Few-fs resolution of a photoactive protein traversing a conical intersection

https://doi.org/10.1038/s41586-021-04050-9

Received: 12 November 2020

Accepted: 23 September 2021

Published online: 3 November 2021

Check for updates

A. Hosseinizadeh[1], N. Breckwoldt[2,3,4], R. Fung[1], R. Sepehr[1], M. Schmidt[1], P. Schwander[1], R. Santra[2,3,4] & A. Ourmazd[1]✉

The structural dynamics of a molecule are determined by the underlying potential energy landscape. Conical intersections are funnels connecting otherwise separate potential energy surfaces. Posited almost a century ago[1], conical intersections remain the subject of intense scientific interest[2–5]. In biology, they have a pivotal role in vision, photosynthesis and DNA stability[6]. Accurate theoretical methods for examining conical intersections are at present limited to small molecules. Experimental investigations are challenged by the required time resolution and sensitivity. Current structure-dynamical understanding of conical intersections is thus limited to simple molecules with around ten atoms, on timescales of about 100 fs or longer[7]. Spectroscopy can achieve better time resolutions[8], but provides indirect structural information. Here we present few-femtosecond, atomic-resolution videos of photoactive yellow protein, a 2,000-atom protein, passing through a conical intersection. These videos, extracted from experimental data by machine learning, reveal the dynamical trajectories of de-excitation via a conical intersection, yield the key parameters of the conical intersection controlling the de-excitation process and elucidate the topography of the electronic potential energy surfaces involved.

The quantum mechanical energies of molecular electrons as a function of molecular geometry give rise to effective potential energy surfaces for the motion of atomic nuclei. When there are d nuclear degrees of freedom, the potential energy surface (PES) is d dimensional. In the so-called Born–Oppenheimer (BO) approximation, the electronic and nuclear degrees of freedom are treated separately. When two PESs come into contact, the BO approximation is no longer valid.

A conical intersection is a region of such potential energy degeneracy, forming a $(d-2)$-dimensional manifold with divergent, non-BO coupling between the participating electronic states. The resultant strong mixing of electronic and vibrational degrees of freedom opens a pathway by which dynamical changes in molecular geometry can cause a transition from one electronic state to another. As this gives rise to ultrafast, non-radiative relaxation of the excited state, conical intersections have an important role in numerous processes in nature. *Trans*-to-*cis* isomerizations of the *p*-coumaric acid chromophore in photoactive yellow protein (PYP)[9,10] (Extended Data Fig. 1a) and retinal[6,11] are prime examples.

Accurate theoretical methods for treating coupled electronic and vibrational dynamics are currently restricted to small molecules. The quality of such simulations—using either a quantum or a classical treatment of nuclear motions—depends on the precise characterization and complexity of the PESs involved. PYP, for example, is composed of 2,289 atoms[12], exhibiting 6,861 vibrational degrees of freedom. This level of complexity renders rigorous, first-principles electronic structure calculations unfeasible for the foreseeable future. State-of-the-art density functional theory can be applied to molecules of comparable size[13], but does not yet provide reliable chemical accuracy, particularly for conical intersections. Even if it were possible to solve the electronic Schrödinger equation for a single molecular geometry accurately, the total number of molecular geometries needed for adequate sampling of the potential energy landscape as a whole grows exponentially with the number of degrees of freedom.

Experimentally, conclusive observation of the structure-dynamical modes involved in electronic switching via conical intersections has remained elusive because high temporal and spatial resolutions must be combined to resolve the ultrafast dynamics with sufficient acuity. Optical pump–probe spectroscopy provides information on the electronic-state population dynamics with femtosecond time resolution, but it does not offer direct access to the structural properties of the system[14]. Similarly, time-resolved diffraction techniques, such as ultrafast electron diffraction (UED) and time-resolved X-ray diffraction, have, up to now, lacked the temporal resolution needed to follow de-excitation via a conical intersection[7,15]. Recently, the combination of UED experiments with extensive, sophisticated ab initio simulations accomplished the structural characterization of conical intersection-induced dynamics in an 11-atom molecule with a time resolution of 150 fs (ref. [7]).

Here we report four key advances. First, structure-dynamical collective modes and trajectories of ultrafast de-excitation can be extracted with atomic spatial resolution and few-femtosecond time resolution from existing time-resolved crystallographic data. Second, in combination with tractable and accurate computational methods, the topography of the electronic states involved in de-excitation via a

[1]University of Wisconsin Milwaukee, Milwaukee, WI, USA. [2]Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany. [3]Department of Physics, Universität Hamburg, Hamburg, Germany. [4]The Hamburg Centre for Ultrafast Imaging, Hamburg, Germany. ✉e-mail: ourmazd@uwm.edu

conical intersection can be determined. Third, our approach can be used to determine the key collective variables and boundary conditions controlling the de-excitation dynamics of molecules consisting of thousands of atoms. Finally, the combination of data-driven machine learning with existing experimental and theoretical techniques offers the time resolution of spectroscopy and the spatial resolution of structural methods.

In the illustrative example of PYP, our approach provides the following new insights.

1. High-frequency charge oscillations are involved in the *trans*-to-*cis* isomerization process in PYP. These oscillations involve previously ignored 'peripheral' regions of the chromophore (Extended Data Fig. 1a).
2. The presence of the above-mentioned oscillations in PYP has been independently corroborated in the spectroscopically accessible range of 3–30 THz, but with no direct structural information. Our work extends the experimentally accessible range to 100 THz, and provides high-resolution spatial information on the structural elements involved in these ultrafast oscillations.
3. Photo-excited PYP proceeds towards the conical intersection via one of five conduits, revealing the structure-dynamical trajectories of approach to the conical intersection.
4. Structure-dynamical trajectories of passage through the conical intersection have been determined, allowing measurement of the key parameters of the conical intersection in PYP. These parameters are essential for a quantitative understanding of the PES manifold and the isomerization process in PYP.
5. Our work establishes an important bridge between structural and spectroscopic studies. This link is essential for a complete understanding of PYP.

The above insights represent a clear advance in understanding of PYP and, by extension, a wide range of 'ultrafast' structure-dynamical systems.

The experimental data were obtained in a time-resolved (optical pump, X-ray probe) serial femtosecond crystallographic study of PYP, as reported in detail elsewhere[10]. This protein is known to undergo a rapid *trans*-to-*cis* isomerization reaction via a conical intersection[10]. The data consist of a time series of two-dimensional (2D) diffraction snapshots, each stemming from a different random central slice through the three-dimensional (3D) diffraction volume. Each 'light' 2D snapshot was recorded after optical excitation at a time point known to an accuracy of ~100 fs owing to unavoidable 'timing jitter' between the optical pump and X-ray probe pulses[10,16]. In addition, 'dark' snapshots were recorded without any optical excitation. Conventionally, enough light 2D snapshots from the same nominal time point are indexed and combined (merged) to obtain the 3D diffraction volume and, from there, the difference between the light and dark atomic structures at each time point (see, for example, ref. [10]). The timing jitter limits the time resolution of the merged 3D volumes to ~100 fs. This is a severe limitation because de-excitation via a conical intersection is often complete within that timeframe (see below and ref. [8]).

Analysis by machine learning

As outlined in Supplementary Information sections 1–4, we circumvent this problem by applying manifold-based machine learning[16–20] to the same dataset of 2D diffraction snapshots, to reconstruct a time series of 3D diffraction volumes, each pertaining to a time point determined with an accuracy of about 1 fs (for details, see Methods sections 'Overview of algorithmic approach', 'Data preprocessing', 'Data representation' and 'Manifold-based machine learning', Supplementary Fig. 2 and ref. [16]). In essence, our approach rests on the celebrated realization by Takens[21] and Packard[22] that dynamics tightly constrain the time evolution of a system. This means that much fewer data are needed to reconstruct dynamics than conventionally thought necessary. As an extreme example, Newton's laws of motion require only one snapshot of the initial conditions (positions and momenta) and the forces acting on a system to predict the dynamical evolution of a non-chaotic system forever. In a similar vein, the time evolution of the diffraction signal is highly constrained by the charge dynamics of the photo-excited system under observation. This allows an essentially jitter-free time series of 3D diffraction volumes to be recovered from a time series of 2D central slices, each recorded with substantial timing uncertainty. This algorithmic approach has been validated with experimental data[16,23,24] and with data from synthetic models, where the actual 'ground truths' are known[16,23,24] (see also Methods sections 'Validating the time resolution by comparison with spectroscopic results' and 'Validation with synthetic data', Extended Data Table 1, Extended Data Figs. 2–8 and Supplementary Figs. 1–5).

Armed with a series of accurately timestamped 3D diffraction volumes, standard time-resolved crystallographic approaches[25,26] can be used to compile jitter-free difference electron density (DED) videos, revealing the dynamics of the photo-excited charge distribution. As described in detail in ref. [16], using time-lagged embedding[21,22], our data-analytical pipeline 'learns' the Riemannian manifold on which the dynamics unfold, and conducts all analysis, including (nonlinear) singular value decomposition, on that curved manifold[16,19]. This approach yields the characteristic collective modes of the charge distribution ('topos') and their respective time evolutions ('chronos'). In essence, each topo represents a characteristic DED map, evolving in time as prescribed by its corresponding chrono (Fig. 1). Each topo–chrono pair thus represents a characteristic structure-dynamical mode of the charge distribution. These modes constitute the empirical basis functions, which, in combination, describe the dynamical trajectories ('the reaction paths') of the system. Videos of the structure-dynamical modes with few-femtosecond time resolution are shown in Supplementary Videos 1–4. As shown in Extended Data Fig. 1b, c, these modes can be combined to describe the structural dynamics in terms of the more intuitive torsional angle as a reaction coordinate.

Validating the time resolution

Fourier analysis of the chronos by multi-taper methods[27,28] reveals the clear presence of frequencies of up to 95 THz (10.5 fs) at signal-to-noise ratios of ~5 or higher (Extended Data Fig. 3). As the observation of a frequency component in an initially non-uniform set of time points requires a time resolution ~5–10 times shorter than the period of the component[29], the clear observation of a signal of 10.5 fs validates the few-femtosecond time resolution of our approach. This high time resolution is particularly remarkable because the data were obtained with a 140-fs optical pump pulse[10]. This suggests that, in the experiment, a temporal gating effect may have had a role, as a consequence of, for example, nonlinear multiphoton processes[30] or light-induced structural disorder[31].

Fourier analysis of the chronos before 615 fs, that is, before the encounter with the conical intersection brings multiple species into play, reveals three prominent peaks at 4, 21 and 33.5 THz in the spectroscopically well-explored region spanning the range from 3–30 THz and its vicinity. These peaks have been previously observed by ultrafast time-domain Raman spectroscopy of PYP before the conical intersection[32]. These spectroscopically measured PYP peak frequencies match those we observe to within 7% or better (see Methods section 'Comparison with spectroscopically determined PYP frequency spectrum' and Extended Data Table 1). Such close agreement with independently known ground truths is the ultimate test of any machine learning approach.

