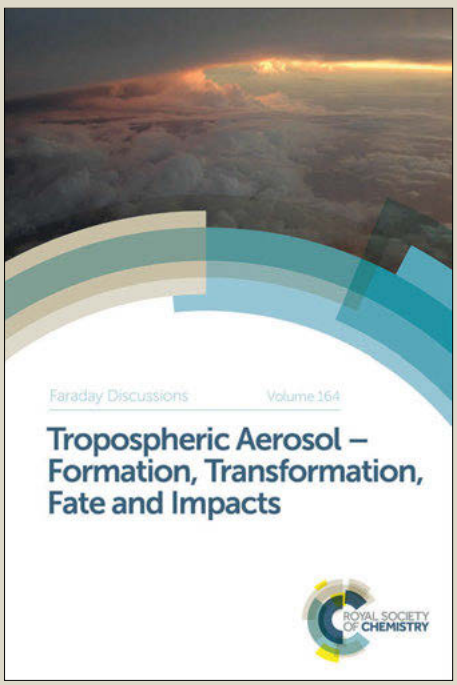
# Faraday Discussions

Accepted Manuscript


This manuscript will be presented and discussed at a forthcoming Faraday Discussion meeting. All delegates can contribute to the discussion which will be included in the final volume.

**Register now to attend!** Full details of all upcoming meetings: http://rsc.li/fd-upcoming-meetings




This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

*Accepted Manuscripts* are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.


ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/faraday_d

CREATED USING THE RSC LaTeX PCCP ARTICLE TEMPLATE – SEE www.rsc.org/electronicfiles FOR DETAILS
PAPER www. rsc.org/faraday_d | Faraday Discussions

# Non-negative matrix analysis for effective feature extraction in x-ray spectromicroscopy

**Rachel Mak,[a] Mirna Lerotic,[b] Holger Fleckenstein,[c] Stefan Vogt,[d] Stefan M. Wild,[e] Sven Leyffer,[e] Yefim Sheynkin,[f] and Chris Jacobsen*[d,a,g]**

***Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX***
***First published on the web Xth XXXXXXXXXX 200X***
**DOI: 10.1039/c000000x**

X-ray absorption spectromicroscopy provides rich information on the chemical organization of materials down to the nanoscale. However, interpretation of this information in studies of "natural" materials such as biological or environmental science specimens can be complicated by the complex mixtures of spectroscopically complicated materials present. We describe here the shortcomings that sometimes arise in previously-employed approaches such as cluster analysis, and we present a new approach based on non-negative matrix approximation (NNMA) analysis with both sparseness and cluster-similarity regularizations. In a preliminary study of the large-scale biochemical organization of human spermatozoa, NNMA analysis delivers results that nicely show the major features of spermatozoa with no physically erroneous negative weightings or thicknesses in the calculated image.

## 1 Introduction

Images let us see what is present in a material, while spectra let us understand what we see. Combining the two in spectromicroscopy (also known as spectrum imaging, or hyperspectral imaging) provides rich data on the composition of complex materials, whether applied to electron energy loss in electron microscopy[1,2], x-ray emission spectroscopy with x-ray excitation[3], or electron ex-

[a] *Department of Physics & Astronomy, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA. E-mail: rachel.mak@u.northwestern.edu*
[b] *2nd Look Consulting, Room 1702, 17/F, Tung Hip Commercial Building, 248 Des Voeux Road, Hong Kong*
[c] *Center for Free-Electron Laser Science (CFEL), DESY, Notkestrasse 85, 22607 Hamburg, Germany*
[d] *Advanced Photon Source, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA*
[e] *Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA*
[f] *Department of Urology, Stony Brook Medicine, Stony Brook, NY 11794-8093, USA*
[g] *Chemistry of Life Processes Institute, Northwestern University, 2170 Campus Drive, Evanston, IL 60208, USA*

citation[4], infrared microscopy[5,6], or x-ray absorption microscopy[7,8]. The challenge we address here involves the interpretation of these data, which is required in order to go from observation to understanding. With spectroscopy of pure, uniform substances there exists a long and rich tradition of understanding observed spectra based on calculations of various electron or phonon interactions in the substance (see, for example, Stöhr[9]). However, microscopy is used to study materials including heterogeneous mixtures and reactive phases on fine spatial scales, and in images of $10^5$–$10^7$ pixels. It is clearly impractical to carry out a painstaking investigation of the spectrum of each pixel on its own. Instead, one can hope to find a reduced set $\mathcal{S}$ of spectra that, when combined, can reproduce the spectrum observed in any one pixel. One can then carry out analysis on this smaller set of spectra or compare them to spectral "standards" of materials expected to be present in the specimen. We describe here an approach to carrying out this analysis based on a non-negative matrix approximation (NNMA, also referred to in the literature as NMF)[10], comparing it with previous methods we have developed (e.g., cluster analysis), and showing its utility for imaging chemical states in complex materials such as human sperm.

In x-ray spectromicroscopy, one obtains transmission images $I(x, y, E)$ at a series of positions $(x, y)$ and $N$ different photon energies $E$. By knowing the incident flux $I_0(E)$, one can determine an optical density $D(x, y, E) = -\ln[I(x, y, E)/I_0(E)]$, which is linear in the thickness $t$ of the absorbing material in the beam direction because of the Lambert-Beer law of $I = I_0 \exp[-\mu(E)t]$. In this expression, $\mu(E)$ is a photon-energy-dependent linear absorption coefficient, which in principle can be calculated from quantum mechanics, and which in practice is often obtained from tabulations of absorption per element and per energy[11]. Missing from these tabulations are the details of $\mu(E)$ in the vicinity of an x-ray absorption edge: rather than reaching the threshold energy to excite and remove a core-level electron from an isolated atom, one instead reaches an energy where an atom's electron can be promoted into a state with an energy only a few electron volts away from the Fermi energy. Since these near-vacuum energy states are strongly affected by the nature of the atom's chemical bonds[9], spectromicroscopy using near-edge x-ray absorption fine structure (NEXAFS) or x-ray absorption near-edge structure (XANES) provides a way to image the element-specific chemical bond distributions in a complex material.

