A machine learning route between band mapping and band structure

Rui Patrick Xian (xian@fhi-berlin.mpg.de)
Fritz Haber Institute of the Max Planck Society https://orcid.org/0000-0001-9895-6956
Vincent Stimper
Max Planck Institute for Intelligent Systems
Marios Zacharias
Fritz Haber Institute of the Max Planck Society
Shuo Dong
Fritz Haber Institute of the Max Planck Society
Maciej Dendzik
Fritz Haber Institute of the Max Planck Society
Samuel Beaulieu
Université de Bordeaux
Bernhard Schoelkopf
Max Planck Institute for Intelligent Systems
Martin Wolf
Fritz-Haber-Institut der May-Planck-Gesellschaft
Laurenz Rettig
Fritz-Haber-Institut der Max-Planck-Gesellschaft https://orcid.org/0000-0002-0725-6696
Christian Carbogno
Fritz Haber Institute of the Max Planck Society
Stefan Bauer
Max Planck Institute for Intelligent Systems
Ralph Ernstorfer
Fritz Haber Institute of the Max Planck Society

Resource

Keywords:

Posted Date: March 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1407122/v1

A machine learning route between band mapping and band structure

R. Patrick Xian,[1,†,*,a] Vincent Stimper,[2,†,*] Marios Zacharias,[1,b] Shuo Dong,[1] Maciej Dendzik,[1,c] Samuel Beaulieu,[1,d] Bernhard Schölkopf,[2] Martin Wolf,[1] Laurenz Rettig,[1] Christian Carbogno,[1] Stefan Bauer,[2,*,e] and Ralph Ernstorfer[1,*]

[1]Fritz Haber Institute of the Max Planck Society, 14195 Berlin, Germany.
[2]Department of Empirical Inference, Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany.
[†]These authors contributed equally to this work.
[*]Correspondence authors: xian@fhi-berlin.mpg.de, vstimper@tuebingen.mpg.de, baue@kth.se, ernstorfer@fhi-berlin.mpg.de.

[a]Current Address: Department of Mechanical Engineering, University College London, WC1E 7JE London, UK.
[b]Current Address: Department of Mechanical and Materials Science Engineering, Cyprus Institute of Technology, 3603 Limassol, Cyprus.
[c]Current Address: Department of Applied Physics, KTH Royal Institute of Technology, 114 19 Stockholm, Sweden.
[d]Current Address: Université de Bordeaux—CNRS—CEA, CELIA, UMR5107, F33405, Talence, France.
[e]Current Address: Division of Decision and Control Systems, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden.

Electronic band structure (BS) and crystal structure are the two complementary identifiers of solid state materials. While convenient instruments and reconstruction algorithms have made large, empirical, crystal structure databases possible, extracting quasiparticle dispersion (closely related to BS) from photoemission band mapping data is currently limited by the available computa-

tional methods. To cope with the growing size and scale of photoemission data, we develop a pipeline including probabilistic machine learning and the associated data processing, optimization and evaluation methods for band structure reconstruction, leveraging theoretical calculations. The pipeline reconstructs all 14 valence bands of a semiconductor and shows excellent performance on benchmarks and other materials datasets. The reconstruction uncovers previously inaccessible momentum-space structural information on both global and local scales, while realizing a path towards integration with materials science databases. Our approach illustrates the potential of combining machine learning and domain knowledge for scalable feature extraction in multidimensional data.

The modelling and characterization of the electronic BS of materials play an essential role in materials design [1] and device simulation [2]. The BS lives in the momentum space, $\Omega(k_x, k_y, k_z, E)$ and imprints the multidimensional and multi-valued functional relations between energy (E) and momenta (k_x, k_y, k_z) of periodically confined electrons [3]. Photoemission band mapping [4] (see Fig. 1a) using momentum- and energy-resolved photoemission spectroscopy (PES), including angle-resolved PES (ARPES) [5] and multidimensional PES [6, 7] measures the BS as an intensity-valued multivariate probability distribution directly in Ω. The proliferation of band mapping datasets and their public availability brought about by recent hardware upgrades [6–9] have ushered in the possibilities of comprehensive benchmarking between theories and experiments, which become especially challenging for multiband materials with complex band dispersions [10–12]. The available methods for interpreting the photoemission spectra fall into two categories: Physics-based methods require least-squares fitting of 1D lineshapes, named energy or momentum distribution curves (EDCs or MDCs), to analytical models [5, 13, 14]. Although physics-informed data models guarantee high accuracy and interpretability, upscaling the pointwise fitting (or estimation) to large, densely sampled regions in the momentum space (e.g. including 10^4 or more momentum locations) presents challenges due to limited numerical stability and efficiency. Therefore, their use is limited to selected momentum locations determined heuristically from physical knowledge of the materials and the experimental settings. Image processing-based methods apply data transformations to improve the visibility of dispersive features [15–18]. They are more computationally efficient and can

operate on entire datasets, yet offer only visual enhancement of the underlying band dispersion. Their outcomes don't achieve reconstruction and therefore are insufficient for truly quantitative benchmarking or archiving.

A method balancing the two sides will extract the band dispersion with sufficiently high accuracy and be scalable to multidimensional datasets, therefore providing the basis for distilling structural information from complex band mapping data and for building efficient tools for annotating and understanding spectra. In this regard, we propose a computational framework (see Fig. 1b) for global reconstruction of the photoemission (or quasiparticle) BS as a set of energy (or electronic) bands, formed by energy values (i.e. band loci) connected along momentum coordinates. Because the local maxima of photoemission intensities are not always good indicators of band loci [19], we exploit the connection between theory and experiment in our framework, based on a probabilistic machine learning [20, 21] model to approximate the intensity data from band mapping experiments. The gist of the model is rooted in Bayes rule,

$$p(\mathbf{\Theta}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{\Theta})p(\mathbf{\Theta}), \tag{1}$$

where the model parameters $\mathbf{\Theta} = \{\theta_i\}$ and the data $\mathcal{D}$ are mapped directly onto unknowns and experimental observables. We assign the energy values of the photoemission BS as the model parameters to extract from data, and a nearest-neighbor (NN) Gaussian distribution as the prior, $p(\mathbf{\Theta})$, to describe the proximity of energy values at nearby momenta. The EDC at every momentum grid point relates to the likelihood, $p(\mathcal{D}|\mathbf{\Theta})$, when we interpret the photoemission intensity probabilistically. The optimal parameters are obtained via *maximum a posteriori* (MAP) estimation in probabilistic inference [20] (see Methods and Supplementary Fig. 2). Given the form of the NN prior, the posterior, $p(\mathbf{\Theta}|\mathcal{D})$, in the current setting forms a Markov random field (MRF) [20, 22, 23], which encapsulates the energy band continuity assumption and the measured intensity distribution of photoemission in a probabilistic graphical model. For one benefit, the probabilistic formulation can incorporate imperfect physical knowledge algebraically in the model or numerically as initialization (i.e. warm start, see Methods) of the MAP estimation, without requiring *de facto* ground truth and training as in supervised machine learning [24]. For another, the graphical model representation allows convenient optimization and extension to other dimensions (see Supplementary Fig. 1 and section S1).

To demonstrate the effectiveness of the method, we have first reconstructed the entire 3D dispersion surfaces, $E(k_x, k_y)$, of all 14 valence bands within the projected first Brillouin zone

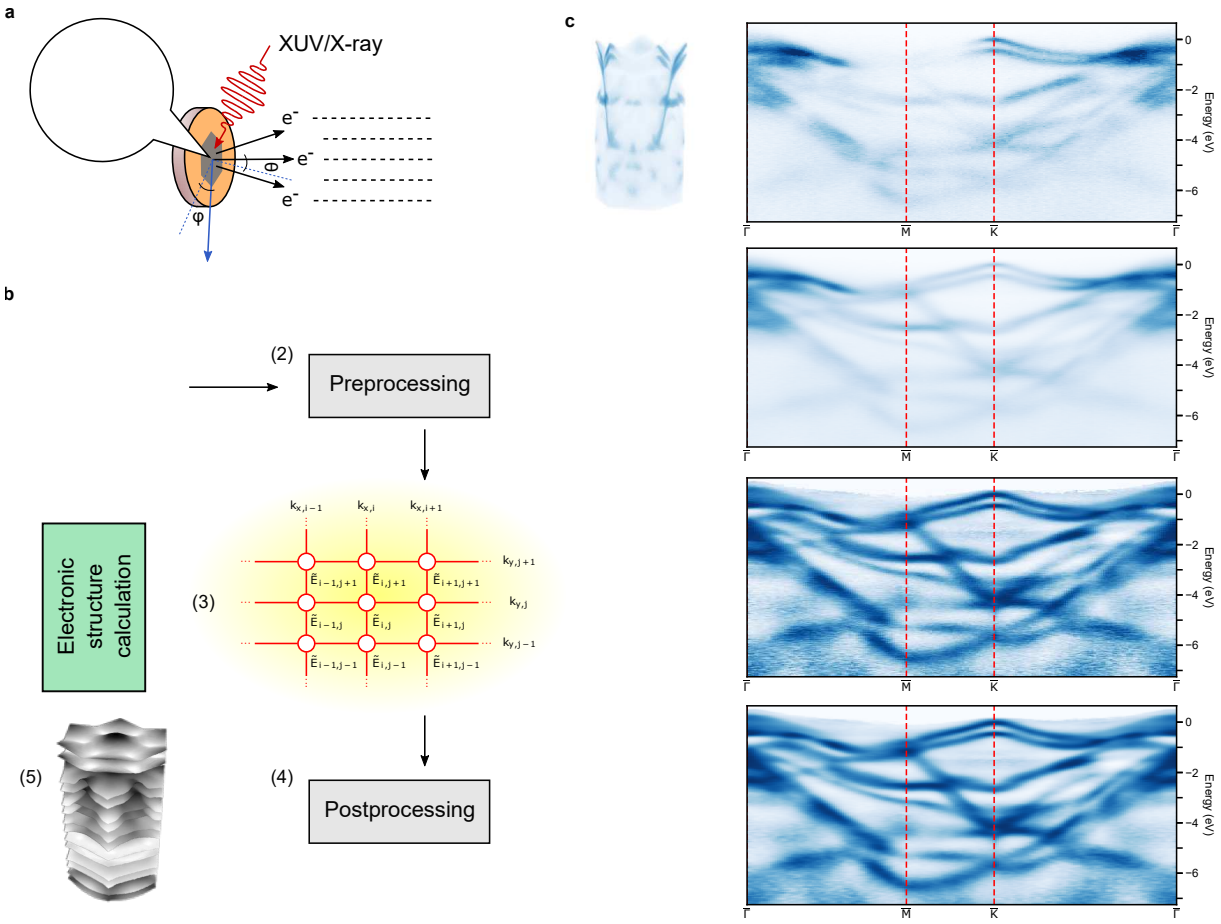

Figure 1: **From band mapping to band structure. a**, Schematic of a photoemission band mapping experiment. The electrons from a crystalline sample's surface are liberated by extreme UV (XUV) or X-ray pulses and collected by a detector through either angular scanning or time-of-flight detection schemes. **b**, Overview of the computational framework for reconstruction of the photoemission (or quasiparticle) band structure: (1) The volumetric data obtained from a band mapping experiment first (2) go through preprocessing steps, then are (3) fed into the probabilistic machine learning algorithm along with electronic structure calculations as initialization of the optimization. The reconstruction algorithm for volumetric band mapping data is represented as a 2D probabilistic graphical model with the band energies as parameters at each node and tens of thousands of nodes in practice. (4) The outcome of the reconstruction is post-processed (e.g. symmetrization) to (5) yield the dispersion surfaces (i.e. energy bands) of the photoemission band structure ordered by band indices. **c-f**, Effects of the intensity transforms in preprocessing viewed in both 3D and along high-symmetry lines of the projected Brillouin zone (hexagonal as in **b**(1)), from the original data (**c**) through intensity symmetrization (**d**), contrast enhancement [25] (**e**) and Gaussian smoothing of intensities (**f**). The intensity data in **c-f** are normalized individually for visual comparison.

(in (k_x, k_y, E) coordinates) of the semiconductor tungsten diselenide (WSe_2), spanning ~ 7 eV in energy and ~ 3 Å^{-1} along each momentum direction. Furthermore, we adapt informatics tools to BS data to sample and compare the reconstructed and theoretical BSs globally. The accuracy of the reconstruction is validated using synthetic data and the extracted local structural parameters in comparison with pointwise fitting. The available data and BS informatics enable detailed comparison of band dispersion at a resolution of < 0.02 Å^{-1}. Besides, we performed various tests and benchmarking on datasets of other materals and simulated data, where ground truth is available to evaluate accuracy and computational efficiency.

Results

Valence band mapping. The 2D layered semiconductor WSe_2 with 2H interlayer stacking (2H-WSe_2) is a model system for band mapping experiments [10, 26, 27]. Its valence BS contains 14 strongly dispersive energy bands, formed by a mixture of the $5d^4$ and $6s^2$ orbitals of the W atoms and the $4p^4$ orbitals of the Se atoms in its hexagonal unit cell. The strong spin-orbit coupling due to heavy elements produces large momentum- and spin-dependent energy splitting and modifications to the BS [10, 28]. The photoemission band mapping experiment captures the photoelectrons directly in their 3D coordinates, (k_x, k_y, E), by a commercial electron momentum microscope (METIS 1000, SPECS GmbH, see Methods) [6, 7]. Earlier valence band mapping and reconstruction in ARPES experiments on WSe_2 have demonstrated a high degree of similarity between theory and experiments [10, 26, 27], but a quantitative assessment within the entire (projected) Brillouin zone is still lacking. Effects of sample degradation has also been reported [27] during the course of long-duration angular scanning in ARPES measurements. With our high-repetition-rate photon source [**Maklar2020**] and the fast electronics of the momentum microscope, band mapping of WSe_2 achieves sufficient signal-to-noise ratio for valence band reconstruction within only tens of minutes of data acquisition, without the need for angular scanning and subsequent reconstruction from momentum-space slices.

Band structure reconstruction and digitization. We use a 2D MRF to model the loci of an energy band within the intensity-valued 3D band mapping data, regarded as a collection of momentum-ordered EDCs. It is graphically represented by a rectangular grid overlaid on the momentum axes with the indices (i, j) (i, j are nonnegative integers), as shown in step (3)

of Fig. 1b. The undetermined band energy of the EDC at (i, j), with the associated momentum coordinates $(k_{x,i}, k_{y,j})$, is considered a random variable (or model parameter), $\tilde{E}_{i,j}$, of the MRF. Together, the probabilistic model is characterized by a joint distribution, expressed as the product of the likelihood and the Gaussian prior, in Eq. (1). To maintain its simplicity, we don't explicitly account for the intensity modulations of various origins (such as imbalanced transition matrix elements [19]) in the original band mapping data, which cannot be remediated by upgrading the photon source or detector. Instead, we preprocess the data to minimize their effects on the reconstruction (see Fig. 1c-f). The preprocessing steps include (1) intensity symmetrization, (2) contrast enhancement [25], followed by (3) Gaussian smoothing (see Methods), whereafter the continuity of band-like features is restored. The EDCs from the preprocessed data, $\tilde{I}$, are used effectively as the likelihood to calculate the MRF joint distribution,

$$p(\{\tilde{E}_{i,j}\}) = \frac{1}{Z}\prod_{ij}\tilde{I}(k_{x,i}, k_{y,j}, \tilde{E}_{i,j}) \cdot \prod_{(i,j)(l,m)|\mathrm{NN}} \exp\left[-\frac{(\tilde{E}_{i,j} - \tilde{E}_{l,m})^2}{2\eta^2}\right]. \tag{2}$$

Here, Z is a normalization constant, η is a hyperparameter defining the width of the Gaussian prior, $\prod_{ij}$ denotes the product over all discrete momentum values sampled in the experiment and $\prod_{(i,j)(l,m)|\mathrm{NN}}$ the product over all the NN terms. Detailed derivation of Eq. (2) is given in Supplementary Information section S1. Reconstruction of the bands in the photoemission BS is carried out sequentially and relies on local optimization of the MRF model parameters, $\{\tilde{E}_{i,j}\}$. To optimize over large graphical models, we adopt multiple parallelization schemes to achieve efficient operations on scalable computing hardware. A single band reconstruction involving optimization over 10^4 model parameters is achieved within seconds and hyperparameter tuning within tens of minutes (see Methods, Supplementary Fig. 3 and section S1). In comparison, pointwise fitting often requires hand-tuning individually and therefore difficult to scale up to whole bands accordingly within a meaningful timeframe. To correctly resolve band crossings and nearly degenerate energies, we further inject relevant physical knowledge in the optimization by using density functional theory (DFT) band structure calculation with semi-local approximation [29] as a starting point for the reconstruction. The calculation qualitatively entails such physical symmetry information for WSe_2, albeit not quantitatively reproducing the experimental quasiparticle BSs at all momentum coordinates. As shown with four DFT calculations with different exchange-correlation functionals [29] to initiate the reconstruction for WSe_2 and in various cases using synthetic data with known ground truth (see Methods, Supplementary

Table 3 and Supplementary Figs. 4-8), the reconstruction algorithm is not particularly sensitive to the initialization as long as the information about band crossings is present. The current framework can also support the initialization from more advanced electronic structure methods, such as GW [30] or that including electronic self-energies renormalized by electron-phonon coupling [31], when semi-local approximation yields not only quantitatively, but also qualitatively wrong quasiparticle BSs compared with experiment. However, a systematic benchmark of theory and experiment goes beyond the scope of this work.

The reconstructed 14 valence bands of WSe_2 initialized by LDA-level DFT are shown in Fig. 2b-d and Supplementary Videos. To globally compare the computed and reconstructed bands at a consistent resolution, we expand the BS in orthonormal polynomial bases [32], which are global shape descriptors and unbiased by the underlying electronic detail. The geometric featurization of band dispersion allows multiscale sampling and comparison using the coefficient (or feature) vectors [33]. We choose Zernike polynomials (ZPs) to decompose the 3D dispersion surfaces (see Fig. 3 and Methods) because of their existing adaptations to various boundary conditions [34]. In Fig. 3a-b, the band dispersions show generally decreasing dependence (seen from the magnitude of coefficients) on basis terms with increasing complexities (see Fig. 3a), and the majority of dispersion is encoded into a subset of the terms (see Fig. 3b). This observation implies that moderate smoothing may be applied to remove high-frequency features to improve the reconstruction in case of limited-quality data (acquired without sufficient accumulation time), which is often unavoidable when materials exhibit vacuum degradation, or during experimental parameter tuning. The example in Fig. 3b and more numerical evidence in Supplementary Fig. 14 illustrate the approximation capability of the hexagonal ZPs. Concisely, these coefficients act as geometric fingerprints of the energy band dispersion, which enable the use of similarity or distance metrics (see Methods) for their classification and comparison [33], as in Fig. 3c. In this context, the positive cosine similarity confirms the strong shape (or dispersion) resemblance of the 7 pairs of spin-split energy bands in the reconstructed BS of WSe_2, while the low negative values, such as those between bands 1-2 and 13-14, reflect the opposite directions of their respective dispersion (see Fig. 2d). These observations are consistent with the outcome obtained from DFT calculations (see Supplementary Fig. 13).

In addition to WSe_2, we have run BS reconstruction on two other photoemission datasets from diverse classes of materials: (1) For the gold (Au) dataset measured at a synchrotron X-

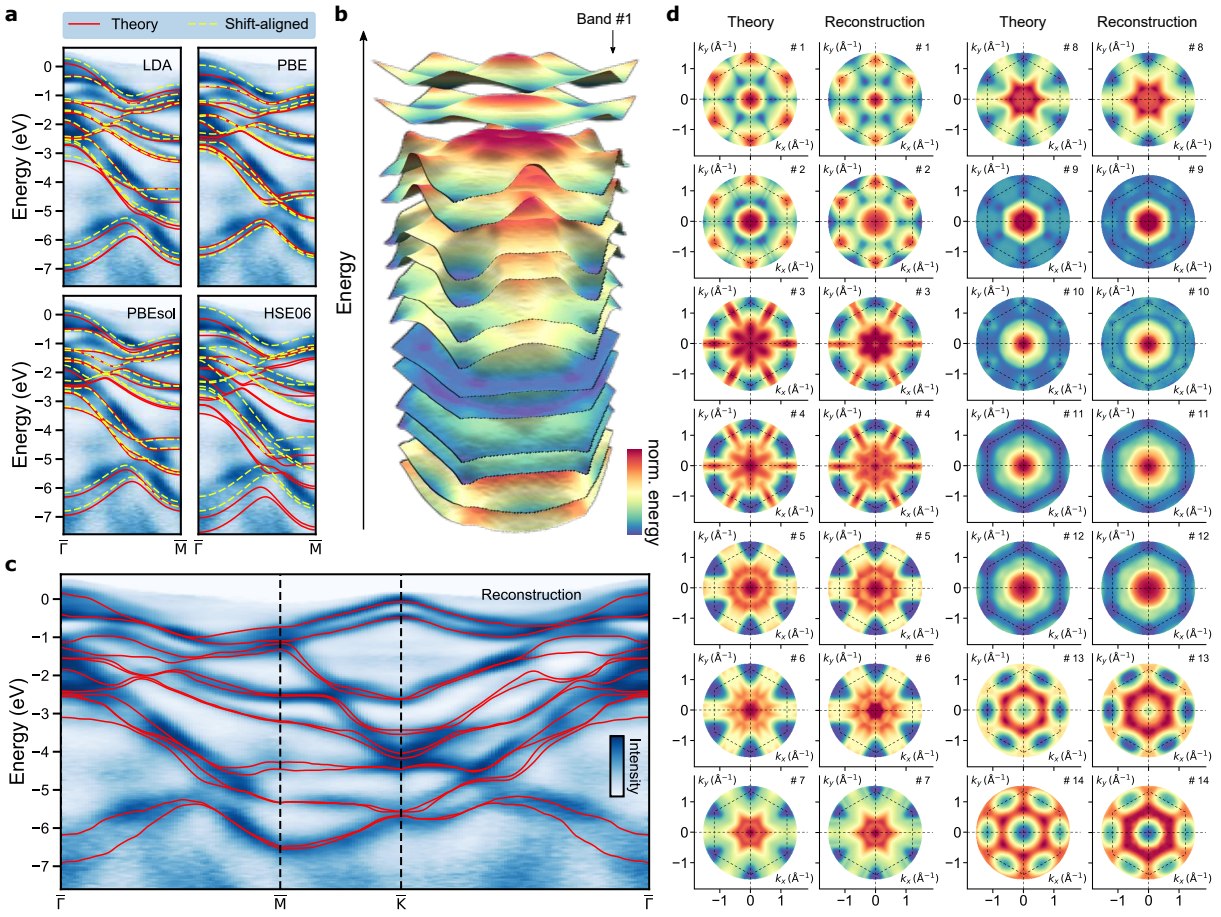

Figure 2: **Reconstructed band structure from WSe_2 photoemission data.** **a**, Comparison between the preprocessed WSe_2 valence band photoemission data along $\overline{\Gamma}$-$\overline{M}$ direction, DFT band structure calculated with different exchange-correlation functionals (solid red lines), and their final positions after band-wise rigid-shift alignment (dashed yellow lines) as part of hyperparameter tuning. The energy zero of each DFT calculation is set at the $\overline{K}$ point (see also Supplementary Fig. 4). **b**, Exploded view (with enlarged spacing between bands for better visibility) of reconstructed energy bands of WSe_2. **c**, Overlay of reconstructed band dispersion (red lines) on preprocessed photoemission band mapping data cut along the high-symmetry lines in the hexagonal Brillouin zone of WSe_2. The residual intensities on the low-energy end are from contrast-amplified background signals unrelated to the band structure. **d**, Band-wise comparison between LDA-level DFT (LDA-DFT) calculation used to initialize the optimization and the reconstructed 14 valence bands of WSe_2 (symmetrized in postprocessing). The boundaries of the first Brillouin zone are traced out by the dashed hexagons. The band indices on the upper right corners in **d** follow the ordering of the electronic orbitals in this material obtained from LDA-DFT. The color scale of band energies in **b** and **d** are normalized within each band to improve visibility.

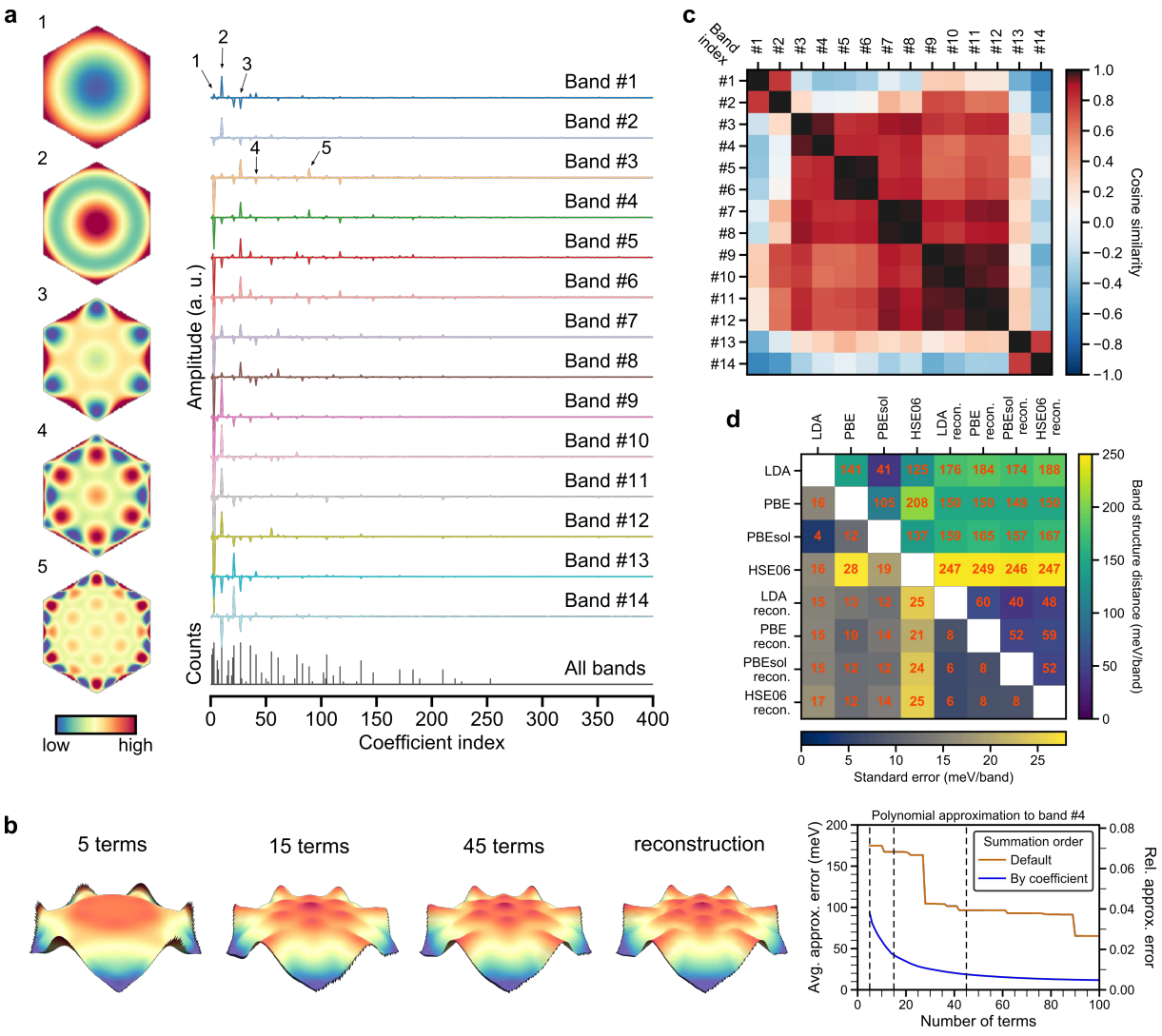

Figure 3: **Digitization and comparison of WSe_2 band structures.** **a**, Decomposition of the 14 energy bands of WSe_2 into hexagonal Zernike polynomials (ZPs) with selected major terms displayed on the left. The zero spatial frequency term in the decomposition is subtracted for each band. The counts of large ($> 10^{-2}$ by absolute value) coefficients of all 14 bands are accumulated at the bottom row of the decomposition to illustrate their distribution, which decrease in value towards higher-order terms. **b**, Approximation of the shape (or dispersion) of the fourth energy band using different numbers of hexagonal ZPs. Both of the average and relative approximation errors (see Methods) shown on the far right drop as more terms are included, while summation in the coefficient-ranked order (used to generate the intermediate results in **b**) achieves faster convergence than in the default polynomial order. **c**, Cosine similarity matrix for pairwise comparison of the reconstructed band dispersion in Fig. 2. The band indices follow those in Fig. 2d. **d**, Two-part similarity matrix showing band structure distances (in the upper triangle) and their corresponding standard errors (in the lower triangle) between the computed and reconstructed band structures of WSe_2. The abbreviation "LDA recon." denotes reconstruction with LDA-level DFT band structure as the initialization.

ray source, we used DFT calculations as the initialization to reconstruct four of the bulk energy bands, which are usually very challenging to extract by hand tracing or parametric function fitting, due in part to the blurring (k_z dispersion) from the 3D characteristics of the electrons in the metallic bulk. (2) For the dataset on a topological insulator, bismuth tellurium selenide (Bi_2Te_2Se), measured using the same laboratory setup as for the WSe_2 dataset, although we used only simple numerical functions (Gaussian and paraboloid) to initialize the MRF reconstruction, the outcome demonstrates correct discrete momentum-space symmetry and details of energy dispersion down to the concave-shaped hexagonal warping in the band energy contours around the Dirac point [35]. Four energy bands, including the two low-energy valence bands, a surface-state energy band, and a partially occupied conduction band, were recovered using our approach for Bi_2Te_2Se. Further discussions on these two materials and their band reconstructions are provided in Supplementary Information section S3.

Computational metrics and performance. To quantify the computational advantages of the machine learning-based reconstruction approach, we examine the outcome from diverse perspectives in consistency, accuracy and cost. To assess the consistency of reconstruction in its entirety, we introduce a BS distance metric (see Methods), invariant to the global energy shift frequently used to adjust the energy zero, to quantify the differences in band dispersion and the relative spacing between bands, which are the two major sources of variation between theories and experiments. The distance is calculated using the geometric fingerprints to bypass interpolation errors while reconciling the coordinate spacing difference between reconstructed and theoretical BSs, essential for differentiating BS data from heterogeneous sources in materials science databases [36, 37]. The results in Fig. 3d refer to the valence BS of WSe_2 discussed in this work, where the distances and their spread (i.e. standard errors) are displayed in the upper and lower triangles, respectively. A high degree of consistency exists among the reconstructions (pairwise distance no larger than 60 ± 8 meV/band) regardless of the level of DFT calculation used for initialization, indicating the robustness of the probabilistic reconstruction algorithm, whereas the distances between the DFT calculations are much larger, both in energy shifts and their spread. As shown in Fig. 3d and Supplementary Fig. 4, the learning algorithm can effectively reduce the epistemic uncertainty [38] between theories to obtain a consistent reconstruction.

For materials scientists, the band dispersions recovered from photoemission data are often examined locally near dispersion extrema. We show in Supplementary Fig. 15 that, besides providing global structural information, the reconstruction improves the robustness of traditional pointwise lineshape fitting in extended regions of the momentum space, when used as initial guess, because BS calculations may exhibit appreciable momentum-dependent deviations from experimental data that prevent them from being a sufficiently good starting point. Pointwise fitting in turn acts as refinement of local details not explicitly included in the probabilistic reconstruction model, which prioritizes efficiency. A compendium of local parameters are retrieved using this approach (see Supplementary Table 4). We obtain the trigonal warping parameter of the first two valence bands around $\overline{\mathrm{K}}$-point, 5.8 eV·Å^3 and 3.9 eV·Å^3, respectively, confirming the magnitude difference between these spin-split bands predicted by theory [28]. Fitting around $\overline{\mathrm{M}'}$ (and $\overline{\mathrm{M}}$) reveals that the gap opened by spin-orbit interaction extends beyond the saddle point in the dispersion surface with the minimum gap at 338 meV, markedly larger than DFT results. Overall, the reconstruction yields local structural information consistent with the more laborious pointwise fitting.

To demonstrate the computational advantage of the MRF reconstruction over traditional line fitting methods, we benchmarked the outcome over selected regions in synthetic photoemission data. The regions are chosen based on their importance and we limit the size in order to have a manageable computing time (about an hour on our computing cluster at maximum for a single run), determined by the slower method, and allow for hyperparameter tuning, which requires tens of runs. The line fitting approach uses the Levenberg-Marquardt least squares optimization [40] with bound constraints for multicomponent photoemission spectra composed of a series of lineshape functions. We used the benchmark established in [39] for pointwise line fitting employing high-performance computing and two synthetic datasets with known ground truth dispersion, representing the local and global settings of band structure reconstruction problem (see section S2.5). To level the hardware requirements, we used only distributed multicore-CPU computing for performance benchmarking. The estimated computing times are normalized to the per-band per-spectrum level [39]. The accuracy of the reconstruction is calculated using root-mean-squared (RMS) error, while the stability is quantified by the standard deviation of the residuals, which measures surface roughness [41]. The benchmarking results are compiled in Fig. 4 and Supplementary Table 2. They show that, compared with pointwise line fitting, the

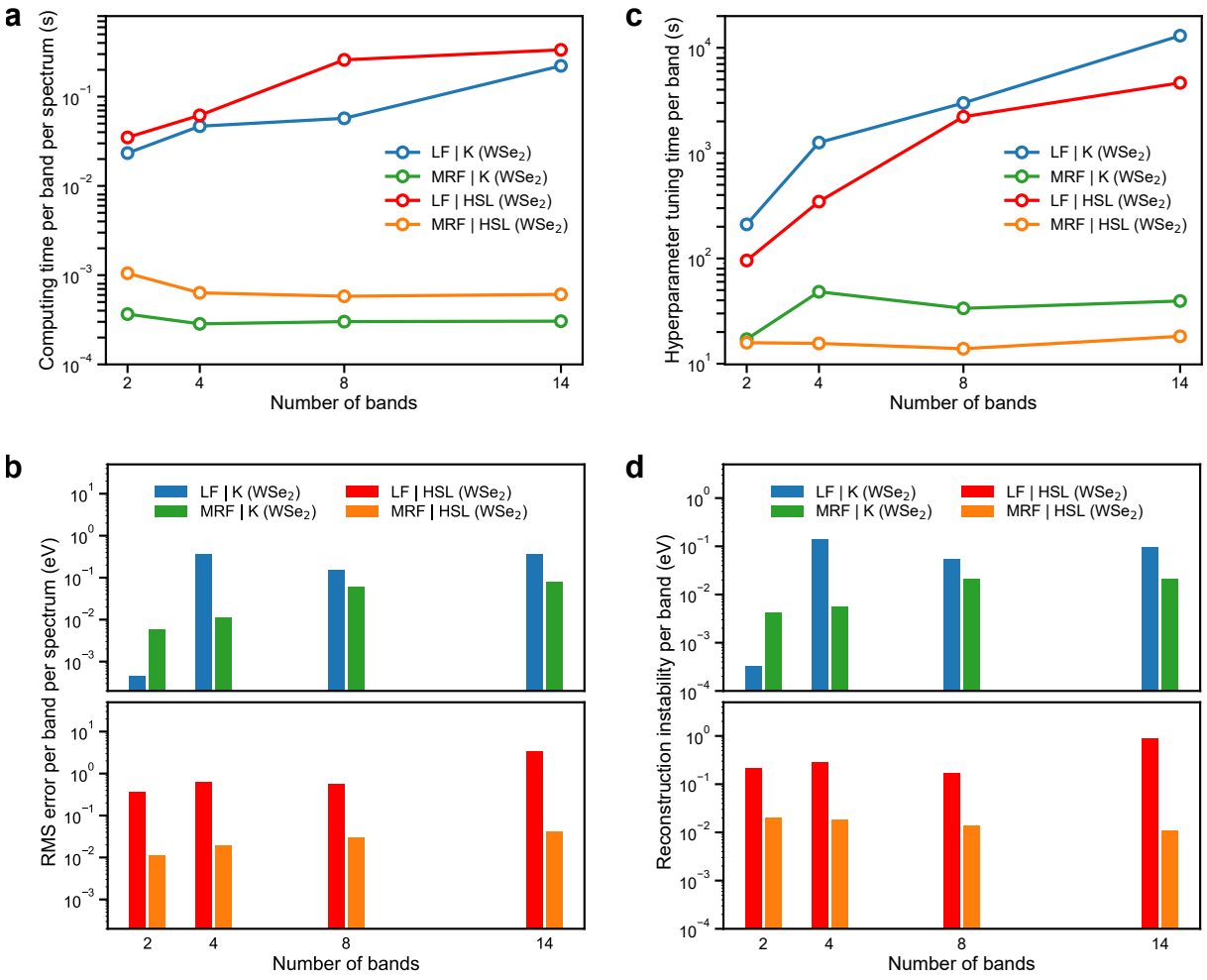

Figure 4: **Performance evaluation on benchmarks.** Visual summary of the benchmarking outcomes for band structure reconstruction using normalized metrics that are able to compare across datasets. These include **a**, the computing time and **b**, root-mean-square error (reconstruction error), both normalized to the per-band, per-spectrum level [39]. The other metrics, including **c** the hyperparameter tuning time and **d**, the reconstruction instability are normalized to the per-band level. The datasets are synthesized based on the calculated band structure of WSe_2 using LDA-DFT around the K point and along the high-symmetry line (HSL) of the Brillouin zone. The methods used in reconstruction include pointwise line fitting (LF) and the Markov random field (MRF) approach presented in this work. The benchmarks were run with synthetic datasets terminated at fixed energy ranges that contain the specified number of bands (2, 4, 8 and 14) shown in **a**-**d**.

MRF reconstruction offers significant reduction in both normalized computing time and hyperparameter tuning time, while achieving consistently higher accuracy and stability in all but the two-band case. The combination of accuracy and stability in MRF reconstruction is due to the connectivity built into the prior, whereas in the pointwise fitting approach, information is not explicitly shared among neighbors. Since the number of bands reflects the complexity of multicomponent spectra, a near-constant normalized computing time and hyperparameter tuning time (see Fig. 4a-b) in MRF reconstruction regardless of the number of bands (or spectral components) allow us to scale up the computation to datasets comprising 10^4-10^5 or more spectra. The significant gain in computational efficiency is a result of the inherent divide-and-conquer strategy in our BS reconstruction problem formulation and the associated distributed optimization method in the algorithm design. Comparatively, the distributed pointwise fitting exhibits a quasi-linear computational scaling with respect to the number of bands. When hyperparameter tuning is taken into account, in practice, it is only feasible for fitting small datasets with up to 10^3 multicomponent spectra [39].

Discussion

We have formulated the band structure reconstruction task, ubiquitous in photoemission band mapping, in a Bayesian inference framework and described an efficient reconstruction procedure by combining probabilistic machine learning with the physical knowledge embedded in electronic structure calculations, as demonstrated for the energy-dispersive, multiband material of WSe_2. The reconstruction method presents significant computational advantages and achieves a complete reconstruction of 3D dispersion surfaces, filling the gap in existing approaches that either focus primarily on visual enhancement [15–17] or on line fitting with heuristic physical models [5, 13, 14] that demand high computational cost. By balancing these two aspects, our method achieves high accuracy and efficiency simultaneously for complex band dispersion and is extendable to multiple materials and compatible with diverse experimental settings. It lends valuable insights to automating data analysis in materials science and the construction of end-to-end frameworks balancing physical constraints and computational efficiency to achieve desired outcome.

In the meantime, we provide a systematic framework for evaluating and benchmarking algorithms for band structure reconstruction in the future. We emphasize that although our ap-

proach is currently limited by the need for hyperparameter tuning and a good starting point from achieving fully automated reconstruction, it represents a first significant step in this direction. The initialization may also make use of the expanding and maturing computational databases [36, 37], either directly or as guidance for constructing crude proxies using analytical functions. Already at the current accuracy and stability, the reconstruction reveals global and local structural information challenging to access previously, and should assist interpretation of deep-lying bands, parametrizing multiband Hamiltonian models [42], simulation of realistic devices [2], and complement theoretical data in materials science databases [36, 37]. We envision that further improvements on the accuracy and computational efficiency aspects as well as integration into existing laboratory workflows [43] will facilitate scientific discovery. Our expected use cases include (i) online monitoring [44] of band mapping experiments in the study of materials phase transitions [45] or functioning devices [46], where changes in atomic structure or carrier mobility are often accompanied by detectable changes in the electronic structure (including band dispersion), resulting in $I(\mathbf{k}, E, t)$ with time (t) dependence in addition to momentum ($\mathbf{k}$) and energy. (ii) A similar scenario occurs in spatial mapping of electronic structure variations for electronic devices via scanning photoemission measurements [47, 48], resulting in $I(\mathbf{k}, E, \mathbf{x})$ with spatial ($\mathbf{x}$) dependence. In cases (i)-(ii), a fast reconstruction and evaluation framework may be used in a feedback loop to steer or optimize experimental conditions and gain real-time insights as opposed to the offline analysis typically deployed at present. (iii) Implementation of the reconstruction across various materials and to band-mapping data [6] conditioned on external parameters, including temperature, photon energy, dynamical time delay and spin as resolved quantities, will generate comprehensive knowledge about the (non)equilibrium electronic structure of materials to benchmark theories.

Moreover, the reconstruction method is (iv) transferable to extracting the band dispersion of other quasiparticles (e.g. phonons [49], polaritons [50], etc [51]) in periodic systems, given the availability of corresponding multidimensional datasets. In fact, the application of our approach goes beyond band mapping data. (iv) Owing to the analogy between band mapping and spatially-resolved spectral imaging, which produces spatially-varying spectra, or $I(x, y, E)$, the reconstruction algorithm may be used to efficiently tease out the spatial (x, y) variation of the spectral shifts, which is complementary to the outcome of clustering algorithms [52]. Besides, the reconstruction effectively reduces the data size by over 5000 times from 3D band

mapping data to geometric features vectors (see Methods), facilitating database integration. Overall, the multidisciplinary methodology provides an example for building next-generation high-throughput materials characterization toolkits combining learning algorithms with physical knowledge [53] to arrive at a comprehensive understanding of materials properties unattainable before.

References

1. Isaacs, E. B. & Wolverton, C. Inverse Band Structure Design via Materials Database Screening: Application to Square Planar Thermoelectrics. *Chemistry of Materials* **30,** 1540–1546 (2018).
2. Marin, E. G., Perucchini, M., Marian, D., Iannaccone, G. & Fiori, G. Modeling of Electron Devices Based on 2-D Materials. *IEEE Transactions on Electron Devices* **65,** 4167–4179 (2018).
3. Bouckaert, L. P., Smoluchowski, R. & Wigner, E. Theory of Brillouin Zones and Symmetry Properties of Wave Functions in Crystals. *Physical Review* **50,** 58–67 (1936).
4. Chiang, T.-C. & Seitz, F. Photoemission spectroscopy in solids. *Annalen der Physik* **10,** 61–74 (2001).
5. Damascelli, A., Hussain, Z. & Shen, Z.-X. Angle-resolved photoemission studies of the cuprate superconductors. *Reviews of Modern Physics* **75,** 473–541 (2003).
6. Schönhense, G., Medjanik, K. & Elmers, H.-J. Space-, time- and spin-resolved photoemission. *Journal of Electron Spectroscopy and Related Phenomena* **200,** 94–118 (2015).
7. Medjanik, K. *et al.* Direct 3D mapping of the Fermi surface and Fermi velocity. *Nature Materials* **16,** 615–621 (2017).
8. 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).
9. Gauthier, A. *et al.* Tuning time and energy resolution in time-resolved photoemission spectroscopy with nonlinear crystals. *Journal of Applied Physics* **128,** 093101 (2020).
10. Riley, J. M. *et al.* Direct observation of spin-polarized bulk bands in an inversion-symmetric semiconductor. *Nature Physics* **10,** 835–839 (2014).
11. Bahramy, M. S. *et al.* Ubiquitous formation of bulk Dirac cones and topological surface states from a single orbital manifold in transition-metal dichalcogenides. *Nature Materials* **17,** 21–28 (2018).
12. Schröter, N. B. M. *et al.* Chiral topological semimetal with multifold band crossings and long Fermi arcs. *Nature Physics* (2019).
13. Valla, T. *et al.* Evidence for Quantum Critical Behavior in the Optimally Doped Cuprate $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Science* **285,** 2110–2113 (1999).

14. Levy, G., Nettke, W., Ludbrook, B. M., Veenstra, C. N. & Damascelli, A. Deconstruction of resolution effects in angle-resolved photoemission. *Physical Review B* **90,** 045150 (2014).

15. Zhang, P. *et al.* A precise method for visualizing dispersive features in image plots. *Review of Scientific Instruments* **82,** 043712 (2011).

16. He, Y., Wang, Y. & Shen, Z.-X. Visualizing dispersive features in 2D image via minimum gradient method. *Review of Scientific Instruments* **88,** 073903 (2017).

17. Peng, H. *et al.* Super resolution convolutional neural network for feature extraction in spectroscopic data. *Review of Scientific Instruments* **91,** 033905 (2020).

18. Kim, Y. *et al.* Deep learning-based statistical noise reduction for multidimensional spectral data. *Review of Scientific Instruments* **92,** 073901 (2021).

19. 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).

20. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* 1104 (MIT Press, 2012).

21. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521,** 452–459 (2015).

22. Wang, C., Komodakis, N. & Paragios, N. Markov Random Field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding* **117,** 1610–1627 (2013).

23. Comer, M. & Simmons, J. The Markov Random Field in Materials Applications: A synoptic view for signal processing and materials readers. *IEEE Signal Processing Magazine* **39,** 16–24 (2022).

24. Kaufmann, K. *et al.* Crystal symmetry determination in electron diffraction using machine learning. *Science* **367,** 564–568 (2020).

25. Stimper, V., Bauer, S., Ernstorfer, R., Scholkopf, B. & Xian, R. P. Multidimensional Contrast Limited Adaptive Histogram Equalization. *IEEE Access* **7,** 165437–165447 (2019).

26. Traving, M. *et al.* Electronic structure of WSe_2: A combined photoemission and inverse photoemission study. *Physical Review B* **55,** 10392–10399 (1997).

27. Finteis, T. *et al.* Occupied and unoccupied electronic band structure of WSe_2. *Physical Review B* **55,** 10400–10411 (1997).

28. Kormányos, A. *et al.* $\mathrm{k} \cdot \mathrm{p}$ theory for two-dimensional transition metal dichalcogenide semiconductors. *2D Materials* **2,** 022001 (2015).

29. Perdew, J. P. & Schmidt, K. *Jacob's ladder of density functional approximations for the exchange-correlation energy* in *AIP Conference Proceedings* **577** (AIP, 2001), 1–20.

30. Golze, D., Dvorak, M. & Rinke, P. The GW Compendium: A Practical Guide to Theoretical Photoemission Spectroscopy. *Frontiers in Chemistry* **7:377** (2019).

31. Zacharias, M., Scheffler, M. & Carbogno, C. Fully anharmonic nonperturbative theory of vibronically renormalized electronic band structures. *Physical Review B* **102,** 045126 (2020).

32. Zhang, D. & Lu, G. Review of shape representation and description techniques. *Pattern Recognition* **37,** 1–19 (2004).

33. Khotanzad, A. & Hong, Y. Invariant image recognition by Zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12,** 489–497 (1990).

34. Mahajan, V. N. & Dai, G.-m. Orthonormal polynomials in wavefront analysis: analytical solution. *Journal of the Optical Society of America A* **24,** 2994 (2007).

35. Heremans, J. P., Cava, R. J. & Samarth, N. Tetradymites as thermoelectrics and topological insulators. *Nature Reviews Materials* **2,** 17049 (2017).

36. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science,* 1900808 (2019).

37. Horton, M. K., Dwaraknath, S. & Persson, K. A. Promises and perils of computational materials databases. *Nature Computational Science* **1,** 3–5 (2021).

38. Kiureghian, A. D. & Ditlevsen, O. Aleatory or epistemic? Does it matter? *Structural Safety* **31,** 105–112 (2009).

39. Xian, R. P., Ernstorfer, R. & Pelz, P. M. Scalable multicomponent spectral analysis for high-throughput data annotation. *arXiv,* 2102.05604 (2021).

40. Nocedal, J. & Wright, S. J. *Numerical Optimization* 2nd ed. (Springer New York, 2006).

41. Smith, M. W. Roughness in the Earth Sciences. *Earth-Science Reviews* **136,** 202–225 (2014).

42. *Multi-Band Effective Mass Approximations* (eds Ehrhardt, M. & Koprucki, T.) (Springer, 2014).

43. Xian, R. P. *et al.* An open-source, end-to-end workflow for multidimensional photoemission spectroscopy. *Scientific Data* **7,** 442 (2020).

44. Noack, M. M. *et al.* Gaussian processes for autonomous data acquisition at large-scale synchrotron and neutron facilities. *Nature Reviews Physics* **3,** 685–697 (2021).

45. Beaulieu, S. *et al.* Ultrafast dynamical Lifshitz transition. *Science Advances* **7** (2021).

46. Curcio, D. *et al.* Accessing the Spectral Function in a Current-Carrying Device. *Physical Review Letters* **125,** 236403 (2020).

47. Wilson, N. R. *et al.* Determination of band offsets, hybridization, and exciton binding in 2D semiconductor heterostructures. *Science Advances* **3** (2017).

48. Ulstrup, S. *et al.* Nanoscale mapping of quasiparticle band alignment. *Nature Communications* **10,** 3283 (2019).

49. Ewings, R. *et al.* Horace : Software for the analysis of data from single crystal spectroscopy experiments at time-of-flight neutron instruments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **834,** 132–142 (2016).

50. Whittaker, C. E. *et al.* Exciton Polaritons in a Two-Dimensional Lieb Lattice with Spin-Orbit Coupling. *Physical Review Letters* **120,** 097401 (2018).

51. Frölich, A., Fischer, J., Wolff, C., Busch, K. & Wegener, M. Frequency-Resolved Reciprocal-Space Mapping of Visible Spontaneous Emission from 3D Photonic Crystals. *Advanced Optical Materials* **2,** 849–853 (2014).

52. Amenabar, I. *et al.* Hyperspectral infrared nanoimaging of organic samples based on Fourier transform infrared nanospectroscopy. *Nature Communications* **8,** 14402 (2017).

53. Von Rueden, L. *et al.* Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering,* 1–1 (2021).

Methods

Band mapping measurements. Photoemission band mapping of WSe_2 using multidimensional photoemission spectroscopy were conducted using a laser-driven, high harmonic generation-based extreme UV light source [8] operating at 21.7 eV and 500 kHz and a METIS 1000 (SPECS GmbH) momentum microscope featuring a delay-line detector coupled to a time-of-flight drift tube [7, 54]. Single crystal samples of WSe_2 ($> 99.995\%$ pure) were purchased from HQ graphene and were used directly for measurements without further purification. Before measurements, the WSe_2 samples were attached to the Cu substrate by conductive epoxy resin (EPO-TEK H20E). The samples were cleaved by cleaving pins attached to the sample surface upon transfer into the measurement chamber, which operates at an ambient pressure of 10^{-11} mbar during photoemission experiments. No effect of surface termination has been observed in the measured WSe_2 photoemission spectra, similar to previous experimental observations [10, 26]. For the valence band mapping experiments, the energy focal plane of the photoelectrons within the time-of-flight drift tube was set close to the top valence band.

Data processing and reconstruction. The raw data, in the form of single-electron events recorded by the delay-line detector, were preprocessed using home-developed software packages [43]. The events were first binned to the (k_x, k_y, E) grid with a size of $256\times256\times470$ to cover the full valence band range in WSe_2 within the projected Brillouin zone, which amounts

to a pixel size of ~ 0.015 Å^{-1} along the momentum axes and ~ 18 meV along the energy axis. The bin sizes are within the limits of the momentum resolution (< 0.01 Å^{-1}) and energy resolution (< 15 meV) of the photoelectron spectrometer [55].

Data binning is carried out in conjunction with the necessary lens distortion correction [56] and calibrations as described in [43]. The outcome provides a sufficient level of granularity in the momentum space to resolve the fine features in band dispersion while achieving higher signal-to-noise ratio than using single-event data directly. Afterwards, we applied intensity symmetrization to the data along the sixfold rotation symmetry and mirror symmetry axes [10] of the photoemission intensity pattern in the (k_x, k_y) coordinates, followed by contrast enhancement using the multidimensional extension of the contrast limited adaptive histogram equalization (MCLAHE) algorithm, where the intensities in the image are transformed by a look-up table built from the normalized cumulative distribution function of local image patches [25]. Finally, we applied Gaussian smoothing to the data along the k_x, k_y and E axes with a standard deviation of 0.8, 0.8 and 1 pixels (or about 0.012 Å^{-1}, 0.012 Å^{-1}, and 18 meV), respectively.

After data preprocessing, we sequentially reconstructed every energy band of WSe_2 from the photoemission data using the *maximum a posteriori* (MAP) approach described in the main text. The reconstruction requires tuning of three hyperparameters: (1) the momentum scaling and (2) the rigid energy shift to coarse-align the computed energy band, e.g. from density functional theory (DFT), to the photoemission data, and (3) the width of the nearest-neighbor Gaussian prior (η in Eq. (2)). The hyperparameter tuning is also carried out individually for each band to adapt to their specific environment. An example of hyperparameter tuning is given in Supplementary Fig. 4. The MAP reconstruction method involves optimization of the model parameters (i.e. band energy random variables), $\{\tilde{E}_{i,j}\}$ to maximize the posterior probability $p = p(\{\tilde{E}_{i,j}\})$ or to minimize the negative log-probability loss function, $\mathcal{L} := -\log p$, obtained from Eq. (2) as is used in our actual implementation.

$$\mathcal{L}(\{\tilde{E}_{i,j}\}) = -\sum_{i,j} \log I(k_{x,i}, k_{y,j}, \tilde{E}_{i,j}) + \sum_{(i,j),(l,m)|\mathrm{NN}} \frac{(\tilde{E}_{i,j} - \tilde{E}_{l,m})^2}{2\eta^2} + \text{const.} \qquad (3)$$

We implemented the optimization using a parallelized version of the iterated conditional mode (ICM) [57] method in Tensorflow [58] in order to run on multicore computing clusters and GPUs. The parallelization involves a checkerboard coloring scheme (or coding method) of the graph nodes [59] and subsequent hierarchical grouping of colored nodes, which allows

alternating updates on different subgraphs (i.e. subsets of the nodes) of the Markov random field during optimization. Typically, the optimization process in the reconstruction of one band converges within and therefore is terminated after 100 epochs, which takes ~ 7 seconds on a single NVIDIA GTX980 GPU for the above-mentioned data size. Details on the parallelized implementation are provided in section S1 of the Supplementary Information. In addition, because symmetry information is not explicitly included in the MRF model, the reconstructed bands generally requires further symmetrization as refinement or post-processing to be ready for database integration.

We described our approach of using band structure calculations to initialize the MAP optimization as a warm start. The term "warm start" in the context of numerical optimization generally refers to the initialization of an optimization using the outcome of an associated and yet more solvable problem (e.g. surrogate model) obtained beforehand that yields an approximate answer, instead of starting from scratch (i.e. cold start). Warm-starting an optimization improves the effective use of prior knowledge and its convergence rate [40]. In the current context, we regard the band structure reconstruction from photoemission band mapping data as the optimization problem to warm start, and the outcome from an electronic structure calculation can produce a sufficiently good approximate to the solution of the optimization problem. For WSe_2, straightforward DFT calculations with semi-local approximation (which in itself involves explicit optimizations such as geometric optimization of the crystal structures) are sufficient, but our approach is not limited to DFT. Therefore, the use of "warm start" in our application is conceptually well-aligned with the origin of the term.

To validate the MAP reconstruction algorithm in a variety of scenarios, we used synthetic photoemission data where the nominal ground-truth band structures are available. The band structures are constructed using analytic functions, model Hamiltonians or DFT calculations. The initializations are generated by tuning the numerical parameters used to generate the ground-truth band structures. The procedures and results are presented in section S2 of the Supplementary Information. In simple cases, such as single or well-isolated bands, the reconstruction yields a close solution to the ground truth even with a flat band initialization. In the more general multiband scenario with congested bands and band crossings (or anti-crossings), an approximate dispersion (or shape) of the band and the crossing information is required in the initialization (i.e. warm start) in order to converge to a realistic solution. We further tested the

robustness of the initializations by (1) scaling the energies of the ground truth and by (2) using DFT calculations with different exchange-correlation (XC) functionals, in order to capture sufficient variability of available band structure calculations in the real world. We quantify the variations in the initializations and the performance of the reconstruction using the average error (Eq. (8), or Fig. 3b), calculated with respect to the ground truth. Among the different numerical experiments, we find that the optimization converges consistently to a set of bands that better matches the experimental data than the initialization. This is manifest in that the average errors of the initializations are reduced to a similar level in the corresponding reconstruction outcomes, a trend seen over all bands regardless of their dispersion. In the synthetic data with an energy spacing of $\sim$ 18 meV, the average error in the reconstruction is on the order of 40-50 meV for each band, which amounts to an average inaccuracy of < 3 bins along the energy dimension at a momentum location. The inaccuracy is, however, dependent on the bin sizes used in the preprocessing and the fundamental resolution in the experiment. We have made the code for the MAP reconstruction algorithm and the synthetic data generation publicly accessible from the online repository fuller [60] for broader applications.

Band structure calculations. Electronic band structures were calculated within (generalized) DFT using the local density approximation (LDA) [61, 62], the generalized-gradient approximation (GGA-PBE) [63] and GGA-PBEsol [64]), and the hybrid XC functional HSE06 [65], which incorporates a fraction of the exact exchange. All calculations were performed with the all-electron, full-potential numeric-atomic orbital code, FHI-aims [66]. They were conducted for the geometries obtained by fully relaxing the atomic structure with the respective XC-functional to keep the electronic and atomic structures consistent. Spin-orbit coupling was included in a perturbational fashion [67]. The momentum grid used for the calculation was equally sampled with a spacing of 0.012 Å^{-1} in both k_x and k_y directions that covers the irreducible part of the first Brillouin zone at k_z = 0.35 Å^{-1}, estimated using the inner potential of WSe_2 from a previous measurement [10]. The calculated band structure is symmetrized to fill the entire hexagonal Brillouin zone to be used to initialize the band structure reconstruction and synthetic data generation. We note here that for the MAP reconstruction, the momentum grid size used in theoretical calculations (such as DFT at various levels used here) need not be identical to that of the data (or instrument resolution) and in those cases an appropriate upsampling (or downsampling) should be applied to the calculation to match their momentum resolution.

Further details are presented in section S4 of the Supplementary Information.

Band structure informatics. The shape feature space representation of each electronic band is derived from the decomposition,

$$E_b(\mathbf{k}) = \sum_l a_l \phi_l(\mathbf{k}) = \mathbf{a} \cdot \Phi \tag{4}$$

Here, $\mathbf{k} = (k_x, k_y)$ represents the momentum coordinate, $E_b(\mathbf{k})$ is the single-band dispersion relation (e.g. dispersion surface in 3D), a_l and $\phi_l(\mathbf{k})$ are the coefficient and its associated basis term, respectively. They are grouped separately into the feature vector, $\mathbf{a} = (a_1, a_2, ...)$, and the basis vector, $\Phi = (\phi_1, \phi_2, ...)$. The orthonormality of the basis is guaranteed within the projected Brillouin zone (PBZ) of the material.

$$\int_{\mathbf{k}\in\Omega_{\text{PBZ}}} \phi_m(\mathbf{k})\phi_n(\mathbf{k})\, d\mathbf{k} = \delta_{mn} \tag{5}$$

For the hexagonal PBZ of WSe_2, the basis terms are hexagonal Zernike polynomials (ZPs) constructed using a linear combination of the circular ZPs via Gram-Schmidt orthonormalization within a regular (i.e. equilateral and equiangular) hexagon [34]. A similar method can be used to generate ZP-derived orthonormal basis adapted to other boundary conditions [34]. The representation in feature space [33] provides a way to quantify the difference (or distance) d between energy bands or band structures at different resolutions or scales without additional interpolation. To quantify the shape similarity between energy bands E_b and $E_{b'}$, we calculate the cosine similarity using the feature vectors,

$$d_{\cos}(E_b, E_{b'}) = \frac{\mathbf{a} \cdot \mathbf{a}'}{|\mathbf{a}| \cdot |\mathbf{a}'|}, \tag{6}$$

The cosine similarity is bounded within $[-1, 1]$, with a value of 0 describing orthogonality of the feature vectors and a value of 1 and -1 describing parallel and anti-parallel relations between them, respectively, both indicating high similarity. The use of cosine similarity in feature space allows comparison of dispersion while being unaffected by their magnitudes. In comparing the dispersion between single energy bands using Eq. (6), the first term in the polynomial expansion, or the hexagonal equivalent of the Zernike piston [68], is discarded as it only represents a constant energy offset (with zero spatial frequency) instead of dispersion, which is characterized by a combination of finite and nonzero spatial frequencies.

The electronic band structure is a collection of energy bands $E_B = \{E_{b_i}\}$ $(i = 1, 2, ...)$. To quantify the distance between two band structures, $E_{B_1} = \{E_{b_{1,i}}\}$ and $E_{B_2} = \{E_{b_{2,i}}\}$, containing the same number of energy bands while ignoring their global energy difference, we first subtract the energy grand mean (i.e. mean of the energy means of all bands within the region of the band structure for comparison). Then, we compute the Euclidean distance, or the ℓ^2-norm, for the ith pair of bands, $d_{b,i}$.

$$d_{b,i}(E_{b_{1,i}}, E_{b_{2,i}}) = \|\tilde{\mathbf{a}}_{1,i} - \tilde{\mathbf{a}}_{2,i}\|_2 = \sqrt{\sum_l (\tilde{a}_{1,il} - \tilde{a}_{2,il})^2}. \tag{7}$$

Here, $\tilde{\mathbf{a}}$ denotes the feature vector after subtracting the energy grand mean so that any global energy shift is removed. We define the band structure distance as the average distance over all N_b pairs of bands, or $d_B(E_{B_1}, E_{B_2}) = \sum_i^{N_b} d_{b,i}(E_{b_1,i}, E_{b_2,i})/N_b$. The values of $d_B(E_{B_1}, E_{B_2})$ are shown in the upper triangle of Fig. 3d and their corresponding standard errors (over the 14 valence bands of WSe_2) in the lower triangle. The distance in Eq. (7) is independent of basis and allows energy bands calculated on different resolutions or from different materials with the same symmetry (e.g. differing only by Brillouin zone size) to be compared.

We use same-resolution error metrics to evaluate the approximation quality of the expansion basis and to quantify the reconstruction outcome with a known ground-truth band structure. Specifically, we define the average approximation error (with energy unit), η_{avg}, for each energy band using the energy difference at every momentum location,

$$\eta_{\text{avg}}(E_{\text{approx}}, E_{\text{recon}}) = \sqrt{\frac{1}{N_k} \sum_{\mathbf{k} \in \Omega_{\text{PBZ}}} (E_{\text{approx},\mathbf{k}} - E_{\text{recon},\mathbf{k}})^2}, \tag{8}$$

where N_k is the number of momentum grid points and the summation runs over the projected Brillouin zone. In addition, we construct the relative approximation error, η_{rel}, following the definition of the normwise error [69] in matrix computation,

$$\eta_{\text{rel}}(E_{\text{approx}}, E_{\text{recon}}) = \frac{\|E_{\text{approx}} - E_{\text{recon}}\|_2}{\|E_{\text{recon}}\|_2}. \tag{9}$$

Eq. (8)-(9) are used to compute the curves in Fig. 3b as a function of the number of basis terms included in the approximation. The relevant code for the representation using hexagonal ZPs and the computation of the metrics is also accessible in the public repository fuller [60].

Data reduction. The raw data and intermediate results are stored in the HDF5 format [43]. The file sizes quoted here for reference are calculated from storage as double-precision floats or integers (for indices). The photoemission band mapping data of WSe_2 ($256\times256\times470$ bins) have a size of about 235 MB (240646 kB) after binning from single-event data (7.8 GB or 8176788 kB). The reconstructed valence bands at the same resolution occupy about 3 MB (3352 kB) in storage, and the size further decreases to 46 kB when we store the shape feature vector associated with each band. If only the top-100 coefficient (ranked by the absolute values of their amplitudes) and their indices in the feature vectors are stored, the data amounts to 24 kB. For the case of WSe_2, the top-100 coefficients can approximate the band dispersion with a relative error (see Eq. (9)) of $< 0.8\%$ for every energy band, as shown in Supplementary Fig. 14.

References

54. Oelsner, A. *et al.* Microspectroscopy and imaging using a delay line detector in time-of-flight photoemission microscopy. *Review of Scientific Instruments* **72,** 3968–3974 (2001).
55. SPECS GmbH. *METIS 1000 Brochure* https://www.specs-group.com/fileadmin/user_upload/products/brochures/SPECS_Brochure-METIS_RZ_web.pdf. 2019.
56. 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).
57. Kittler, J. & Föglein, J. Contextual classification of multispectral pixel data. *Image and Vision Computing* **2,** 13–29 (1984).
58. Martín Abadi *et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* Software available from tensorflow.org. 2015.
59. Li, S. *Markov Random Field Modeling in Image Analysis* 3rd ed. (Springer, 2009).
60. Stimper, V. & Xian, R. P. *fuller* https://github.com/mpes-kit/fuller.
61. Ceperley, D. M. & Alder, B. J. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.* **45,** 566–569 (7 1980).
62. Perdew, J. P. & Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **45,** 13244–13249 (23 1992).
63. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **77,** 3865–3868 (1996).
64. Perdew, J. P. *et al.* Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Physical Review Letters* **100,** 136406 (2008).

65. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *The Journal of Chemical Physics* **118,** 8207–8215 (2003).

66. Blum, V. *et al.* Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* **180,** 2175–2196 (2009).

67. Huhn, W. P. & Blum, V. One-hundred-three compound band-structure benchmark of post-self-consistent spin-orbit coupling treatments in density functional theory. *Physical Review Materials* **1,** 033803 (2017).

68. Wyant, J. C. & Creath, K. in *Applied Optics and Optical Engineering* 1–53 (Academic Press, Inc., 1992).

69. Watkins, D. S. *Fundamentals of matrix computations* 3rd, 644 (Wiley, 2010).

Acknowledgments

We thank M. Scheffler for fruitful discussions and S. Schülke, G. Schnapka at Gemeinsames Netzwerkzentrum (GNZ) in Berlin and M. Rampp at Max Planck Computing and Data Facility (MPCDF) in Garching for support on the computing infrastructure. The work was partially supported by BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials-Science, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant No. 740233 and Grant No. ERC-2015-CoG-682843), the German Research Foundation (DFG) through the Emmy Noether program under grant number RE 3977/1, the SFB/TRR 227 "Ultrafast Spin Dynamics" (projects A09 and B07), and the NOMAD pillar of the FAIR-DI e.V. association. We thank M. Bremholm for providing the Bi_2Te_2Se samples, Ph. Hofmann and M. Bianchi for their support in obtaining Au(111) photoemission data. M. Dendzik acknowledges support from the Göran Gustafssons Foundation. S. Beaulieu acknowledges the financial support of the Banting Fellowship from the Natural Sciences and Engineering Research Council (NSERC) in Canada.

Authors contributions

R.P.X. and R.E. conceived the project. S.D., M.D. and Sa.B. performed the photoemission band mapping experiment. M.Z., M.D. and C.C. performed the DFT band structure calculations. R.P.X. processed the raw data, devised the band structure digitization, algorithm validation schemes and metrics, and performed computational benchmarking. V.S. designed and implemented the machine learning algorithm under the supervision of St.B. and B.S. with inputs from

R.P.X.. R.P.X., V.S. and M.Z. co-wrote the first draft of the manuscript. All authors contributed to discussion and revision of the manuscript to its final version.

Data availability

The electronic structure calculations are available from the NOMAD repository (DOI: 10.17172/NOMAD/2020.03.28-1). The photoemission dataset used in this work is currently available at https://fhi-cloud.gnz.mpg.de/index.php/s/w7FBWHipCp2mDrQ and will be made publicly available on Zenodo.

Code availability

The code developed for band structure reconstruction has been made available at https://github.com/mpes-kit/fuller.

Competing interests

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryVideo1WSe23DReconkx.mp4
- SupplementaryVideo2WSe23DReconky.mp4
- SupplementaryVideo3WSe23DReconviews.mp4
- nreditorialpolicychecklistNATCOMPUTSCI220208.pdf
- XianetalMLrouterevisedSI.pdf