An open-source, end-to-end workflow for multidimensional photoemission spectroscopy

R. Patrick Xian[1,*],   Yves Acremann[2],   Steinn Y. Agustsson[3],   Maciej Dendzik[1],   Kevin Bühlmann[2],   Davide Curcio[4],   Dmytro Kutnyakhov[5,6],   Federico Pressacco[6],   Michael Heber[5],   Shuo Dong[1],   Tommaso Pincelli[1],   Jure Demsar[3], Wilfried Wurth[5,6,†],   Philip Hofmann[4],   Martin Wolf[1],   Markus Scheidgen[1,7], Laurenz Rettig[1,*], and Ralph Ernstorfer[1,*]

[1]Fritz Haber Institute of the Max Planck Society, 14195 Berlin, Germany

[2]Laboratory for Solid State Physics, ETH Zurich, 8093 Zurich, Switzerland

[3]Department of Physics, University of Mainz, 55128 Mainz, Germany

[4]Department of Physics and Astronomy, Interdisciplinary Nanoscience Center (iNANO), Aarhus University, 8000 Aarhus C, Denmark

[5]DESY Photon Science, 22607 Hamburg, Germany

[6]Department of Physics, University of Hamburg, 22761 Hamburg, Germany

[7]Department of Physics, Humboldt University of Berlin, 12489 Berlin, Germany

[†]Deceased

[*]Corresponding authors: xian@fhi-berlin.mpg.de, rettig@fhi-berlin.mpg.de, ernstorfer@fhi-berlin.mpg.de

Characterization of the electronic band structure of solid state materials is routinely performed using photoemission spectroscopy. Recent advancements in short-wavelength light sources and electron detectors give rise to multidimensional photoemission spectroscopy, allowing parallel measurements of the electron spectral function simultaneously in energy, two momentum components and additional physical parameters with single-event detection capability.

Efficient processing of the photoelectron event streams at a rate of up to tens of megabytes per second will enable rapid band mapping for materials characterization. We describe an open-source workflow that allows user interaction with billion-count single-electron events in photoemission band mapping experiments, compatible with beamlines at 3rd and 4th generation light sources and table-top laser-based setups. The workflow offers an end-to-end recipe from distributed operations on single-event data to structured formats for downstream scientific tasks and storage to materials science database integration. Both the workflow and processed data can be archived for reuse, providing the infrastructure for documenting the provenance and lineage of photoemission data for future high-throughput experiments.

Introduction

Many disciplines in the natural sciences are increasingly dealing with densely sampled multidimensional datasets. The scientific workflows to obtain and process them are becoming increasingly complex due to the provenance and structure of the data and the information needed to be extracted and analyzed [1, 2]. In materials science and condensed matter physics, various spectroscopic and structural characterization techniques produce experimental data of distinct formats and characteristics. Their creation and understanding require customized processing and analysis pipelines designed by specialists in the respective fields. The growing incentive for building experimental materials science databases [3] that complement established theoretical counterparts [4] calls for open-source and reusable workflows for data processing [5, 6] that transform raw data to shareable formats for downstream query, analysis and comparison by non-specialists of the experimental techniques [7, 8]. Among the various properties associated with materials, the electronic band structure (EBS) of condensed matter systems is of vital importance to the understanding of their electronic properties in and out of equilibrium. Multidimensional photoemission spectroscopy (MPES) [9, 10, 11] is an emerging technique that bears the potential of high-throughput EBS characterization through band mapping experiments and holds promise as an enabling technology for building experimental EBS databases, where data integration requires traceable knowledge of the processing steps

between the archived and the raw format. Here we present an open-source workflow that focuses on band mapping data from MPES. In the following, we briefly introduce the technology of MPES and the associated data processing, before providing details on the workflow from raw data to database integration.
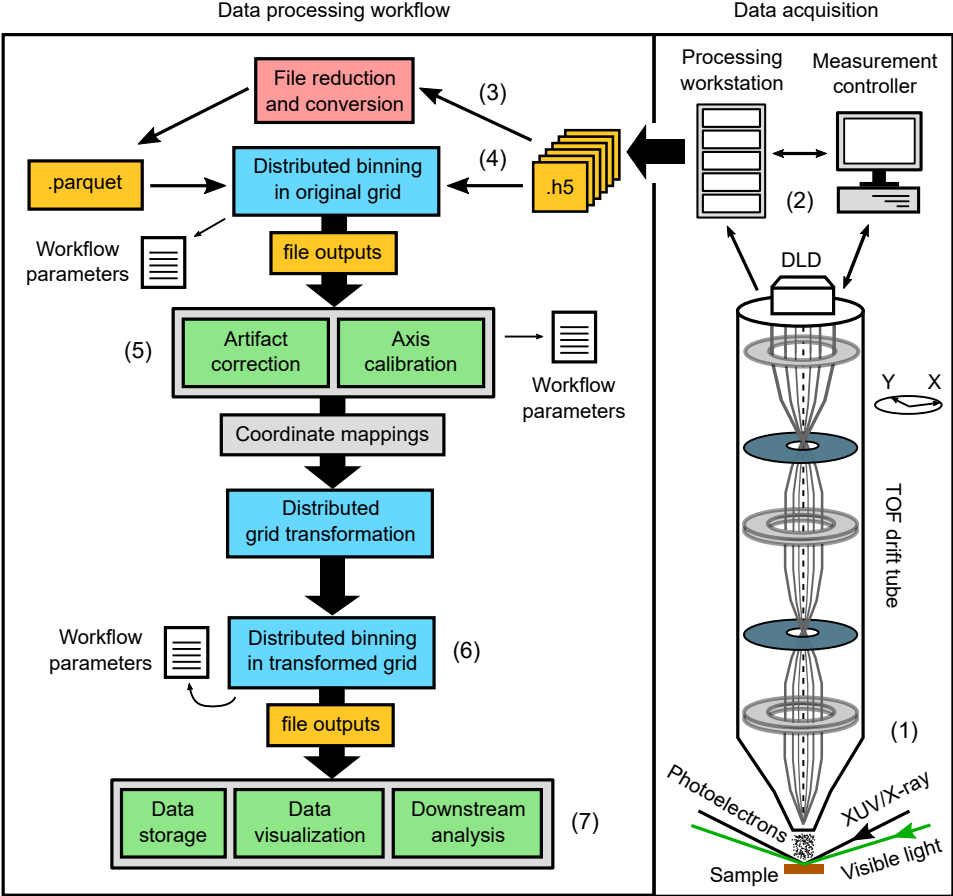

Figure 1: **Schematic of the workflow in MPES.** The data acquisition in MPES starts from (1) photoelectrons liberated by the extreme UV (XUV) or X-ray photons travelling through the lens systems and the TOF tube to trigger detection events on the delay-line detector (DLD). (2) Single-event data acquisition is monitored and controlled by the measurement controller computer. The raw data are first streamed and stored onto a hard drive in HDF5 format (.h5) and subsequently processed in the workflow through (3) file reduction (optional), (4)(6) distributed binning, (5) artifact correction and axis calibrations, carried out at the single-event or the binned data levels. At the end of the workflow, other data formats are generated (such as HDF5, MAT or TIFF) for (7) storage, visualization or downstream analysis for extracting relevant physical parameters. Critical parameters within the workflow may be exported (as workflow parameters files), shared and reused for processing other datasets.

