and equal to the intensity of the feature-peak combination otherwise.

Peak parameters are obtained using a curve-fitting procedure that relies on Scargle's Bayesian Block algorithm to segment each measurement and then performs iterative peak-fitting on every block through a nonlinear least squares optimization on a sum of peak profiles. A Voigt profile is used in this work as it is appropriate for powder diffraction data.

Peak profiles are added to the fit curve until the fit residual satisfies

$|R \oslash \eta| < s$,

where s is a chosen threshold of order unity and $\oslash$ denotes elementwise division.

maybe talk about regularization here

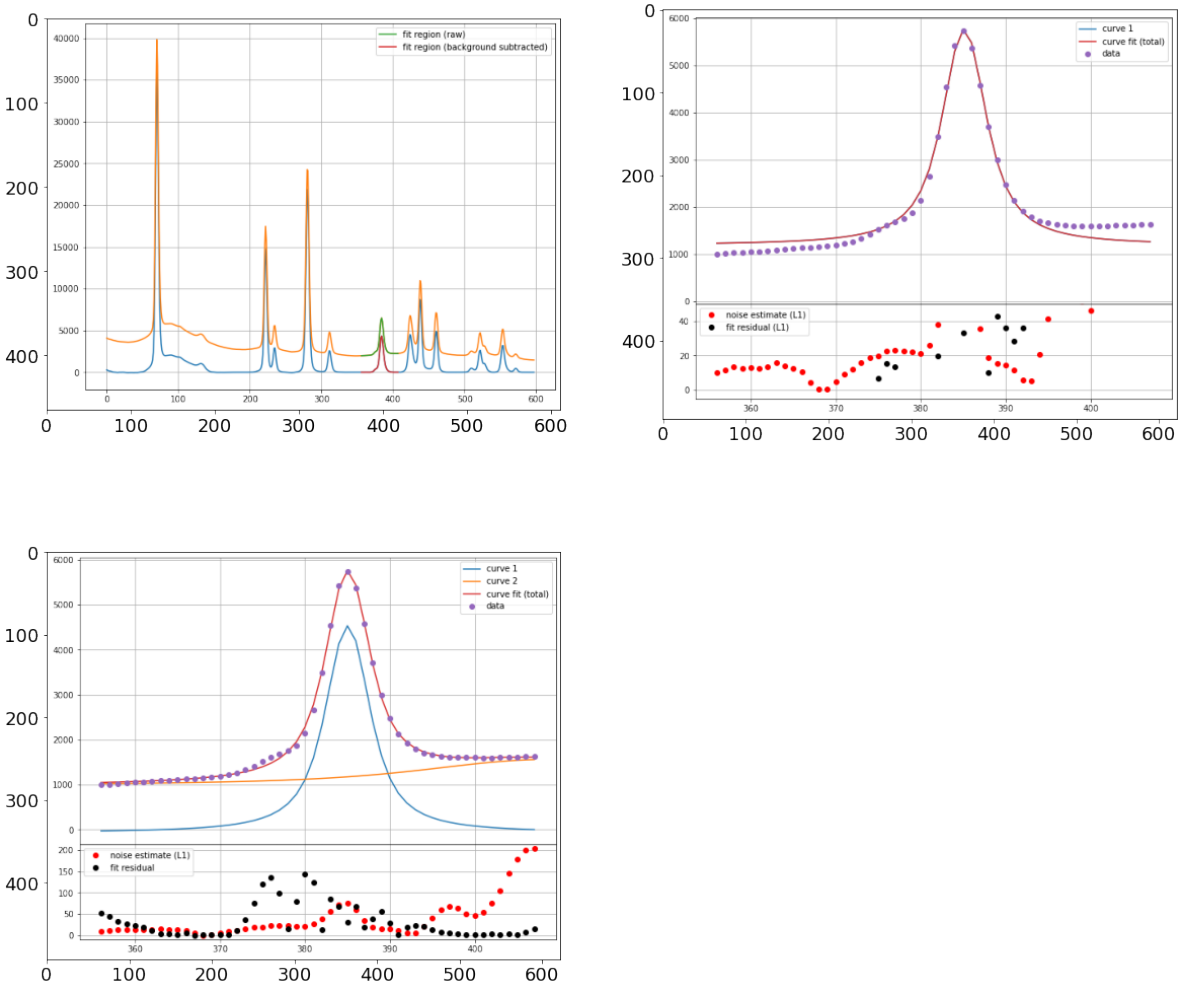


Figure 4. Example of the effect of background subtraction and noise estimation on peak fitting for one BBA block. This figure needs to be reformatted (font size, layout)

