Structural Dynamics

Dynamics retrieval from stochastically weighted incomplete data by low-pass spectral analysis

Cite as: Struct. Dyn. **9**, 044101 (2022); https://doi.org/10.1063/4.0000156
Submitted: 09 April 2022 • Accepted: 12 July 2022 • Published Online: 16 August 2022

Cecilia M. Casadei, Ahmad Hosseinizadeh, Gebhard F. X. Schertler, et al.

COLLECTIONS

F This paper was selected as Featured

View Online Export Citation CrossMark

Dynamics retrieval from stochastically weighted incomplete data by low-pass spectral analysis

Cite as: Struct. Dyn. **9**, 044101 (2022); doi: 10.1063/4.0000156
Submitted: 9 April 2022 · Accepted: 12 July 2022 ·
Published Online: 16 August 2022

Cecilia M. Casadei,[1,2,a)] Ahmad Hosseinizadeh,[3] Gebhard F. X. Schertler,[1,2] Abbas Ourmazd,[3] and Robin Santra[4,5,a)]

AFFILIATIONS

[1]Institute of Molecular Biology and Biophysics, Department of Biology, ETH Zürich, Zürich, Switzerland
[2]Laboratory of Biomolecular Research, Biology and Chemistry Division, Paul Scherrer Institute, Villigen PSI, Switzerland
[3]University of Wisconsin Milwaukee, Milwaukee, Wisconsin 53201, USA
[4]Center for Free-Electron Laser Science CFEL, Deutsches Elektronen-Synchrotron DESY, 22607 Hamburg, Germany
[5]Department of Physics, Universität Hamburg, 22607 Hamburg, Germany

[a)]Authors to whom correspondence should be addressed: cecilia.casadei@psi.ch and robin.santra@cfel.de

ABSTRACT

Time-resolved serial femtosecond crystallography (TR-SFX) provides access to protein dynamics on sub-picosecond timescales, and with atomic resolution. Due to the nature of the experiment, these datasets are often highly incomplete and the measured diffracted intensities are affected by partiality. To tackle these issues, one established procedure is that of splitting the data into time bins, and averaging the multiple measurements of equivalent reflections within each bin. This binning and averaging often involve a loss of information. Here, we propose an alternative approach, which we call low-pass spectral analysis (LPSA). In this method, the data are projected onto the subspace defined by a set of trigonometric functions, with frequencies up to a certain cutoff. This approach attenuates undesirable high-frequency features and facilitates retrieving the underlying dynamics. A time-lagged embedding step can be included prior to subspace projection to improve the stability of the results with respect to the parameters involved. Subsequent modal decomposition allows to produce a low-rank description of the system's evolution. Using a synthetic time-evolving model with incomplete and partial observations, we analyze the LPSA results in terms of quality of the retrieved signal, as a function of the parameters involved. We compare the performance of LPSA to that of a range of other sophisticated data analysis techniques. We show that LPSA allows to achieve excellent dynamics reconstruction at modest computational cost. Finally, we demonstrate the superiority of dynamics retrieval by LPSA compared to time binning and merging, which is, to date, the most commonly used method to extract dynamical information from TR-SFX data.

I. INTRODUCTION

Time-resolved serial femtosecond crystallography (TR-SFX) has emerged as a prominent technique for investigating the dynamics of light-sensitive macromolecules with atomic spatial resolution at ultrafast timescales.[1–4] In a typical experiment, a laser pulse pumps the molecules into an excited state. An x-ray pulse from an X-ray Free-Electron Laser (X-FEL) probes the system a certain time delay—typically in the femtosecond to picosecond range—after photo-excitation. Microcrystals of the protein of interest, embedded in a viscous medium, are delivered into the interaction region through a continuous flow. The experiment is carried out in a serial fashion. Since the interaction with the X-FEL beam is destructive, each microcrystal gives rise to up to one diffraction pattern, from a specific, random orientation (Fig. 1).

Under typical experimental conditions, only a small fraction of the reciprocal lattice points within the accessible resolution range fulfill the elastic scattering condition for a specific crystal's orientation (Fig. 1). The diffraction signal recorded in one frame is therefore highly incomplete [Fig. 2(a)]. Because crystals have a finite size and some extent of lattice disorder, and the spectral distribution of the X-FEL beam is limited, the recorded intensities are in general, smaller—by factors that are orientation dependent—than the corresponding diffraction intensities from an infinite and perfectly ordered crystal [Fig. 2(b)]. This effect is commonly referred to as partiality.[5] Variations in crystal size and beam fluence distribution can be accounted for by estimating and applying frame-related scale factors. Such estimates can be computed using standard SFX software.[6]

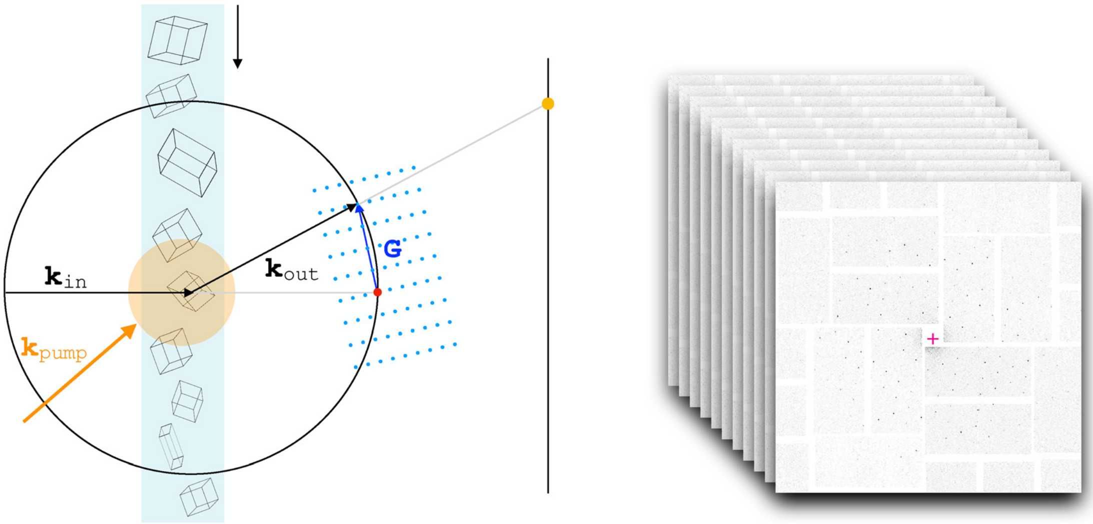

FIG. 1. Schematic representation of the time-resolved serial crystallography experiment. The microcrystals are brought into the interaction region by a continuous flow of a viscous medium. Individual crystals are probed by an X-FEL pulse at a certain time delay after optical pumping. The diffraction condition $\boldsymbol{k}_{\text{out}} - \boldsymbol{k}_{\text{in}} = \boldsymbol{G}$, with $\boldsymbol{G}$ a reciprocal lattice vector, is represented graphically by the diffraction sphere construction.[10] A dataset is composed of an ensemble of frames, each recording a diffraction pattern from an individual crystal in a random orientation.

The uncertainty in timing between pump and probe pulses (timing jitter)[7,8] and photon counting errors[9] are other factors that affect time-resolved serial crystallography data.