On the basis of the argument outlined above, the clear observation of the oscillatory signal at 33.5 THz (30 fs) already indicates a time resolution of 3–6 fs (ref. [29]). The signal at 95 THz may be associated with a C–H bond stretch, typically observed in the band at 100 THz (10–12 fs) at moderate to high intensities. It is also possible that this

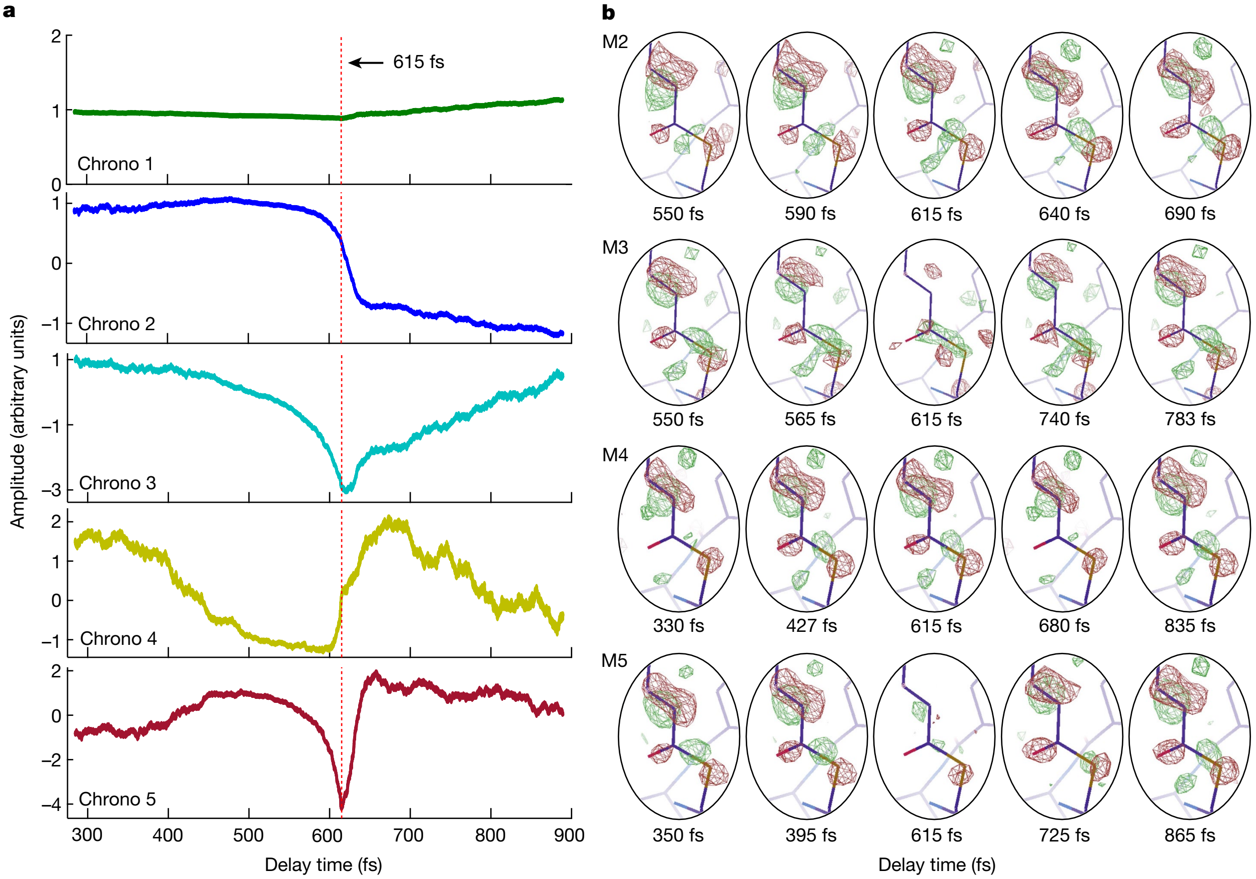


Fig. 1 | Evolution of dynamical modes as a function of pump–probe delay time. **a**, Chronos. Note the sharp turning point at 615 fs in all chronos. **b**, Evolution of DED with time. In each case, mode 1 (not shown) represents the moving average, and has been added to allow comparison with DEDs obtained by conventional means. The DED maps near the top of the oval region have been associated with *trans*-to-*cis* isomerization. The patterns near the bottom of the oval region show strong oscillatory charge dynamics. M, mode. Contour level, 3σ.

oscillatory signal stems from an N–H vibrational bond stretch, which typically results in an absorption band in the range of 9.5–10.4 fs (ref. [33]).

Structure-dynamical modes

As mode 1 represents the moving average of the signal, its topo is added to each of the subsequent modes to facilitate comparison with DEDs obtained by conventional means. Supplementary Videos 1–4 reveal the nature of the structure-dynamical motifs involved in the relaxation of PYP. Supplementary Videos 5–9 display the combinations of these motifs, which reveal the actual isomerization trajectories at work (see also Extended Data Fig. 2 and below).

Mode 2, with a time evolution (chrono) resembling a step function (Fig. 1), represents the only structure-dynamical evolution not reversed within the picosecond timespan of the dataset. This mode also captures strong DED features in the vicinity of the C2–C3 and C3–C4 bonds in PYP (Extended Data Fig. 1a) associated with the *trans*-to-*cis* isomerization in PYP[34]. The other chronos return roughly to their initial values in the course of the experimental timespan (Fig. 1). All modes reveal previously unreported rapid charge oscillations.

Structure-dynamical trajectories

Although the characteristic structure-dynamical modes (basis functions) described above display key features of the system, they must be appropriately combined to reconstruct the trajectories (reaction pathways) associated with de-excitation via the conical intersection. As in standard singular value decomposition, additional information is needed to accomplish this task[16,26,35]. Away from the conical intersection, we determine the appropriate mode combinations as previously described in ref. [36], and extract structures from the resulting DEDs as outlined in ref. [34].

Near the conical intersection, the mode space is 2D, and the structure-dynamical trajectories of passage through the conical intersection can be determined as follows. We compare the six possible experimental trajectories obtained from pairwise combinations of experimental modes with simulated trajectories contained in a dictionary of half a million theoretically calculated de-excitation trajectories.

Specifically, using the 500,000 different combinations of the six model parameters, we determine, as a function of time, the expectation values of the position operators associated with each mode. In this way, we obtain dynamical trajectories in the 2D space spanned by two collective modes, denoted x and y for short.

A dynamical trajectory in this 2D space consists of a time-ordered sequence of points given by $(x(t), y(t))$. Each instance of $x(t)$ and $y(t)$ is associated with a characteristic DED map (a topo), whose time-varying contribution is determined by the associated chrono. Thus, in the space spanned by x and y, a potential dynamical trajectory is obtained by plotting two chronos against each other. (Of course, the choice of characteristic modes is not unique; any linear combination, with or without sign inversion and scaling, provides an equally good basis set.) Using the bank of 500,000 quantum-dynamical de-excitation trajectories in the vicinity of a conical intersection, we identified the experimental trajectories leading to high-probability de-excitation of PYP via a conical intersection (see Methods section 'Overview of theoretical model used to simulate dynamical trajectories').

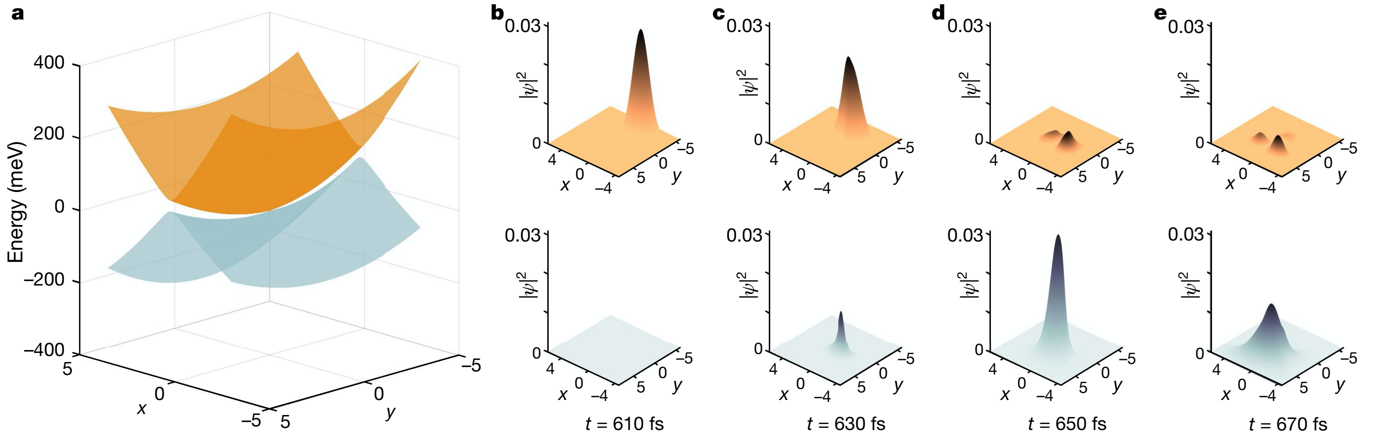


Fig. 2 | Topography of the conical intersection and the associated population dynamics in PYP, as deduced from five experimental dynamic trajectories. **a**, Reconstructed potential energy landscape in the close vicinity of the conical intersection. Touching of the two PESs constitutes the conical intersection. The topography of the landscape is determined by the vibrational frequencies γ_1 and γ_2 and the coupling λ between the two electronic states of the underlying vibronic coupling model. The numerical values of the model parameters are deduced from the best fit between simulated trajectories in the collective-mode space and their experimental counterparts. Both normal mode coordinates x and y are specified in dimensionless coordinates normalized with respect to the frequency ω. **b**–**e**, Time evolution of a Gaussian wave packet on the electronically excited state (top panels) and electronic ground state (bottom panels) in the vicinity of the conical intersection. The simulation stems from the best fit to the experimental trajectories of PYP.

PYP de-excitation trajectory

Of the six possible pairwise combinations of experimentally determined dynamical modes, five can be identified with simulated trajectories of de-excitation via the conical intersection (Extended Data Figs. 2, 4–7) (the sixth combination does not correspond to any simulated trajectory in our databank). Identification of these experimental trajectories and their simulated counterparts yields direct information on the structure-dynamical changes involved in the approach to, and the passage through, the conical intersection and the key parameters governing the properties of the PYP conical intersection itself, including its topography (Extended Data Table 2 and Fig. 2).

The very similar parameters obtained from all five pairwise combinations of experimentally determined structure-dynamical modes indicate that, in the vicinity of the conical intersection, the five dynamical trajectories can be described in terms of the same underlying conical intersection. Further away from the conical intersection, however, our analysis reveals the presence of at least five distinct de-excitation trajectories in PYP. We interpret these segments as different conduits to and from the vicinity of the PYP conical intersection. In this vicinity, the trajectories represent high-probability de-excitation routes on the same 2D PES manifold. These insights are essential to understanding the structural dynamics of PYP relaxation.

We now summarize the primary conclusions of our work. First, our results demonstrate a novel data-driven approach that combines the superb spatial resolution of structural methods, such as crystallography, with the exquisite time resolution of spectroscopy. In essence, this approach is tantamount to structure-dynamical spectroscopy with atomic spatial resolution and femtosecond timing acuity. Second, our results on the ultrafast atomic-level changes associated with the femtosecond de-excitation of PYP via a conical intersection reveal previously unobserved oscillatory charge dynamics involving often ignored regions surrounding the chromophore. Third, our results corroborate independent spectroscopic results on PYP in the usually accessed regime of 3–30 THz, provide direct structural information and extend the amenable range to ~100 THz. Fourth, our results reveal the structure-dynamical trajectories leading to the vicinity of, and through, the PYP conical intersection, and elucidate the properties of the PES manifold involved in PYP de-excitation.

More generally, by combining machine learning analysis of experimental data with simple and numerically accurate quantum-dynamical simulations, we have demonstrated a powerful data-driven route to studying a wide variety of important processes in complex molecular systems inaccessible by first-principles calculations, and established a bridge between spectroscopic and structural techniques for investigating ultrafast processes.

Of course, future tasks remain. These include investigating the possible effects of crystallinity on the observed relaxation modes and trajectories and whether the small number of important collective variables revealed by our approach offers a route to more accurate theoretical calculations than hitherto possible. These future tasks notwithstanding, our present results already reveal the unanticipated trove of information that can be extracted from existing experimental data by a combination of data-driven machine learning and physically based theory.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04050-9.

1. von Neumann, J. & Wigner, E. P. *Über das Verhalten von Eigenwerten bei adiabatischen Prozessen* Vol. A1 (Springer, 1993).
2. Chang, K. F. et al. Revealing electronic state-switching at conical intersections in alkyl iodides by ultrafast XUV transient absorption spectroscopy. *Nat. Commun.* **11**, 4042 (2020).
3. Cerullo, G. & Garavelli, M. A novel spectroscopic window on conical intersections in biomolecules. *Proc. Natl Acad. Sci. USA* **117**, 26553–26555 (2020).
4. Yang, J. et al. Imaging CF3I conical intersection and photodissociation dynamics with ultrafast electron diffraction. *Science* **361**, 64–67 (2018).
5. Tenboer, J. et al. Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science* **346**, 1242–1246 (2014).
6. Nogly, P. et al. Retinal isomerization in bacteriorhodopsin captured by a femtosecond X-ray laser. *Science* **361**, eaat0094 (2018).
7. Yang, J. et al. Simultaneous observation of nuclear and electronic dynamics by ultrafast electron diffraction. *Science* **368**, 885–889 (2020).
8. Zinchenko, K. S. et al. Sub-7-femtosecond conical-intersection dynamics probed at the carbon K-edge. *Science* **371**, 489–494 (2021).
9. Perman, B. et al. Energy transduction on the nanosecond time scale: early structural events in a xanthopsin photocycle. *Science* **279**, 1946–1950 (1998).
10. Pande, K. et al. Femtosecond structural dynamics drives the *trans*/*cis* isomerization in photoactive yellow protein. *Science* **352**, 725–729 (2016).
11. Polli, D. et al. Conical intersection dynamics of the primary photoisomerization event in vision. *Nature* **467**, 440–443 (2010).