Our challenge is that what has been measured is simply the optical density $D(x, y, E)$, but we would like to interpret it as a product of absorption spectra $\mu(E, \mathcal{S})$ from a set $\mathcal{S}$ with $S$ spectroscopically distinguishable components and a set of thickness maps or weighting images $t(\mathcal{S}, x, y)$ that show how much of each spectrum is present at each pixel. If we do not seek to find any spatial correlation of spectral responses (i.e., we do not assume the spectral response of any given pixel to be correlated with that of its neighbors), we can flatten the two-dimensional $(x, y)$ coordinates and use a one-dimensional coordinate $p$ to represent the position of each pixel. This is also generalizable to 3D tomographic spectromicroscopy data[12]. We are therefore left with a matrix equation for our desired analysis of

$$\mathbf{D}_{N\times P} = \boldsymbol{\mu}_{N\times S}\, \mathbf{t}_{S\times P}, \tag{1}$$

where $N$ denotes the number of photon energy indices and $P$ the number of pixels[13]. Our goal is to find the set of spectra $\boldsymbol{\mu}_{N\times S}$ that describes all the signifi-

cant variations in the data. The absorption spectra $\boldsymbol{\mu}_{N\times S}$ should be non-negative (since negative absorption would imply that the material is adding energy to the transmitted beam instead of absorbing energy from it); the thickness or weighting maps $\mathbf{t}_{S\times P}$ should likewise be non-negative because of the additive nature of the densities of the materials in the sample. Because $\mathbf{D}_{N\times P}$ measures the optical density $-\ln(I/I_0)$, which is always non-negative (barring errors in the incident flux $I_0$ normalization), it should be possible, in theory, to find non-negative $\boldsymbol{\mu}_{N\times S}$ and $\mathbf{t}_{S\times P}$ such that Eq. 1 is satisfied.

The problem of analyzing the measured data $\mathbf{D}_{N\times P}$ in terms of a set of spectra $\boldsymbol{\mu}_{N\times S}$ has been the subject of numerous multivariate statistical analysis approaches in energy loss electron microscopy[14,15] and in infrared spectromicroscopy[5,16]. In x-ray spectromicroscopy, approaches using spectral standards or hand-defined regions assumed to be of uniform, pure composition have allowed one to obtain a set of $S$ spectra $\tilde{\boldsymbol{\mu}}_{N\times S}$ from which thickness maps $\mathbf{t}_{N\times P}$ can be obtained[17,18] by using the singular value decomposition (SVD) for matrix inversion, yielding

$$\mathbf{t}_{S\times P} = \tilde{\boldsymbol{\mu}}^{\dagger}_{S\times N}\,\mathbf{D}_{N\times P}, \tag{2}$$

where $\tilde{\boldsymbol{\mu}}^{\dagger}_{S\times N}$ is the pseudo-inverse of $\tilde{\boldsymbol{\mu}}_{S\times N}$.

Approaches for understanding more complex samples in x-ray microscopy have included the use of principal component analysis (PCA)[19,20] to identify a limited or significant basis set $\bar{S}$ of orthonormal spectral signatures. However, SVD inversion does not guarantee a non-negative thickness map $\mathbf{t}_{N\times P}$, and PCA can produce a basis set $\tilde{\boldsymbol{\mu}}_{N\times S}$ that includes both positive and negative spectral values. Therefore, these approaches do not satisfy the non-negative condition of our desired solution described in Eq. 1.

## 2 Cluster analysis and negative values

Although PCA does not provide a set of spectra that are individually interpretable as positive absorption spectra of separate materials present in the specimen, it *does* provide a well-organized and reduced-dimensionality search space for cluster analysis[13,21] as a way of finding pixels with similar spectra. Once the clusters

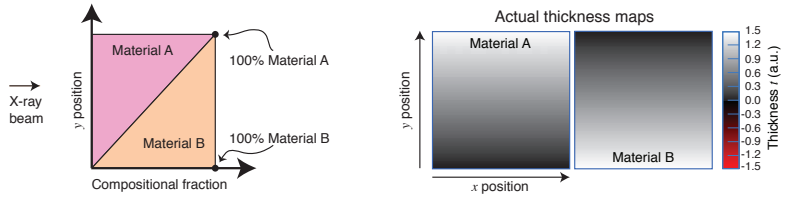



**Fig. 1** Two-material test specimen to illustrate compositional mapping approaches. We assume that the specimen is a three-dimensional block comprises two separate materials A and B with a continuous variation between the two. The compositional variation is shown at left: the view is along the $x$ direction (into the page), while the x-ray beam is traveling along the $z$ direction. The thickness maps associated with each of the separate materials are shown at right.

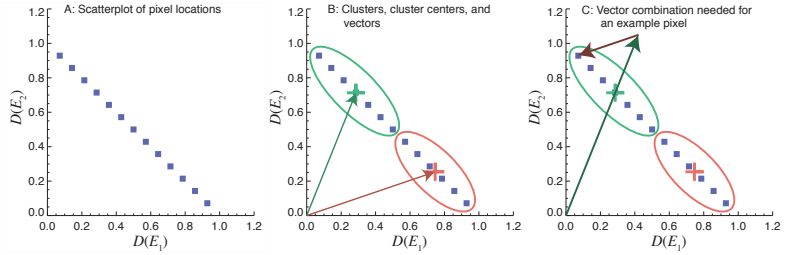



**Fig. 2** Illustration of how cluster analysis can give rise to negative (and thus unphysical) values of $\mathbf{D}_{N\times P} = \boldsymbol{\mu}_{N\times S}\,\mathbf{t}_{S\times P}$. The figure at the left (A) shows a scatterplot of pixels from a continuously varying material combination as shown in Fig. 1, assuming that the two materials have opposing responses at energies $E_1$ and $E_2$. The middle figure (B) illustrates how these pixels will be grouped into two clusters; without other information, it would be natural to describe each cluster by the spectrum corresponding to the cluster center (marked with the red and green $+$). The figure at the right (C) shows how recreation of the spectrum of a pixel near one of the axes in this illustration would require a combination of positive weighting of one cluster spectrum (green in this case) but a *negative* weighting of the other cluster spectrum (red in this case); this would give rise to negative values in the thickness or weighting matrix $\mathbf{t}_{S\times P}$, implying negative absorption of the x-ray beam, which is unphysical.

are found, the spectra calculated from each cluster center provide a set $\tilde{\boldsymbol{\mu}}_{N\times\bar{S}}$ for calculation of thickness weightings $\mathbf{t}_{S\times P}$ according to Eq. 2. Cluster analysis has proven useful for a variety of applications including soil and environmental analysis[22,23]; however, it is also observed to yield some regions with slightly negative values in the thickness maps $\mathbf{t}_{S\times P}$, which are unphysical and thus represent limitations in the analysis.

To understand the way in which non-negative errors can arise in cluster analysis, we consider a simple example of a specimen with uniform thickness and a continuum of composition starting with 100% of material A, which is strongly absorptive at energy $E_1$, and ending with 100% of material B, which is strongly absorptive at energy $E_2$ (see Fig. 1). A scatterplot of the location of individual pixels based on their responses at the energies $\{E_1, E_2\}$ is shown schematically in Fig. 2A. If these pixels are organized into two clusters, the groupings shown in Fig. 2B will be the result, where the vectors shown point to the center of the respective cluster centers; this will give rise to a set of spectra $\tilde{\boldsymbol{\mu}}_{N\times S}$ (with $S = 2$ in this example) from which one can calculate thickness maps according to Eq. 2. However, consider the case of a pixel that is far from the median in composition, such as the upper left one in Fig. 2C. The only possible thickness map $\mathbf{t}_{S\times P}$ or weighting map of the cluster spectra $\boldsymbol{\mu}_{N\times S}$ able to reach that point is one that involves a negative weighting of one of the cluster spectra; that is, one that produces negative values in the thickness map $\mathbf{t}_{S\times P}$, which are unphysical in our desired interpretation of the measured positive optical densities of Eq. 1. Of course, if the variation among spectral response of the pixels assigned to a cluster is small, these errors can be negligibly small; however, as Fig. 2 shows, there is no guarantee that cluster analysis will produce a thickness map with few negative pixels. Indeed, this negative thickness error is exactly what is observed

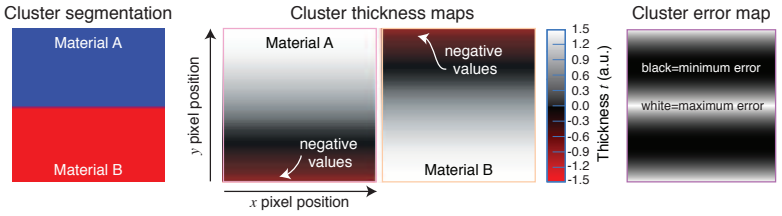



**Fig. 3** Cluster analysis applied to the $\mathbf{D}_{N\times P}$ simulated data of Fig. 1 using two different spectra. The "Cluster segmentation"' image at left shows how the simulated specimen is correctly classified as being made up of $S = 2$ distinct spectra $\boldsymbol{\mu}_{S\times N}$, while the "Cluster thickness maps" in the center show the weightings or thickness maps $\mathbf{t}_{S\times P}$ that result. Careful examination of these cluster thickness maps reveals that they include (unphysical) negative values, and the "Cluster error map" at right (which represents the error $|\mathbf{D}_{N\times P} - \boldsymbol{\mu}_{N\times S}\,\mathbf{t}_{S\times P}|$ as a grey-scale image) shows that both the extrema and midpoint mixtures are not reconstructed with full accuracy using cluster analysis.

in an actual cluster analysis of data of the form of Fig. 1, as shown in Fig. 3.

A better solution to the simple example shown in Fig. 2 would locate the component vectors not in the center of clusters, but closer to extrema points so that the full range of spectroscopic variations can be represented; that is, in this example, the component vectors would point along (or close to) the respective axes rather than at the cluster centers. This is simple to arrive at for the deliberately constructed example of Fig. 2; but with nontrivial data distributions in multiple dimensions and in the presence of noise, the problem becomes much more challenging. We therefore present an alternative spectromicroscopy analysis approach in Sec. 4.

## 3 Application: biochemical analysis of spermatozoa

As an example application of x-ray spectromicroscopy analysis to a complicated, real-world specimen, we consider the case of human spermatozoa. Sperm are compact cells with tightly-packed and well-segregated materials in their head and a long flagellum that allows them to move through fluid (Fig. 4). Their density and total thickness make them difficult to study in electron microscopy without sectioning, and their small size means that the sub-50 nm spatial resolution of x-ray spectromicroscopy is helpful for resolving compositional details[24].

One in four couples experiences difficulties in conceiving[25], with a male factor contributing in more than 50% of these cases[26]. The assessment of male infertility relies mostly on conventional criteria of sperm quality, such as concentration, motility, and morphology. While threshold values of these metrics can be used to classify men as subfertile, of indeterminate fertility, or fertile, none of the measures are reliable diagnostics of infertility[27]. This fact indicates that the sperm of infertile men may have hidden abnormalities in the composition of their nuclei. DNA damage above a certain threshold appears to impair fertilization and embryo development[28,29], but little is known about the etiologies of sperm DNA damage and its full impact on human reproduction. Light microscopy does not deliver valuable information on sperm DNA or chromatin abnormalities, while

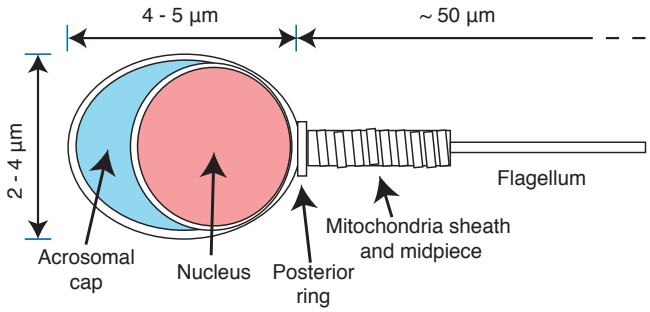



**Fig. 4** Diagram of a human spermatozoa. The enzymes involved in penetrating the egg are in the acrosomal cap, while the nucleus contains the DNA tightly packed with histone proteins. The flagellar motor is within the posterior ring.

bulk chemical measurements average over many morphologies and are not sensitive to individual spermatozoa. Flow cytometry can correlate sperm morphology with total DNA content[30], but it is still useful to visualize overall biochemical organization at higher resolution and without relying on a single biochemical marker. X-ray spectromicroscopy insights into the correlation between sperm morphology and abnormal DNA or protein distributions could lead to a better understanding of the basis for light microscopy selection of one abnormality over another for *in vitro* fertilization in cases where no sperm are present with normal morphology.

Several investigators have carried out high-resolution soft x-ray microscopy studies of sperm[31–34]. Zhang et al. have used carbon near-absorption-edge x-ray spectromicroscopy for compositional mapping of hamster, rat, and bull sperm[17]. They acquired spectra of thin film standards of proteins protamine 1 and 2 and of calf thymus DNA; a species-weighted ratio of the protamine spectra was used along with the DNA spectrum to form a two-spectrum matrix $\boldsymbol{\mu}$, which was then inverted by using the SVD in order to yield thickness maps (Eq. 2) and estimate protein-to-DNA ratios. The results suggested that protamine content is independent of protamine 2 gene expression, but they did not allow one to discover other variations in biochemical organization because the analysis assumed a composition consisting of just the three targeted biochemicals. We therefore wish to consider analysis methods that are not based on such limiting assumptions.

We have used ejaculated sperm obtained via masturbation from randomly selected unidentified donors at the Andrology Lab at Stony Brook University Hospital. Fresh ejaculate was washed in phosphate-buffered saline to dilute the optically thick semen and then was imaged wet in a special sample holder[35], air dried, or freeze dried. Wet sperm suffered some degradation during x-ray microscopy measurements, so an air-dried sperm was selected for the data shown here[24] since in images of dozens of sperm this preparation method seemed to preserve sperm morphology better than what we observed when plunge-freezing in liquid ethane followed by freeze drying (perhaps because of ice crystal formation in the dense sperm head during plunge-freezing). Images were taken with a scanning transmission x-ray microscope developed by us at Stony Brook University[36] and formerly operated on an undulator beamline X1A1 at the National

Synchrotron Light Source at Brookhaven National Laboratory (the version of microscope used in these studies was upgraded to include laser interferometer control of the scanning stage[24,37,38]). Images were acquired with 100 nm pixel size and at 133 photon energies across the carbon x-ray absorption near-edge spectroscopy region from 283.8 to 291.6 eV.

## 4 Non-negative matrix approximation methods

Cluster analysis based on data orthogonalized and reduced by using PCA is rapid and useful for analyzing complex data[13,21–23], although it can return negative values as described above. There is also a wide range of other productive approaches for spectromicroscopy analysis[4,5,14,15,39]. However, we restate our fundamental requirement, which is to find an approach that is constrained by the physics of x-ray absorption to yield only non-negative values for the matrices in the expression of Eq. 1 of $\mathbf{D}_{N\times P} = \boldsymbol{\mu}_{N\times S}\,\mathbf{t}_{S\times P}$. This is precisely the requirement satisfied by non-negative matrix factorization, an analysis approach first explored by Paatero and Tapper[40] and later implemented with considerable notice by Lee and Seung[10]. We describe our implementation of NNMA analysis for x-ray absorption spectromicroscopy, realizing that the same approaches can be used for other types of spectral analysis[41,42], spectrum imaging[43], and hyperspectral imaging methods[44,45].

The approach of Lee and Seung[10] for face recognition was to use a multiplicative update algorithm for non-negative matrix factorization of data in the form of Eq. 1, with minimization of the basic data-matching cost function $F_0(\boldsymbol{\mu}, \mathbf{t})$ of (dropping matrix subscripts for simplicity):

$$F_0(\boldsymbol{\mu}, \mathbf{t}) = ||\mathbf{D} - \boldsymbol{\mu}\mathbf{t}||_2^2. \tag{3}$$

The Lee and Seung algorithm in our notation initializes with random non-negative values for the matrices $\boldsymbol{\mu}$ and $\mathbf{t}$ and then applies iterative updates[46] using multiplicative rules of the form

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}\,\frac{(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D})}{(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\mathbf{t})} \tag{4}$$

$$\mathbf{t} \leftarrow \mathbf{t}\,\frac{(\mathbf{D}\mathbf{t}^{\mathrm{T}})}{(\boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}})}, \tag{5}$$

where the multiplications and divisions not in parentheses are taken componentwise, until a minimum of the data-matching cost function $F_0$ of Eq. 3 is reached (or, in practice, until $F_0$ falls below some predetermined threshold).

Minimizing the cost function to make the NNMA factorization of $\boldsymbol{\mu}\mathbf{t}$ as close as possible to the optical density data $\mathbf{D}$ is necessary but not sufficient for achieving a clear, easy-to-interpret analysis of x-ray spectromicroscopy data. With this basic cost function as the only consideration, one can miss several desired features of a useful solution; furthermore, the minimizer will not be unique (since any positive scaling between $\boldsymbol{\mu}$ and $\mathbf{t}$ would achieve the same cost function value). One approach is to introduce other considerations such as spectral smoothness as constraints, but this mixed strategy of optimization for some criteria, and constraints for others, can lead to very slow convergence[24].

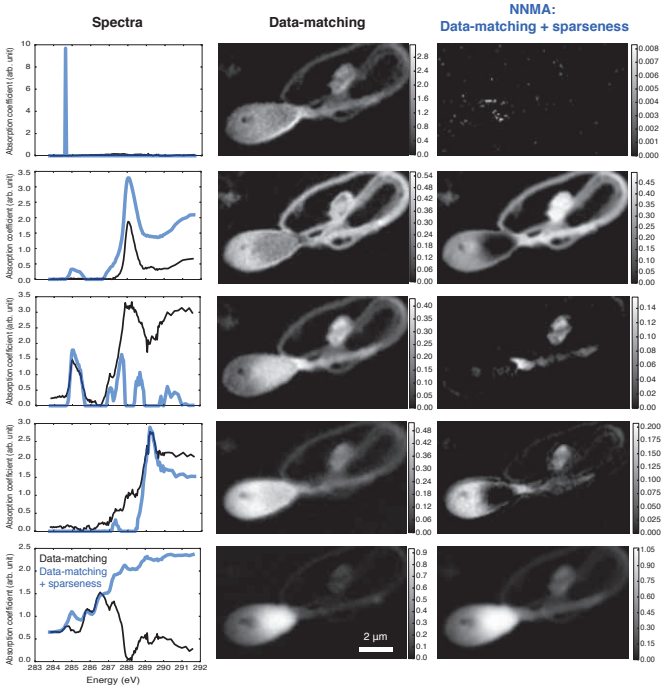



**Fig. 5** Comparison between NNMA without and with sparseness regularization. The first column compares the reconstructed spectra $\boldsymbol{\mu}$. The second column shows the NNMA reconstructed thickness maps $\mathbf{t}$ of the sperm dataset without any sparseness regularization, while the third column shows NNMA reconstructed maps with sparseness regularization ($\lambda_{\mathbf{t}} = 0.7$). In the last column, the components appear better separated than without sparseness regularization. However, some of the reconstructed spectra do not resemble observed x-ray absorption spectra; approaches to address this are illustrated in Figs. 6 and 7.

### 4.1 NNMA regularization

One method to narrow and refine the search space for NNMA is to introduce regularizations in addition to the basic cost function minimization. Regularization is one way to incorporate additional information we might have about the data into the overall cost function to be minimized. In this way, we find a balance between the error minimization from data-matching, and a good fit to data-modeling. Each regularization is controlled by a continuous regularization parameter $\lambda$ in the cost function

$$F(\boldsymbol{\mu}, \mathbf{t}) = F_0(\boldsymbol{\mu}, \mathbf{t}) + \lambda_{\boldsymbol{\mu}}\, J_{\boldsymbol{\mu}}(\boldsymbol{\mu}) + \lambda_{\mathbf{t}}\, J_{\mathbf{t}}(\mathbf{t}), \tag{6}$$

where $J_{\boldsymbol{\mu}}$, $J_{\mathbf{t}}$ are the regularizers and $\lambda_{\boldsymbol{\mu}}$, $\lambda_{\mathbf{t}}$ are the regularization parameters applied to the spectral $\boldsymbol{\mu}$ and thickness or weighting $\mathbf{t}$ matrices, respectively.

From a machine-learning perspective, $\lambda$ represents the trade-off between errors in data-matching and the complexity of the model[47]. For small $\lambda$, the errors become smaller at the cost of not accurately modeling the data. For larger $\lambda$, the data-matching errors become reduced in importance relative to other desired

characteristics of the data model. We consider here two regularization schemes for desired characteristics of our solution (our model): sparseness and similarity to cluster spectra.

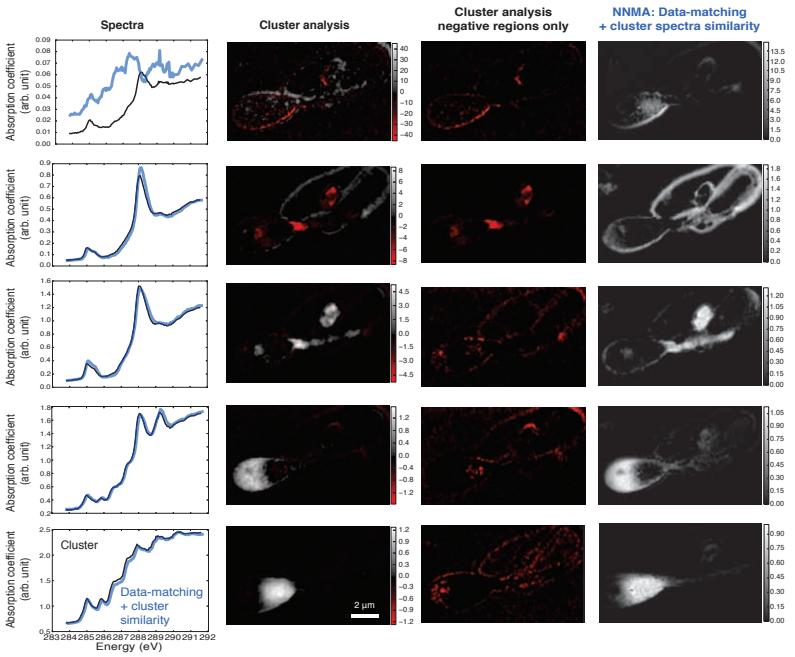



**Fig. 6** Comparison between cluster analysis and NNMA with a cluster spectra similarity regularization scheme as described in Section 4.1. By tuning the regularization parameter $\lambda_{\boldsymbol{\mu}_{\mathrm{sim}}}$ to be high (in this case, $\lambda_{\boldsymbol{\mu}_{\mathrm{sim}}} = 100$), we can obtain NNMA reconstructed spectra that are similar to those from cluster analysis (first column). At the same time, the negative regions from the cluster thickness maps (second column) are eliminated, as seen in the NNMA reconstructed maps (fourth column). To highlight the negative regions in the cluster analysis thickness maps, only these regions are shown in the third column.

**Sparseness:** In many x-ray spectromicroscopic datasets, the $\mathbf{t}$ matrix is expected to be sparse—each pixel would contain at most a few components (column sparseness), and each component would be favored to show up in only a small subset of pixels (row sparseness)—so that many entries in $\mathbf{t}$ would be zero or close to zero. The typical regularizer to model the sparseness of $\mathbf{t}$ is the one-norm[48]: $||\mathbf{t}||_1 = \sum_{k,p} \mathbf{t}_{k,p}$ . The cost function to be minimized becomes

$$F(\boldsymbol{\mu}, \mathbf{t}) = ||\mathbf{D} - \boldsymbol{\mu}\,\mathbf{t}||_2^2 + \lambda_{\mathbf{t}}\,||\mathbf{t}||_1, \tag{7}$$

and the addition to the update rule for $\mathbf{t}$ is

$$\lambda_{\mathbf{t}} \frac{\partial J_{\mathbf{t}}}{\partial \mathbf{t}} = \lambda_{\mathbf{t}}\,\mathbf{ones}_{S\times P}, \tag{8}$$

so that Eq. 5 becomes

$$\mathbf{t} \leftarrow \mathbf{t}\,\frac{\left(\mathbf{D}\mathbf{t}^{\mathrm{T}}\right)}{\left(\boldsymbol{\mu}\mathbf{t}\mathbf{t}^{\mathrm{T}} + \lambda_{\mathbf{t}}\right)}. \tag{9}$$

With spectral imaging, one of the consequences of incorporating sparseness into the model would be to create more clearly separable components, as seen in the reconstructed thickness maps $\mathbf{t}$. Figure 5 compares the results of applying NNMA without any regularization and one with sparseness regularization ($\lambda_{\mathbf{t}} = 0.7$). Although we now have more distinguishable thickness maps, the reconstructed spectra $\boldsymbol{\mu}$ do not resemble observed x-ray absorption spectra. This result is not surprising given that NNMA has no expectation of what the reconstructions should look like as long as the cost function is minimized.

**Cluster similarity:** Based on the above, we wish to also include a regularization to increase the similarity of solutions to observed x-ray absorption spectra. Since cluster analysis yields a set of spectra $\boldsymbol{\mu}_{\text{cluster}}$ averaged from spectroscopically similar pixels, its spectra provide a good basis both for the starting solution of an optimization procedure and for a "similarity regularizer" $J_{\boldsymbol{\mu}_{\text{sim}}}$ to penalize reconstructions that deviate far from the input cluster spectra:

$$\lambda_{\boldsymbol{\mu}_{\text{sim}}} J_{\boldsymbol{\mu}_{\text{sim}}} = \lambda_{\boldsymbol{\mu}_{\text{sim}}} ||\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}}||_2^2 \tag{10}$$

$$= \lambda_{\boldsymbol{\mu}_{\text{sim}}} \sum_{k=1}^{S} \sum_{n=1}^{N} (\boldsymbol{\mu}_{n,k} - \boldsymbol{\mu}_{\text{cluster}\,n,k})^2. \tag{11}$$

The addition to the update rule for $\boldsymbol{\mu}$ is

$$\lambda_{\boldsymbol{\mu}_{\text{sim}}} \frac{\partial J_{\boldsymbol{\mu}_{\text{sim}}}}{\partial \boldsymbol{\mu}} = \lambda_{\boldsymbol{\mu}_{\text{sim}}} 2(\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}}), \tag{12}$$

so that Eq. 4 becomes

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} \frac{(\boldsymbol{\mu}^{\mathrm{T}} \mathbf{D})}{\left(\boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\mu} \mathbf{t} + 2\lambda_{\boldsymbol{\mu}_{\text{sim}}} (\boldsymbol{\mu} - \boldsymbol{\mu}_{\text{cluster}})\right)}. \tag{13}$$

Figure 6 compares reconstructions from cluster analysis against NNMA with cluster spectra similarity regularization. By tuning the regularization parameter $\lambda_{\boldsymbol{\mu}_{\text{sim}}}$ to be high, we can obtain reconstructed spectra that match closely with those from cluster analysis, while eliminating the negative regions in the thickness maps.

While cluster similarity adds an important bias toward the properties of actual observed x-ray absorption spectra, sparseness is still a desirable property for our solutions since it maximizes chemical separability in the weighting or thickness maps. Fortunately, one can seek a simultaneous minimum of the three cost components together: the data-matching cost of Eq. 3, the sparseness regularizer of Eq. 7, and the cluster similarity regularizer of Eq. 10 by using the combined updates of Eqs. 9 and 13. As shown in Fig. 7, this at last gives solutions that satisfy our desired properties simultaneously, both for the sperm spectromicroscopy data shown here and in other studies to be described elsewhere.

### 4.2 Selection of the number of spectroscopically distinguishable components $S$

An important parameter in NNMA analysis (as well as in cluster analysis) is the selection of the number of spectroscopically distinguishable components $S$ to

seek. If $S$ is too small, we will arrive at a basis set that is not able to reproduce all the important variations in the data; if $S$ is too large, we may have simply analyzed variations due to noise from photon statistics or other sources. In PCA, the eigenvalues of the covariance matrix $Z = DD^T$ can provide a good estimate, since they often drop sharply from the first eigenvalue down to a point where subsequent eigenvalues decrease only slightly, indicating a transition from variations in significant signals to small variations due to different measures of the same noise factors[49]. In cluster analysis, we have used this "knee" in the eigenvalue curve to estimate the number of significant components $\bar{S}$ in the sample[21], although in practice it has been found to be important to manually examine the analysis result. The number of clusters to seek would then be $\bar{S}$.

Since NNMA analysis involves the cost function $F_0(\boldsymbol{\mu}, \mathbf{t})$ of Eq. 3 that measures how well the solution $\boldsymbol{\mu}\mathbf{t}$ matches the data $\mathbf{D}$, we have a good basis for evaluating the effect of decreasing or increasing the number $S$ of spectroscopically distinguishable components. By carrying out NNMA analysis with a range of values for $S$, we can see when the error $F_0(\boldsymbol{\mu}, \mathbf{t})$ no longer decreases as a function of $S$; we can similarly examine when decreases to $S$ are insufficient to capture all the important spectroscopic variations in the sample. This topic will be explored further in future work.

## 5 Discussion

Our goal in x-ray spectromicroscopy analysis is to find a "basis set" of spectra that allow us to describe the intrinsic distribution (thickness or weighting maps) of spectroscopically resolvable components in the specimen. While cluster analysis does this rapidly and delivers spectra that closely resemble those observed from individual pixels in the spectral image set (because they simply average a subset of observed spectra together), we have shown that the resulting basis set can lead to negative values in the thickness maps, which are unphysical. Non-negative matrix approximation analysis techniques provide a path out of this dilemma and also allow us to incorporate other characteristics desired from the data in a combined cost function approach for optimization. For our example x-ray spectromicroscopy data, NNMA delivers results consistent with the known large-scale biochemical organization of human spermatozoa when we simultaneously add two regularizers to the basic data-matching condition of Eq. 3: sparseness (Eq. 7), and cluster similarity (Eq. 10). With this combined regularizer approach, we are able to obtain thickness maps and spectra (Fig. 7) that highlight the expected large-scale biochemical organization of spermatozoa as shown schematically in Fig. 4: image $t_3$ highlights the acrosomal cap, flagellar motor, and mitochondrion; image $t_4$ highlights the nucleus where histones are bound to DNA; image $t_2$ highlights the lipid membrane and flagellum; and image $t_1$ isolates a high-density area in the flagellar motor with some combination of chemical sensitivity and experimental absorption saturation limits. The spectroscopic peaks observed in Fig. 7 can be interpreted by careful comparison between theoretical calculations and experimental measurements of organic molecular assemblies such as amino acids[50] and manufactured polymers[51], although detailed discussion is beyond the scope of the present work.

An exploratory version of the NNMA analysis approach described here is

implemented in a Python open source code* called MANTiS[52] developed by 2nd Look Consulting; a more refined interface to NNMA analysis is planned for an upcoming release of MANTiS. For the data shown here, the combined cost function converged to a minimum over about 10,000 iterations taking about 10 minutes on a laptop computer. These results show the potential of NNMA analysis on complicated data.

## Acknowledgments

We thank the U.S. Department of Energy (DOE) Office of Science for support of this work under contract DE-AC02-06CH11357 to Argonne National Laboratory. Earlier exploration of NNMA as applied to sperm spectromicroscopy analysis was supported by the National Institutes of Health under grant R01 EB-000479; RM was supported in part by grant R01 GM-104530.

## References

1 C. Jeanguillaume and C. Colliex, *Ultramicroscopy*, 1989, **28**, 252–257.
2 J. A. Hunt and D. B. Williams, *Ultramicroscopy*, 1991, **38**, 47–73.
3 C. J. Sparks, Jr, in *Synchrotron Radiation Research*, ed. H. Winick and S. Doniach, Plenum Press, New York, 1980, pp. 459–512.
4 P. G. Kotula, M. Keenan and J. R. Michael, *Microscopy and Microanalysis*, 2003, **9**, 1–17.
5 P. Lasch, W. Wäsche, W. J. McCarthy, G. J. Müller and D. Naumann, *Proceedings SPIE*, 1998, **3257**, 187–198.
6 M. Diem, S. Boydston-White and L. Chiriboga, *Applied Spectroscopy*, 1999, **53**, 148A–161A.
7 H. Ade, X. Zhang, S. Cameron, C. Costello, J. Kirz and S. Williams, *Science*, 1992, **258**, 972–975.
8 C. Jacobsen, G. Flynn, S. Wirick and C. Zimba, *Journal of Microscopy*, 2000, **197**, 173–184.
9 J. Stöhr, *NEXAFS Spectroscopy*, Springer-Verlag, 1992, vol. 25.
10 D. D. Lee and H. S. Seung, *Nature*, 1999, **401**, 788–791.
11 B. L. Henke, E. Gullikson and J. C. Davis, *Atomic Data and Nuclear Data Tables*, 1993, **54**, 181–342.
12 G. A. Johansson, T. Tyliszczak, G. E. Mitchell, M. H. Keefe and A. P. Hitchcock, *Journal of Synchrotron Radiation*, 2007, **14**, 395–402.
13 M. Lerotic, C. Jacobsen, J. Gillow, A. Francis, S. Wirick, S. Vogt and J. Maser, *Journal of Electron Spectroscopy and Related Phenomena*, 2005, **144–147**, 1137–1143.
14 N. Bonnet, *Journal of Microscopy*, 1998, **190**, 2–18.
15 N. Bonnet, N. Brun and C. Colliex, *Ultramicroscopy*, 1999, **77**, 97–112.
16 P. Lasch, M. Boese, A. Pacifico and M. Diem, *Vibrational Spectroscopy*, 2002, **28**, 147–157.
17 X. Zhang, R. Balhorn, J. Mazrimas and J. Kirz, *Journal of Structural Biology*, 1996, **116**, 335–344.
18 I. N. Koprinarov, A. P. Hitchcock, C. T. McCrory and R. F. Childs, *Journal of Physical Chemistry B*, 2002, **106**, 5358–5364.
19 P. L. King, R. Browning, P. Pianetta, I. Lindau, M. Keenlyside and G. Knapp, *Journal of Vacuum Science and Technology A*, 1989, **7**, 3301–3304.
20 A. Osanna and C. Jacobsen, X-ray Microscopy: Proceedings of the Sixth International Conference (American Institute of Physics Conference Proceedings), 2000, pp. 350–357.
21 M. Lerotic, C. Jacobsen, T. Schäfer and S. Vogt, *Ultramicroscopy*, 2004, **100**, 35–57.
22 G. Mitrea, J. Thieme, P. Guttmann, S. Heim and S. Gleber, *Journal of Synchrotron Radiation*, 2007, **15**, 26–35.
23 J. Lehmann, D. Solomon, J. Kinyangi, L. Dathe, S. Wirick and C. Jacobsen, *Nature Geoscience*, 2008, **1**, 238–242.

*http://code.google.com/p/spectromicroscopy

24 H. Fleckenstein, *Ph.D. thesis*, Stony Brook University, 2008.
25 L. Schmidt, K. Munster and P. Helm, *British Journal of Obstetric Gynecology*, 1995, **102**, 978–984.
26 D. S. Irvine, *Human Reproduction*, 1996, **13**, 33–44.
27 D. S. Guzick, J. W. Overstreet, P. Factor-Litvak *et al.*, *New England Journal of Medicine*, 2001, **345**, 1388–1393.
28 A. Ahmadi and S. C. Ng, *Journal Of Experimental Zoology*, 1999, **284**, 696–704.
29 C. Cho, H. Jung-Ha, W. D. Willis, E. H. Goulding, P. Stein, Z. Xu, R. M. Schultz, N. B. Hecht and E. M. Eddy, *Biology of Reproduction*, 2003, **69**, 211–217.
30 U. B. Hacker-Klom, W. Göhde, E. Nieschlag and H. M. Behre, *Human Reproduction*, 1999, **14**, 2506–2512.
31 B. W. Loo, Jr., S. Williams, S. Meizel and S. S. Rothman, *Journal of Microscopy*, 1992, **166**, RP5–RP6.
32 R. Balhorn, R. E. Braun, B. Breed, J. T. Brown, D. Evenson, J. M. Heck, J. Kirz, I. McNulty, W. Meyer-Ilse and X. Zhang, X-ray Microscopy and Spectromicroscopy, Berlin, 1998, pp. II–29–46.
33 T. Vorup-Jensen, T. Hjort, J. V. Abraham-Peskir, P. Guttmann, J. C. Jensenius, E. Uggerhøj and R. Medenwaldt, *Human Reproduction*, 1998, **14**, 880–884.
34 J. Abraham-Peskir, E. Chantler, C. McCann, R. Medenwaldt and E. Ernst, *Medical Science Research*, 1998, **26**, 663–667.
35 U. Neuhäusler, C. Jacobsen, D. Schulze, D. Stott and S. Abend, *Journal of Synchrotron Radiation*, 2000, **7**, 110–112.
36 M. Feser, T. Beetz, C. Jacobsen, J. Kirz, S. Wirick, A. Stein and T. Schäfer, Soft X-Ray and EUV Imaging Systems II, Bellingham, Washington, 2001, pp. 146–153.
37 M. Lerotić, *Ph.D. thesis*, Department of Physics and Astronomy, Stony Brook University, 2005.
38 B. Hornberger, *Ph.D. thesis*, Department of Physics and Astronomy, Stony Brook University, 2007.
39 A. de Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault and M. Maeder, *Trends in Analytical Chemistry*, 2004, **23**, 70–79.
40 P. Paatero and U. Tapper, *Environmetrics*, 1994, **5**, 111–126.
41 V. P. Pauca, J. Piper and R. J. Plemmons, *Linear Algebra and its Applications*, 2006, **416**, 29–47.
42 G. Buchsbaum and O. Bloch, *Vision Research*, 2002, **42**, 559–563.
43 J. S. Lee, D. D. Lee, S. Choi and D. S. Lee, In 3rd International Conference on Independent Component Analysis and Blind Signal Separation, 2001, pp. 629–632.
44 S. A. Robila and L. G. Maciak, Proceedings SPIE, 2006, p. 63840F.
45 C.-Y. Liou and K.-D. Ou Yang, International Conference on Neural Information Processing, 2005, pp. 280–285.
46 D. D. Lee and H. S. Seung, Advances in Neural Information Processing Systems, 2001, pp. 556–562.
47 E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2nd edn., 2010.
48 R. Tibshirani, *Journal of the Royal Statistical Society B*, 1996, **58**, 267–288.
49 E. R. Malinowski, *Factor Analysis in Chemistry*, John H. Wiley & Sons, New York, 2nd edn., 1991.
50 K. Kaznacheyev, A. Osanna, C. Jacobsen, O. Plashkevych, O. Vahtras, H. Ågren, V. Carravetta and A. Hitchcock, *Journal of Physical Chemistry A*, 2002, **106**, 3153–3168.
51 O. Dhez, H. Ade and S. Urquhart, *Journal of Electron Spectroscopy and Related Phenomena*, 2003, **128**, 85–96.
52 M. Lerotic, R. Mak, S. Wirick, F. Meirer and C. Jacobsen, *Journal of Synchrotron Radiation*, 2014, **(submitted)**, 0.

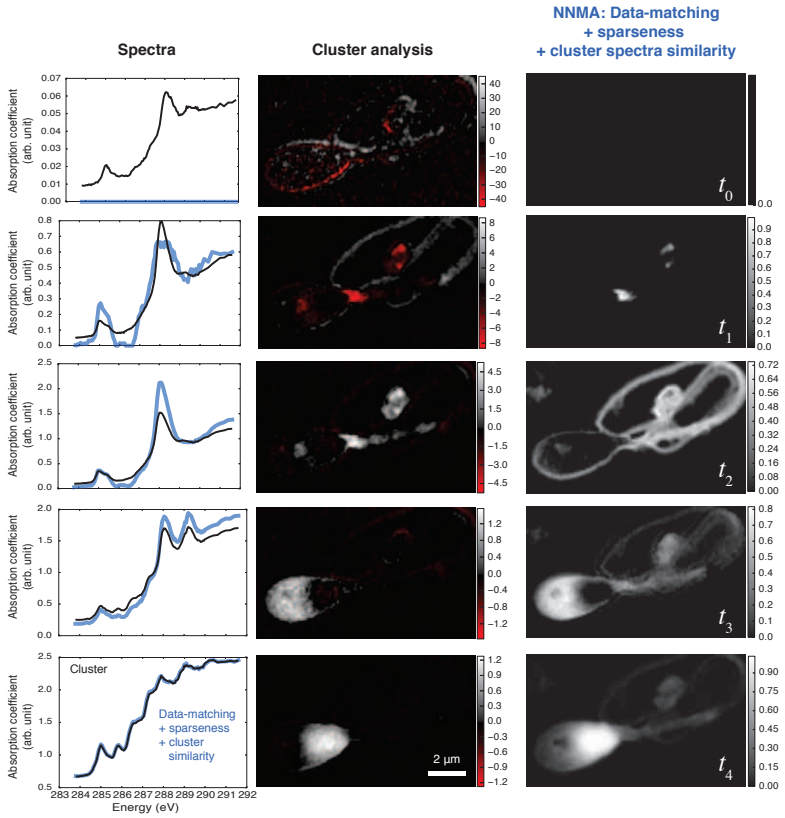



**Fig. 7** X-ray spectromicroscopy analysis results obtained by the cluster analysis approach[21] and NNMA with data-matching (Eq. 3), sparseness (Eq. 7), and cluster similarity (Eq. 10) regularizers used in combination. As suggested in Fig. 2 and shown for this data set in Fig. 6, cluster analysis produces maps with negative weightings for some regions, which are not allowed by the physics of x-ray absorption, whereas the NNMA solution using data-matching, sparseness, and cluster similarity delivers an analysis result with recognizable x-ray absorption spectra and positive weightings or thickness maps, which nicely illustrate the large-scale biochemical organization of sperm. Image $t_3$ highlights the acrosomal cap, flagellar motor, and mitochondrion; image $t_4$ highlights the nucleus where histones are bound to DNA; image $t_2$ highlights the lipid membrane and flagellum; and image $t_1$ isolates a high-density area in the flagellar motor with some combination of chemical sensitivity and experimental absorption saturation limits. The regularization weightings used for the NNMA analysis were $\lambda_{\boldsymbol{\mu}_{\mathrm{sim}}} = 10$ and $\lambda_{\mathbf{t}} = 0.5$.