To retrieve a complete set of reciprocal-space intensities and mitigate the effects of partiality, a binning-and-merging procedure is routinely adopted (see, for example, Refs. 12–18). This approach involves dividing the pump–probe delay window into time bins, and merging the measurements by averaging the equivalent reflections within each bin. The number of frames required in the time-binning approach is commonly of the order of the tens of thousands for each time point, often leading to broad bins and a consequent deterioration of the timing information.

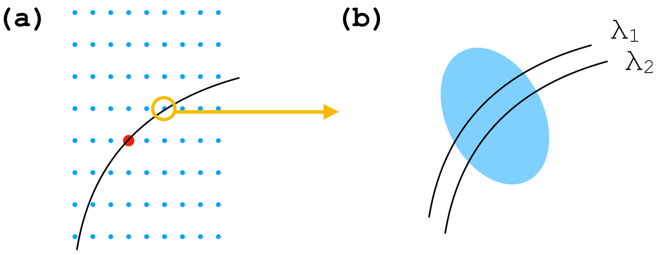

FIG. 2. Schematic representation of data incompleteness and partiality. (a) The diffraction condition is satisfied at the intersection between the reciprocal lattice, whose orientation is determined by the crystal's, and the diffraction sphere, with radius $|\boldsymbol{k}_{\text{in}}| = |\boldsymbol{k}_{\text{out}}|$. Most reciprocal lattice points are unmeasured in an individual frame. This determines the incompleteness of the data. (b) Due to the finite size of the crystal and lattice disorder, the reciprocal lattice regions that can give rise to constructive interference are not vanishingly small, but can rather be modeled as three-dimensional ellipsoids. The spectral distribution of the incident X-FEL beam can be accounted for by attributing a finite thickness to the scattering sphere (scattering shell). In this model, diffraction arises from the volumes at the intersection between reciprocal space ellipsoids and the scattering shell,[11] so that the resulting partial intensities are, in general, smaller and not representative of the full intensities that would arise from an infinite and perfect crystal.

Here, we analyze alternative strategies to the binning-and-merging approach, with the purpose of improving the quality of the reconstructed dynamics and limiting information losses. We specifically tackle the issues of data incompleteness and partiality. To demonstrate our findings while isolating these effects, we employ a synthetic dataset. We present a new method, which we call low-pass spectral analysis (LPSA), to extract accurate dynamics from extremely incomplete and partial data. We compare the performance of LPSA, in terms of reconstruction quality and computational effort, to that of time binning as well as a range of other dynamics-retrieval techniques.

II. OUTLINE OF THE PROBLEM

Consider m reciprocal lattice points in the resolution range of interest. Let $x_i \geq 0$ be the diffraction intensity related to the lattice point i. The m-tuple $\boldsymbol{x} = (x_1, \ldots, x_m)^T$ for a given set of non-negative numbers $x_1, \ldots, x_m$ may be viewed as a possible diffraction pattern. If the physical system giving rise to the diffraction pattern undergoes dynamical evolution, then, in the absence of stochasticity, the dynamics are governed by a differential equation with respect to time. This implies that the associated diffraction pattern as a function of time $\boldsymbol{x}(t)$ must be at least singly differentiable with respect to t. Hence, in the immediate vicinity of any time point t_0, we can approximate $\boldsymbol{x}(t)$ as follows:

$$\boldsymbol{x}(t) \sim \boldsymbol{x}(t_0) + \dot{\boldsymbol{x}}(t_0)(t - t_0). \tag{1}$$

This means that locally, the $\boldsymbol{x}(t)$ points lie on a straight line. Local linearity renders $\boldsymbol{x}(t)$ a one-dimensional manifold.[19]

For a given orthonormal basis of $\mathbb{R}^m$, $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$, the system's trajectory in reciprocal space can be expanded as follows:

$$\boldsymbol{x}(t) = \sum_{j=1}^{m} \alpha_j(t)\boldsymbol{u}_j, \tag{2}$$

with the expansion coefficients,

$$\alpha_j(t) = \boldsymbol{u}_j^T \boldsymbol{x}(t). \tag{3}$$

The basis vectors associated with nonnegligible expansion coefficients span the linear subspace of $\mathbb{R}^m$ explored by the trajectory $\boldsymbol{x}(t)$. As a consequence of experimental reality, the one-dimensional manifold that underlies the system's dynamical evolution may be completely unrecognizable. In practice, the stochastic effects of incompleteness and partiality alter the trajectory of the system in data space and artificially increase the apparent dimension of the subspace explored by the dynamics.

The task at hand is then twofold. First, we have to retrieve the one-dimensional manifold described by the dynamical system, that is, mitigate the stochastic effects introduced by data incompleteness and partiality. Second, we need to identify the linear space of minimal dimension in which the recovered manifold can be embedded. There are various strategies to accomplish the first of our tasks, which will be detailed in Secs. III and V. Subsequent singular value decomposition (SVD)[20,21] allows to identify the linear subspace explored by the system.

Since the recovered dynamics are expected to describe a locally linear manifold, to help the analysis of our results, we introduce a measure of the deviation from local linearity. We denote with $\{\boldsymbol{x}_j\}$ the sequence of time-ordered data vectors related to the time points $\{t_j\}$, with $\boldsymbol{x}_j = \boldsymbol{x}(t_j)$. From any pair of temporally neighboring points, $\boldsymbol{x}_{j-1}$ and $\boldsymbol{x}_j$, we can construct a local linear approximation to $\boldsymbol{x}(t)$, which we call $\boldsymbol{x}^{(j)}(t)$,

$$\boldsymbol{x}^{(j)}(t) = \boldsymbol{x}_{j-1} + \frac{t - t_{j-1}}{t_j - t_{j-1}}(\boldsymbol{x}_j - \boldsymbol{x}_{j-1}). \tag{4}$$

Local linearity implies that the two immediate temporal neighbors of $\boldsymbol{x}_{j-1}$ and $\boldsymbol{x}_j$, i.e., $\boldsymbol{x}_{j-2}$ and $\boldsymbol{x}_{j+1}$, lie close to the points $\boldsymbol{x}^{(j)}(t_{j-2})$ and $\boldsymbol{x}^{(j)}(t_{j+1})$, respectively. We, therefore, define

$$L^{(j)} = \frac{1}{2}\left[|\boldsymbol{x}_{j-2} - \boldsymbol{x}^{(j)}(t_{j-2})| + |\boldsymbol{x}_{j+1} - \boldsymbol{x}^{(j)}(t_{j+1})|\right]. \tag{5}$$

The average over all $L^{(j)}$ represents our measure of deviation from local linearity L.

III. LOW-PASS SPECTRAL ANALYSIS

We map the incomplete data to a sparse-matrix representation, that is, we set any unmeasured component of $\boldsymbol{x}(t)$ equal to zero.[22] With this choice, the fundamental issue that we need to address and mitigate is the stochastic weighting (by factors comprised between zero and one) of the underlying intensities, introduced by sparsity and partiality. To alleviate the effects of randomness, we project the data onto the subspace spanned by a set of trigonometric functions. The frequencies of these functions are defined by integer multiples of the first harmonic corresponding to one period of oscillation in the time range covered by the (time-lagged embedded) data points and up to a certain cutoff. This procedure effectively removes undesired high-frequency features and allows to recover the system's underlying trajectory in data space. Time-lagged embedding of the data[23–25] can be performed prior to subspace projection, to improve the stability of the reconstructed signal, but at an increased computational cost (Secs. IV and VI). Subsequent modal decomposition allows to represent the dynamical evolution of the system in the subspace of minimal dimension. We present the details of the method hereafter.