12. Van Beeumen, J. J. et al. Primary structure of a photoactive yellow protein from the phototrophic bacterium *Ectothiorhodospira halophila*, with evidence for the mass and the binding site of the chromophore. *Protein Sci.* **2**, 1114–1125 (1993).
13. Jones, R. O. Density functional theory: its origins, rise to prominence, and future. *Rev. Mod. Phys.* **87**, 897–923 (2015).
14. Calegari, F., Sansone, G., Stagira, S., Vozzi, C. & Nisoli, M. Advances in attosecond science. *J. Phys. B* **49**, 062001 (2016).
15. Wolf, T. J. A. et al. The photochemical ring-opening of 1,3-cyclohexadiene imaged by ultrafast electron diffraction. *Nat. Chem.* **11**, 504–509 (2019).
16. Fung, R. et al. Dynamics from noisy data with extreme timing uncertainty. *Nature* **532**, 471–475 (2016).
17. Coifman, R. R. et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl Acad. Sci. USA* **102**, 7426–7431 (2005).
18. Giannakis, D., Schwander, P. & Ourmazd, A. The symmetries of image formation by scattering. I. Theoretical framework. *Opt. Express* **20**, 12799–12826 (2012).
19. Giannakis, D. & Majda, A. J. Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proc. Natl Acad. Sci. USA* **109**, 2222–2227 (2012).
20. Sauer, T., Yorke, J. A. & Casdagli, M. Embedology. *J. Stat. Phys.* **65**, 579–616 (1991).
21. Takens, F. in *Lecture Notes in Mathematics* Vol. 898 (ed. Warwick) 366–381 (Springer-Verlag, 1981).
22. Packard, N., Crutchfield, J., Farmer, J. & Shaw, R. Geometry from a time series. *Phys. Rev. Lett.* **45**, 712–716 (1980).
23. Fung, R. et al. Achieving accurate estimates of fetal gestational age and personalised predictions of fetal growth based on data from an international prospective cohort study: a population-based machine learning study. *Lancet Digit. Health* **2**, e368–e375 (2020).
24. Dashti, A. et al. Retrieving functional pathways of biomolecules from single-particle snapshots. *Nat. Commun.* **11**, 4734 (2020).
25. Moffat, K. The frontiers of time-resolved macromolecular crystallography: movies and chirped X-ray pulses. *Faraday Discuss.* **122**, 65–77; discussion 79–88 (2003).
26. Schmidt, M., Rajagopal, S., Ren, Z. & Moffat, K. Application of singular value decomposition to the analysis of time-resolved macromolecular X-ray data. *Biophys. J.* **84**, 2112–2129 (2003).
27. Prieto, G. A., Parker, R. L. & Vernon Iii, F. L. A Fortran 90 library for multitaper spectrum analysis. *Comput. Geosci.* **35**, 1701–1710 (2009).
28. Mitra, P. & Bokil, H. *Observed Brain Dynamics* (Oxford Univ. Press, 2008).
29. Jerri, A. J. The Shannon sampling theorem—its various extensions and applications: a tutorial review. *Proc. IEEE* **65**, 1565–1596 (1977).
30. Grunbein, M. L. et al. Illumination guidelines for ultrafast pump-probe experiments by serial femtosecond crystallography. *Nat. Methods* **17**, 681–684 (2020).
31. Barty, A. et al. Self-terminating diffraction gates femtosecond X-ray nanocrystallography measurements. *Nat. Photonics* **6**, 35–40 (2012).
32. Kuramochi, H. et al. Probing the early stages of photoreception in photoactive yellow protein with ultrafast time-domain Raman spectroscopy. *Nat. Chem.* **9**, 660–666 (2017).
33. Socratese, G. *Infrared and Raman Characteristic Group Frequencies: Tables and Charts* 3rd edn (Wiiley, 2004).
34. Pandey, S. et al. Time-resolved serial femtosecond crystallography at the European XFEL. *Nat. Methods* **17**, 73–78 (2020).
35. Henry, E. & Hofrichter, J. Singular value decomposition: application to analysis of experimental data. *Methods Enzymol.* **210**, 129–192 (1992).
36. Fung, R. et al. Dynamics from noisy data with extreme timing uncertainty. *Nature* **532**, 471–475 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Overview of algorithmic approach

We use manifold-based machine learning to extract accurate 3D structure-dynamical information from 2D snapshots (central slices through the diffraction volume). The approach exploits time-lagged embedding, a powerful strategy for extracting accurate dynamical information from measurements of a subset ('projections') of the system variables[16,19,21–23].

Manifold-based machine learning recognizes that data reside on curved manifolds, which can be learned from the data[17]. All analytical operations, such as singular value decomposition, are then performed on the learned manifold, which reflects the information content of the entire dataset[19]. We have previously shown that such algorithmic approaches are highly robust against noise[37] and timing uncertainty[16]. The work reported in this paper addresses the case where, in addition to substantial noise and timing uncertainty, the measurements are incomplete. Specifically, only a single 2D central slice is experimentally measured at each (inaccurately known) time point, but the 3D diffraction volume is required at accurately known time points.

The Diffusion Map embedding algorithm expresses the curved data manifold in terms of the eigenfunctions of the Laplace–Beltrami operator[17]. In the presence of data incompleteness, one can no longer assume that the manifold eigenfunctions pertain to the Laplace–Beltrami operator.

After preprocessing the experimental data as described below, data vectors are formed, each representing a 3D (crystal) diffraction volume at a given time point. As only one central slice of the diffraction volume is measured at each (approximately known) time point, the data vectors are incomplete, with many missing elements. A well-defined manifold can nonetheless be obtained by time-lagged embedding (time-ordered concatenation) of many snapshots ordered according to their nominal timestamps. As demonstrated previously[16,37], this algorithmic approach is highly robust to noise and timing uncertainty.

Data preprocessing

Diffracted intensity data collected at the Linac Coherent Light Source were indexed, integrated, rescaled and detwinned as described in ref. [34]. In total, 337,852 light snapshots, each containing about 600 Bragg reflections, were recorded after exposure to a 140-fs optical pulse. With assistance from a timing tool, the pump–probe delay was recorded with an accuracy of ~100 fs owing to unavoidable timing uncertainty (jitter) between the optical pump and X-ray probe pulses.

In contrast to standard practice[10], no time averaging or binning of snapshots was performed. Supplementary Fig. 1a shows the histogram of pump–probe delay times for light data. To ensure that all time points are equally weighted, 190,053 light snapshots were randomly removed, mostly from the highly populated regions at ~200 fs and ~900 fs (Supplementary Fig. 1b). This resulted in a dataset of 147,799 light snapshots for further analysis. The data collected without exposure to a pump pulse (the dark dataset) consist of 79,937 snapshots whose reflections were mapped to the same asymmetric unit as the light data.

PYP forms crystals with $P6_3$ symmetry[10]. Mapping all collected reflections (Miller indices; $k_{\max} = 36$, $l_{\max} = 26$) to the asymmetric unit gives 21,556 unique reflections. Reflections beyond the resolution window of $0.0667\ \text{Å}^{-1} < |q| < 0.667\ \text{Å}^{-1}$, as well as those that were their own twins, were removed, resulting in 15,498 unique reflections with an average of 500 reflections per 2D snapshot.

Data representation

In the representation used in this study, a data vector contains as many components as the number of reflections in a 3D volume. As only one central slice is accessed in each 2D snapshot, a data vector is highly incomplete. The data matrix consists of as many data vectors as 2D snapshots in the dataset. Specifically, the matrix has $D = 15{,}498$ rows and $N = 147{,}799$ columns, with each row corresponding to a unique reflection and each column corresponding to the preprocessed intensity data recorded in a 2D snapshot. The columns were ordered according to the experimentally measured (inaccurate) timestamps. With each snapshot containing only ~500 reflections, the data matrix is highly sparse (sparsity $= \frac{\sim 500}{15{,}498} = \sim 3.2\%$).

Applying the same preprocessing steps to the dark snapshots results in a data matrix of $D = 15{,}487$ rows and $N = 79{,}937$ columns. As no optical pump, and hence no pump–probe time delay, was involved in recording dark snapshots, these data were lexicographically sorted according to run numbers followed by event numbers.

Manifold-based machine learning

Preprocessed snapshots were analysed by time-lagged manifold embedding, nonlinear spectral analysis (NLSA)[19] and standard techniques for compiling DED maps[10] (Supplementary Information). Manifold embedding (Diffusion Map) was performed on supervectors with concatenation parameter $c = 32{,}768$, number of nearest neighbours $n_N = 15{,}000$ and a Gaussian kernel $\sigma = 1{,}420$. This embedding results in four noise-reduced orthogonal eigenfunctions. NLSA on the manifold reveals five modes above the noise plateau. As in standard singular value analysis, the first mode represents the moving average, with subsequent modes representing deviations from the average. Reconstruction using the modes above the noise plateau yields single frames containing full diffraction volumes at uniformly spaced time points. The same procedure was applied to the dark dataset with $N = 79{,}937$, $c = 32{,}768$, $n_N = 1{,}000$ and $\sigma = 3{,}380$. In this case, NLSA yielded only two identical modes differing only in scale, as expected from a single-parameter process (Extended Data Fig. 8).

To compute DED maps, we subtract the average reconstructed 3D diffraction volume of the dark data from the reconstructed 3D diffraction volume of the light data at each time point. The CCP4 package[38] was used to scale the light data to the dark data. The dark phases were obtained from the Protein Data Bank model of PYP deposited under accession code 5HD3 (ref. [10]). The Coot toolbox[39] was used to compile DED maps at 1.5-fs intervals and, from there, DED videos.

Validating the time resolution by comparison with spectroscopic results

The pump and probe pulses used in serial femtosecond crystallography are typically tens of femtoseconds long. In principle, simulation can be used to investigate the effect of the pulse characteristics on the achievable time resolution. However, extensive studies of the spatial and temporal characteristics of incident pulses[30] have shown that it is extremely difficult, if not impossible, to determine the actual pulse characteristics in time-resolved serial femtosecond crystallography. We therefore follow a data-driven machine learning approach, whereby the veracity of our results is determined by the extent to which our algorithm reproduces independently known ground truths. Using experimental XFEL pump–probe data obtained with optical pulses as long as 75 fs, we have previously shown that the vibrational frequencies revealed by our approach accurately match those of well-known systems such as N_2, even when the incident pulse lengths are long when compared with the vibrational frequencies[16]. Similarly, in the spectroscopically examined frequency range, each of the PYP vibrational frequencies revealed by our present work has been independently observed by time-resolved Raman techniques[32] (Extended Data Table 1). This clearly demonstrates that reliable dynamical information can be extracted with few-femtosecond time resolution.

Validation with synthetic data

We have previously demonstrated the efficacy of our approach by reference to synthetic models[16]. To confirm the validity of our approach for the present case, we generated a 3D diffraction volume whose structure and time evolution closely resembled the outcome of the analysis of the experimental data. The simulated and experimental data shared the same diffraction space sampling and level of data sparsity.

2D snapshots were generated by taking Ewald cuts (central slices) from a jitter-free, noise-reduced model. The same set of indexed Bragg reflections and timestamps were used in analysis of the experimental data. For instance, let h_j and t_j be the set of indexed Bragg reflections and the measured (and jittered) timestamp of the jth experimental snapshot, respectively. The jth simulated snapshot consists of the same set of indexed Bragg peaks h_j extracted from the full diffraction volume in the jitter-free, noise-reduced video extracted from the experimental data with timestamp t_j. In other words, on a snapshot-to-snapshot basis, the synthetic and experimental data have the same set of indexed Bragg peaks and the same nominal timestamps. Hence, the synthetic and experimental data have the same timing structure, the same diffraction space sampling and the same level of data incompleteness.

The resulting synthetic data, consisting of highly incomplete data vectors, were then passed through the same analytical pipeline as that used for experimental data, and the recovered structural dynamics were compared with the known ground truths pertaining to the input synthetic model. This comparison was quantified in terms of the Pearson correlation coefficient between the synthetic model and the output of the analytical pipeline. As shown in the Supplementary Information, the correlation coefficients typically exceed 0.99 (Supplementary Fig. 3a). In addition, the *R* factor between the synthetic input structural dynamics and the output of our data-analytical pipeline was used to validate the veracity of the outcome of our analysis as a function of spatial frequency. As shown in Supplementary Fig. 3b, the *R* factor is below 15%, even at the highest spatial frequency (1.6 Å).