MPES, also called momentum microscopy (MM), is born out of the recent integration of time-of-flight (TOF) electron spectrometers with delay-line detectors (DLDs) and improved

electron-optic lens designs [12, 13, 14, 15]. Compared with the earlier generations of angle-resolved photoemission spectroscopy (ARPES) [16, 17] using hemispherical analyzers to measure the 2D energy-momentum distribution of the photoemitted electrons [18], MPES is capable of recording single-electron events simultaneously sorted into the (k_x, k_y, E) coordinates (E: electron energy, k_x, k_y: parallel momentum components) in band mapping experiments, obviating the need for scanning across sample orientations and subsequent data merging as is the case for similar experiments using a single hemispherical analyzer. Operation of the TOF DLD in MPES requires a pulsed photon source and is directly compatible with 3rd and 4th generation light sources [19] as well as laboratory-based table-top setups [20, 21, 22, 23], harnessing their high repetition rates in the range of multi-kilohertz to megahertz to drastically improve the detection speed and efficiency. Mapping of the 3D band structure with sufficient signal-to-noise ratio (SNR) may be achieved on the timescale of minutes. The technological convergence opens up the possibilities to record 3D datasets in dependence of one or more additional parameters, such as spatial location $I(x, y, k_x, k_y, E)$, probe photon energy, $I(k_x, k_y, E, k_z)$ [10], spin-polarization, $I(k_x, k_y, E, S)$ [9], or pump-probe time in time-resolved MM, $I(k_x, k_y, E, t)$ [24] within a reasonable time frame.

From the data perspective, the pulsed sources with high repetition rates generate densely sampled data at rates of multiple megabytes per second (MB/s), which has brought about challenges in data processing and management compared with conventional ARPES experiments. The raw data in MPES are single photoelectron events registered by the DLD and the physical quantities related to the detected events are streamed in parallel to the storage files in a hierarchical file format (e.g. HDF5 [25]). A typical dataset involves $10^7 - 10^{10}$ detected events with a total size of up to a few hundred of gigabytes (GBs), depending on the number of coordinates measured (3D or 4D) and the required SNR. Unlike the large 2D or 3D image-based datasets, such as those obtained in various forms of optical [26, 27] and electron microscopy techniques [28, 29], processing and conversion of tabulated single-event data requires additional steps of statistical computing for conversion into standard images. This motivates the current workflow development for efficient data processing and analysis. In data processing and calibration, experiments performed at different facilities share similar procedures going from the raw events to the multidimensional hypervolume with calibrated axes, which is the basis for archiving and downstream analysis. To maintain reproducibility

for the particular data source and structure, we have summarized the workflow (see Fig. 1) into two open-source software packages (hextof-processor [30] and mpes [31]), with similar design principles for coping with large-scale facility and table-top experiments, respectively. The core of our approach includes distributed statistical processing at the single-event level using parameters calibrated and determined from preprocessed volumetric datasets, which enables effective instrument diagnostics, artifact correction, and sample condition monitoring. The algorithms involved balance physical knowledge and existing methods in image processing and computer vision. The workflow is illustrated next with data obtained at some of the electron momentum microscopes currently in operation, such as the HEXTOF (high energy X-ray time-of-flight) measurement system [24] at the free-electron laser source FLASH [32] at DESY, and the table-top high harmonic generation-based setup at the Fritz Haber Institute (FHI) [21] involving a commercial TOF and DLD (METIS 1000, SPECS GmbH). We expect the workflow described here to serve as a blueprint for upcoming software platforms in similar setups to be installed in other facilities or laboratories worldwide.

Results

Workflow description. The workflow schematic shown in Fig. 1 starts with raw single-event data from measurements. The data are (i) binned in a distributed fashion in the measurement coordinates, including each of the photoelectrons' position on the detector (X, Y), its TOF, a digital encoder (ENC) axis, and others, if more than four dimensions are acquired in parallel. The binned histogram is (ii) used to estimate the numerical transforms for distortion correction and axis calibration. Next, these transforms are (iii) applied to the raw single-event data to convert the measurement coordinates to the physical axes, (k_x, k_y, E, t_{pp}) and others for higher dimensions (see also Fig. 2). Finally, the single-event data are (iv) binned in the transformed grid to yield 3D, 3D+t or other higher-dimensional data with the correct axis values. The outcome may be exported in different formats for storage, visualization and downstream analysis.

Tasks and software infrastructure. Processing billion-count single-event data requires user interaction for data checking and distributed processing to reduce the time consumption. The general tasks in the workflow include the transformation of data streams to multidimen-

sional histograms, artifact correction and axis calibration. These operations can be efficiently decomposed into column-wise operations of the distributed dataframe format offered by the dask package [33] in Python. While the use of dask dataframes provide the common foundation for interactivity with single events of hextof-processor and mpes, they distinguish themselves by the experimental requirements. At large-scale facilities, experiments often record a large number of machine parameters that need to be stored, though only a small number of relevant parameters are needed for downstream processing. Therefore, the hextof-processor package includes a parameter sampling step to retrieve intermediate tabulated data in the Apache Parquet format (https://parquet.apache.org/), a column-based data structure optimized for computational efficiency. This approach reduces the processing overhead in searching through the raw data files every time when data are queried during the subsequent processing. As an open-source project, other beamtime-specific functionalities are added by users to the existing framework at every new experimental run. The mpes package adapts to the much simpler file structure produced at table-top experimental setups and makes direct use of the HDF5 raw data. It comes with added functionalities motivated by the existing issues encountered in data acquisition and downstream processing such as axis calibration, masking, alignment and different forms of artifact correction.