A. Time-lagged embedding

For S sampled time points, the columns of $\boldsymbol{x} \in \mathbb{R}^{m\times S}$ give a discretized representation of the trajectory of interest $\boldsymbol{x}(t)$. Hence, for a given time point t_j, $\boldsymbol{x}_j = \boldsymbol{x}(t_j)$, where $\boldsymbol{x}_j$ is the jth column of $\boldsymbol{x}$. With the values of missing entries set equal to zero, $\boldsymbol{x}$ is typically highly sparse. The time-lagged embedding procedure, with concatenation parameter q, consists in the delayed-coordinate mapping defined by

$$\boldsymbol{x}(t_j) \mapsto \boldsymbol{X}(t_j) = \begin{pmatrix} \boldsymbol{x}_j \\ \boldsymbol{x}_{j-1} \\ \vdots \\ \boldsymbol{x}_{j-q+1} \end{pmatrix}. \tag{6}$$

B. Low-pass filtering

With the purpose of denoising the data by removal of the high-frequency components, and given the (time-lagged embedded) data $\boldsymbol{X} \in \mathbb{R}^{mq\times s}$, with $s = S - q + 1$, we define a time-domain projector $\boldsymbol{\Phi\Phi}^T \in \mathbb{R}^{s\times s}$, with $\boldsymbol{\Phi} \in \mathbb{R}^{s\times k}$. We consider the fundamental oscillation period $T = 2\pi/\omega$ corresponding to the time window spanned by the ensemble of the time-lagged embedded points. We define $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k)$, where $\boldsymbol{\psi}_j$ is the jth column of $\boldsymbol{\Psi}$. The matrix entries are obtained by sampling a series of trigonometric functions at discrete time points as follows:

$$\psi_{i,2j} = \cos(j\omega t_i), \tag{7}$$

$$\psi_{i,2j+1} = \sin(j\omega t_i), \tag{8}$$

for $i = 1, \ldots, s$; $j = 1, \ldots, j_{\max}$ and $\boldsymbol{\psi}_1$ is a constant vector. The columns of $\boldsymbol{\Phi}$, $(\phi_1, \ldots, \phi_k)$, are obtained by orthonormalization of the vectors $(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k)$ to fulfill the condition,

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \mathbb{I}_{k\times k}, \tag{9}$$

with $k = 2j_{\max} + 1$. Only if k equals s,

$$\boldsymbol{\Phi\Phi}^T = \mathbb{I}_{s\times s}, \tag{10}$$

holds. Typical choices of k satisfy $k \ll s$ to (i) remove the undesired high-frequency features and (ii) make the subsequent calculations more affordable. The low-pass filtered data $\boldsymbol{X\Phi\Phi}^T$ retain frequency components up to the cutoff $j_{\max}\omega$.

C. Modal decomposition

The linear subspace of $\mathbb{R}^{mq}$ explored by the system's dynamics can be identified by modal decomposition of the linear mapping $\boldsymbol{A} \in \mathbb{R}^{mq\times k}$ given by the subspace projection,

$$\boldsymbol{A} = \boldsymbol{X\Phi}. \tag{11}$$

The SVD of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \tag{12}$$

with $\boldsymbol{U} \in \mathbb{R}^{mq\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, $\boldsymbol{V} \in \mathbb{R}^{k\times r}$, and $r \le k$ is the rank of $\boldsymbol{A}$. Using $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$, $\Sigma_{ij} = \sigma_i\delta_{ij}$, and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$, with $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively, the SVD gives the modal decomposition,

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T. \tag{13}$$

D. Reconstruction in time-lagged embedding space

The reconstructed time-lagged embedded data points are

$$\tilde{\boldsymbol{X}} = \boldsymbol{A}\boldsymbol{\Phi}^T \sim \sum_{i=1}^{\tilde{r}} \sigma_i \boldsymbol{u}_i (\boldsymbol{\Phi}\boldsymbol{v}_i)^T, \tag{14}$$

where only $\tilde{r} \le r$ dominant modes are retained.

E. Signal retrieval in data space

The reconstructed time-lagged embedded points $\tilde{\boldsymbol{X}}(t_i)$ have the structure described in Eq. (6). Hence, the data point $\tilde{\boldsymbol{x}}(t)$ can be retrieved by averaging the q reconstructed copies extracted from time-lagged embedded vectors in the range from $\tilde{\boldsymbol{X}}(t_i)$ to $\tilde{\boldsymbol{X}}(t_{i+q-1})$.

IV. RESULTS OF LPSA

To investigate the capabilities of LPSA and compare it to other dynamics retrieval methods, we employ the synthetic model,

$$\begin{aligned}\boldsymbol{x}(t) = {} & (1 - e^{(-t/t_c)})[\boldsymbol{A} + \boldsymbol{B}\cos(3\omega t) + \boldsymbol{C}\sin(10\omega t)] \\ & + e^{(-t/t_c)}[\boldsymbol{D} + \boldsymbol{E}\sin(7\omega t) + \boldsymbol{F}\sin(11\omega t + \pi/10)],\end{aligned} \tag{15}$$

shown in Fig. 3(a), with t_c corresponding to the middle of the time interval considered, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, $\boldsymbol{E}$ and $\boldsymbol{F}$ noncollinear vectors $\in \mathbb{R}^m$, with components,

$$\begin{aligned} A_i &= \cos[0.6\chi i], \\ B_i &= \sin[3.0\chi i + \pi/5], \\ C_i &= \sin[0.8\chi i + \pi/7], \\ D_i &= \cos[2.1\chi i], \\ E_i &= \cos[1.2\chi i + \pi/10], \\ F_i &= \sin[1.8\chi i + \pi/11], \end{aligned} \tag{16}$$

for $i = 1, \ldots, m$ and $\chi = 2\pi/m$. Fundamental molecular physics dictates that, particularly on ultrafast time scales, structural dynamics are dominated by vibrations and, occasionally, quasi-irreversible transitions between local minima of the respective potential energy surfaces. Our model [Eq. (15)] is designed to reflect both of these effects. To mimic the extent of incompleteness affecting TR-SFX datasets, we set equal to zero 98.2% of the $x_i(t)$ values, chosen at random (and matching the incompleteness of the dataset in Ref. 16). In addition, we multiply each data point $x_i(t)$ by a random number extracted from a constant distribution between zero and one [Fig. 3(b)], to model data partiality. Because the signal is generated by a linear combination of six linearly independent vectors, the dimension of the linear subspace of $\mathbb{R}^m$ explored by the underlying dynamics is six. However, the dimension of the subspace of $\mathbb{R}^{qm}$ explored by the time-lagged embedded data manifold is not constrained to six, which rather represents a lower limit.

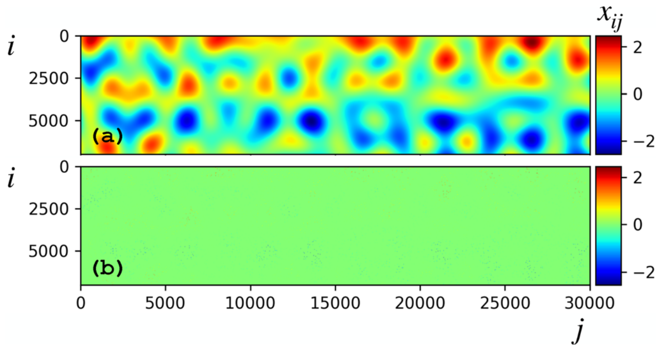

FIG. 3. (a) Underlying dynamics $\boldsymbol{x}(t)$ [Eq. (15)], with x_{ij} the ith component of data vector $\boldsymbol{x}_j = \boldsymbol{x}(t_j)$. (b) Incomplete and partial input data. Missing entries are assigned to zero values generating a sparse input data matrix.