Outcome of machine learning compared with results from conventional analysis

To validate our analytical pipeline further, we compare the DED at 3 ps obtained by our approach and the corresponding DED obtained by the standard methods[10,25]. The time point of 3 ps was chosen to be close to the femtosecond regime without being substantially impacted by timing uncertainty. As shown in Supplementary Fig. 4, the two DEDs are highly similar, with a correlation coefficient of 0.998 between the respective volumes and crystallographic *R* factors below 20% at the highest frequency. This validates the robustness of our algorithm to noise and data incompleteness.

Frequency content of the observed dynamical modes

The spectral features of the chronos were examined by Fourier and multi-taper analysis, the latter of which uses the Chronux package in MATLAB (http://chronux.org/). First, using mode 4 as an example, FFT analysis was performed with zero padding and a Hann window over the span of chrono 4 used in the NLSA reconstruction of mode 4. The Fourier spectrum shows clear peaks at frequencies exceeding 95 THz (shorter than 10.5 fs) (Extended Data Fig. 3). To verify the reliability of these FFT results, multi-taper *F*-test analysis was performed with padding parameter pad = 3 (default value) and time half-bandwidth products $tw = 2, 3, 4, 5$ (the number of tapers is $2 \times tw - 1$). The results for pad = 3 and $tw = 3$ are shown in Extended Data Fig. 3. The vertical axis represents, in essence, the signal-to-noise ratio of each Fourier component. Unless otherwise stated, we consider only peaks with *F*-ratio values above 5 in at least one set of taper parameters.

Each chrono displays a characteristic frequency spectrum, which sometimes includes a subset of the peaks observed in another chrono (Extended Data Table 1). The exact peak position can change by a few terahertz as multi-taper parameters are varied. Such closely separated peaks are grouped together as one peak at the average frequency position.

Comparison with spectroscopically determined PYP frequency spectrum

Using time-resolved Raman spectroscopy, Kuramochi et al.[32] have investigated the frequency range of ~3–30 THz in PYP. As shown in Extended Data Table 1, all frequency peaks in this frequency range revealed by multi-taper analysis of the chronos before the encounter with the conical intersection can be identified with a peak observed by Kuramochi et al. with a frequency accuracy of 7% or better. However, the time-resolved Raman spectra contain additional peaks (not shown) not observed in our analysis of the chronos. This suggests that not all spectroscopically observed frequencies pertain to the structure-dynamical collective variables we have extracted from time-resolved scattering data.

Overview of theoretical model used to simulate dynamical trajectories

A prerequisite for the theoretical description of an *N*-particle system is, in general, a set of $(3N - 6)$ generalized coordinates, also referred to as normal modes, which correspond to the system's internal degrees of freedom. However, to model the potential energy landscape in the vicinity close to the conical intersection in PYP, we take into consideration only two normal modes that contribute to formation of the conical intersection. On the one hand, this is motivated by identifying pairs of collective modes in the experimental data and at least two degrees of freedom being required for a conical intersection to occur. On the other hand, we want to keep the number of model parameters as small as possible to avoid overfitting. The two modes considered, in turn, are related to $3N - 6$ Cartesian real-space coordinates by a linear, yet unknown, transformation. Therefore, distinct structural modifications and the resultant effects on extrinsic quantities, for example, the electron density, are also expected to emerge in the two-mode description if these modifications are caused by the conical intersection.

The calculations use an effective model Hamiltonian in a compressed collective-mode space to capture the properties of a conical intersection and its vicinity. This allows numerically exact calculation of the nuclear quantum dynamics as a function of a small number of molecule-specific model parameters. By comparing the resulting simulated structure-dynamical trajectories with the experimentally determined trajectories provided by our data-analytical pipeline, we determine the numerical values that the model parameters assume in PYP. This provides detailed insight into the topography of this photochemically exemplary conical intersection (see next section).

Our theoretical model is tantamount to the simplest possible realization of a conical intersection, implicitly assuming dissipation into additional modes is negligible on the timescale of 100 fs. The PESs are approximated by a second-order Taylor expansion, known as the vibronic coupling model[40], and are assumed to be symmetric with respect to both normal modes, possibly differing in their respective vibrational frequencies. Inter-state coupling is mediated via one mode only.

To determine the dynamics in the space spanned by two modes, we numerically solve the time-dependent Schrödinger equation for a wave packet initially occupying the excited electronic state. We take into account a total of six model parameters, including the reference ground-state frequency related to the kinetic energy of the wave packet, the respective vibrational frequencies of both PESs, the coupling strength and the initial position of the wave packet relative to the conical intersection.

We describe the conical intersection and the potential energy landscape in the close vicinity of the conical intersection by a Taylor expansion of the diabatic PESs up to second order[40,41]. This model is well established for the description of molecular dynamics, and has been applied to the theoretical description of photo-excitation and photo-dissociation processes in molecules (see, for example, refs. [42,43]). Recently, a two-mode vibronic coupling model was used to study electronic coherences in the presence of a conical intersection[44].

The normal modes denoted x and y are considered in dimensionless coordinates normalized with respect to a reference vibrational ground-state frequency ω. The corresponding Hamiltonian, including two electronic states, reads $H = T + W$, with the kinetic energy operator $T = -\frac{\omega}{2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)$ and the diabatic potential energy matrix of the general form

$$W = \begin{pmatrix} V_1(x,y) & W_{12}(y) \\ W_{12}^*(y) & V_2(x,y) + \Delta E \end{pmatrix}$$

Here $V_{1,2}(x,y)$ and $W_{12}(y)$, respectively, denote the diabatic PESs and the diabatic inter-state coupling between the two electronic states. By construction, the coupling depends solely on mode y, referred to as coupling mode, while x is known as the tuning mode. ΔE describes a relative energy offset at the origin of the coordinate system, $(x,y) = (0,0)$.

The choice of the normal modes x and y is, to some extent, arbitrary as they are determined by the structure of the Hamiltonian. The inter-state coupling, for instance, is mediated only via the coupling mode y, by definition. In general, the normal modes of the vibronic coupling model do not coincide with the normal modes of motion that are assessed by experiments. However, the configuration spaces are connected via an orthogonal transformation. In principle, one would also have to individually rescale the modes to go back from dimensionless coordinates to physical units. However, in our model, we assume equal scaling of the modes, that is, only one normalization frequency ω is used instead of $\omega_{x,y}$. This approach is used to reduce the total number of parameters. To transform from the normal modes of the model to the normal modes of motion, a possible 90° rotation angle (x, y swap) was used as a globally applied fitting parameter.

The PESs in adiabatic representation are given by the eigenvalues of the diabatic potential energy matrix W, that is:

$$V_{1,2}^{\mathrm{ad}} = \frac{V_1 + V_2 + \Delta E}{2} \pm \frac{1}{2}\sqrt{(V_1 - V_2 - \Delta E)^2 + 4|W_{12}|^2}$$

Following from the defining degeneracy of the adiabatic PESs, that is, $\Delta E = V_1 - V_2$, the conditions $W_{12} = 0$ and $\mathrm{V}_1^{\mathrm{ad}} - \mathrm{V}_2^{\mathrm{ad}} = 0$ must be fulfilled for a conical intersection to occur. Because these conditions are solved independently, a conical intersection constitutes a $(d-2)$-dimensional manifold (with d equal to the total number of degrees of freedom), requiring at least two normal modes for modelling a conical intersection.

In our description of a conical intersection, we neglect intra-state coupling of the modes to further reduce the total number of parameters, and assume harmonic diabatic potentials of the general form

$$V_{1,2}(x,y) = \frac{1}{2}\left(\gamma_{1,2}^{(x)}x^2 + \gamma_{1,2}^{(y)}y^2\right) + \kappa_{1,2}^{(x)}x + \kappa_{1,2}^{(y)}y$$

with potential minima at

$$\left(x_{1,2}^{(0)}, y_{1,2}^{(0)}\right) = \left(\frac{-\kappa_{1,2}^{(x)}}{\gamma_{1,2}^{(x)}}, \frac{-\kappa_{1,2}^{(y)}}{\gamma_{1,2}^{(y)}}\right)$$

whereas the inter-state coupling between the diabatic electronic states is assumed to be a linear function, that is, $W_{12}(y) = \lambda y$.

For simplicity, we assume the diabatic potentials to be symmetric with respect to both modes, that is, $\gamma_{1,2}^{(x,y)} = \gamma_{1,2} > 0$. The coefficients $\kappa_{1,2}^{(x,y)}$ are assumed to depend on the respective vibrational frequencies $\gamma_{1,2}$, and the coupling strength λ is regarded as an independent parameter.

Without loss of generality, we fix the conical intersection coordinates at $(x_{\mathrm{CI}}, y_{\mathrm{CI}}) = (0, 0)$ and, instead, vary the initial position of the wave packet relative to the conical intersection. Therefore, the energy offset is $\Delta E = 0$. To determine the remaining four parameters $\kappa_{1,2}^{(x,y)}$, we make the following assumptions.

1. For $\gamma_1 = \gamma_2$, the adiabatic potentials are symmetric with respect to x and y:

$$V_{1,2}^{\mathrm{ad}}(-x,y) = V_{1,2}^{\mathrm{ad}}(x,y) \wedge V_{1,2}^{\mathrm{ad}}(x,-y) = V_{1,2}^{\mathrm{ad}}(x,y)$$

2. The positions of the diabatic potential minima are fixed and independent of the vibrational frequencies $\gamma_{1,2}$.
   Incorporating these assumptions, we obtain

$$\kappa_1^{(x)} = \gamma_1 x_0$$

$$\kappa_2^{(x)} = -\gamma_2 x_0$$

$$\kappa_1^{(y)} = \kappa_2^{(y)} = 0$$

with $\pm x_0$ denoting the x coordinate of the diabatic potential minima. The final diabatic potentials read

$$V_1(x,y) = \frac{1}{2}\gamma_1(x^2 + y^2) + \gamma_1 x_0 x$$

$$V_2(x,y) = \frac{1}{2}\gamma_2(x^2 + y^2) - \gamma_2 x_0 x$$

We chose $x_0 = 10$ for all simulations. Although the diabatic potentials share the y coordinate of the minima ($y = 0$), the coupling parameter λ causes the adiabatic potential minima to be symmetrically shifted along the y axis.

The initial nuclear wave packet is chosen to be a Gaussian with an initial width $\langle\Delta x\rangle = \langle\Delta y\rangle = 1$ (with $\langle\Delta x\rangle^2 = \langle x^2\rangle - \langle x\rangle^2$), given by

$$\chi_0(x,y) = \frac{1}{\sqrt{\pi}}\exp\left(-\frac{(x - r_0\cos\theta_0)^2 + (y - r_0\sin\theta_0)^2}{2}\right)$$

Here r_0 and θ_0, respectively, denote the radius and polar angle of the initial wave packet position. Because the adiabatic potentials are symmetric with respect to y, that is, $V_{1,2}^{\mathrm{ad}}(x,-y) = V_{1,2}^{\mathrm{ad}}(x,y)$, we restrict the initial positions to the lower semicircle. In all simulations, the wave packet is placed on the second diabatic potential V_2. However, we take into consideration only those situations where the initial diabatic population of V_2 corresponds to a population of the excited adiabatic state of at least 50%.

Our model is based on the six parameters listed in Extended Data Table 2 to generate a library of 500,000 different configurations. The simulations were carried out using the multi-configuration time-dependent Hartree method[45–47] in its multiset implementation of the Heidelberg package, which allows efficient wave packet propagation. The wave function is represented on a grid of Hermite functions with 175 grid points between −35.0 and +35.0. We use 25 single-particle functions per degree of freedom and electronic state, such that the natural population of the highest single-particle function is below 10^{-4}. The integration is carried out with a variable mean field scheme with an accuracy of 10^{-8}.

Identifying each experimental trajectory with a simulated counterpart

The purpose of identifying simulated trajectories with experimental ones is to determine the set of parameters best able to describe each of the experimental trajectories. Identification proceeds by comparing simulated dynamical trajectories with experimental ones. In principle, one could simply select the simulated trajectory most closely resembling (having the smallest χ^2 value with respect to) a given experimental trajectory. In practice, the axes describing the experimental and simulated trajectories may be rotated and/or rigidly shifted with respect to each other. Attenuation due to noise and timing jitter may also change the 'unit length' of each experimental axis by an unknown amount. The best-fit search must therefore allow rigid shifts in the origin, axis swap and scaling of the simulated trajectories. (Allowing frame rotations results in axis swaps.)

More specifically, we model the *trans*-to-*cis* isomerization of PYP as a wave packet de-exciting via a conical intersection. To extract model parameters, we numerically solve the time-dependent Schrödinger equation for this system, taking into account a total of six model parameters, as described above. Dynamical trajectories in the 2D space spanned by two collective modes are then compared with pairs of chronos recovered in our analytical pipeline.

