Exploiting proximal similarity for improved interpretation of scientific measurements

Abstract. The inability to control for unknown uncertainty, base-line drifts (background) and other artifacts are often makes a meaningful comparison between datasets collected under different conditions difficult. The underlying manifold can change not only with experimental variables, but also with uncontrollable or unforeseen variables. To achieve statistical significance in a dataset, effort is made to reproduce the measurements under identical conditions to minimize variability due to uncontrolled conditions. It is commonly assumed that the underlying manifold changes negligibly between repeated measurements. One under-utilized axiom is that the sampling frequency is often greater than the rate of change. Here, we extend the assumption of local similarity to all the other dimensions of the manifold to devise a strategy to assess background, noise, and boundaries from the local rate of variation. We illustrate this strategy on temperature and composition-dependent wide angle X-ray scattering measurements to locate phase changes without human curation. Furthermore, we show that the proposed strategy naturally suggests an approach for capturing rare events in a data-frugal manner for a wide range of scientific problems,and could have a large impact on energy efficient remote-sensing, space telescopy and experiments at high repetition rate, and high brightness light sources.

1. Introduction

It is generally important in science for experimental measurements to be accompanied by uncertainties, yet it is quite common for those uncertainties–both aleatoric and epistemic–to be partially or entirely unavailable. This may be from a combination of factors: e.g. poorly characterized backgrounds, lack of sufficient models for detector characteristics, and stochastic variations in properties of the observed system itself that are unrelated to the physics of interest yet confound the analysis. At the same time, there may be significant reduncancy across measurements with respect to the underlying signal.

Diffraction and spectroscopy experiments undertaken at synchrotron and XFEL light sources are an example of such a setting: they generate large volumes of data that are rich in information but plagued by contributions such as shot noise, beam energy and intensity jitter, sample nonuniformity and detector artifacts. However there is often an intrinsic connectedness between successive measurements taken on a progression of different physical states: for example, values of the state variables in a thermodynamic study. 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 a uncertainty whose underlying distribution is *a prior* unknown but that is discontinuous and substantially uncorrelated from one measurement to the next.

Taking a series of 1d measurements, the relevant underlying physical parameters can be treated as ordinal variables to the dataset into a tensor of the form

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

where the first i_1 to i_N index the ordinal variables and iq indexes the 1d measurement grid (q denotes momentum transfer in 1d x-ray diffraction, which for concreteness we will assume is the experimental modality).

The connectedness of X with respect to indices $i_1, i_2, \ldots i_N$ describes both the sample-derived signal and any smoothly-varying background. 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 types of 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 physical parameter 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 importance to (2) and any other downstream analysis where one would want to use signal-to-noise heuristics.