We process the sparse and partial dataset by LPSA and measure the quality of the retrieved signal as a function of the number of modes $\tilde{r}$ employed in the reconstruction [Eq. (14)]. We examine the evolution of the results as we vary the two parameters involved: the concatenation parameter q and the cutoff frequency $j_{\max}$. The quality of the reconstructed signal is quantified by calculating the linear correlation coefficient to the underlying dynamics from Eq. (15). Since this metric is generally unavailable, we analyze the corresponding evolution of two indicators that can be used to guide the choice of the number of modes to be employed in the reconstruction. The singular value spectrum of the matrix $\boldsymbol{A}$ shows the relative weight of the terms of the modal decomposition in Eq. (13). We expect noise terms to have a relatively low weight. In addition, we compute a measure of the deviation from local linearity of the reconstructed signal, which we call L (see Sec. II). We expect the retrieved dynamics to deviate significantly from local linearity when the number of modes employed exceeds the optimal one, and noise from the input data is reintroduced in the reconstructed signal.

Figure 4 shows the evolution of the quantities mentioned above with varying $\tilde{r}$, for various values of q and fixed $j_{\max} = 100$. We observe that the best linear correlation of the reconstructed signal to the ground truth is obtained with six modes and $q = 1$, matching our expectation that the dimension of the subspace explored by the data manifold is six [Fig. 4(a)]. We also notice that for $q = 1$ the correlation coefficient does not converge to its maximal value with increasing number of modes, but rather deteriorates progressively when $\tilde{r}$ increases beyond six. The choice of the number of modes is then critical to the quality of the results. Such a choice can be guided by identifying the end of the plateau section in $L(\tilde{r})$ [Fig. 4(b)]. This shows that considerable noise is added to the reconstructed signal as $\tilde{r}$ is increased beyond six modes. A concomitant sharp decline in the singular value spectrum is observed [Fig. 4(c)]. With large values of the concatenation parameter, $q \gg 1$, a robust convergence of the correlation coefficient with increasing $\tilde{r}$ is observed, at the cost of higher computational effort. The minimal number of modes required to obtain the maximal

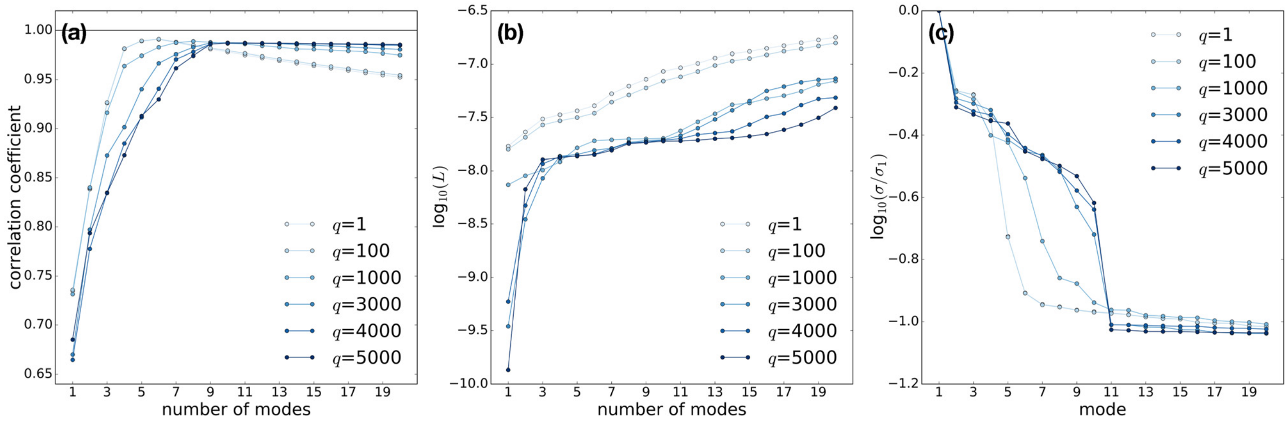

FIG. 4. LPSA of the sparse and partial dataset shown in Fig. 3(b), for various values of q, and with $j_{\max} = 100$. (a) Linear correlation coefficient between the reconstructed signal and the underlying dynamics. (b) Measure of deviation from local linearity. (c) Singular value spectrum.

correlation is ten, in accordance with our expectation that in time-lagged embedding space, the dimension of the explored subspace can be larger than six.

We now analyze the evolution of the results with varying $j_{\max}$, for $q = 1$ (Fig. 5) and $q = 4000$ (Fig. 6). The results converge toward the optimal reconstruction with increasing $j_{\max}$, with $q = 1$ and $\tilde{r} = 6$; and with $q = 4000$ and $\tilde{r} \geq 10$. We observe a degradation of the reconstruction quality as the number of modes exceeds the optimal one for $q = 1$, but a robust convergence with increasing number of modes for $q = 4000$. The measure L of deviation from local linearity [Figs. 5(b) and 6(b)] and the singular value spectrum [Figs. 5(c) and 6(c)] can guide the choice of $\tilde{r}$. Relatively high noise levels in the reconstructed signal, and relatively low singular values are observed when the optimal number of modes is exceeded. In particular, with $q = 4000$ and sufficiently high $j_{\max}$, the local linearity measure allows to detect a sharp increase in noise reconstruction beyond 10 modes, in agreement with the abrupt decline of the singular value.

V. COMPARISON WITH OTHER DYNAMICS RETRIEVAL METHODS

A. Singular spectrum analysis

Time-lagged embedding followed by modal decomposition by SVD is known as singular spectrum analysis (SSA).[26] With concatenation parameter $q = 1$, SSA is equivalent to SVD.

B. Nonlinear Laplacian spectral analysis

Nonlinear Laplacian spectral analysis (NLSA) was introduced in Ref. 27 and used in the context of time resolved experiments in Refs. 22 and 28. Similar to LPSA, the overarching framework is that of a subspace projection preceded by time-lagged embedding and followed by modal decomposition. The difference between the two methods resides in the choice of the subspace basis set. In NLSA, a data-driven basis set is employed, specifically a set of functions derived from the diffusion map algorithm.[29] In this work, we propose a modified version of the diffusion map, which allows to obtain a set of orthonormal