cover the utility of noise estimates and background subtraction for better peak fitting The separation of

4. Conclusions and future work

5. References

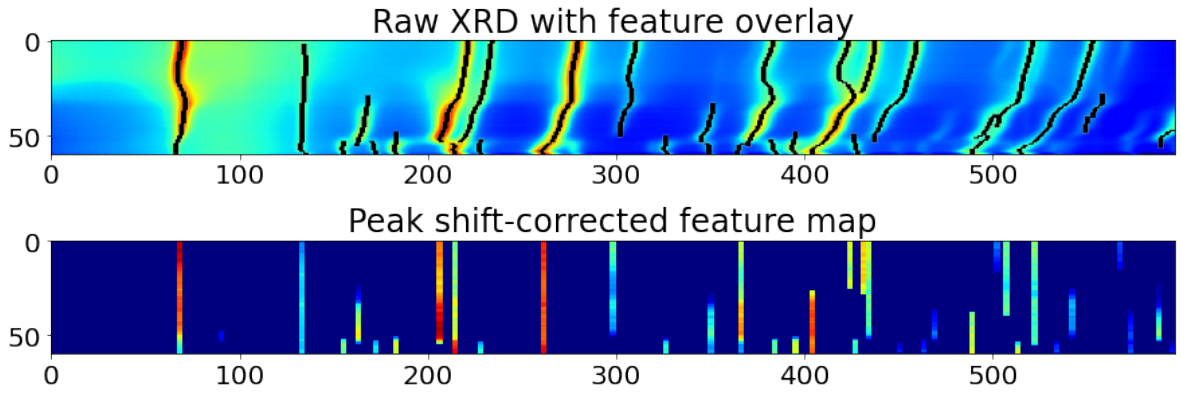


Figure 5. (a) Features identified from peaks connected in i, q space for XRD dataset X_{iq} corresponding to a temperature scan of XRD patterns for xxxxxx system. Index i denotes temperature.

(b) Intensity profile along each of the features in (a).