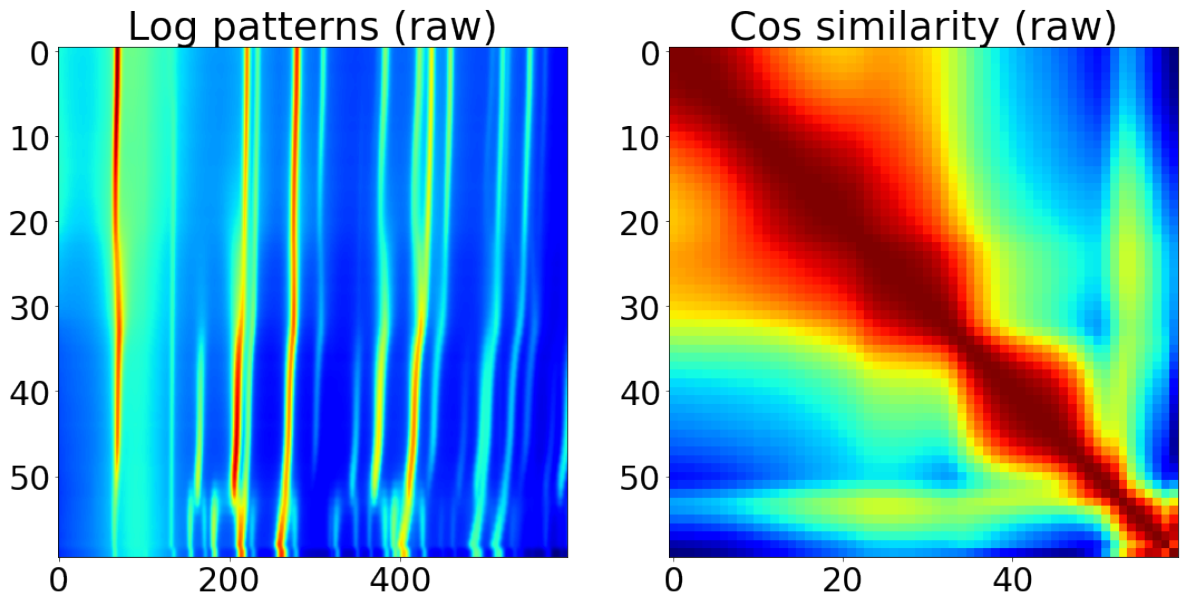

Figure 1. (a) Heatmap of 60 experimental diffraction patterns of the XXX system at diffferent temperatures, ordered vertically in decreasing temperature. (b) Heatmap of cosine similarity between pairs of those diffraction patterns (a)

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 with the following transformation:

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

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

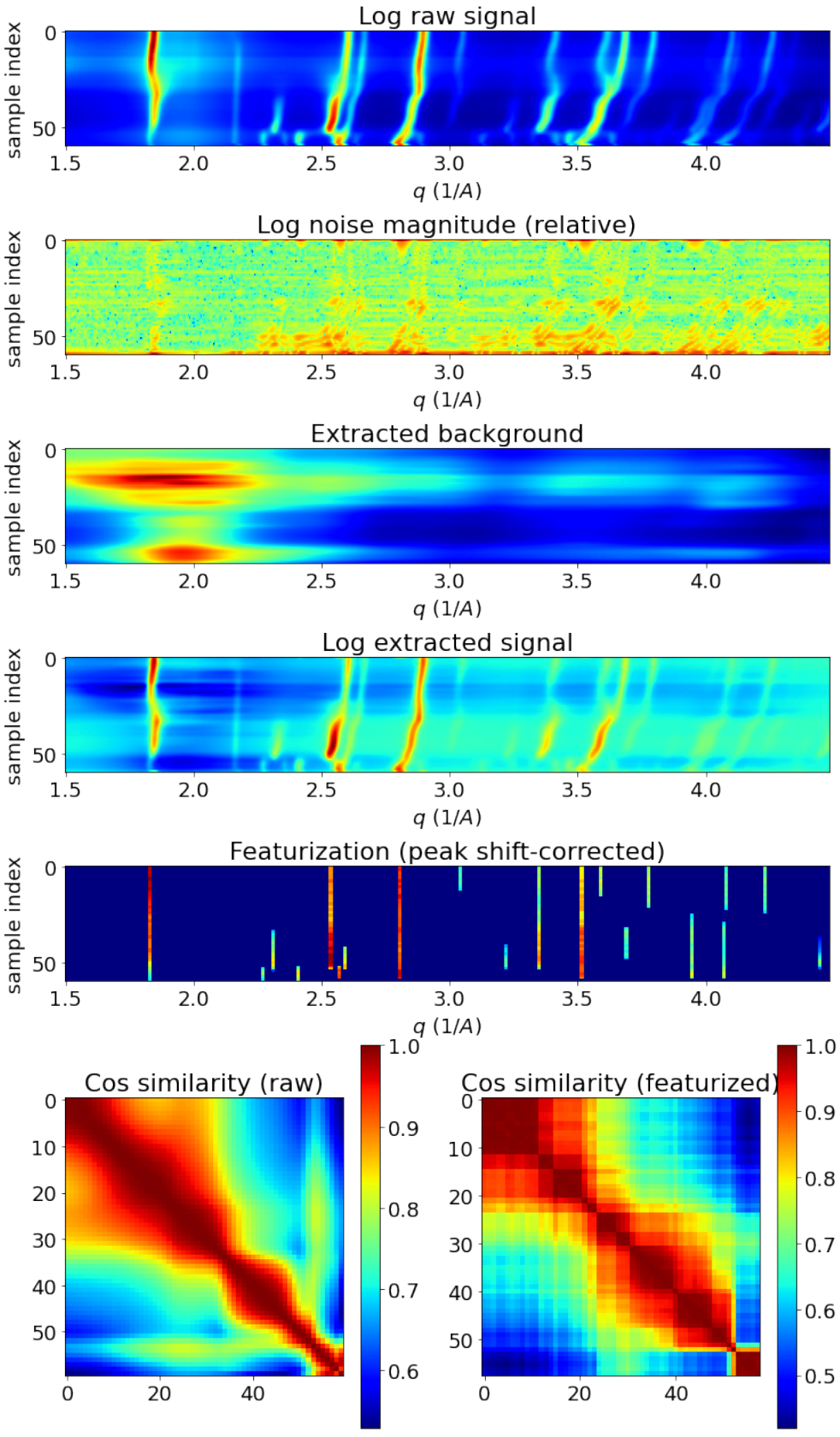

Figure 2. (a) Heatmap of XRD dataset X_{iq} corresponding to a temperature scan of XRD patterns for xxxxx system. The vertical index i and horizontal index q index temperature and momentum transfer, respectively.

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

The background is then estimated by smoothed nearest-neighbor interpolation in multiple dimensions, wherein data from non-peak regions is used to fill in background intensities within the peak regions (Fig. 1(c)). An adjustable threshold parameter t specifies the fraction of pixels belonging to peak regions; favorable values of t depend mainly on the noise level and density of diffraction peaks in a particular dataset, and are determined manually. Finally, the background estimate 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 a 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 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 1d measurement into signal regions (i.e. blocks) 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.

For an individual block comprised by a set B of momentum transfers, peak profiles are added to the fit curve until the fit residual R satisfies

$$(\Sigma_{q \in B} \frac{1}{N} \frac{R_q^2}{\eta_q^2}))^{1/2} < s, \tag{3}$$

where s is a threshold parameter of order unity.

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

4. Supplemental figures

5. Conclusions and future work

In this section, we propose using this type of approach for event detection in a high-data rate environment. We can also talk about the prospect that our peak shift-correction will help solve the phase-mapping problem.

6. References

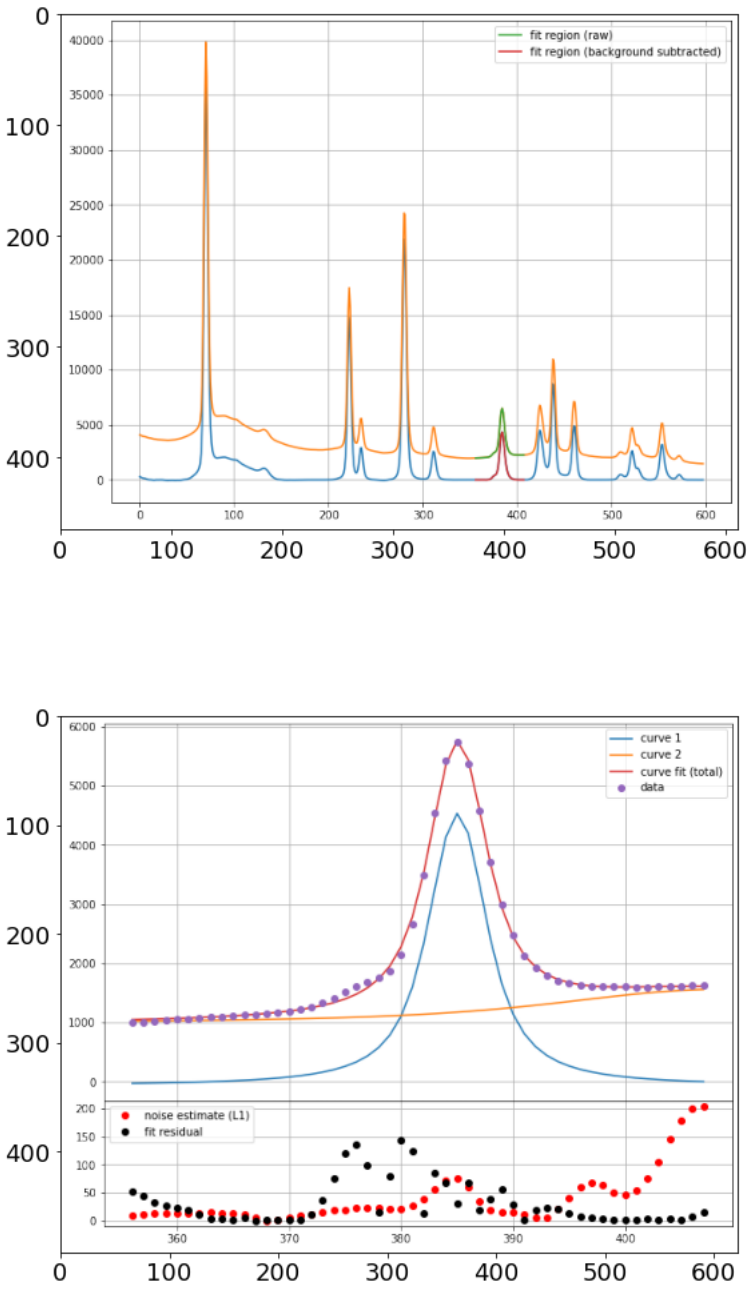

Figure 3. (a) Powder diffraction pattern of the xxx system showing background subtraction and a single Bayesian block in red. (b) Iterative peak fitting of the single block using estimated noise relative to curve fit residual to determine the number of peak profiles to include.

This figure needs to be reformatted (font size, layout). I will add a third panel showing the effect of background subtraction on the number of peak profiles found.

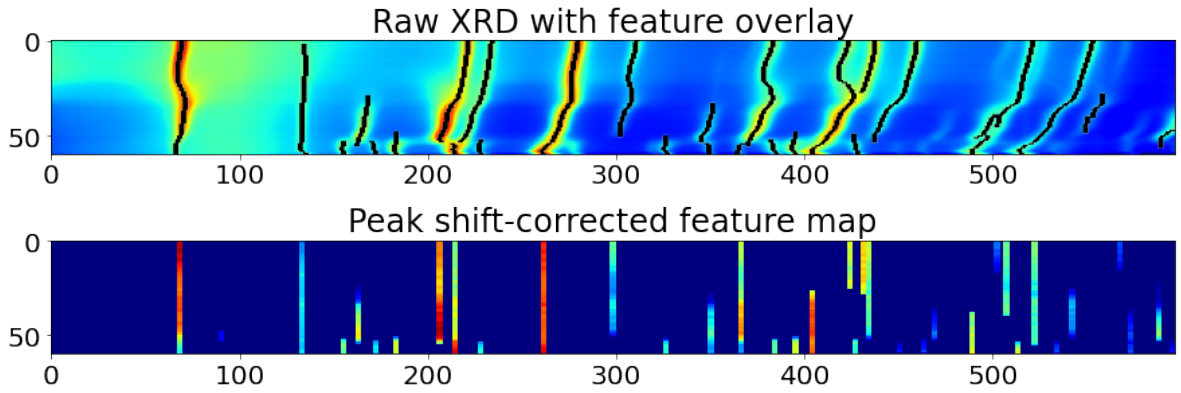

Figure 4. (a) Features identified from peaks connected in i, q space for XRD dataset X_{iq} corresponding to a temperature scan of XRD patterns for xxxxxx system. Index i denotes temperature.

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

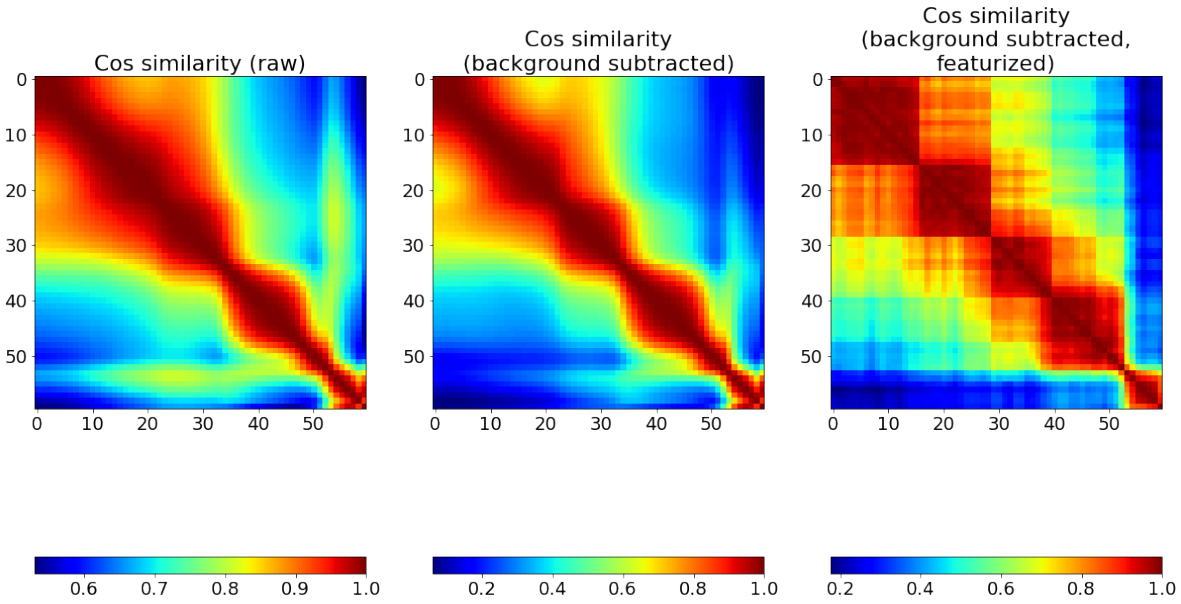

Figure 5. Comparison of cosine similarity squares for (a) raw powder diffraction of the xxx system, (b) post background subtraction, and (c) peak shift-corrected feature vectors derived from the same dataset.

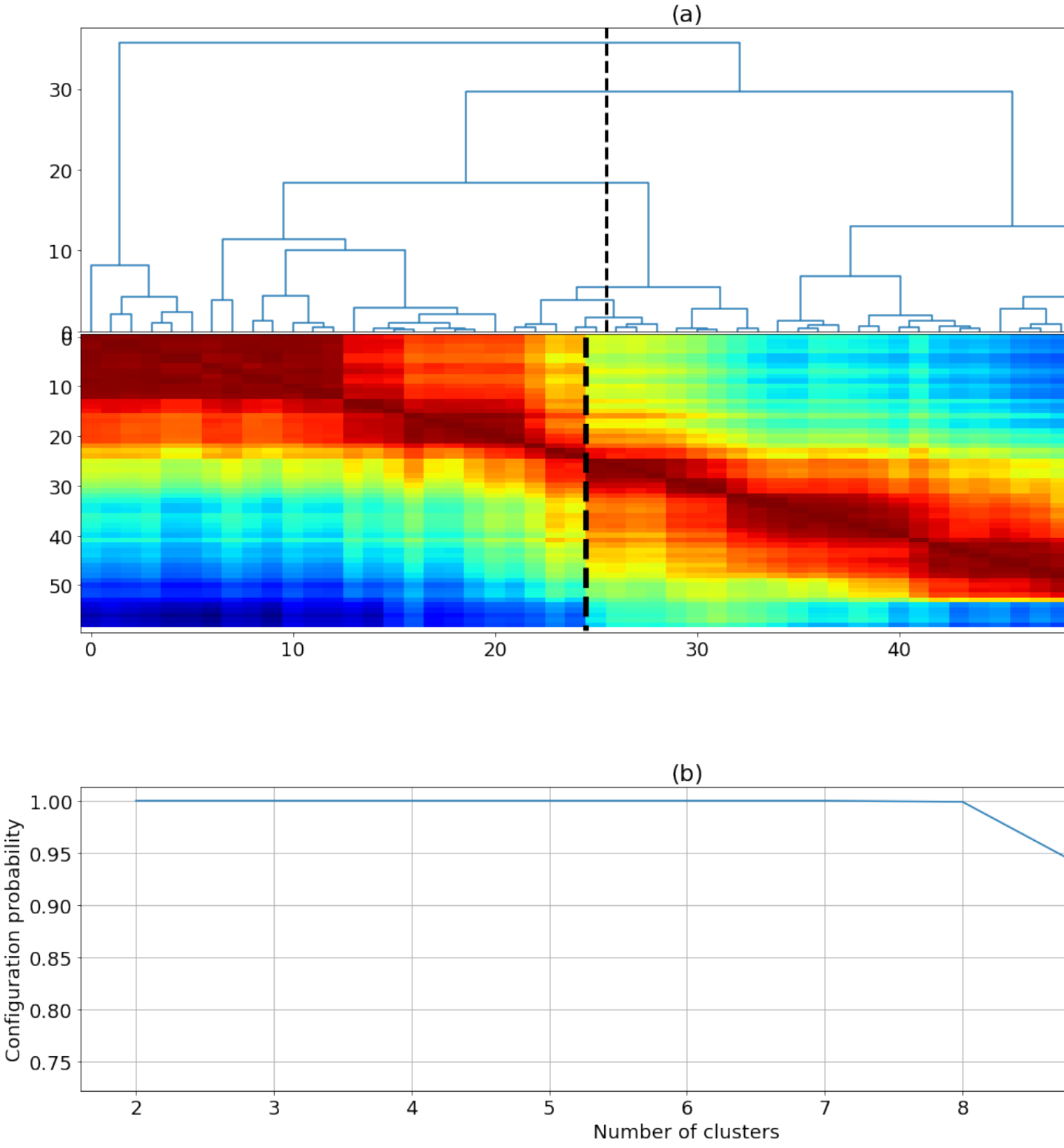

Figure 6. (a) Optimal clustering for a set of XRD patterns demonstrating a temperature progression in xxxx system. The optimal number of clusters is determined by clustering stability over an ensemble of XRD patterns generated by draws from a according to the simple noise model described in section xxxx.

(b) Ensemble probability of the most probable cluster configuration for agglomerative clustering with Ward linkage criterion, for the same dataset. Horizontal axis varies the number of clusters.

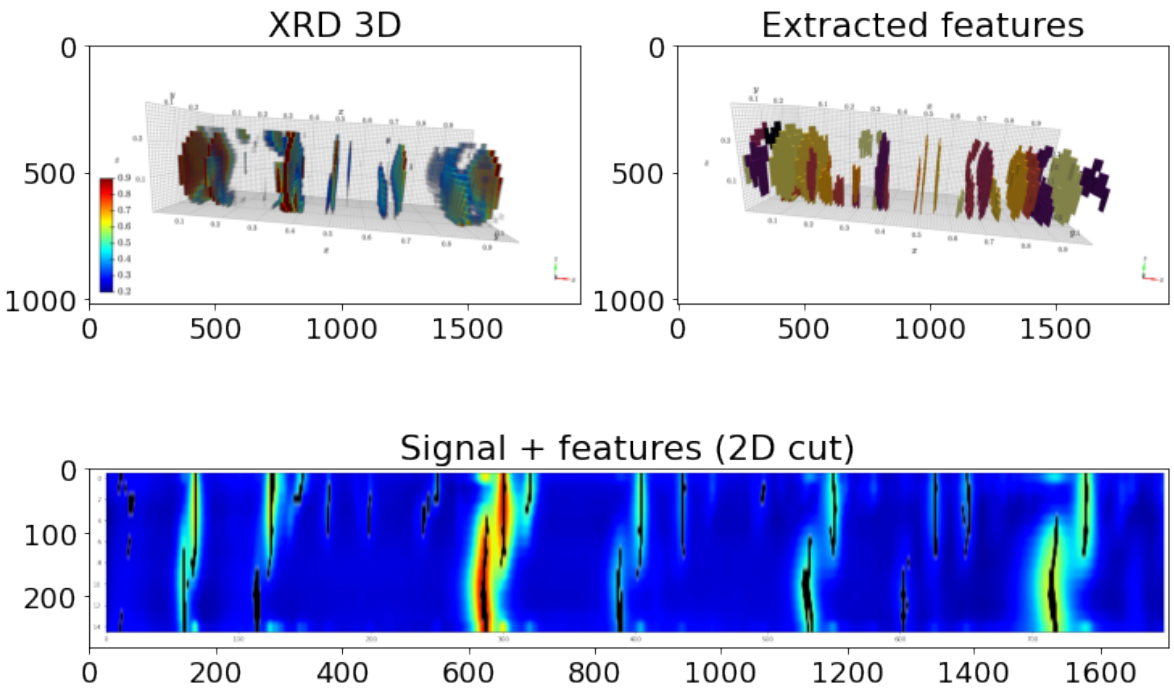

Figure 7. (a) volumetric heatmap for a three-dimensional diffraction study of the CoNiTi ternary system. Dimensions x any y are coordinates on the compositional simplex, while dimension z indexes momentum transfer. (b) Three-dimensional rendering of peak features identified by the shift-correcting procedure described in section xx. (c) Two-dimensional cut of the dataset displayed in (a), with peak features overlayed in black.

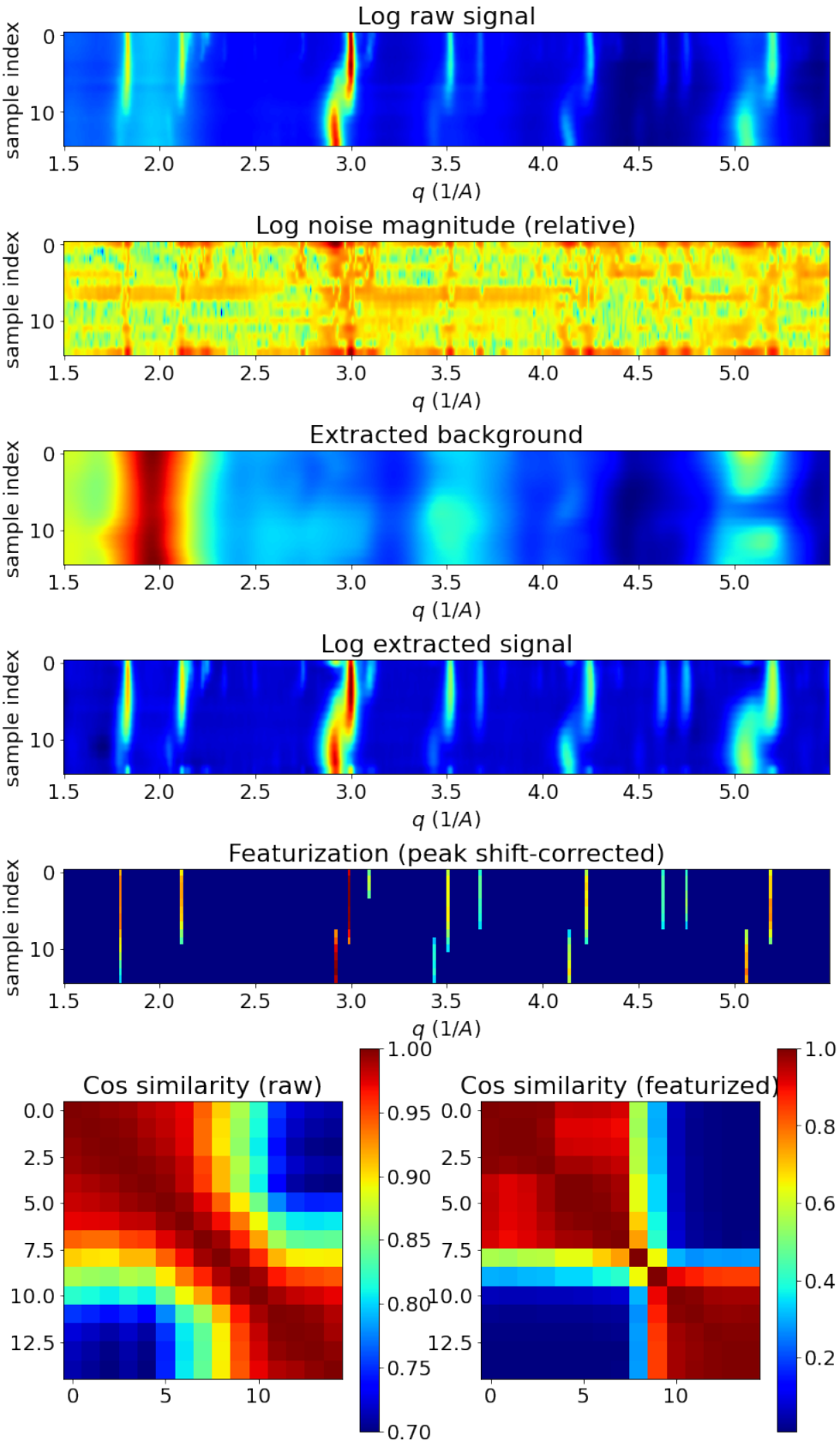

Figure 8. Separation and feature extraction for a 2d slice of the CoNiTi ternary system.