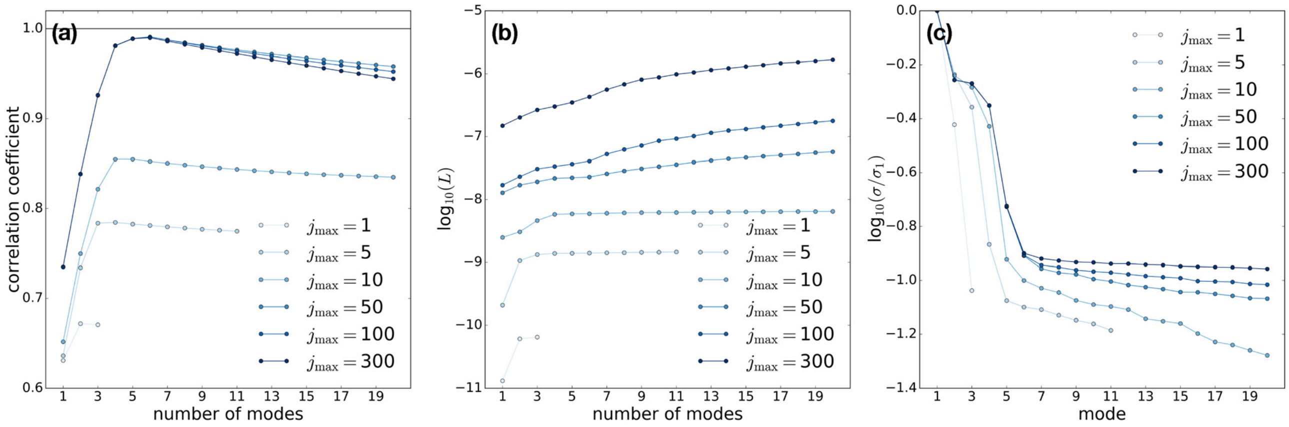

FIG. 5. LPSA of the sparse and partial dataset shown in Fig. 3(b), for various values of $j_{\max}$ and with $q = 1$. (a) Linear correlation coefficient between the reconstructed signal and the underlying dynamics. (b) Measure of deviation from local linearity. (c) Singular value spectrum.

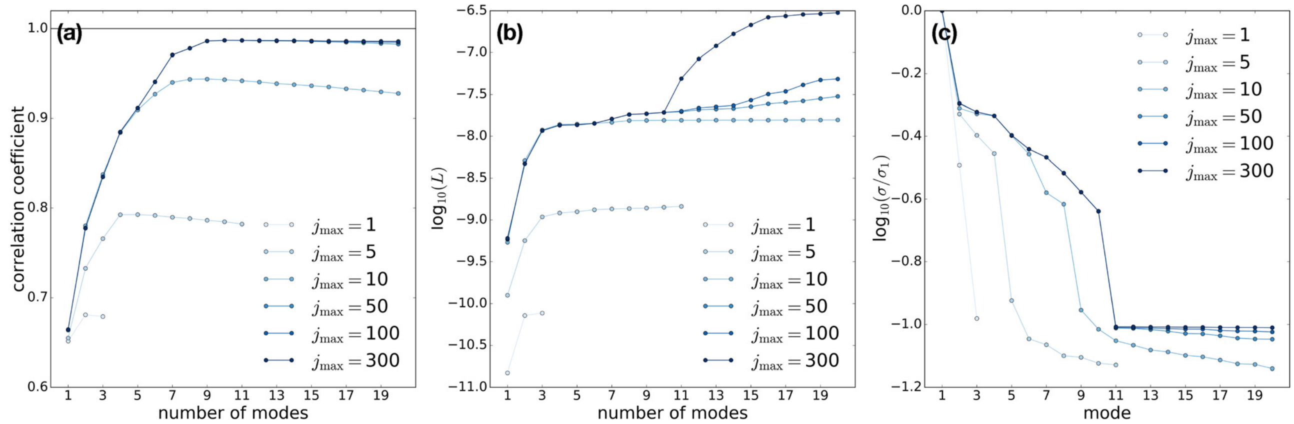

FIG. 6. LPSA of the sparse and partial dataset shown in Fig. 3(b), for various values of $j_{\max}$ and with $q = 4000$. (a) Linear correlation coefficient between the reconstructed signal and the underlying dynamics. (b) Measure of deviation from local linearity. (c) Singular value spectrum.

vectors to use directly as a subspace basis set. We also consider two different formulations of NLSA. In the standard version (E-NLSA), each data point's Euclidean nearest neighbors are considered. With the purpose of using all available information, we propose a procedure whereby time nearest neighbors are retained instead (T-NLSA). Here, timing information is used to guide the nearest-neighbor selection. As described in Sec. III, missing observations are set equal to zero.

1. Distance calculation

Euclidean distances between highly sparse time-lagged embedded vectors are calculated by retaining only common terms (i.e., the set of reflections that are present in both time-lagged embedded vectors), and are normalized by the number of retained components. This approach appears to better represent underlying distances (those pertaining to the underlying dynamics), compared to distances calculated by including all components.

2. Diffusion map

We use the diffusion map kernel,

$$K_{ij} = \exp\left[-D_{ij}^2/(2\epsilon)\right], \tag{17}$$

as a measure of similarity between time-lagged embedded points. In this expression, D_{ij} are Euclidean distances in $\mathbb{R}^{mq}$ and ϵ refers to the extent of the local neighborhood. An estimate of this parameter is obtained as described in Ref. 30. For each time-lagged embedded point, $b \ll s$ nearest neighbors are considered. In standard NLSA, b Euclidean nearest neighbors are retained (E-NLSA). Alternatively, we consider b nearest neighbors in time (T-NLSA). After symmetrization, the diffusion kernel is normalized to consider local densities:[31]

$$\tilde{K}_{ij} = \frac{K_{ij}}{\sum\limits_i K_{ij} \sum\limits_j K_{ij}}. \tag{18}$$

With $Q_i = \sum_j \tilde{K}_{ij}$ and $Q_{ij} = Q_i \delta_{ij}$, we define the diagonal and positive-definite matrix $\boldsymbol{Q}$. We solve the eigenvalue problem,

$$\boldsymbol{W\Phi} = \boldsymbol{\Phi}\Lambda, \tag{19}$$

for the real and symmetric matrix,

$$\boldsymbol{W} = \boldsymbol{Q}^{-1/2}\tilde{\boldsymbol{K}}\boldsymbol{Q}^{-1/2}. \tag{20}$$

The orthonormal eigenvectors $\boldsymbol{\Phi}$ are closely related to the (in general, nonorthogonal) eigenvectors of the probability matrix,

$$\boldsymbol{P} = \boldsymbol{Q}^{-1}\tilde{\boldsymbol{K}}, \tag{21}$$

obtained by row-normalization of $\tilde{\boldsymbol{K}}$.

3. Modal decomposition and reconstruction

We use a subset of the orthonormal eigenvectors $\boldsymbol{\Phi}$ to construct a data-driven projector to a subspace of $\mathbb{R}^s$. A number l of eigenvectors, related to the eigenvalues Λ_{ii} with largest absolute value, are retained and used as a basis set for the subspace projection, analogous to Eq. (11). Modal decomposition and signal reconstruction are carried out as described in Sec. III.

C. Time binning

To compare the dynamics-retrieval results to those from time binning and merging, we compute the running average of $\boldsymbol{x}(t)$ for various values of the time bin size.

VI. DISCUSSION

We compare the performance of LPSA to that of the methods presented in Sec. V on the task of reconstructing the synthetic signal presented in Sec. IV [Eq. (15)], from input data with extreme incompleteness and partiality. Figure 7 shows the evolution of the linear correlation between the recovered signal and the underlying dynamics, as a function of the number of modes employed, for the various techniques considered.