Artifact correction. Artifacts in MPES data come from mechanical imperfections, stray fields (electric and magnetic), uncertainties in the alignment of the sample, light beams and the multistage electron-optic lens systems as well as the data digitization process. Minimizing and correcting instrumental imperfections plays an important role in the validity of downstream analysis. We carry out artifact correction sequentially at the level of single photoelectron events or the data hypervolume obtained from multidimensional histogramming (see Fig. 2). The outcomes are illustrated using the correction of (1) digitization artifact (see Fig. 3) and (2) spherical timing aberration artifact (see Fig. 4), with technical details in Methods.

Axis calibration. To transform the measurement axes of the DLD into physically relevant axes for electronic band mapping, calibrations are required, as shown in Fig. 2. The calibration functions are constructed with parameters derived from comparing physical knowledge of the materials (e.g. Brillouin zone size, Fermi level position) with the corresponding scales in data. They are applied either to the binned data hypervolume, or to the single-electron events
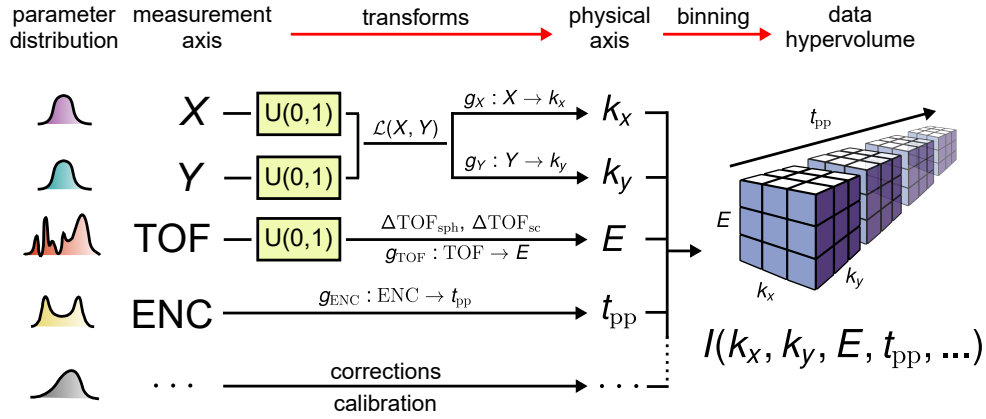
Figure 2: **Examples of workflow components.** Illustrations are given for artifact correction and axis calibration. Characteristic 1D distributions of the measured X, Y, TOF, ENC and an arbitrary axis are shown on the very left. $U(0,1)$ represent uniformly distributed random noise added to suppress digitization artifacts (jittering or dithering). The transforms (g's) are calibration functions that convert the values in the measurement axes to the physical ones. The transform $\mathcal{L}(X, Y)$ corrects the symmetry distortion, while the spherical timing aberration and space charge are compensated for by $\Delta\mathrm{TOF}_{\mathrm{sph}}$ and $\Delta\mathrm{TOF}_{\mathrm{sc}}$, respectively. Binning of the corrected single-event data over the calibrated physical axes yields a multidimensional hypervolume (right picture) of photoemission intensity data along with the physical axes values.
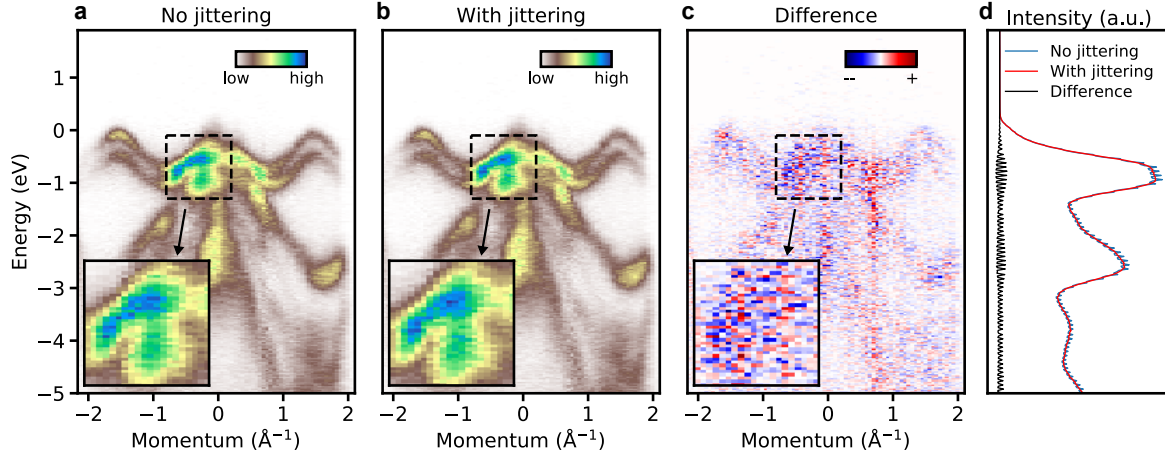

Figure 3: **Digitization artifact correction by histogram jittering.** Removal of the digitization artifact is illustrated with a 2D k-E slice across the Brillouin zone center of the band mapping dataset measured at FHI on WSe_2. The images before and after histogram jittering and their difference are shown in **a**, **b** and **c** respectively. A zoomed-in section of the data are shown in the insets in **a-c**. The effective removal of the digitization artifact is further demonstrated in the momentum-integrated energy distribution curves in **d**. The traces in **d** are computed by averaging horizontally over their corresponding 2D images in **a-c**.

7

raw data individually in a distributed fashion before binning. Details on the calibration data transforms are provided in Methods.
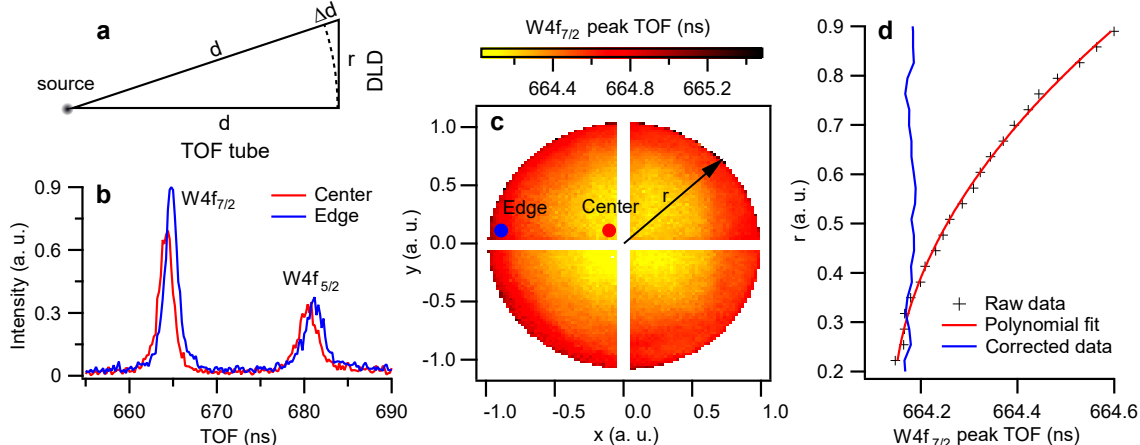

Figure 4: **Spherical timing aberration correction.** The correction is demonstrated using W4f core-level data measured at FLASH. The energy spacing between the $W4f_{7/2}$ and $W4f_{5/2}$ levels is about 2.1 eV [34]. **a.** Illustration of the geometric origin of the spherical timing aberration in the time-of-flight (TOF) drift tube. **b.** Comparison of the W4f spectra at the center and on the edge of the detector plane. The energy spectra are extracted from the corresponding regions, marked by the dots in the same blue and red colors, respectively, in **c.** The white stripes crossing at the detector center block the exposed edges of the four-quadrant detector quadrants. **d.** The uncorrected and corrected radial-averaged peak TOF positions for the $W4f_{7/2}$ core level.

Data storage and format. The simplistic form of the output data hypervolume derived from single-electron events includes non-negative scalar values of the photoemission intensity and the calibrated real-valued axes coordinates, including k_x, k_y, E, and other parameter dependencies such as the pump-probe time delay t_{pp}. These values are exported as HDF5, MAT or TIFF, with the metadata included as attributes of the files.

Workflow archiving and reuse. Computational workflows are valued by their the reproducibility [35]. Archiving and sharing the workflow parameters among users of the beamlines or facilities allow comparison between experimental runs and reuse for the simultaneous benefits of machine diagnostics and experimental efficiency. To achieve this, we store critical parameters generated within the workflow in a separate file as *workflow parameters* (see Fig. 1) during each step, including the numerical values used in binning, the intermediate parameters and coefficients of the correction and calibration functions, etc. They can be reused when loading into the processing of other datasets.

Data visualization. The adaptation of established scientific visualization methods in the physical sciences [36, 37] to band mapping data should incorporate the requirements and knowledge of the data characteristics in this field of research. The band mapping data in 3D (multi-megavoxel) and 3D+t (multi-gigavoxel) include the inherent symmetries from the electronic band structure of the material, but the intensity modulations in the photoemission process [38], dynamics and sample condition disrupt the original symmetry. The overall goal is to emphasize the features of interest while exploiting the symmetry to simplify the visualization (see Methods). The output files from the processing pipeline are compatible with open-source visualization software such as matplotlib [39], ParaView [36] and Blender [40].
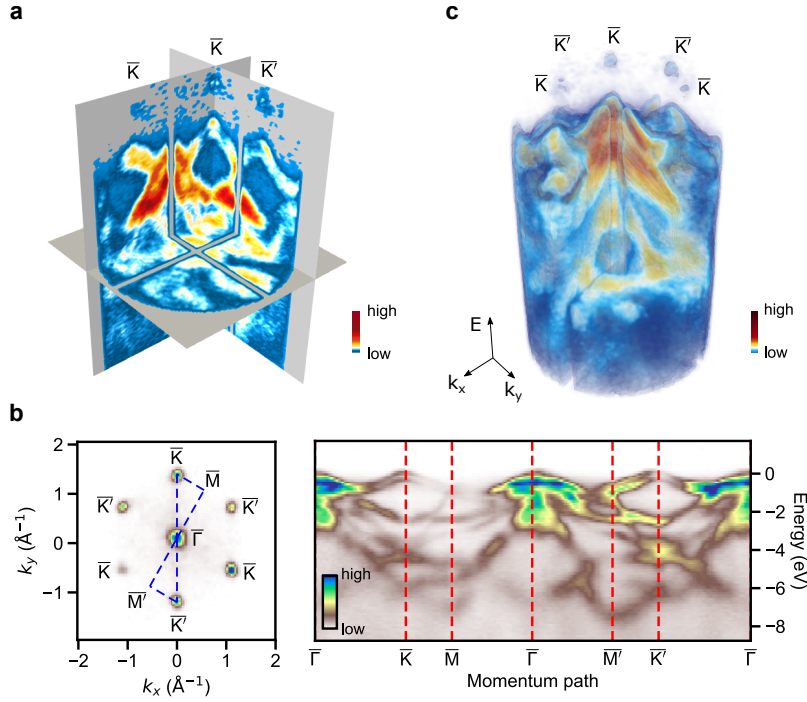

Figure 5: **Typical visual representations of the volumetric band mapping data.** The examples are illustrated using band mapping data of the layered semiconductor WSe_2, measured with the HEXTOF instrument at FLASH and the METIS detector at the FHI (see Methods). The visualizations are **a.** the orthoslice representation, **b.** the band-path diagram (right) with the momentum path labelled in dashed blue line in the momentum k_x-k_y plane (left), and **c.** the cut-out view. All color scales represent photoemission intensity. The letters label the high symmetry points of the hexagonal Brillouin zone of WSe_2 [41].

Downstream analysis integration. Typical photoemission data analysis involves extracting electronic band structure parameters, physical coupling constants and lifetimes via fitting

of lineshapes [16] or dynamical models [42], often carried out specific to the material under study. At the end of our distributed workflow, the data size is on the order of a few to tens of gigabytes, which can be directly loaded into memory on users' local machines for downstream data analysis with custom routines.

Experimental metadata. The metadata of the data files have a tree structure and contain information of the experimental setting, parameters of the pulsed light source, the detector and the sample under study. A list of top-level metadata parameters is presented in Table 1. A full and current list of all metadata parameters, including the top-level parameters and their constituent lower-level parameters, along with their definitions, units and values, is provided in Supplementary Tables 1-4. For database integration, an accompanying data parser (parser-mpes, see Code Availability) for MPES data has been written in accordance with existing standards [43] for computational materials science in NOMAD [8], featuring an electronic version of the metadata parameter list in the file mpes.nomadmetainfo.json online. The metadata parameter list and the data parser are versioned and are updated based on the corresponding changes in the data structure for photoemission spectroscopy experiments. The existing WSe_2 photoemission data have been integrated into the experimental section of the materials science database NOMAD (see Data Availability).

Table 1: Top-level metadata parameters

Category name	Description
General parameters	Descriptive information of the experiment and facility
Source parameters	Technical parameters relating to the photon source
Detector parameters	Technical parameters relating to the photoelectron detector
Sample parameters	Parameters relating to the material sample in experiment

Discussion

We have designed and implemented an open-source, end-to-end workflow for processing single-event data produced in multidimensional photoemission spectroscopy, linking to downstream tasks, providing guidelines and software for integrating processed data into the NOMAD experimental materials science database. The distributed processing takes full advantage of the single-event data streams directly accessible from the TOF delay-line detector for event-

wise correction and calibration and converts the raw events to the calibrated data hypervolume for project-specific downstream analysis. The functionalities within the workflow are publicly accessible through the software packages we have developed (hextof-processor [30] and mpes [31]). The processing workflow is archived at each step of operation and the processed data may be integrated into experimental database with user-specified metadata. The methods described here are applicable to all existing types of multidimensional photoemission band mapping measurements beyond the static and time-dependent settings described here.

Our end-to-end workflow from raw data to processed data to database integration provides a fast-track and all-in-one solution to the demands for open experimental data and reproducible research in the materials science community [7, 8]. The public repositories for the software packages are the foundations for phased future extension and integration with existing analytical tools in the photoemission spectroscopy community. The modular structure of the packages introduced here allows targeted upgrades by both temporary and dedicated maintainers and users. Casting the workflow in the Python programming environment provides the foundation for convenient incorporation of existing image processing and machine-learning resources [44] for further exploration and understanding of the band mapping datasets, which contain rich information owing to the complex nature of the photoemission process [16, 18]. This is especially beneficial for broader adoption of photoemission since the interpretation of photoemission data is often linked to the observed or extracted outstanding features such as local intensity extrema, dispersion kinks and satellites, lineshape parameters and pattern symmetry [16], therefore, the access to experimental data and the potential integration with existing electronic structure-related software [5, 45, 46, 47] will facilitate method developments and the direct comparison between experimental results and theoretical band structure calculations within the same programming platform.

Methods

Sample preparation. Single-crystalline samples of 2D bulk WSe_2 were purchased from HQ Graphene. Crystals of size around 5 mm $\times$ 5 mm $\times$ 1 mm were used directly for the measurements. To prepare a clean surface by cleaving, we attached a cleaving pin upright to the sample surface using conducting epoxy (EPOTEC H20) outside the vacuum chamber and

removed the pin by mechanical force in ultrahigh vacuum.

Photoemission experiments. The measurements were conducted using the HEXTOF instrument [24] at the DESY FLASH PG-2 beamline [48] with the free-electron laser (FEL) as well as a laboratory source [21] with a METIS electron momentum microscope (SPECS METIS 1000) installed at the FHI. In all measurements at FLASH, the FEL was tuned to 36.5 eV (or 34.0 nm) and 109 eV, the optical pump pulse had a center wavelength of 775 nm. The measurements at the FHI used a 21.7 eV home-built extreme ultraviolet source based on high harmonic generation and driven by an optical parametric chirped-pulse amplifier operating at 500 kHz repetition rate [49].

Digitization artifact. The time-to-digital converter (TDC) outputs digitized data according to the binning width of the on-board electronics. Data conversion from one digitized format to another in a rebinning process often creates a picket fence-like effect (see Fig. 3). This phenomenon originates from the incommensurate bin size in the two rounds of sampling processes (binning and rebinning). To solve the problem, one introduces a slight amount of uniformly distributed noise, with an amplitude equal to half of the original bin size, to the single-event values when carrying out the bin counts. This is similar to the histogram jittering (or dithering) technique [50, 51] used in statistical visualization and computer graphics. Mathematically, the uniformly distributed noise $U(0,1)$ bounded in the range $[0,1]$ is added before binning a univariate data stream, $S = \{S_i\}$ via,

$$S_i' = S_i + \frac{w_b}{2} \times U(0,1). \tag{1}$$

Here, w_b is the bin width. For binning of multivariate data streams, such as the detector X position (or k_x), Y position (or k_y), and the photoelectron TOF (or E), we adopt the same approach individually for each dimension. The effect of jittering in reducing the digitization artifact is demonstrated in Fig. 3.

Spherical timing aberration. Electrons entering the TOF tube at different lateral positions travel through different path lengths to reach the detector, which is the origin of the spherical timing aberration as illustrated in Fig. 4. The lateral position-dependent time delay may be

expressed as,

$$\Delta \text{TOF}_{\text{sph}}(r) = (\sqrt{1 + r^2/d^2} - 1)\text{TOF}_0, \tag{2}$$

where r is the radial distance from the center of the DLD and TOF_0 is the TOF normalization constant. For a typical field-free region length of $d \sim 1$ m in the TOF tube and a DLD screen radius of $r = 50$ mm, $\Delta\text{TOF}/\text{TOF}_0 \approx 1.25 \times 10^{-3}$. Assuming $\text{TOF}_0 = 0.5~\mu s$, the spherical timing aberration in TOF scale is $\Delta\text{TOF}_{\text{sph}} \approx 0.62$ ns, which is larger than the DLD's temporal resolution of ~ 0.15 ns. The effect of the spherical timing aberration is visible for a few eV energy range with fine bins but quite small on a large energy range. To illustrate this effect, we use the W4f core-level data presented in Fig. 4b. For every (X, Y) position on the detector the peak of W4$f_{7/2}$ was fitted with a Voigt profile and the peak positions are shown in Fig. 4c. As the spectra from deep core levels typically do not show dispersion, the deviation from fitting corresponds to the spherical timing aberration of the electron optics. In order to compensate for the spherical timing aberration, we first transform the data from Cartesian to the polar coordinates (see Fig. 4c), and then fit the radial-averaged peak position to a polynomial function of the radius,

$$\Delta\text{TOF}_{\text{sph}}(r) = \frac{r^2 \text{TOF}_0}{2d^2} - \frac{r^4 \text{TOF}_0}{8d^4} + \text{O}(r^6). \tag{3}$$

The fitting results together with the corrected radial distribution are presented in Fig. 4d.

Symmetry distortion. Photoemission patterns in the (k_x, k_y) plane (i.e. an energy slice) may exhibit distorted symmetry due to the influence of various factors from the instrument, the sample and the experimental geometry on the trajectory of low-energy photoelectrons. Correction of the symmetry distortion yet preserving the intensity features requires the use of symmetry-related landmarks to solve for the symmetrization coordinate transform in the framework of nonrigid image registration [52]. In typical situations with an excellent electron lens alignment, the energy dependence of the momentum distortion within the focused phase space volume covering an energy range of several eV is negligible, so the same coordinate transform can be applied to all energy slices in the volumetric data (including both valence and conduction bands) or simultaneously to all single events.

Other single-experiment artifacts. (1) Momentum center shift: The momentum center

of the emergent photoelectrons travelling through the electron-optic system may experience an energy-dependent shift owing to the slight misalignment in the system or the influence of stray fields. Correction of the center shift requires an energy-dependent center alignment of energy slices. The shift along the energy (or TOF) axis may be estimated using phase correlation [53] or mutual information-based [54] sequential image registration methods, in which the series of energy slices are treated as an image sequence. In a well-shielded and well-aligned electron-optic lens system, generally, the momentum center shift is negligible in the focused photoelectron energy range. (2) Space-charge effect (SCE): The secondary photoelectron clouds originating from the probe and pump pulses cause a "doming effect" of the photoemission intensity distribution around the momentum center of the band structure. This is especially visible in systems with a clear Fermi edge [9, 11] or non-dispersing shallow core levels, which may be used as references for calibrating the parameters used for the flattening transform by applying a momentum-dependent shift $\Delta\mathrm{TOF}_{\mathrm{sc}}(k_x, k_y)$ in the TOF (or the calibrated energy) coordinate of the single-event data.

Momentum calibration. The scaling factors for momentum calibration are computed by comparing the positions of known high symmetry points in the band structure with their corresponding locations in an energy slice. Suppose A and B are two high symmetry points identifiable (e.g. as local extrema) from the experimental data with pixel positions (X_A, Y_A) and (X_B, Y_B), and momentum positions, (k_x^A, k_y^A) and (k_x^B, k_y^B), respectively. We calculate the pixel-to-momentum scaling ratios, f_X and f_Y, along the X (column) and Y (row) directions of a 2D k-space image, respectively. Then, the momentum coordinate (k_x, k_y) at each pixel position (X, Y) may be derived.

$$f_D = (k_d^A - k_d^B)/(D_A - D_B) \tag{4}$$

$$k_d = f_D \times (D - D_A) \quad (D, d = X, x \text{ or } Y, y) \tag{5}$$

Energy calibration. The calibration requires a set of band mapping data measured at different bias voltages (applied between the material sample and the ground), usually sampled with a spacing of 0.5 V in a range of $\pm$3-5 V around the normally applied bias voltage for a particular sample. The calibration proceeds by finding the TOF feature (e.g. local extrema)

14

correspondences in the 1D energy distribution curves (EDCs) at different biases using the dynamic time warping algorithm [55]. The transformation from the TOF to the photoelectron energy E is approximated as a polynomial function,

$$E(\text{TOF}) = \sum_{i=0}^{n} a_i \text{TOF}^i \tag{6}$$

The approximation is sufficiently accurate within a range of ~ 20 eV, sufficient to cover the entire valence band and some low-lying conduction bands of typical materials. The polynomial coefficients are determined using nonlinear least squares by solving $\Delta T \cdot \mathbf{a} = \Delta E$, in which $\mathbf{a} = (a_1, a_2, ...)^T$ is the coefficient vector while the constant offset a_0 is determined by manual alignment to an energy reference, such as the Fermi level or valence band maximum. The vector ΔE and the matrix ΔT contain, respectively, the pairwise differences of the bias voltages and the polynomial terms of differential TOF values. To calibrate a large energy range including multiple core levels, a piecewise polynomial may be used [11].

Pump-probe delay calibration. The time origin ('time zero') in time-resolved photoemission spectroscopy, i.e. the temporal overlap of pump and probe pulses, is determined by fitting of a characteristic trace extracted from the data. Since the readings of the digital encoder (see Fig. 2) are sampled linearly, equally-spaced pump-probe delays are directly convertible from the readings using linear interpolation, given the boundary values of the translation stage positions and the corresponding delay times. For unequally-spaced delays, a delay marker is first added to each data point as a separate column after data acquisition to group together the encoder reading ranges that correspond to the specific time delays. The data binning is carried out over the delay marker column instead of the equally-sampled encoder readings.

Visualization strategies. We discuss here three methods for the display of volumetric band mapping data, which are, at the same time, the basis for visualizing 3D+t data with time as an animated axis. (1) The orthoslice representation includes orthogonal 2D planes selected in specific regions in the 3D volume [36], which highlights specific slices deep within the data less visible in a volumetrically rendered view (see Fig. 5a). Along this line, we have developed a software package, 4Dview [56], to explore 4D data using simultaneously linked orthoslices, which also features contrast adjustment and data integration within a hypervolume of interest.

15

(2) The band-path plot (see Fig. 5b) is a 2D representation of the 3D band mapping volume generated by combining a series of 2D cuts along selected momentum paths (or k-paths) traversing a list of so-called high-symmetry points [57, 58]. This representation captures the largest dispersion within the band structure. For volumetric data, the same path may be sampled from all the full energy range to produce the plot shown in Fig. 5b. The analysis and visualization modules in the mpes package include functionalities to compose customized band-path plots. (3) The cut-out view (see Fig. 5c) exposes a specific part of interest in the volumetric data, while not losing the rest [36]. The analysis module in the mpes package provides ways to generate precise cut-outs using position landmarks (e.g. high-symmetry points labelled in Fig. 5) and inequalities.

Acknowledgements

Author contributions

Y.A., K.B., S.Y.A., D.C., R.P.X. and M.D. wrote the hextof-processor package. R.P.X. and L.R. wrote the mpes package. D.K., Y.A., F.P., R.P.X., S.Y.A., D.C., M.D., M.H., S.D., P.H., L.R., R.E. and W.W. participated in the experiments at the FLASH PG-2 beamline using the HEXTOF instrument in Hamburg. S.D. and L.R. conducted the experiment at the Fritz Haber Institute using the METIS electron momentum microscope. R.P.X., M.D., T.P., L.R., R.E., M.S. T.P. constructed the metadata format, R.P.X. and M.S. implemented them into parser-mpes. R.P.X. wrote the initial manuscript with contributions from M.D. and Y.A. All authors contributed to discussions to bring the manuscript to its final form.

Data availability

The single-event photoemission data used for demonstrating the workflow is available on the Zenodo platform at 10.5281/zenodo.2704787 and 10.5281/zenodo.3987303. The preprocessed data are being integrated into the NOMAD database in the domain for experimental materials science data accessible at https://nomad-lab.eu/prod/rae/gui/search?domain=ems.

Code availability

The code for the data transformation in the workflow is available on GitHub, under the names hextof-processor (https://github.com/momentoscope/hextof-processor) and mpes (https://github.com/mpes-kit/mpes). The parser for integrating preprocessed experimental data to the NOMAD database is parser-mpes (https://gitlab.mpcdf.mpg.de/rpx/parser-mpes).

Conflict of interests

The authors declare no conflict of interests in the content of the article.

References

[1] Pruneau, C. *Data Analysis Techniques for Physical Scientists* (Cambridge University Press, 2017).

[2] Deelman, E. *et al.* The future of scientific workflows. *The International Journal of High Performance Computing Applications* **32**, 159–175 (2018). URL http://journals.sagepub.com/doi/10.1177/1094342017704893.

[3] Zakutayev, A. *et al.* An open experimental database for exploring inorganic materials. *Scientific Data* **5**, 180053 (2018). URL http://www.nature.com/articles/sdata201853.

[4] Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science* 1900808 (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.201900808.

[5] Pizzi, G., Togo, A. & Kozinsky, B. Provenance, workflows, and crystallographic tools in materials science: AiiDA, spglib, and seekpath. *MRS Bulletin* **43**, 696–702 (2018). URL https://www.cambridge.org/core/product/identifier/S0883769418002038/type/journal{_}article.

[6] Perkel, J. M. Workflow systems turn raw data into scientific knowledge. *Nature* **573**, 149–150 (2019). URL http://www.nature.com/articles/d41586-019-02619-z.

[7] Hill, J. *et al.* Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bulletin* **41**, 399–409 (2016). URL https://www.cambridge.org/core/product/identifier/S0883769416000932/type/journal{_}article.

[8] Draxl, C. & Scheffler, M. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bulletin* **43**, 676–682 (2018). URL https://www.cambridge.org/core/product/identifier/S0883769418002087/type/journal{_}article.

[9] Schönhense, G., Medjanik, K. & Elmers, H.-J. Space-, time- and spin-resolved photoemission. *Journal of Electron Spectroscopy and Related Phenomena* **200**, 94–118 (2015). URL http://linkinghub.elsevier.com/retrieve/pii/S0368204815001243.

[10] Medjanik, K. *et al.* Direct 3D mapping of the Fermi surface and Fermi velocity. *Nature Materials* **16**, 615–621 (2017). URL http://www.nature.com/doifinder/10.1038/nmat4875.

[11] Schönhense, B. *et al.* Multidimensional photoemission spectroscopy—the space-charge limit. *New Journal of Physics* **20**, 033004 (2018). URL http://stacks.iop.org/1367-2630/20/i=3/a=033004?key=crossref.d5618dec9ebafc233e6eff1b7ce89ee0.

[12] Krömker, B. *et al.* Development of a momentum microscope for time resolved band structure imaging. *Review of Scientific Instruments* **79**, 053702 (2008). URL http://aip.scitation.org/doi/10.1063/1.2918133.

[13] Ovsyannikov, R. *et al.* Principles and operation of a new type of electron spectrometer – ArTOF. *Journal of Electron Spectroscopy and Related Phenomena* **191**, 92–103 (2013). URL https://linkinghub.elsevier.com/retrieve/pii/S0368204813001357.

[14] Damm, A. *et al.* Application of a time-of-flight spectrometer with delay-line detector for time- and angle-resolved two-photon photoemission. *Journal of Electron Spectroscopy and Related Phenomena* **202**, 74–80 (2015). URL https://linkinghub.elsevier.com/retrieve/pii/S0368204815000663.

[15] Tusche, C., Krasyuk, A. & Kirschner, J. Spin resolved bandstructure imaging with a high resolution momentum microscope. *Ultramicroscopy* **159**, 520–529 (2015). URL https://linkinghub.elsevier.com/retrieve/pii/S0304399115000698.

[16] Damascelli, A., Hussain, Z. & Shen, Z.-X. Angle-resolved photoemission studies of the cuprate superconductors. *Reviews of Modern Physics* **75**, 473–541 (2003). URL https://link.aps.org/doi/10.1103/RevModPhys.75.473.

[17] Yang, H. *et al.* Visualizing electronic structures of quantum materials by angle-resolved photoemission spectroscopy. *Nature Reviews Materials* **3**, 341–353 (2018). URL http://www.nature.com/articles/s41578-018-0047-2.

[18] Suga, S. & Sekiyama, A. *Photoelectron Spectroscopy: Bulk and Surface Electronic Structures* (Springer, 2014). URL https://link.springer.com/content/pdf/10.1007{%}2F978-3-642-37530-9.pdf.

[19] Couprie, M. New generation of light sources: Present and future. *Journal of Electron Spectroscopy and Related Phenomena* **196**, 3–13 (2014). URL https://linkinghub.elsevier.com/retrieve/pii/S0368204813002429.

[20] Chiang, C.-T. *et al.* Boosting laboratory photoelectron spectroscopy by mega-hertz high-order harmonics. *New Journal of Physics* **17**, 013035 (2015). URL http://stacks.iop.org/1367-2630/17/i=1/a=013035?key=crossref. 6ebe926eef9ebdacee9ddaabac19036d.

[21] Puppin, M. *et al.* Time- and angle-resolved photoemission spectroscopy of solids in the extreme ultraviolet at 500 kHz repetition rate. *Review of Scientific Instruments* **90**, 023104 (2019). URL http://aip.scitation.org/doi/10.1063/1.5081938.

[22] Corder, C. *et al.* Ultrafast extreme ultraviolet photoemission without space charge. *Structural Dynamics* **5**, 054301 (2018). URL http://aca.scitation.org/doi/10.1063/1. 5045578.

[23] Buss, J. H. *et al.* A setup for extreme-ultraviolet ultrafast angle-resolved photoelectron spectroscopy at 50-kHz repetition rate. *Review of Scientific Instruments* **90**, 023105 (2019). URL http://aip.scitation.org/doi/10.1063/1.5079677.

[24] Kutnyakhov, D. *et al.* Time- and momentum-resolved photoemission studies using time-of-flight momentum microscopy at a free-electron laser. *Review of Scientific Instruments* **91**, 013109 (2020). URL http://aip.scitation.org/doi/10.1063/1.5118777.

[25] Folk, M., Heber, G., Koziol, Q., Pourmal, E. & Robinson, D. An overview of the HDF5 technology suite and its applications. In *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases - AD '11*, 36–47 (ACM Press, New York, New York, USA, 2011). URL http://portal.acm.org/citation.cfm?doid=1966895.1966900.

[26] Weiler, N. C., Collman, F., Vogelstein, J. T., Burns, R. & Smith, S. J. Synaptic molecular imaging in spared and deprived columns of mouse barrel cortex with array tomography. *Scientific Data* **1**, 140046 (2014). URL http://www.nature.com/articles/ sdata201446.

[27] Ker, D. F. E. *et al.* Phase contrast time-lapse microscopy datasets with automated and manual cell tracking annotations. *Scientific Data* **5**, 180237 (2018). URL http: //www.nature.com/articles/sdata2018237.

[28] Levin, B. D. *et al.* Nanomaterial datasets to advance tomography in scanning transmission electron microscopy. *Scientific Data* **3**, 160041 (2016). URL http://www.nature.com/articles/sdata201641.

[29] Aversa, R., Modarres, M. H., Cozzini, S., Ciancio, R. & Chiusole, A. The first annotated set of scanning electron microscopy images for nanoscience. *Scientific Data* **5**, 180172 (2018). URL http://www.nature.com/articles/sdata2018172.

[30] Acremann, Y. *et al.* hextof-processor. https://github.com/momentoscope/hextof-processor. URL https://github.com/momentoscope/hextof-processor.

[31] Xian, R. P. & Rettig, L. mpes. https://github.com/mpes-kit/mpes. URL https://github.com/mpes-kit/mpes.

[32] Ackermann, W. *et al.* Operation of a free-electron laser from the extreme ultraviolet to the water window. *Nature Photonics* **1**, 336–342 (2007). URL http://www.nature.com/articles/nphoton.2007.76.

[33] Dask Development Team. *Dask: Library for dynamic task scheduling* (2016). URL https://dask.org.

[34] Shallenberger, J. R. 2D tungsten diselenide analyzed by XPS. *Surface Science Spectra* **25**, 014001 (2018). URL http://avs.scitation.org/doi/10.1116/1.5016189.

[35] Stodden, V. *et al.* Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241 (2016). URL https://www.sciencemag.org/lookup/doi/10.1126/science.aah6168.

[36] Hansen, C. D. & Johnson, C. R. (eds.) *The Visualization Handbook* (Elsevier Butterworth-Heinemann, 2005).

[37] Lipşa, D. R. *et al.* Visualization for the Physical Sciences. *Computer Graphics Forum* **31**, 2317–2347 (2012). URL http://doi.wiley.com/10.1111/j.1467-8659.2012.03184.x.

[38] Moser, S. An experimentalist's guide to the matrix element in angle resolved photoemission. *Journal of Electron Spectroscopy and Related Phenomena* **214**, 29–52 (2017). URL https://linkinghub.elsevier.com/retrieve/pii/S0368204816301724.

[39] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

[40] Community, B. O. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018). URL http://www.blender.org.

[41] Riley, J. M. *et al.* Direct observation of spin-polarized bulk bands in an inversion-symmetric semiconductor. *Nature Physics* **10**, 835–839 (2014). URL http://www.nature.com/articles/nphys3105.

[42] Weinelt, M. Time-resolved two-photon photoemission from metal surfaces. *Journal of Physics: Condensed Matter* **14**, R1099–R1141 (2002). URL http://stacks.iop.org/0953-8984/14/i=43/a=202?key=crossref.95d2a41303a5d98bdb7544d34eeba966.

[43] Ghiringhelli, L. M. *et al.* Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *npj Computational Materials* **3**, 46 (2017). URL http://www.nature.com/articles/s41524-017-0048-5.

[44] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018). URL http://www.nature.com/articles/s41586-018-0337-2.

[45] Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314–319 (2013). URL https://linkinghub.elsevier.com/retrieve/pii/S0927025612006295.

[46] Hjorth Larsen, A. *et al.* The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **29**, 273002 (2017). URL http://stacks.iop.org/0953-8984/29/i=27/a=273002?key=crossref.20f9751653d872507bf6c0cc5737032c.

[47] M. Ganose, A., J. Jackson, A. & O. Scanlon, D. sumo: Command-line tools for plotting and analysis of periodic ab initio calculations. *Journal of Open Source Software* **3**, 717 (2018). URL http://joss.theoj.org/papers/10.21105/joss.00717.

[48] Gerasimova, N., Dziarzhytski, S. & Feldhaus, J. The monochromator beamline at FLASH: performance, capabilities and upgrade plans. *Journal of Modern Optics* **58**, 1480–1485 (2011). URL http://www.tandfonline.com/doi/abs/10.1080/09500340.2011.588344.

[49] Puppin, M. *et al.* 500 kHz OPCPA delivering tunable sub-20 fs pulses with 15 W average power based on an all-ytterbium laser. *Optics Express* **23**, 1491 (2015). URL https://www.osapublishing.org/abstract.cfm?URI=oe-23-2-1491.

[50] Chambers, M., Cleveland, S., Tukey, A. & Kleiner, B. *Graphical Methods for Data Analysis* (Wadsworth International Group, 1983). URL http://www.amazon.com/Graphical-Analysis-Wadsworth-Statistics-Probability/dp/053498052X.

[51] Novo, D. & Wood, J. Flow cytometry histograms: Transformations, resolution, and display. *Cytometry Part A* **73A**, 685–692 (2008). URL http://doi.wiley.com/10.1002/cyto.a.20592.

[52] Xian, R. P., Rettig, L. & Ernstorfer, R. Symmetry-guided nonrigid registration: The case for distortion correction in multidimensional photoemission spectroscopy. *Ultramicroscopy* **202**, 133–139 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S0304399118303474.

[53] Guizar-Sicairos, M., Thurman, S. T. & Fienup, J. R. Efficient subpixel image registration algorithms. *Optics Letters* **33**, 156 (2008). URL https://www.osapublishing.org/abstract.cfm?URI=ol-33-2-156.

[54] P. Viola & W.M.Wells. Alignment by Maximisation of Mutual Information. *International Journal of Computer Vision* **24**, 137–154 (1997). URL http://link.springer.com/10.1023/A:1007958904918.

[55] Salvador, S. & Chan, P. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* **11**, 561–580 (2007).

[56] Dendzik, M. 4Dview (2019). URL https://zenodo.org/record/3360817.

[57] Setyawan, W. & Curtarolo, S. High-throughput electronic band structure calculations: Challenges and tools. *Computational Materials Science* **49**, 299–312 (2010). URL http://linkinghub.elsevier.com/retrieve/pii/S0927025610002697.

[58] Hinuma, Y., Pizzi, G., Kumagai, Y., Oba, F. & Tanaka, I. Band structure diagram paths based on crystallography. *Computational Materials Science* **128**, 140–184 (2017). URL https://linkinghub.elsevier.com/retrieve/pii/S0927025616305110.