In greater detail, we fit each of the simulated trajectories to the six experimental trajectories, using the smallest χ^2 to identify the best match in each case (see, for example, Supplementary Fig. 5). In this, we consider both temporal and spatial translations of the simulated trajectories, and allow the simulated trajectories to be linearly scaled. Following this procedure, we are able to reproduce the observed collective-mode behaviour, and determine the geometric properties of the conical intersection and the uncertainties in our determination of these physically important parameters (see Extended Data Table 2 and Methods section 'Extracting parametric values and uncertainties'). pecifically, the parameter ω characterizes the kinetic energy, and (r_0, θ_0) characterizes the initial position of the wave packet. The shape of the PESs is determined mainly by the vibrational frequencies γ_1 and γ_2. The coupling parameter λ defines the probability of a transition between the two electronic states as a result of the wave packet moving through or close to the conical intersection. Vibrational frequencies and coupling strength together define the topography of the conical intersection.

Because of timing jitter, noise and data incompleteness, the experimentally determined dynamical modes may be differently damped, and not appear in the correct order[16]. Also, any linear combination of the experimentally determined dynamical modes is, in principle, a valid dynamical mode. To correctly compare the simulated and experimental trajectories in two dimensions, we allow each coordinate axis to be shifted and stretched, with the two coordinate axes interchanged.

With four continuous degrees of freedom (two for scaling, two for translation and rotation proving unnecessary), the mutual identification of experimental and simulated trajectories is achieved using a linear least-squares fit in the space of experimental data with the following cost function:

$$\chi^2 = \sum_{i=1}^{N} [(\tilde{x}_i - (a_1 + a_2 x_i))^2 + (\tilde{y}_i - (b_1 + b_2 y_i))^2]$$

Here, with N time points per trajectory, (x_i, y_i) and $(\tilde{x}_i, \tilde{y}_i)$ denote the simulated and experimental trajectories at the ith time point, respectively, with (a_1, b_1) and (a_2, b_2) as the fitting parameters. In our analysis, 1,001 time points span a total interval of 100 fs. Fits performed in the space of simulated data were prone to instability. In cases where stable solutions could be found, the extracted parameters agree with those presented in Extended Data Table 2.

Segment search in experimental trajectories

The starting time points for simulated and experimental trajectories are not necessarily identical. The simulated trajectories (each 100 fs long) are, therefore, compared with different 100-fs spans of the experimental trajectories to find the best match. Using t_c to denote the centre of the 100-fs span, we repeated the χ^2 analysis above for 66 values of t_c: 590 fs $\leq t_c \leq$ 655 fs in increments of 1 fs. This time range covers the region where *trans*-to-*cis* isomerization is expected to occur[10], and includes the sharp turning points at 615 fs in the chronos considered in this analysis.

Extracting parametric values and uncertainties

With the fitting procedure described above, experimental trajectories can be identified with their simulated counterparts, from which the key physical parametric values pertaining to each experimental trajectory can be determined. In effect, the simulated trajectories constitute a bank of trajectories calculated for a range of physical parameters characterizing the dynamical trajectories in the vicinity of a conical intersection. From 66 possible segments per experimental trajectory representing a shift of up to 65 fs in time, 500,000 simulated trajectories and swaps between x and y axes, a total of 66 million χ^2 fits per experimental trajectory, were performed, and the physical parameters were extracted. The parameter values obtained from the best fits are summarized in Extended Data Table 2. The uncertainties in the extracted parameters were established by calculating the root mean square (r.m.s.) difference between the best-fit parameters and the parameters corresponding to all the simulated trajectories with $\chi^2 \leq 1.2$ relative to the best fit. For cases in which the r.m.s. difference is zero, the uncertainty is assumed to correspond to the parametric sampling interval used to generate the simulated databank.

Computational resources for manifold-based machine learning

Software. The NLSA pipeline and χ^2 fittings were implemented using MATLAB (R2015b and R2019a, with parallel computing toolbox). DED calculations were performed using CCP4 v7.0, and Coot 0.8.9 was used to provide visualization of DEDs.

Hardware. Parallel computations were performed on an Intel CPU cluster (320 CPU cores, 2.6 GHz, arranged as 16 nodes, each with 128 GB of RAM). All other analyses were performed on a single Linux computer with 24 cores, a 3-GHz Intel Xeon CPU and 256 GB of RAM.

Computational resources for theoretical calculations. The wave packet propagations were carried out on 12-core Intel Xeon X5660 CPUs with 2.8 GHz and 96 GB of memory and 32-core Intel Xeon E5-2630L CPUs with 1.8 GHz and 126 GB of memory.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The structures have been deposited in the Protein Data Bank, together with their respective weighted difference structure factor amplitudes, under accession codes 5HD3, 5HDC, 5HDD, 5HDS and 5HD5. Source data are provided with this paper.

Code availability

The code will be made available on request.

37. Schwander, P., Giannakis, D., Yoon, C. H. & Ourmazd, A. The symmetries of image formation by scattering. II. Applications. *Opt. Express* **20**, 12827–12849 (2012).
38. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
39. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486-501 (2010).
40. Köppel, H., Domcke, W. & Cederbaum, L. S. in *Advances in Chemical Physics* (eds Rice, S. et al.) 59–246 (2007).
41. Domcke, W., Yarkony, D. R. & Köppel, H. *Conical Intersections* (WorldScientific, 2004).
42. Gromov, E. V. et al. Theoretical study of excitations in furan: spectra and molecular dynamics. *J. Chem. Phys.* **121**, 4585–4598 (2004).
43. Faraji, S., Meyer, H. D. & Koppel, H. Multistate vibronic interactions in difluorobenzene radical cations. II. Quantum dynamical simulations. *J. Chem. Phys.* **129**, 074311 (2008).
44. Arnold, C., Vendrell, O., Welsch, R. & Santra, R. Control of nuclear dynamics through conical intersections and electronic coherences. *Phys. Rev. Lett.* **120**, 123001 (2018).
45. The MCTDH package v.8.4.18 (2019).
46. Beck, M. The multiconfiguration time-dependent Hartree (MCTDH) method: a highly efficient algorithm for propagating wavepackets. *Phys. Rep.* **324**, 1–105 (2000).
47. Meyer, H. D., Manthe, U. & Cederbaum, L. S. The multi-configurational time-dependent Hartree approach. *Chem. Phys. Lett.* **165**, 73–78 (1990).

Acknowledgements We acknowledge valuable discussions with E. Lattman, K. Moffat, T. Martinez and past and present members of the UWM data science group. The development of underlying techniques was supported by the US Department of Energy, Office of Science, Basic Energy Sciences, under award DE-SC0002164 (underlying dynamical techniques) and by the US National Science Foundation under awards STC-1231306 (underlying data analytical techniques) and DBI-2029533 (underlying analytical models). N.B. and R. Santra were supported by the Cluster of Excellence 'CUI: Advanced Imaging of Matter' of Deutsche Forschungsgemeinschaft (DFG)–EXC 2056–project no. 390715994.

Author contributions A.O. designed this study. A.O., A.H., P.S., R.F. and R. Sepehr co-defined the algorithm and the data-analytical pipeline. A.H. and R.F. performed analytical and computational work, and tested and validated the algorithm with participation by A.O. N.B. and R. Santra performed the quantum-dynamics simulations. M.S. provided experimental data and expertise in crystallographic data analysis. All authors contributed to the manuscript.

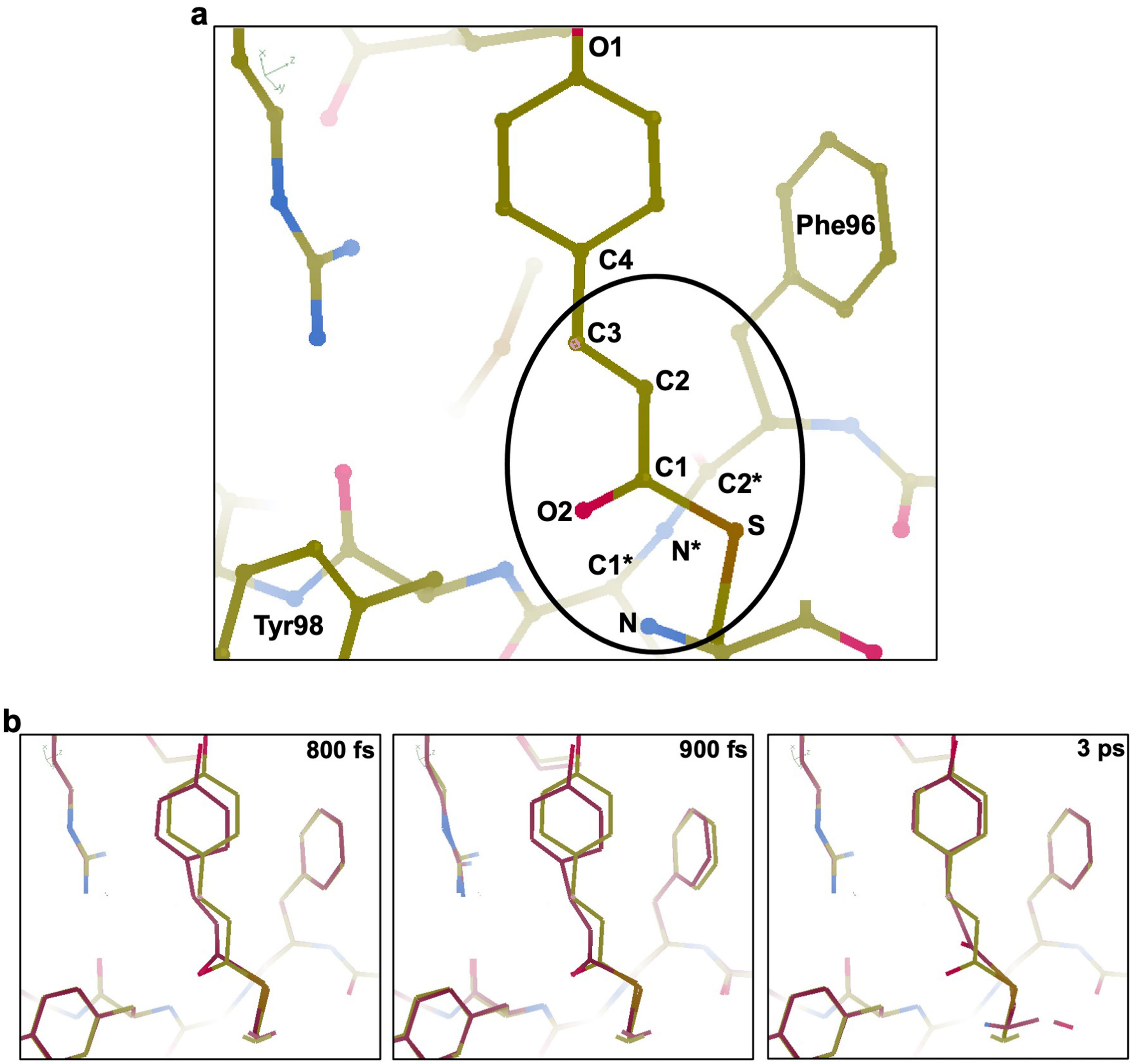
Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04050-9.

Correspondence and requests for materials should be addressed to A. Ourmazd.