With an increasing number of modes, the reconstructed signal from pure SVD reproduces more and more closely the sparse and partial input data. A maximum in the correlation between the reconstructed signal and the underlying dynamics is observed with four modes. As the number of modes exceeds four, the correlation

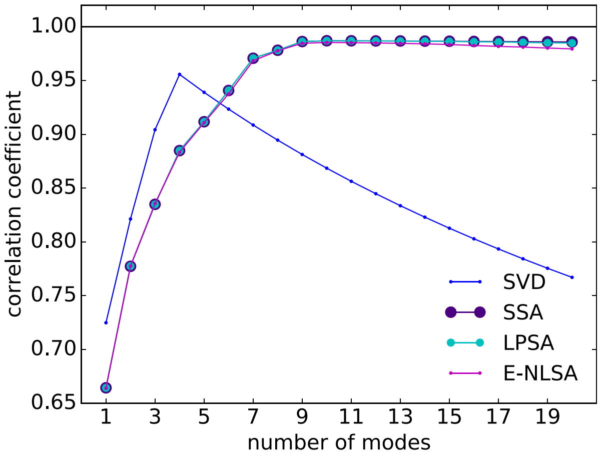

FIG. 7. Comparison between SVD, SSA ($q = 4000$), LPSA ($q = 4000$, $j_{\max} = 100$), E-NLSA ($q = 4000$, $b = 3000$, $\log_{10}\epsilon = 1.0$, $l = 50$). Linear correlation between the reconstructed dynamics from sparse and partial input data and the underlying dynamics.

deteriorates. The maximal correlation achievable by SVD lies well below that from concatenation-based or projection-based methods. By including a time-lagged embedding step (SSA), the reconstruction achieves optimal quality with ten modes, and shows a robust convergence with increasing number of modes. The drawback resides in the large size of the matrix $\boldsymbol{X} \in \mathbb{R}^{mq \times s}$ to be singular-value decomposed.

NLSA produces excellent reconstruction results, similar to SSA. The subspace projection allows to reduce the size of the matrix to be singular-value decomposed ($\boldsymbol{A} \in \mathbb{R}^{mq \times l}$, with $l \ll s$). However, the computation of data-driven subspace basis vectors is expensive, as it involves the calculation of s^2 (in E-NLSA) Euclidean distances between tuples in $\mathbb{R}^{mq}$, and the eigendecomposition of the large (but sparse) matrix $\boldsymbol{W} \in \mathbb{R}^{s \times s}$. A specific drawback of NLSA is that the parameter space to be considered is four dimensional (q, b, ϵ, l).

We compare the results from E-NLSA and T-NLSA, for $b = 1500$ and $b = 3000$. Figure 8 shows that the best reconstruction is obtained with T-NLSA and $b = 1500$. This is due to the fact that T-NLSA effectively uses the time order of the input data as prior knowledge to guide the choice of the nearest neighbors. In addition, compared to E-NLSA, T-NLSA presents the advantage that only bs, rather than s^2, Euclidean distances must be computed. The use of a data-driven subspace basis may be important when dealing with chaotic dynamical systems, whereby the underlying dynamics explores a truly high-dimensional subspace of $\mathbb{R}^{mq}$. NLSA appears in this case to effectively provide a low-rank representation of the dynamics, where SSA fails to do so.[27] Data-driven basis functions were found to approximate a set of periodic functions at large values of the concatenation parameter.[32]

The dynamics retrieval from LPSA is excellent, as shown in Fig. 7. Similar to NLSA, the subspace projection allows to reduce the size of the matrix to be singular-value decomposed ($\boldsymbol{A} \in \mathbb{R}^{mq \times k}$, with $k = 2j_{\max} + 1 \ll s$). In contrast to NLSA, the computation of the subspace basis set is inexpensive in LPSA. In addition, LPSA only involves two parameters $(q, j_{\max})$, which represents a major practical advantage compared to NLSA. LPSA involves particularly cheap computations when no time-lagged embedding is performed ($q = 1$). While excellent results can be achieved in this case, it is important to consider that the number of employed modes plays a major role in determining the quality of the reconstruction, as convergence with respect to the number of modes cannot be assured. In this case, the singular value spectrum, and the deviation from local linearity measured by the indicator L, can be used to guide the choice of the number of modes.

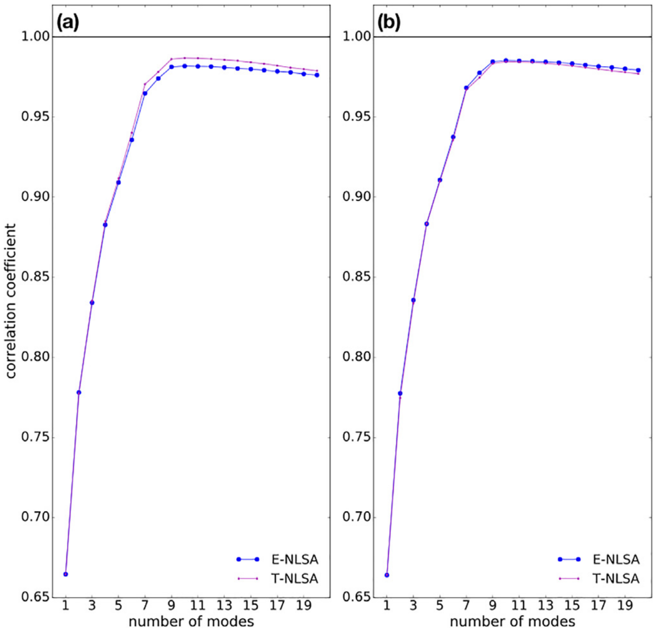

FIG. 8. Comparison between E-NLSA and T-NLSA. Linear correlation between the reconstructed signal and the underlying dynamics as a function of the number of modes employed. The NLSA was computed with concatenation parameter $q = 4000$, subspace dimension $l = 50$, neighborhood size $\log_{10}\epsilon = 1.0$, nearest-neighbor number (a) $b = 1500$ and (b) $b = 3000$.

Typical values in TR-SFX applications are $S \sim 10^5$, $m \sim 10^4$, $q \sim 10^4$, and $s \sim 10^5$. In this context, SSA mandates the SVD of a 10^{13}-element matrix. By contrast, projection-based methods involve the SVD of a much smaller matrix. With the number of basis vectors typically ranging between 10^1 and 10^2, $\boldsymbol{A}$ is substantially smaller than $\boldsymbol{X}$ in both NLSA and LPSA TR-SFX applications. However, the calculation of data-driven basis vectors for the NLSA subspace projection is computationally expensive. To this end, the calculation of distances between time-lagged embedded vectors is particularly challenging. To give an example, the analysis of a 10^5-frame dataset, with $m \sim 10^4$ and $q \sim 10^4$, involves the computation of $\sim 10^{10}$ Euclidean distances between 10^8-element tuples. In addition, to obtain data-driven subspace basis vectors, a (sparse) 10^{10}-element matrix must be eigendecomposed. In this respect, LPSA presents the practical advantage that the subspace basis vectors can be readily computed as a set of orthonormalized trigonometric functions.