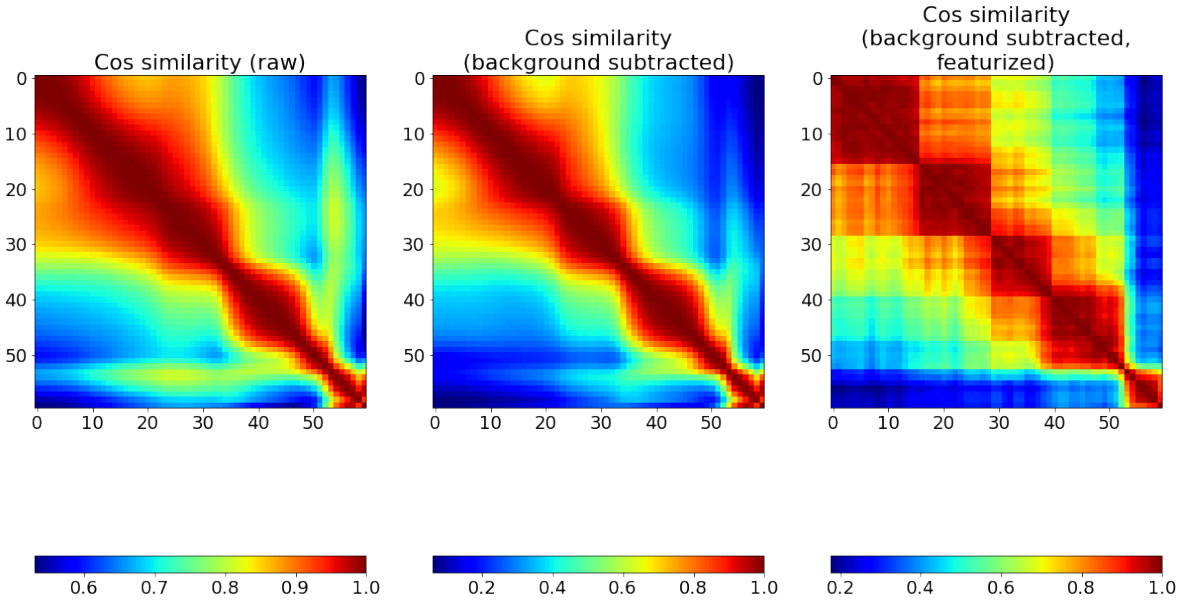


Figure 6.