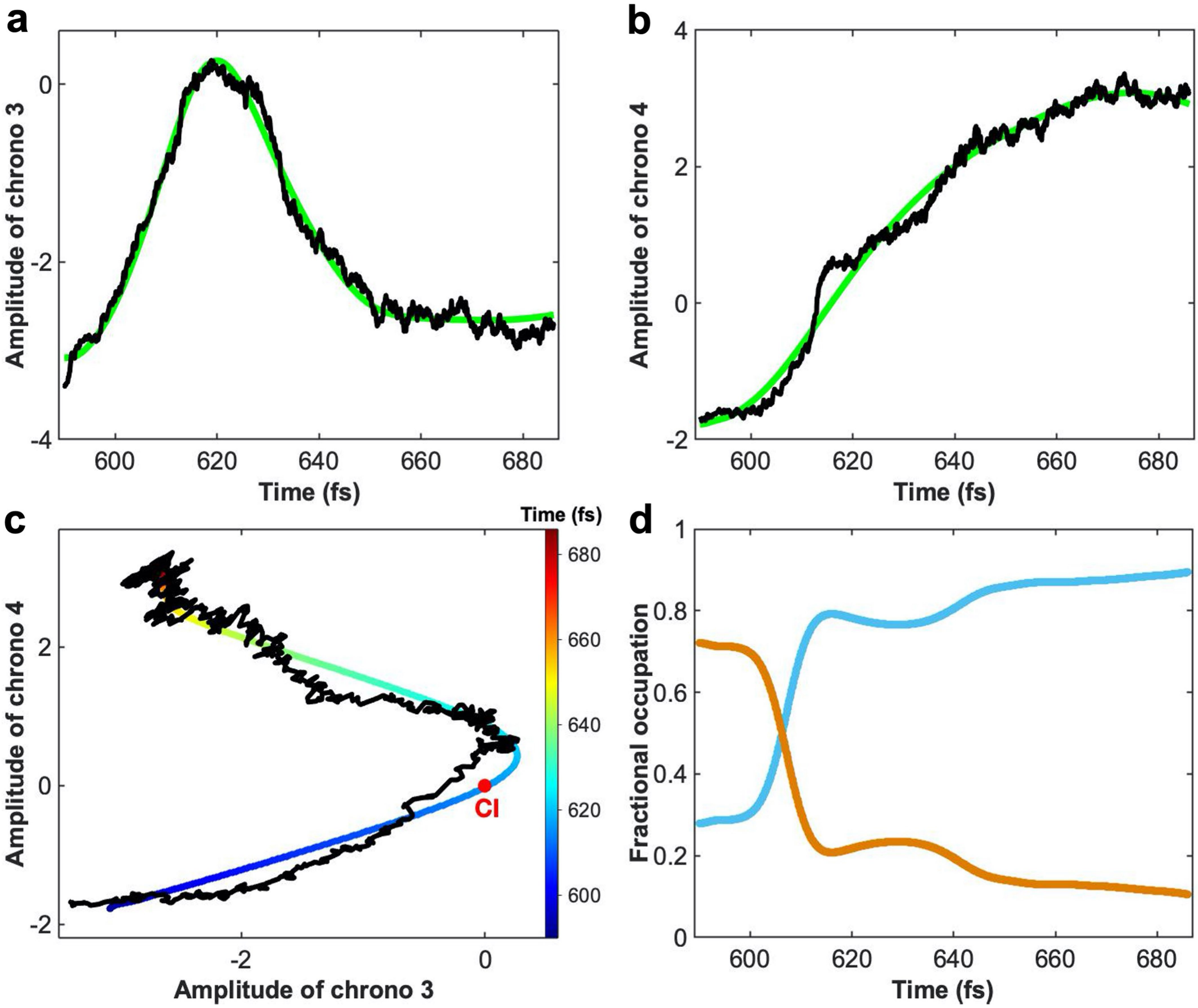
Peer review information *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



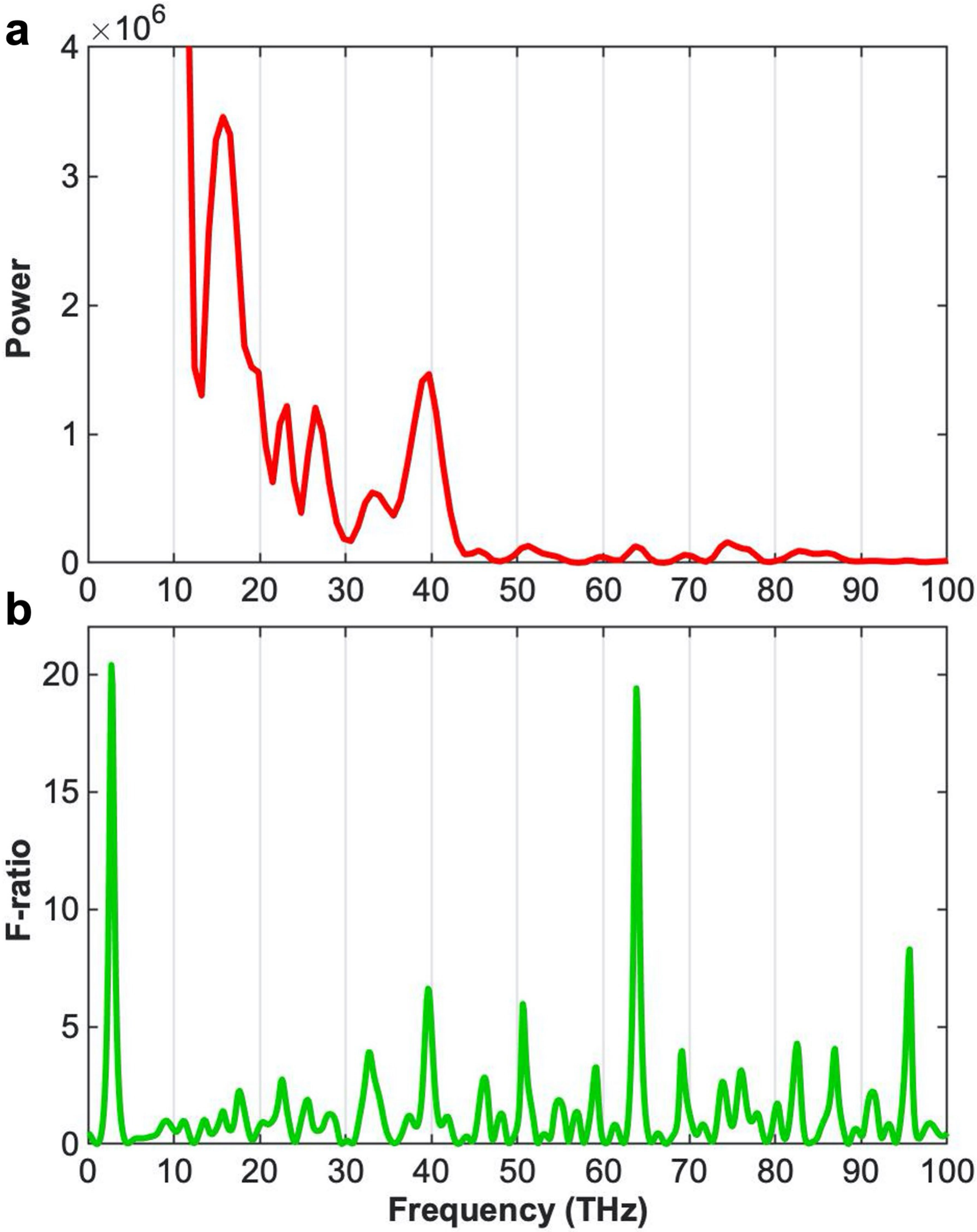
c

Delay time (fs)	0	285	800	900	3000
Torsion angle	172°	156°	93°	67°	44°

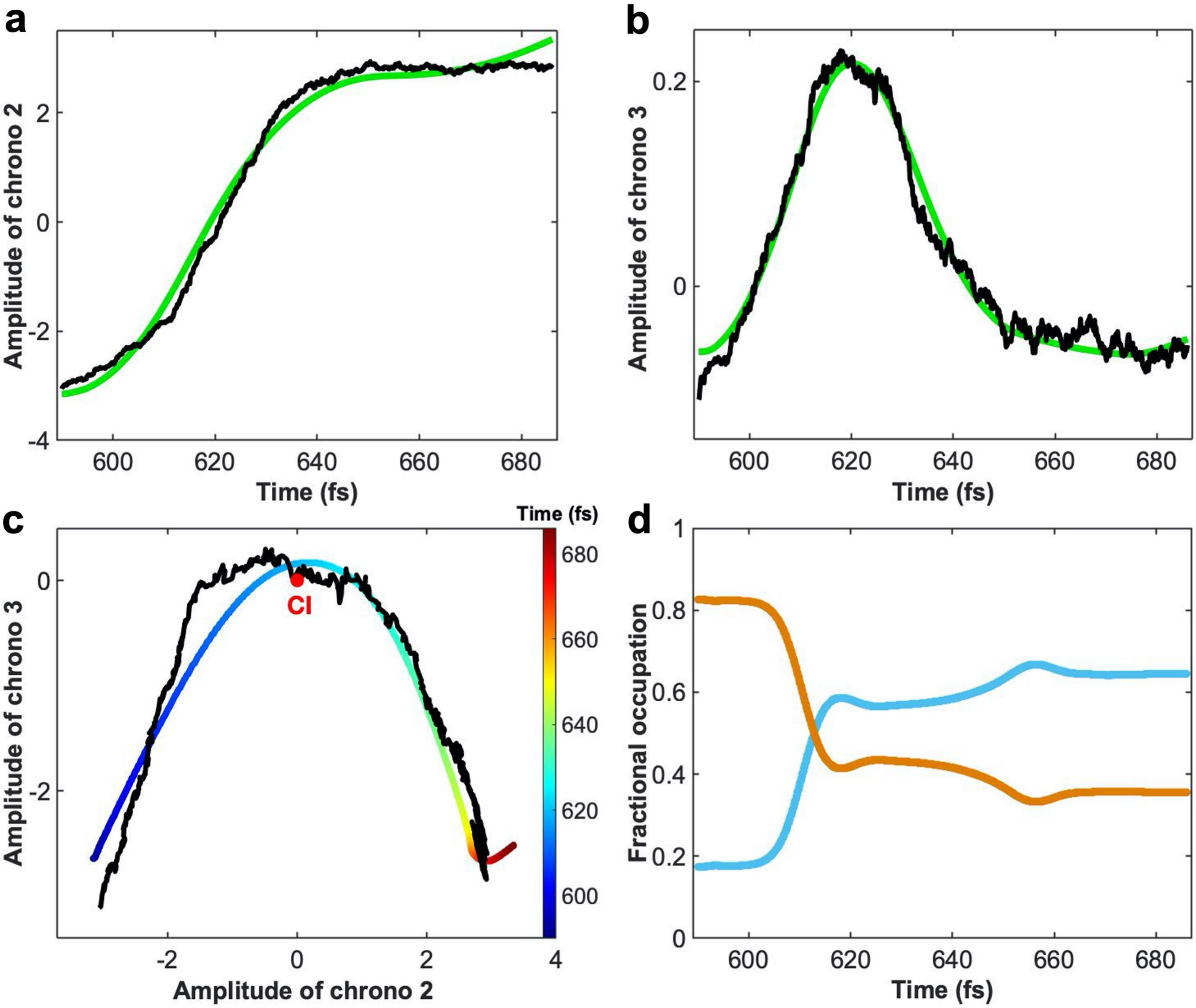
Extended Data Fig. 1 | PYP chromophore in trans configuration and structure dynamical modes obtained by our approach. a, The PYP chromophore in the trans configuration. The oval contains the primary structure-dynamically active region, with the numbered atoms and aromatic structures identified. C: carbon, N: nitrogen, O: oxygen, S: sulfur. **b, c**, The structure dynamical modes obtained by our approach can be combined to yield the more intuitive torsional angle, which is commonly chosen as the primary reaction coordinate for isomerization in PYP. Changes in the torsional angle and the bend of the chromophore axis relative to equilibrium values necessarily increase the energy of the ground state structure. Near the CI the structure on the ground state PES and that on the excited state PES are essentially identical with very similar energies. The structure on the excited state PES determined at 615 fs is therefore an excellent model for the electronic ground state structure near the PYP conical intersection.