Finally, we compare our results to those from time binning and merging, which is to date the most commonly used technique to analyze TR-SFX data. Figure 9 shows the linear correlation between the binned and merged signal and the underlying dynamics, as a function of the size of the time bins. The maximal correlation achieved is 0.952,

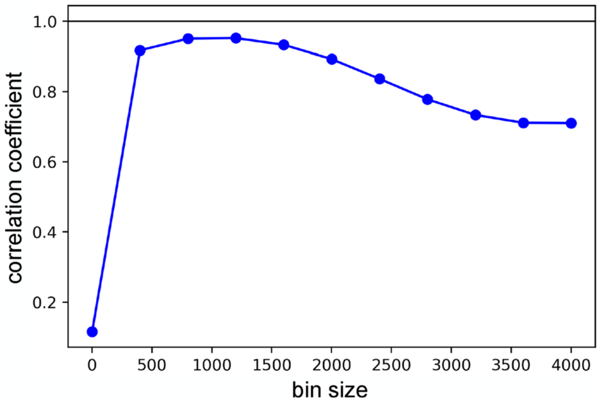

FIG. 9. Linear correlation between the binned and merged signal obtained by computing the running average of $\boldsymbol{x}(t)$, and the underlying dynamics, as a function of the bin size employed.

below the value of 0.987 obtained by LPSA. The difference in the quality of the reconstruction can be appreciated by comparing the binned and merged signal in Fig. 10(c) to the LPSA results in Fig. 10(b). It should be emphasized that, in the standard binning-and-merging approach, there is no intrinsic way to ensure whether the width of each bin has been selected optimally. Here, we present the best-case scenario, in which we choose the optimal bin size by maximizing the correlation with the benchmark, which is generally not available. By contrast, LPSA parameters can be optimized based on the deviation from local linearity and the singular value spectrum, i.e., indicators that do not depend on *a priori* knowledge of the ground truth.

VII. CONCLUSIONS

We have presented a new approach to retrieving dynamical information from highly incomplete and partial data, of the type obtained by TR-SFX experiments. This approach, which we call LPSA, allows an improved signal reconstruction, compared to time binning and merging, which is to date the most common procedure to gain dynamical insight from TR-SFX data. We have also shown that, while achieving the same reconstruction quality as other sophisticated dynamics retrieval techniques (SSA, NLSA), LPSA presents practical advantages, in particular concerning the computational cost of the algorithm, and the number of parameters to be optimized. While being developed in the context of TR-SFX data analysis, LPSA is a general tool for the analysis of time series affected by stochastic weighting and incompleteness, which could be employed in a diverse range of applications in science and engineering.

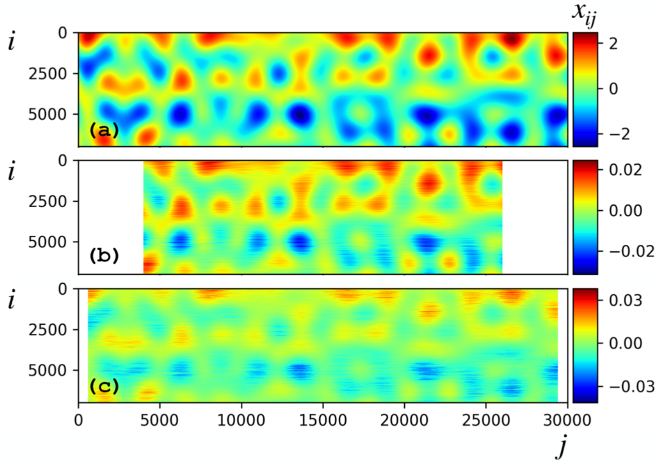

FIG. 10. (a) Underlying dynamics $\boldsymbol{x}(t)$ [Eq. (15)]. (b) Reconstruction by LPSA with $q = 4000$, $j_{\max} = 100$ and 10 modes. The correlation coefficient to the underlying signal is 0.9869. (c) Reconstruction by time-binning and merging with bins comprising 1201 frames. The correlation coefficient to the underlying signal is 0.9519.

ACKNOWLEDGMENTS

G.F.X.S. acknowledges Swiss National Science Foundation Grant Nos. 173335, 192760, and 192780. A.H. and A.O. were supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award No. DE-SC0002164. R.S. acknowledges support from the Cluster of Excellence “CUI: Advanced Imaging of Matter” of the Deutsche Forschungsgemeinschaft (DFG)-EXC 2056-project ID 390715994.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Cecilia Maria Casadei: Formal analysis (lead); Investigation (equal); Methodology (supporting); Software (lead); Visualization (lead); Writing – original draft (lead); Writing – review and editing (equal). **Ahmad Hosseinizadeh:** Methodology (supporting); Writing – review and editing (supporting). **Gebhard F. X. Schertler:** Conceptualization (equal); Funding acquisition (lead); Investigation (supporting); Project administration (lead); Resources (lead); Supervision (equal); Writing – review and editing (equal). **Abbas Ourmazd:** Conceptualization (supporting); Investigation (supporting); Methodology (supporting); Writing – review and editing (supporting). **Robin Santra:** Conceptualization (equal); Formal analysis (supporting); Investigation (equal); Methodology (lead); Supervision (equal); Writing – original draft (supporting); Writing – review and editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

[1] I. Schlichting, “Serial femtosecond crystallography: The first five years,” IUCrJ **2**, 246–255 (2015).
[2] P. Fromme, “Xfels open a new era in structural chemical biology,” Nat. Chem. Biol. **11**, 895–899 (2015).
[3] J. C. H. Spence, “XFELs for structure and dynamics in biology,” IUCrJ **4**, 322–339 (2017).
[4] H. N. Chapman, “X-ray free-electron lasers for the structure and dynamics of macromolecules,” Annu. Rev. Biochem. **88**, 35–58 (2019).
[5] H. M. Ginn, A. S. Brewster, J. Hattne, G. Evans, A. Wagner, J. M. Grimes, N. K. Sauter, G. Sutton, and D. I. Stuart, “A revised partiality model and post-

refinement algorithm for X-ray free-electron laser data," Acta Crystallogr., Sect. D **71**, 1400–1410 (2015).

[6] T. A. White, V. Mariani, W. Brehm, O. Yefanov, A. Barty, K. R. Beyerlein, F. Chervinskii, L. Galli, C. Gati, T. Nakane, A. Tolstikova, K. Yamashita, C. H. Yoon, K. Diederichs, and H. N. Chapman, "Recent developments in CrystFEL," J. Appl. Crystallogr. **49**, 680–689 (2016).

[7] M. R. Bionta, H. T. Lemke, J. P. Cryan, J. M. Glownia, C. Bostedt, M. Cammarata, J.-C. Castagna, Y. Ding, D. M. Fritz, A. R. Fry, J. Krzywinski, M. Messerschmidt, S. Schorb, M. L. Swiggers, and R. N. Coffee, "Spectral encoding of x-ray/optical relative delay," Opt. Express **19**, 21855–21865 (2011).

[8] J. M. Glownia, K. Gumerlock, H. T. Lemke, T. Sato, D. Zhu, and M. Chollet, "Pump–probe experimental methodology at the Linac Coherent Light Source," J. Synchrotron Radiat. **26**, 685–691 (2019).

[9] D. Borek, W. Minor, and Z. Otwinowski, "Measurement errors and their consequences in protein crystallography," Acta Crystallogr., Sect. D **59**, 2031–2038 (2003).

[10] P. Ewald, "Zur theorie der interferenzen der röntgenstrahlen in kristallen," Phys. Z. **14**, 465 (1913).

[11] D. Sherwood and J. Cooper, *Crystals, X-Rays, and Proteins: Comprehensive Protein Crystallography* (Oxford University Press, 2015).

[12] J. Tenboer, S. Basu, N. Zatsepin, K. Pande, D. Milathianaki, M. Frank, M. Hunter, S. Boutet, G. J. Williams, J. E. Koglin, D. Oberthuer, M. Heymann, C. Kupitz, C. Conrad, J. Coe, S. Roy-Chowdhury, U. Weierstall, D. James, D. Wang, T. Grant, A. Barty, O. Yefanov, J. Scales, C. Gati, C. Seuring, V. Srajer, R. Henning, P. Schwander, R. Fromme, A. Ourmazd, K. Moffat, J. J. V. Thor, J. C. H. Spence, P. Fromme, H. N. Chapman, and M. Schmidt, "Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein," Science **346**, 1242–1246 (2014).

[13] T. R. M. Barends, L. Foucar, A. Ardevol, K. Nass, A. Aquila, S. Botha, R. B. Doak, K. Falahati, E. Hartmann, M. Hilpert, M. Heinz, M. C. Hoffmann, J. Köfinger, J. E. Koglin, G. Kovacsova, M. Liang, D. Milathianaki, H. T. Lemke, J. Reinstein, C. M. Roome, R. L. Shoeman, G. J. Williams, I. Burghardt, G. Hummer, S. Boutet, and I. Schlichting, "Direct observation of ultrafast collective motions in co myoglobin upon ligand dissociation," Science **350**, 445–450 (2015).

[14] E. Nango, A. Royant, M. Kubo, T. Nakane, C. Wickstrand, T. Kimura, T. Tanaka, K. Tono, C. Song, R. Tanaka, T. Arima, A. Yamashita, J. Kobayashi, T. Hosaka, E. Mizohata, P. Nogly, M. Sugahara, D. Nam, T. Nomura, T. Shimamura, D. Im, T. Fujiwara, Y. Yamanaka, B. Jeon, T. Nishizawa, K. Oda, M. Fukuda, R. Andersson, P. Båth, R. Dods, J. Davidsson, S. Matsuoka, S. Kawatake, M. Murata, O. Nureki, S. Owada, T. Kameshima, T. Hatsui, Y. Joti, G. Schertler, M. Yabashi, A.-N. Bondar, J. Standfuss, R. Neutze, and S. Iwata, "A three-dimensional movie of structural changes in bacteriorhodopsin," Science **354**, 1552–1557 (2016).

[15] K. Pande, C. D. M. Hutchison, G. Groenhof, A. Aquila, J. S. Robinson, J. Tenboer, S. Basu, S. Boutet, D. P. DePonte, M. Liang, T. A. White, N. A. Zatsepin, O. Yefanov, D. Morozov, D. Oberthuer, C. Gati, G. Subramanian, D. James, Y. Zhao, J. Koralek, J. Brayshaw, C. Kupitz, C. Conrad, S. Roy-Chowdhury, J. D. Coe, M. Metz, P. L. Xavier, T. D. Grant, J. E. Koglin, G. Ketawala, R. Fromme, V. Šrajer, R. Henning, J. C. H. Spence, A. Ourmazd, P. Schwander, U. Weierstall, M. Frank, P. Fromme, A. Barty, H. N. Chapman, K. Moffat, J. J. van Thor, and M. Schmidt, "Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein," Science **352**, 725–729 (2016).

[16] P. Nogly, T. Weinert, D. James, S. Carbajo, D. Ozerov, A. Furrer, D. Gashi, V. Borin, P. Skopintsev, K. Jaeger, K. Nass, P. Båth, R. Bosman, J. Koglin, M. Seaberg, T. Lane, D. Kekilli, S. Brünle, T. Tanaka, W. Wu, C. Milne, T. White, A. Barty, U. Weierstall, V. Panneels, E. Nango, S. Iwata, M. Hunter, I. Schapiro, G. Schertler, R. Neutze, and J. Standfuss, "Retinal isomerization in bacteriorhodopsin captured by a femtosecond x-ray laser," Science **361**, eaat0094 (2018).

[17] N. Coquelle, M. Sliwa, J. Woodhouse, G. Schirò, V. Adam, A. Aquila, T. R. M. Barends, S. Boutet, M. Byrdin, S. Carbajo, E. De la Mora, R. B. Doak, M. Feliks, F. Fieschi, L. Foucar, V. Guillon, M. Hilpert, M. S. Hunter, S. Jakobs, J. E. Koglin, G. Kovacsova, T. J. Lane, B. Lévy, M. Liang, K. Nass, J. Ridard, J. S. Robinson, C. M. Roome, C. Ruckebusch, M. Seaberg, M. Thepaut, M. Cammarata, I. Demachy, M. Field, R. L. Shoeman, D. Bourgeois, J.-P. Colletier, I. Schlichting, and M. Weik, "Chromophore twisting in the excited state of a photoswitchable fluorescent protein captured by time-resolved serial femtosecond crystallography," Nat. Chem. **10**, 31–37 (2018).

[18] P. Skopintsev, D. Ehrenberg, T. Weinert, D. James, R. K. Kar, P. J. M. Johnson, D. Ozerov, A. Furrer, I. Martiel, F. Dworkowski, K. Nass, G. Knopp, C. Cirelli, C. Arrell, D. Gashi, S. Mous, M. Wranik, T. Gruhl, D. Kekilli, S. Brünle, X. Deupi, G. F. X. Schertler, R. M. Benoit, V. Panneels, P. Nogly, I. Schapiro, C. Milne, J. Heberle, and J. Standfuss, "Femtosecond-to-millisecond structural changes in a light-driven sodium pump," Nature **583**, 314–318 (2020).

[19] T. Willmore, *An Introduction to Differential Geometry*, Dover Books on Mathematics Series (Dover Publications, 2012).

[20] N. Aubry, R. Guyonnet, and R. Lima, "Spatiotemporal analysis of complex signals: Theory and applications," J. Stat. Phys. **64**, 683–739 (1991).

[21] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. (The Johns Hopkins University Press, 1996).

[22] A. Hosseinizadeh, N. Breckwoldt, R. Fung, R. Sepehr, M. Schmidt, P. Schwander, R. Santra, and A. Ourmazd, "Few-fs resolution of a photoactive protein traversing a conical intersection," Nature **599**, 697–701 (2021).

[23] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," Phys. Rev. Lett. **45**, 712–716 (1980).

[24] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980*, edited by D. Rand and L.-S. Young (Springer, Berlin, Heidelberg, 1981), pp. 366–381.

[25] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," J. Stat. Phys. **65**, 579–616 (1991).

[26] R. Vautard and M. Ghil, "Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series," Physica D **35**, 395–424 (1989).

[27] D. Giannakis and A. J. Majda, "Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability," Proc. Natl. Acad. Sci. **109**, 2222–2227 (2012).

[28] R. Fung, A. M. Hanna, O. Vendrell, S. Ramakrishna, T. Seideman, R. Santra, and A. Ourmazd, "Dynamics from noisy data with extreme timing uncertainty," Nature **532**, 471–475 (2016).

[29] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," Proc. Natl. Acad. Sci. **102**, 7426–7431 (2005).

[30] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph Laplacian tomography from unknown random projections," IEEE Trans. Image Process. **17**, 1891–1899 (2008).

[31] R. R. Coifman and S. Lafon, "Diffusion maps," Appl. Comput. Harmonic Anal. **21**, 5–30 (2006).

[32] D. Giannakis, "Delay-coordinate maps, coherence, and approximate spectra of evolution operators," Res. Math. Sci. **8**, 8 (2021).