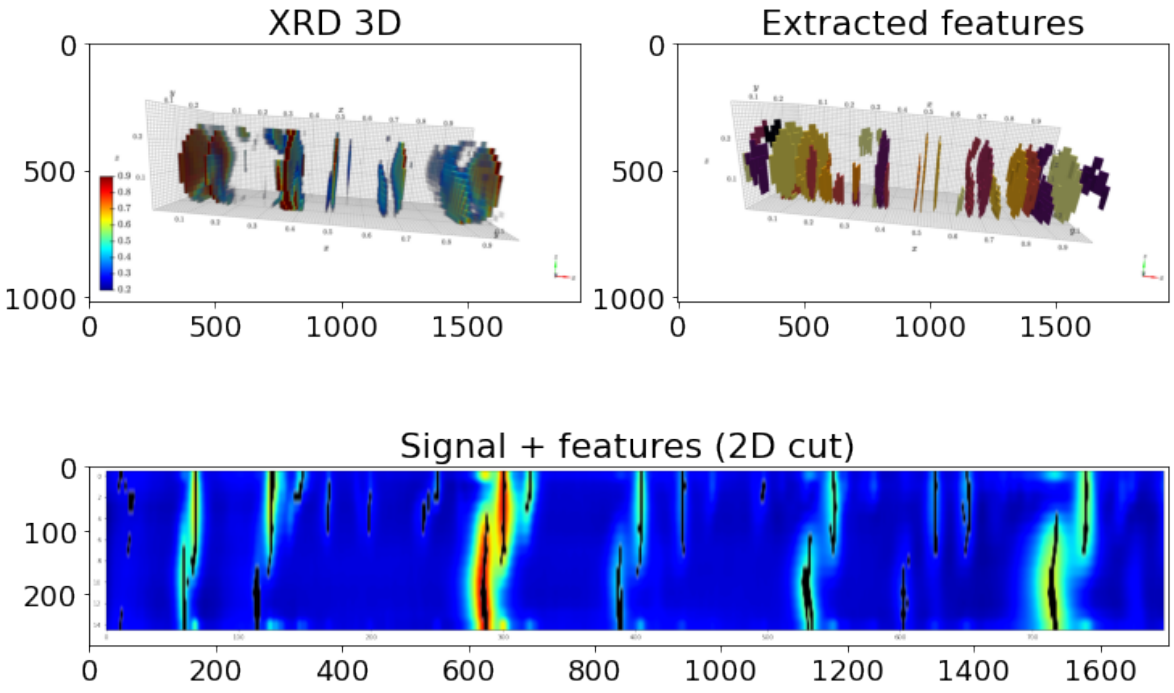
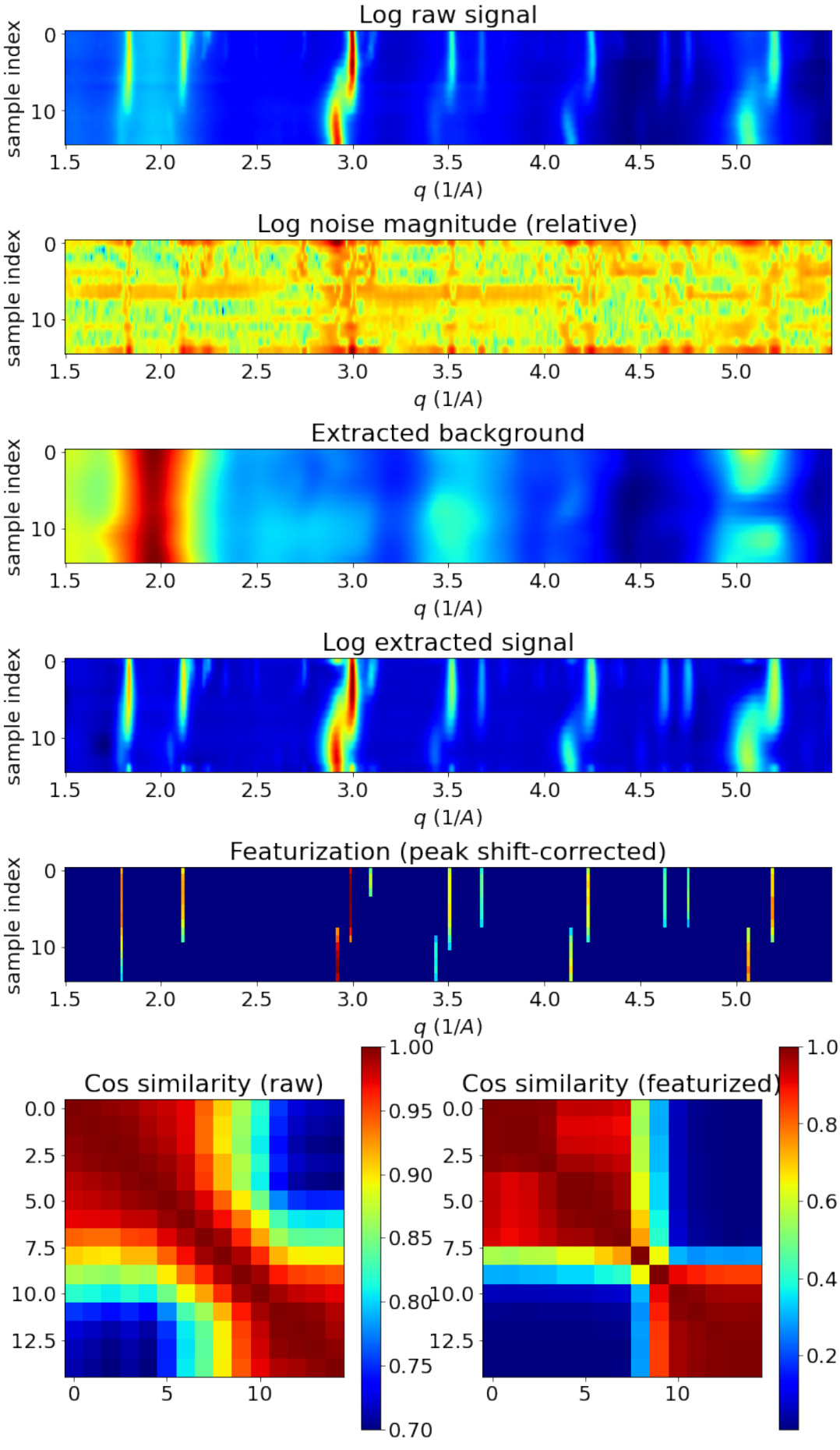


Figure 7. caption


Log raw signal
sample index
0
10
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
q (1/A)
Log noise magnitude (relative)
sample index
q (1/A)
Extracted background
sample index
q (1/A)
Log extracted signal
sample index
q (1/A)
Featurization (peak shift-corrected)
sample index
q (1/A)
Cos similarity (raw)
0.0
2.5
5.0
7.5
10.0
12.5
0
5
10
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Cos similarity (featurized)
1.0
0.8
0.6
0.4
0.2