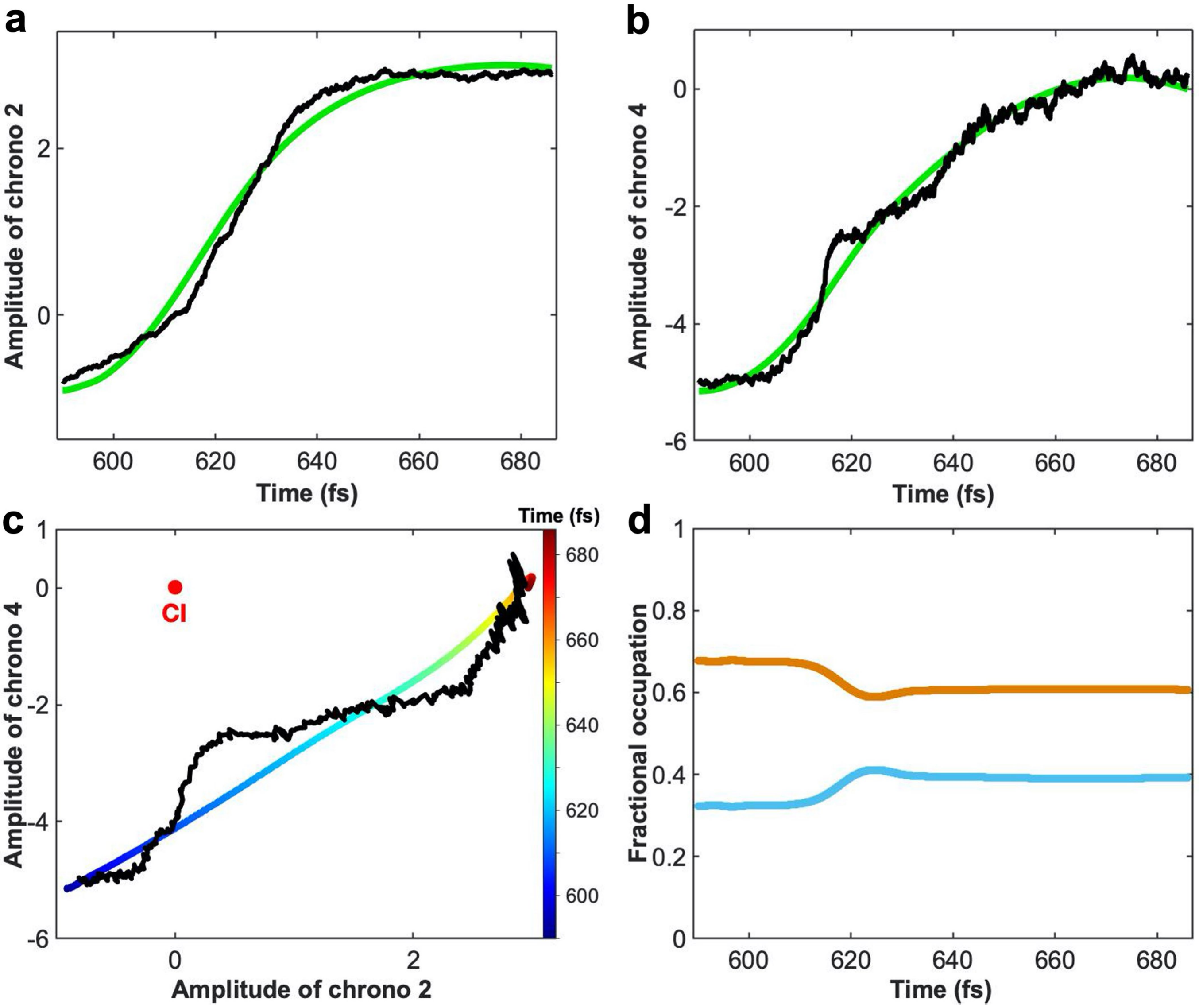
Extended Data Fig. 2 | Dynamical trajectories near the conical intersection. Unless otherwise stated, arbitrary units. **a, b**, Time evolutions (chronos) of modes 3 and 4, respectively. **c**, The experimental dynamical trajectory (in black) obtained from modes 3 and 4 as collective variables *x* and *y*, respectively, and the best-fit simulated trajectory, with color showing the passage of time (see color bar). The red dot indicates the position of the conical intersection. For additional trajectories, see Supplementary Information. **d**, The calculated de-excitation dynamics as reflected in the electronic state population for the trajectory shown in Panel c above. The brown and blue curves represent the populations of the upper and the lower adiabatic electronic states, respectively.



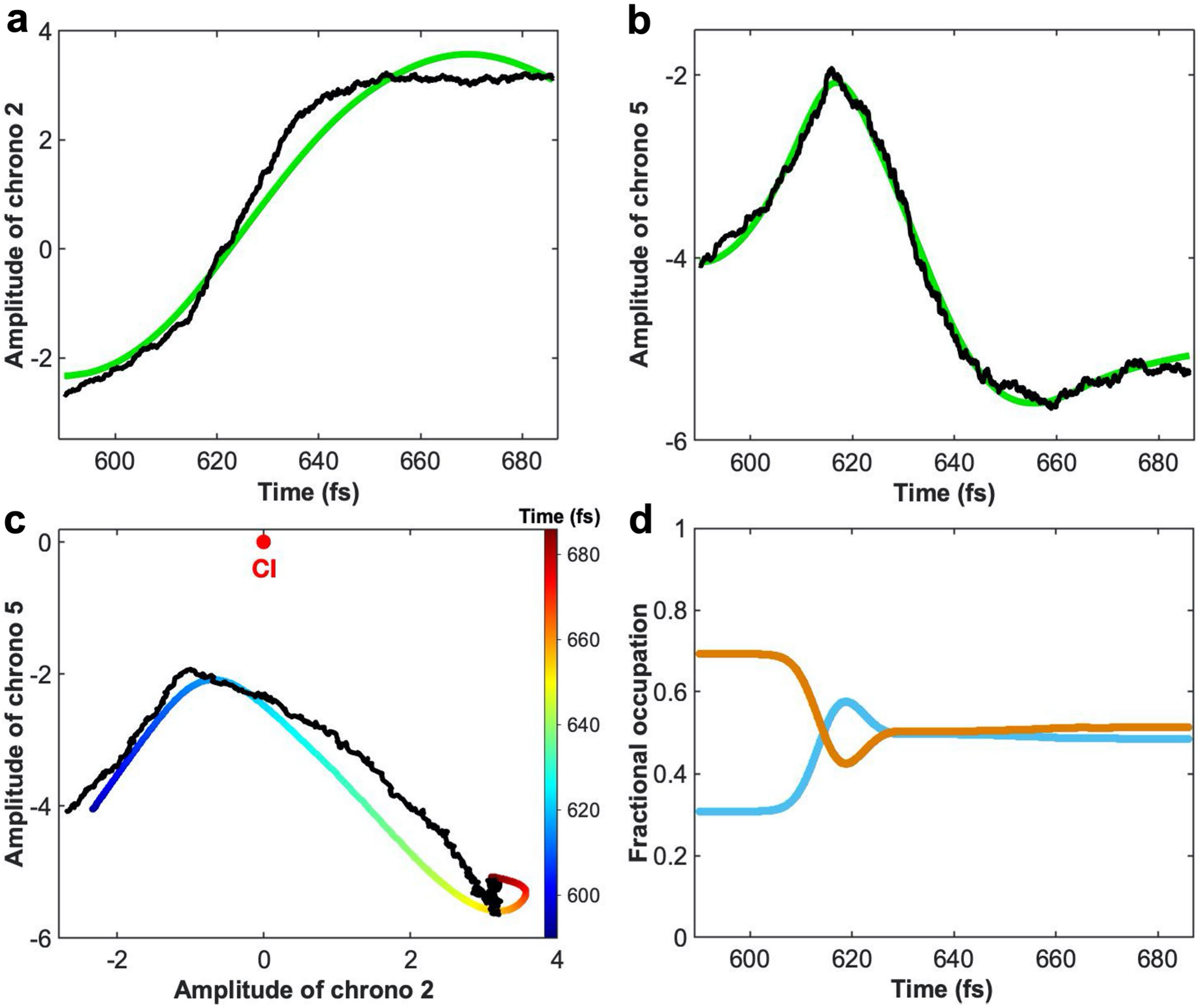
Extended Data Fig. 3 | Frequency content of a typical chrono, in this case chrono-4. a, Fourier power spectrum. **b**, Multi-taper analysis. The vertical axis of the latter essentially represents the signal-to-noise ratio. Each chrono displays a characteristic frequency spectrum.



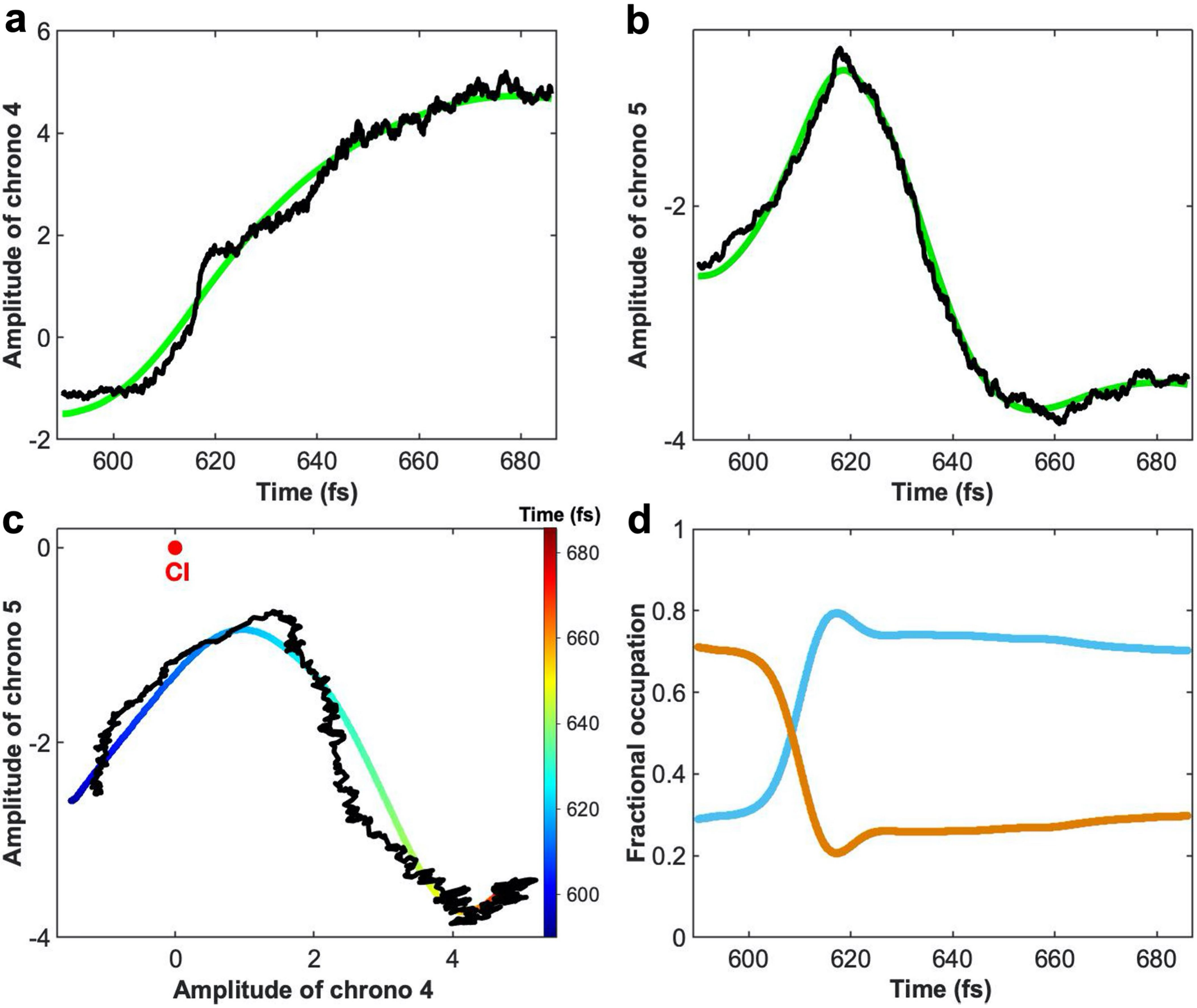
Extended Data Fig. 4 | Dynamical trajectories near the conical intersection. Unless otherwise stated, arbitrary units. **a, b**, Time evolutions (chronos) of modes 2 and 3, respectively. **c**, The experimental dynamical trajectory (in black) obtained from modes 2 and 3 as collective variables *x* and *y*, respectively, and the best-fit simulated trajectory, with color showing the passage of time (see color bar). The red dot indicates the position of the conical intersection. **d**, The calculated de-excitation dynamics as reflected in the electronic state population for the trajectory shown in Panel c above. The brown and blue curves represent the populations of the upper and the lower adiabatic electronic states, respectively.



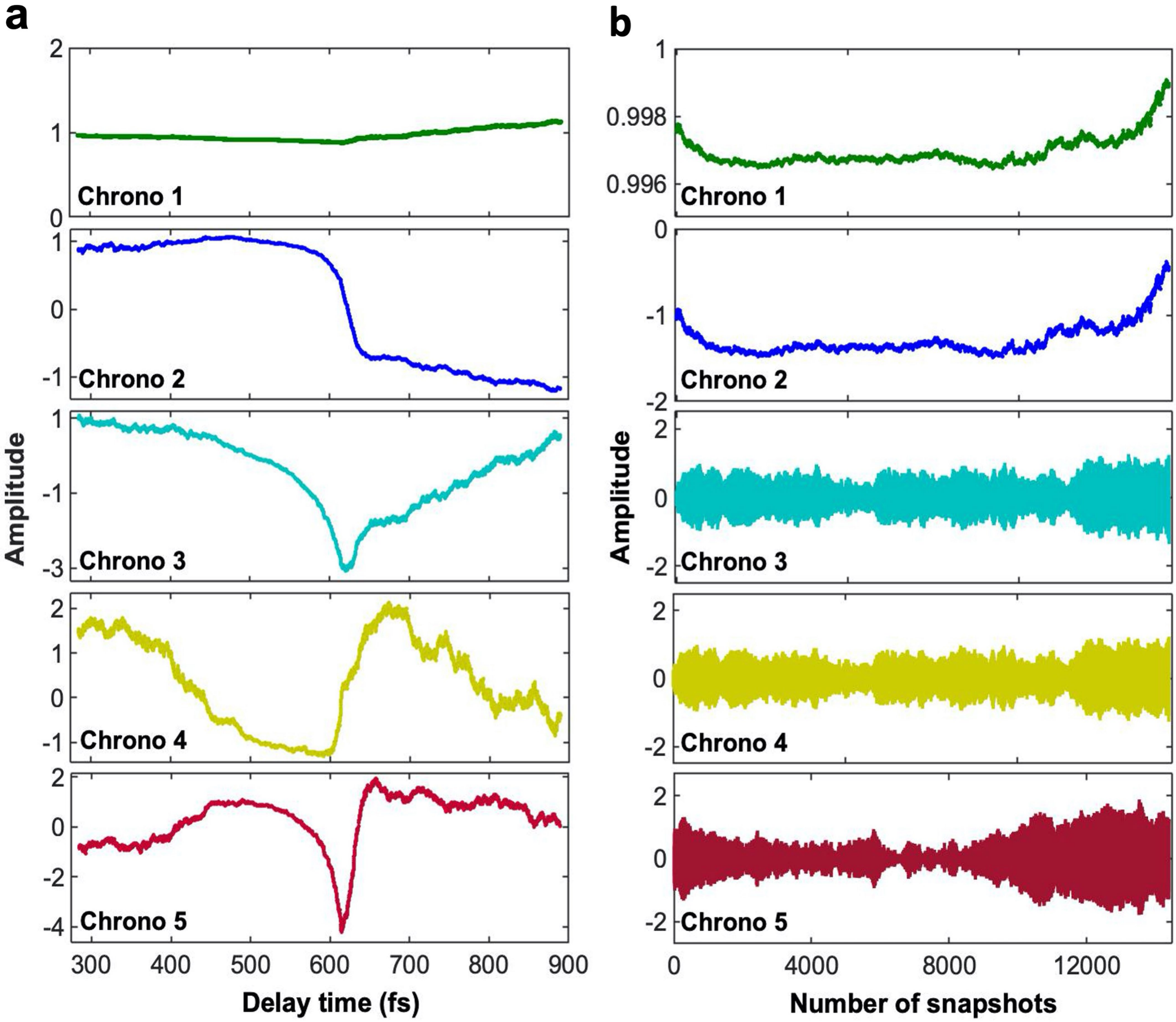
Extended Data Fig. 5 | Dynamical trajectories near the conical intersection. Unless otherwise stated, arbitrary units. **a, b**, Time evolutions (chronos) of modes 2 and 4, respectively. **c**, The experimental dynamical trajectory (in black) obtained from modes 2 and 4 as collective variables *x* and *y*, respectively, and the best-fit simulated trajectory, with color showing the passage of time (see color bar). The red dot indicates the position of the conical intersection. **d**, The calculated de-excitation dynamics as reflected in the electronic state population for the trajectory shown in Panel c. The brown and blue curves represent the populations of the upper and the lower adiabatic electronic states, respectively.



Extended Data Fig. 6 | Dynamical trajectories near the conical intersection. Unless otherwise stated, arbitrary units. **a, b**, Time evolutions (chronos) of modes 2 and 5, respectively. **c**, The experimental dynamical trajectory (in black) obtained from modes 2 and 5 as collective variables *x* and *y*, respectively, and the best-fit simulated trajectory, with color showing the passage of time (see color bar). The red dot indicates the position of the conical intersection. **d**, The calculated de-excitation dynamics as reflected in the electronic state population for the trajectory shown in Panel c above. The brown and blue curves represent the populations of the upper and the lower adiabatic electronic states, respectively.

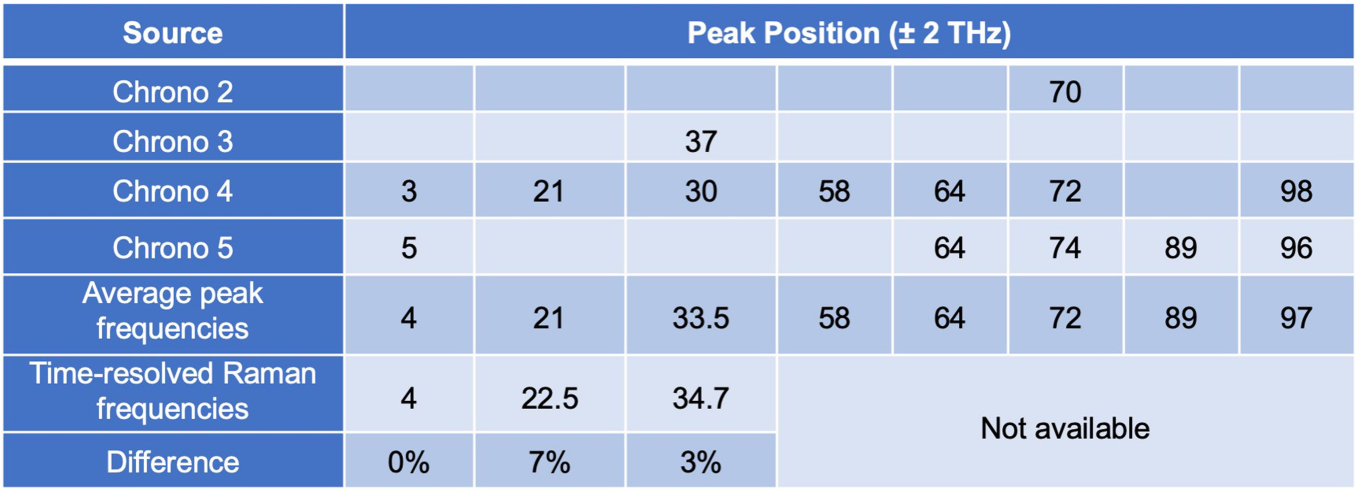


Extended Data Fig. 7 | Dynamical trajectories near the conical intersection. Unless otherwise stated, arbitrary units. **a, b**, Time evolutions (chronos) of modes 4 and 5, respectively. **c**, The experimental dynamical trajectory (in black) obtained from modes 4 and 5 as collective variables *x* and *y*, respectively, and the best-fit simulated trajectory, with color showing the passage of time (see color bar). The red dot indicates the position of the conical intersection. **d**, The calculated de-excitation dynamics as reflected in the electronic state population for the trajectory shown in Panel c above. The brown and blue curves represent the populations of the upper and the lower adiabatic electronic states, respectively.



Extended Data Fig. 8 | Comparing modes from light and dark data. a, The first five chronos obtained from light data ordered according to pump-probe delay. **b**, The first five chronos obtained from dark data lexicographically sorted according to run numbers followed by event numbers. The first two chronos are identical, except for scale. This is a hallmark of a one-parameter process. Correlation analysis shows the single-parameter process correlates with the integrated Bragg spot intensity (Pearson correlation: 0.93), most likely pertaining to drift in the incident beam intensity. The subsequent chronos represent noise.

Extended Data Table 1 | Peak positions obtained from multi-taper Fourier analysis of the chronos before the encounter with the conical intersection vs. peak positions obtained from time-resolved Raman spectra of PYP, all in THz



Source	Peak Position (± 2 THz)							
Chrono 2						70		
Chrono 3			37					
Chrono 4	3	21	30	58	64	72		98
Chrono 5	5				64	74	89	96
Average peak frequencies	4	21	33.5	58	64	72	89	97
Time-resolved Raman frequencies	4	22.5	34.7	Not available				
Difference	0%	7%	3%					

Each chrono displays a characteristic frequency spectrum, which sometimes includes a subset of the peaks observed in another chrono. The exact peak position can change by ~ 2 THz as multi-taper parameters are varied. Closely separated peaks are identified as one peak at the average frequency position. Using time-resolved Raman spectroscopy, Kuramochi et al.[33] have investigated the approximately 3 – 30 THz frequency range in PYP. As shown in the Table, all frequency peaks in this frequency range revealed by multi-taper analysis of the chronos before the encounter with the conical intersection can be identified with a peak observed by Kuramochi et al. with a frequency accuracy of 7% or better. However, the time-resolved Raman spectra contain additional peaks (not shown) not observed in our analysis of the chronos. This suggests that not all spectroscopically observed frequencies pertain to the structure dynamical collective variables we have extracted from time-resolved scattering data.

Extended Data Table 2 | Parameters of the potential energy surface and parametric grid of simulated trajectories near the conical intersection

a

Trajectory Parameter	M2-M3	M2-M4	M2-M5	M3-M4	M4-M5
$\boldsymbol{r_0}$	4.11 ± 2.12	5.22 ± 1.40	4.67 ± 1.24	3.56 ± 0.98	3.00 ± 1.67
$\boldsymbol{\theta_0}$	-140 ± 18	-100 ± 26	$-120 \pm 20^{(Q)}$	-120 ± 6.70	$-120 \pm 20^{(Q)}$
$\boldsymbol{\omega}\ (\times\mathbf{10^{-3}})$	2.50 ± 0.93	1.67 ± 1.06	$2.08 \pm 0.42^{(Q)}$	3.33 ± 0.28	$2.08 \pm 0.42^{(Q)}$
$\boldsymbol{\lambda/\omega}$	1.50 ± 0.27	1.00 ± 0.33	$1.50 \pm 0.50^{(Q)}$	$1.00 \pm 0.50^{(Q)}$	$1.50 \pm 0.50^{(Q)}$
$\boldsymbol{\gamma_1/\omega}$	0.13 ± 0.03	0.08 ± 0.04	0.08 ± 0.01	0.10 ± 0.02	0.07 ± 0.02
$\boldsymbol{\gamma_2/\omega}$	0.10 ± 0.02	0.20 ± 0.10	0.12 ± 0.02	0.08 ± 0.04	0.15 ± 0.03

b

Parameter	$\boldsymbol{\omega}$	$\boldsymbol{\gamma_1}$ / $\boldsymbol{\omega}$	$\boldsymbol{\gamma_2}$ / $\boldsymbol{\omega}$	$\boldsymbol{\lambda}$ / $\boldsymbol{\omega}$	$\boldsymbol{\theta_0}$	$\boldsymbol{r_0}$
Interval	$[0.00125, 0.005]$	$[0.05, 0.2]$	$[0.05, 0.2]$	$[0.5, 5]$	$[-180, 0]$	$[3, 8]$

a, Parameters of the potential energy surface near the conical intersection derived in this work. Radius r_0 (dimensionless), and angle θ_0 (degrees) together specify the initial position of wave packet. λ: coupling strength; ω: normalization frequency (kinetic energy); $\gamma_{1,2}$: frequency on respective PES. Superscript Q indicates the accuracy is limited by the spacing in the parametric grid of simulated data. Unless stated otherwise, where appropriate, parameters in atomic units. **b**, The parametric grid used to simulate dynamical trajectories. ω: normalization frequency; λ: coupling strength; $\gamma_{1,2}$: frequency on respective PES; angle θ_0 (in degrees) and radius r_0 (dimensionless) are the initial position of the wave packet. Except for θ_0 with six different values, parameters take 10 values uniformly distributed in their respective intervals. Unless otherwise specified, all parameters in atomic units.

nature portfolio

Corresponding author(s): Abbas OURMAZD

Last updated by author(s): Jul 27, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☒ □ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☒ □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ □ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ □ A description of all covariates tested

☒ □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☒ □ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ □ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Standard data collection software available at the LCLS was used.
Data analysis	Standard data-analytical software used in time-resolved crystallography was used including CrystFEL (0.9.0), CCP4 (v7.0), and Coot (0.8.9). This was augmented by manifold-based machine learning software written in MATLAB (R2015b, R2019a). As stated in the manuscript, the latter software will be made publicly available upon acceptance of the paper.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The PYPref, PYPfast, PYPslow, PYP3ps, and PYP200ns structures are already deposited in the Protein Data Bank, together with their respective weighted difference structure factor amplitudes under accession codes 5HD3, 5HDC, 5HDD, 5HDS, and 5HD5, respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	337,852 light snapshots, and 79,937 dark snapshots were collected.
Data exclusions	190,053 light snapshots were randomly removed to yield a uniform delay-time histogram.
Replication	No data were replicated. Repeating rows or columns of data in analyses involving matrix manipulation often leads to poor conditioning and errorneous results.
Randomization	The experiment naturally captures snapshots of protein molecules in random orientations. We did not divide the data into random sub-groups. Splitting data into smaller sub-groups would result in lower spatial resolution.
Blinding	We did not divide the data into random sub-groups. Splitting data into smaller sub-groups would result in lower spatial resolution.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging