npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Deep learning approach to genome of two-dimensional materials with flat electronic bands

A. Bhattacharya[1,2]✉, I. Timokhin[1], R. Chatterjee[3], Q. Yang[1]✉ and A. Mishchenko[1]✉

Electron-electron correlations play central role in condensed matter physics, governing phenomena from superconductivity to magnetism and numerous technological applications. Two-dimensional (2D) materials with flat electronic bands provide natural playground to explore interaction-driven physics, thanks to their highly localized electrons. The search for 2D flat band materials has attracted intensive efforts, especially now with open science databases encompassing thousands of materials with computed electronic bands. Here we automate the otherwise daunting task of materials search and classification by combining supervised and unsupervised machine learning algorithms. To this end, convolutional neural network was employed to identify 2D flat band materials, which were then subjected to symmetry-based analysis using a bilayer unsupervised learning algorithm. Such hybrid approach of exploring materials databases allowed us to construct a genome of 2D materials hosting flat bands and to reveal material classes outside the known flat band paradigms.

npj Computational Materials (2023)9:101 ; https://doi.org/10.1038/s41524-023-01056-x

INTRODUCTION

Electrons in some materials have dispersionless spectrum, i.e. their energy $E_\mathbf{k}$ is independent of momentum $\mathbf{k}$, resulting in the formation of flat band, $E_\mathbf{k} \approx$ constant. Vanishing group velocity, $\nabla_\mathbf{k} E \approx 0$, suppresses kinetic energy, making flat bands conducive to electron-electron interactions[1]. Flat bands originate from spatial localization of wavefunctions in, for instance, *f*-orbitals of lanthanides and actinides, lattice sites of atomic insulators, line- and split-graphs of bipartite lattices, or twisted bilayer graphene superlattices[2,3], where controlled engineering of the flat band has been demonstrated. Flat bands often exhibit topological non-triviality, resulting in a plethora of exotic physics like unconventional superconductivity in twisted bilayer graphene[4], quantum anomalous Hall effect[5], anomalous Landau levels[6], strongly correlated Chern insulators[7], Wigner crystallization[8], unusual ferromagnetism[9], chiral plasmons[10], or bulk photovoltaic effect[11], and the list goes on. Most of these were reported in two-dimensional (2D) materials and have triggered intense search for new 2D materials hosting flat bands and strongly-correlated physics. Some have focused on theoretical 2D lattices hosting dispersionless bands, like Lieb, kagome, or dice[12]. Meanwhile, progress in open materials science platforms, such as Materials Project[13] and Aflow[14], provides thousands of possible exfoliable 2D candidates from their 3D counterparts[15–17], forming large databases made solely of 2D materials, like C2DB[18], 2D materials encyclopedia[19], and MC2D[16]. 2D lattices hosting flat bands, mostly theoretically, can then be identified for further investigation. Alternatively, a symmetry-based approach resorting to known flat dispersions in line- and split-graphs of bipartite lattices could also help to predict new 2D lattices[1,20]. However, with a limited number of known 2D flat band materials, current symmetry-based studies lack the statistical significance to make accurate predictions.

In this work, we use a combination of these two approaches – first, identify 2D flat band materials in the vast database using the convolutional neural network deep learning approach; the identified lattices are then classified based on their structural fingerprints, employing symmetry-based clustering. Our deep learning algorithm surveys the entire database with high throughput and precision. The resulting flat band sublattices are presented as clustering charts, serving as a roadmap for 2D flat band lattices in the future. Our work, therefore, offers a comprehensive and efficient path towards 2D flat band materials and, more importantly, the exciting physics they enable.

The search of flat dispersions in 2D materials has been attempted, where band flatness was defined by an arbitrary fixed bandwidth[21,22]. However, the identification of flat bands is not a straightforward task, even for simple band structures. When an electronic band is parametrized, the band index is determined by the order in which it appears in the energy scale. However, crossings between bands can change the band index, which could lead to large bandwidth even when a flat band exists. Hence, using parametrized bands and predefined bandwidth to identify flat bands may largely underestimate their number, therefore, risk overlooking potential flat band materials. To account for this, we put forward an unusual yet more inclusive approach, utilizing the band structure images from a database rather than the parametrized bands themselves. Current databases still lack the complete description of wavefunctions, which is a prerequisite for systematic classification of the electronic bands using machine learning algorithm based on, for example, topology, or other features[23–26]. However, the easily accessible band structure images offer a timely alternative to help identifying flat features in the bands, allowing us to harvest these 2D databases already at such an early stage while more development in data science to enrich their contents is on the way.

Our deep-learning-assisted framework enables high-throughput identification of flat bands from any database with images of band structures. Here we choose 2D Materials Encyclopedia (2Dmatpedia), currently the largest open 2D materials database, as the

[1]Department of Physics and Astronomy, University of Manchester, Manchester, UK. [2]Department of Mechanical Engineering, Indian Institute of Technology Delhi, New Delhi, India. [3]Department of Physics, Indian Institute of Technology Delhi, New Delhi, India. ✉email: anupamcounting@gmail.com; qian.yang@manchester.ac.uk; artem.mishchenko@gmail.com

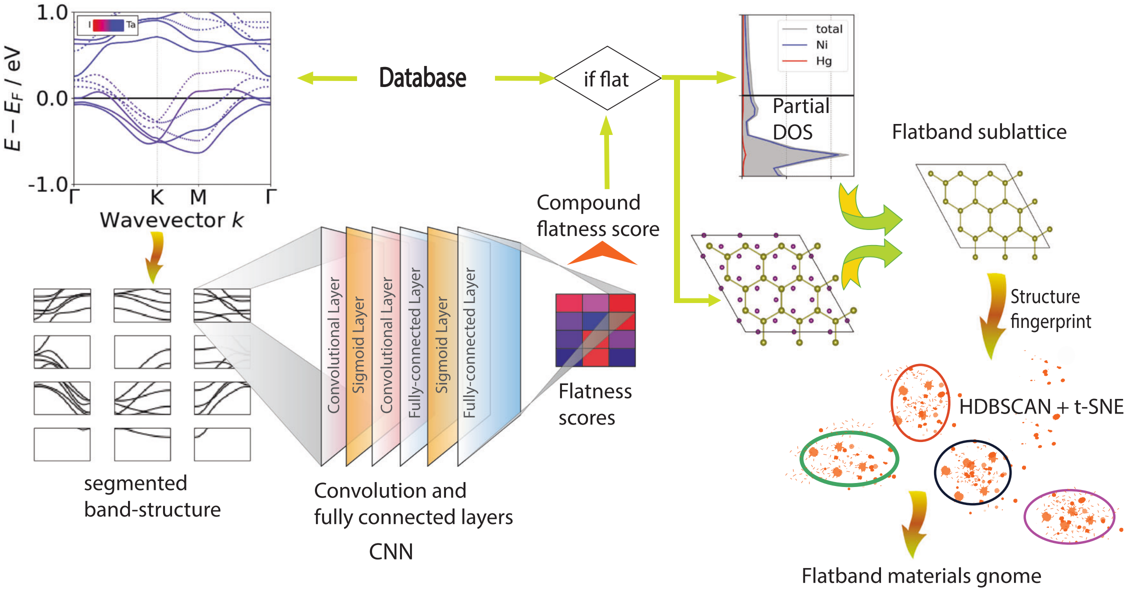

Fig. 1 Architecture of combined supervised and unsupervised machine learning algorithms used in this work. CNN was trained to identify flat band materials using segmented band structure images from the database, followed by identifying sublattices that are responsible for flat dispersion using element projected DOS. Then, density-based clustering combined with t-SNE was used to classify the assigned structural fingerprints, and to identify classes of flat band 2D materials.

source of the band structure images. A convolutional neural network (CNN) was trained and applied to identify a genome of flat band materials across 2Dmatpedia. Afterwards, symmetry-based analysis using unsupervised machine learning algorithm was employed to classify flat-band materials into clusters, based on the structural fingerprints of identified corresponding sublattices. This classification framework helps to retain subtle connections among structural nuances, enabling the identification of hierarchical classes within the flat band 2D materials genome. Perhaps more importantly, it allows the prediction of flat band 2D materials outside the known paradigm, quickly expanding our database of 2D flat band materials. Our work, therefore, presents a path towards the exploration of the rich and exciting interdisciplinary field across data science and physics.

RESULTS

Our protocol for the hybrid approach (supervised deep learning for flat band identification + unsupervised machine learning for symmetry-based clustering) is outlined in Fig. 1. First, CNN is trained and used to identify flat bands by examining all the band structure images across 2Dmatpedia, allowing us to reveal materials with flat bands spanning the whole Brillouin zone (BZ). Second, a crystal sublattice which has the maximum projected density of states (DOS) near the flat band is assigned as the sublattice responsible for the flat dispersion. Third, a structure descriptor algorithm CrystalNNFingerprint[27] is used to calculate structural similarity of the sublattices with predefined local coordination templates and assign a structure fingerprint. Finally, the vectorized fingerprints of the structures were fed into a bilayer unsupervised clustering module, with density-based hierarchical algorithm HDBSCAN[28] as the inner layer and t-distributed Stochastic Neighbourhood Embedding[29] (t-SNE) as the output layer. This module clustered the flat band material population into stratified isostructural groups. All the developed code is available at https://github.com/Anupam-Bh/ML_2D_flat_band.

Identification of flat bands using CNN

For flat band recognition, we used CNN, which was trained on a subset of the Materials Project database, see Methods (Convolutional neural network). Figure 2a highlights one of the essential steps, image segmentation, which provides two main benefits. First, vertical segmentation allows identifying the location of a flat band with respect to the Fermi energy. Secondly, horizontal segmentation of the band structures at high-symmetry points enables us to find whether the dispersion is plane-flat (flat in the whole plane) or line-flat (flat only along one direction) across the BZ.

In this study we focused on plane-flat 2D materials, as their electrons have suppressed kinetic energy in wider momentum space, promoting electron-electron interactions. Figure 3a shows an example of plane-flat band material. For comparison, we show an example of line-flat band in Fig. 2a, where two flat segments (outlined in red boxes) belong to different energy strips. After all the materials with plane-flat bands have been identified (scored more than 0.5 along all the k-paths within an energy bandwidth), each material is assigned a compound flatness score between 0 to 1, which can be used to estimate the relative flatness of bands. The number of materials as a function of their compound flatness score is shown in Fig. 2b, with a major peak near flatness score of 1. We found 2127 plane-flat materials out of total 5270 materials in 2Dmatpedia.

Identification of flat band sublattices

To build a connection between the identified flat band materials and their structural features, we propose a conjecture based on spatial localization of electrons forming a flat band. Let us assume a spatially localized region outside which electronic density approaches zero, formed either by the interference of wavefunctions or through other mechanisms. If this localized state originates from orbitals from multiple types of atoms, to keep the eigenvalue constant throughout the span of the localized state (plane-flat materials), there has to exist an accidental degeneracy between the states of those multiple types of atoms. Therefore, for most of the observed flat bands, the constituent electrons originate from a single element in the compound, assuming that the probability of accidental band degeneracy is low. Our conjecture is supported by several studies showing that it is the elemental sublattices which obey specific lattice and orbital symmetries that lead to flat bands[22,30–32]. For example, in intermetallic CoSn, the flat band is attributed to the kagome sublattice of the transition metal element[30], and in HgF_2, it is

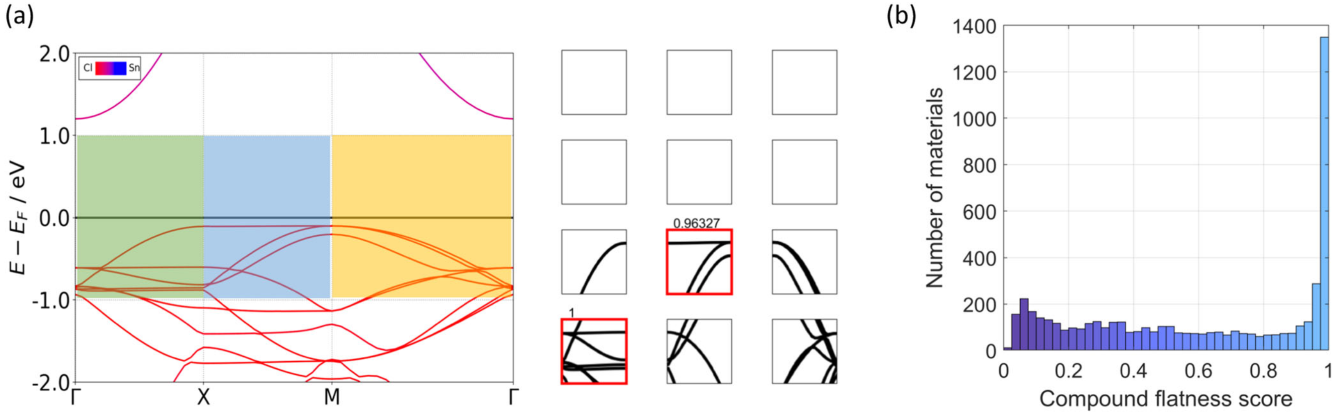

Fig. 2 Identification of flat bands from band structure images using supervised machine learning. **a** Segmentation of band structure images from 2Dmatpedia, and identification of flat band segments. Left panel: segmentation of band structure image of $SnCl_4$ [2dm-13] horizontally into four 0.5 eV energy strips, and vertically along high symmetry points in k-space, as denoted in green, blue, and yellow for Γ → X, X → M, and M → Γ paths, respectively. Right panel: segmented band structure in the [−1,1] eV range and predicted outputs. Segments with flat bands are marked in red frames their corresponding flatness score is shown on top. **b** Histogram of compound flatness scores of all the materials in 2Dmatpedia.

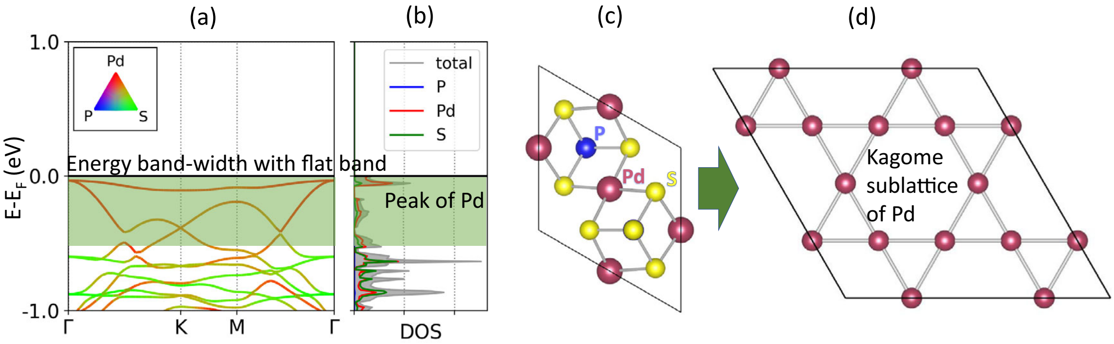

Fig. 3 Identification of flat band sublattices. **a** Band structure of an identified flat band material example $Pd_3P_2S_8$. **b** Element projected DOS from **a** reveals the element responsible for flat band is Pd. **c** The crystal structure of $Pd_3P_2S_8$. **d** Kagome sublattice of Pd remains after stripping off P and S elements from $Pd_3P_2S_8$ structure.

the diamond-octagon sublattice of mercury[21]. Furthermore, supplementary Note IV and Supplementary Figure 3 demonstrates that the statistical distribution of elemental contributions in the segments containing flat bands strongly supports the conjecture presented above.

This conjecture allows us to use flat band element sublattice instead of full crystal structure for further analysis, largely simplifying the process while maintaining high accuracy. To this end, we first extract the element sublattice with the highest orbital contribution to the flat band. Here, element projected DOS (Fig. 3b) is used to identify the element with maximum density contribution, ignoring the orbital mixing for classification purposes. As shown in Fig. 3, using flat band material $Pd_3P_2S_8$ as an example, after the energy segment containing flat band is identified (green segment in Fig. 3a, b), the corresponding bandwidth in the element projected DOS is analyzed to obtain the element which has maximum projected DOS, and its sublattice. In this case, it is element Pd, and its sublattice, as shown in Fig. 3d. The lattice structure of $Pd_3P_2S_8$ (Fig. 3c) is stripped off of P and S atoms to segregate the Kagome sublattice of Pd (Fig. 3d). As we conjecture, the symmetry operations relevant to this elemental sublattice lead to the flat band, thus only the sublattice with the chosen element for each compound is kept for further analysis.

The extracted flat band sublattices are then subjected to a structure descriptor, CrystalNNFingerprint[27] and represented as a 244-dimensional vector to facilitate further classification using unsupervised machine learning algorithms, see Methods (Sublattice extraction and vectorization).

Unsupervised machine learning: bilayer clustering

The vectorized flat band sublattices were further classified based on their structural fingerprints. We use a complementary bilayer classification algorithm to achieve optimal clustering. Density-based algorithm HDBSCAN[28] is used to obtain hierarchical information and to identify closely-packed clusters, while t-SNE[29] was compensating for the tendency of HDBSCAN to overlook local neighborhood information of the clusters. Details are given in the Clustering module section in Methods.

Two parameters of HDBSCAN, minimum cluster size (MS), and sample size for density calculation (SS), were tuned to obtain optimal clustering solutions. The MS parameter trims off clusters which are smaller than its value and marks their members as unclassified, while the density around each point is calculated as a reciprocal relation to the distance of the point from its SS-th nearest neighbor. Both MS and SS were varied in a wide range, from 3 to 20, to obtain optimal clustering solutions. Performance of the clustering algorithm was quantified using two indices, density-based clustering validation (DBCV)[33], and cluster validity index (S_Dbw)[34]. DBCV calculates the intra-cluster and inter-cluster densities to estimate affinity between objects inside a cluster in comparison to connectivity between clusters. A higher

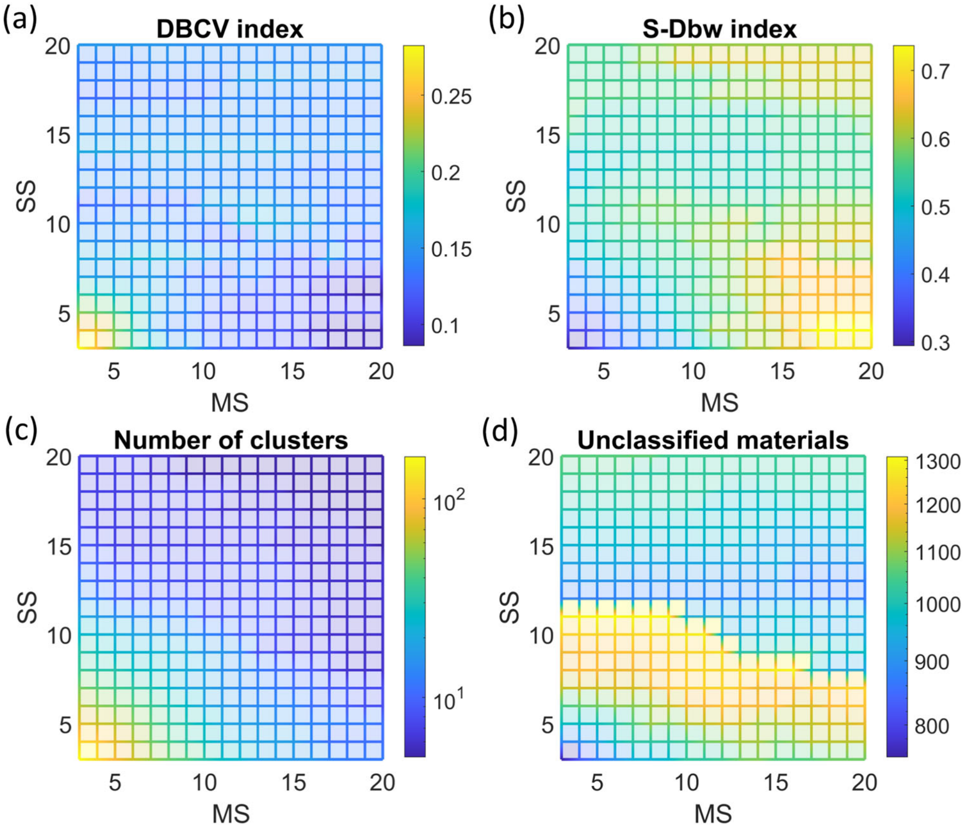

Fig. 4 HDBSCAN algorithm optimization. Graphs showing **a** DBCV index, **b** S_Dbw indices, **c** number of clusters, and **d** number of unclassified materials as functions of minimum cluster size (MS) and sample size for density calculation (SS).

DBCV index indicates a better clustering solution. The S_Dbw, on the other hand, is expressed as the sum of intra-cluster variance as a measure of cluster compactness, and inter-cluster density as a measure of separation. A smaller S_Dbw index marks a better clustering solution. Figure 4a–d shows the DBCV and S_Dbw indices, the number of clusters, and the number of unclassified materials for different values of MS and SS. We notice that smaller values of MS and SS result in better DBCV and S_Dbw, as well as a higher number of clusters. The number of unclassified materials in Fig. 4d, however, demonstrates a non-monotonic behavior with MS and SS: it first increases, caused by the existence of groups of fingerprints that are difficult to classify; then drastically drops, implying that many such dispersed fingerprints suddenly get included into clusters, resulting in bad classification. Therefore, the choice of MS and SS parameters is expected to be smaller than the values at which such sudden change in Fig. 4d occurs, in order to obtain the optimal solution.

For HDBSCAN, we choose MS = 7 and SS = 6 to map the coordination patterns into the structure fingerprint space, yielding a total of 51 clusters and 1448 unclassified materials. Our results are shown as a phylogenetic tree of clusters in Fig. 5a, and as a 2D t-SNE representation in Fig. 5b. A relatively small number of clusters facilitates the identification of major groups of structure fingerprints, although it yields a higher number of unclassified materials. We also tried to classify our results using a finer clustering solution with MS = 4 and SS = 3. Reduced MS and SS values improve the performance of the HDBSCAN and give a much higher number of clusters (131 clusters). Detailed classification of all the flat band sublattice structures for MS = 4 and SS = 3 is provided in Supplementary Note II, which includes Supplementary Fig. 1 and Supplementary Tables II–IV.

Identifying structural groups from the HDBSCAN clusters has been largely possible because of the soft clustering feature of HDBSCAN. This density-based algorithm attaches a probability (representing the chance of being a member of a cluster; more about this membership probability in Methods) to each member of a cluster. The members with perfect probabilities or 'exemplars' help the direct identification of sublattice represented by a cluster. Thus, we find exemplar sublattice structures from each cluster; however, often different clusters which are in close proximity, represent almost identical sublattices. To identify this local neighborhood, we use t-SNE. We combine the sublattice structures from nearly identical cluster exemplars to obtain the total list of lattice structures responsible for flat dispersion. For MS = 7 and SS = 6, we identify 27 such sublattice structures which are given in the Supplementary Table I.

For the optimized t-SNE representation, distance information of the 90 nearest neighbors for each structure fingerprint is preserved using perplexity = 30. The principal component analysis is used to calculate the projection plane to retain maximal global neighborhood information. The learning rate is automatically adjusted according to

$$\text{learning rate} = \frac{\text{sample size}}{\text{early exaggeration factor}} \quad (1)$$

where the early exaggeration factor is 12 for the first 250 iterations[35,36]. The total number of iterations was kept at 10,000 to ensure convergence of t-SNE.

Evolution of the identified coordination patterns and lattice structures

The phylogenetic tree in Fig. 5a shows the hierarchy among identified clusters of structurally similar lattices, which maps into the t-SNE plot in Fig. 5b.

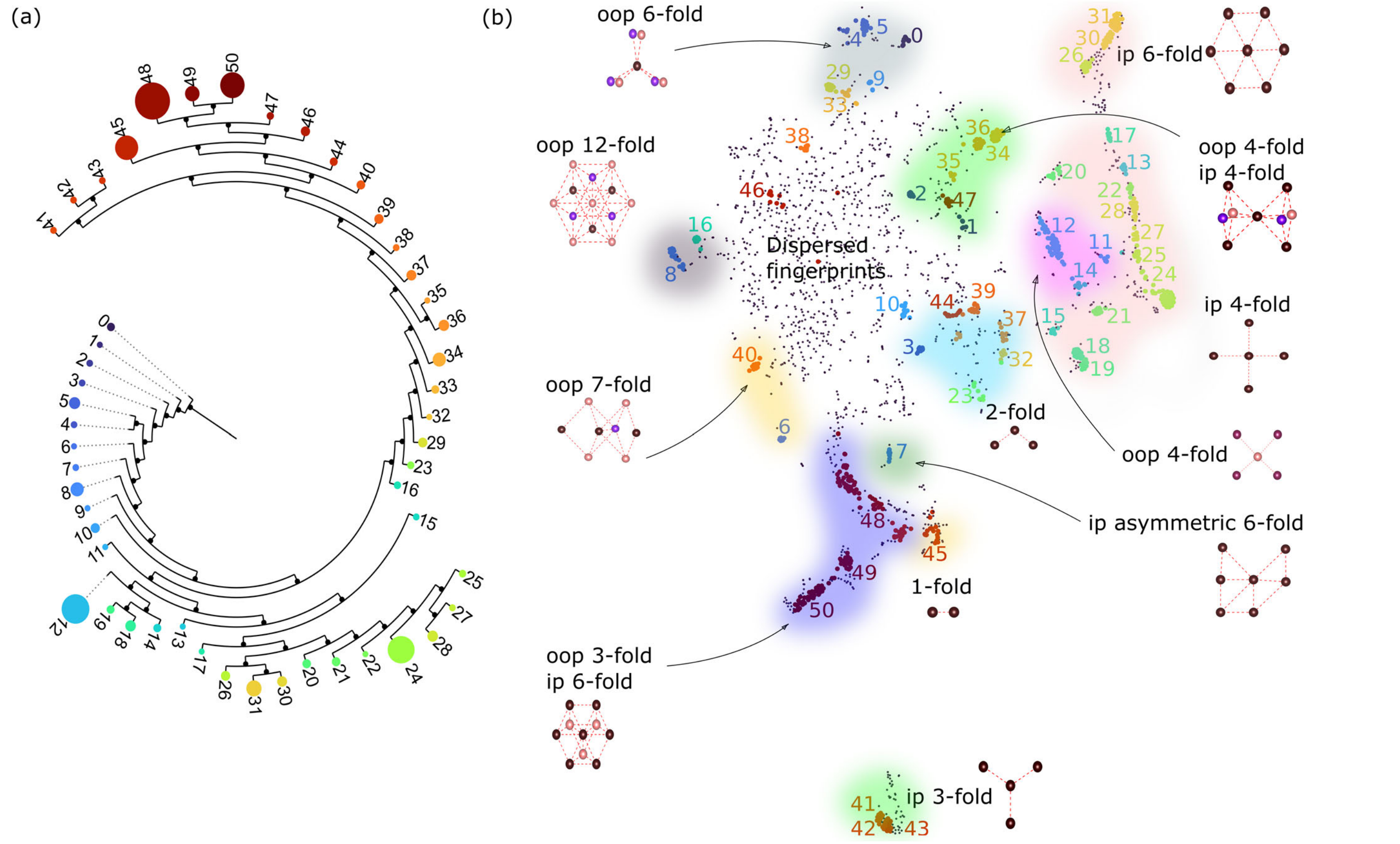

Fig. 5 Clustering solution from HDBSCAN with MS = 7, SS = 6. a Phylogenetic tree expression of the hierarchical relations among clusters represented by colored circles. The size of each circle is proportional to the corresponding cluster's volume, while the cluster index is indicated by the number nearby. **b** t-SNE 2D visualization of the structure fingerprint space, different coordination patterns are color-coded; ip is in-plane and oop is out-of-plane coordination. Insets are representative coordination templates, where atoms in different atomic planes of the templates are colored differently.

By analyzing the clusters, 12 coordination patterns are revealed, shown as shaded regions in Fig. 5b. Fuzzy boundaries between different coordination patterns develop, as the coordination in each sublattice evolves gradually throughout the embedded space. Most well-defined isostructural clusters are found near the edges in the t-SNE plot, where the structure fingerprints provide a satisfactory description of underlying lattice geometries. On the contrary, many dispersed unclassified fingerprints appear closer to the central region in the t-SNE plot. This can be ascribed to lattices with large unit cells, which can be described by multiple coordination patterns (typically, more than 3–4) in their lattice-sites. The average of their multiply-coordinated structure fingerprints results in lattices that struggle to be properly included in any single coordination geometry. Examples of such dispersed fingerprints are seen in $LiSb(PO_3)_4$ with 24 oxygen sites, or $AlPI_8$ with 16 iodine sites. Overall, finding an adequate structure descriptor for such large-unit-cell lattices is still an open problem[37]. Some unclassified examples are also found lying near clusters despite sharing similar structures to those within the well-defined clusters, because of density-based segregation.

Let us highlight the main features of these plots. We first notice that clusters #41-43, corresponding to honeycomb lattice materials with planar 3-fold coordination of atoms, clearly stand out at the end of the phylogenetic tree, Fig. 5a, and are positioned at the bottom of the t-SNE plot, Fig. 5b. Their separation from the rest of the charts marks the distinctness of their structural fingerprints. Moving to the top of Fig. 5b, we see the structure evolution from in-plane hexagonal sublattice (clusters #26, #30 and #31) to out-of-plane hexagonal lattice (#0) on the left, further to the out-of-plane 6-fold coordination sublattices (#4, #5, #9 and #29). Similarly, the transition from in-plane 4-fold coordination (e.g., #22, #27, and #28) to out-of-plane 4-fold coordination (#11, #12, and #14) occurs in the same manner.

Our t-SNE chart also captures the gradual evolution between different coordination patterns, for example, from hexagonal to rectangular sublattices. Moving downwards from the in-plane hexagonal sublattice, we first encounter elongated hexagons (#17), which can also be represented as planar centered orthorhombic lattice. Further distortion of the central atom creates planar monoclinic lattice (#13 and #20) and planar orthorhombic lattice (#22, #27, and #28). From here, evolution into the square lattice (#24 and #25) is straightforward. Even the subtle structural differences, such as differences in stacking distance, within the same coordination are also well reflected in our clustering plot. For example, stacked hexagonal bilayer in 9-fold coordination (6-fold in-plane and 3-fold out-of-plane) with short stacking distance (#45) is excluded from the branch for longer stacking distance coordination (#48-50), as shown in the phylogenetic tree. Even among the clusters #48-50, we could identify slight in-plane distortion in #48 leading to coloring triangle lattices whereas #49 and #50 show no such distortion leading to simple AB stacked hexagonal lattice. Contrary to the well-defined isostructures, clusters that don't show any regular or identifiable coordination patterns (e.g., #10, #33, #38, and #46) lie around the dispersed region in the center of t-SNE plot.

DISCUSSION

Using our hybrid machine learning approach, we identified many sublattice structures responsible for hosting flat bands. Among these, some have already been confirmed in the literature, endorsing the validity of our approach. For example, kagome, breathing kagome, sawtooth, square-chain (Creutz) lattices are known to host flat bands[30,38–40]. But many of the lattices identified in this work are entirely not explored, and the origin of their flat bands remains to be clarified. Some of such monolayer sublattices are listed below, alongside example materials followed by their 2Dmatpedia id. These include Lieb-square lattice (Fe_4S_5 [2dm-4626], Os_4S_5 [2dm-1815]), centered orthorhombic lattice (Li_4VF_8 [2dm-4139], $VZnF_4$ [2dm-4036]), diamond-chain lattice (ReS_2 [2dm-4006], $ReSe_2$ [2dm-3917]), $(3^2, 4, 3, 4)$ Archimedean lattice (TiP [2dm-2672]), vertex shared square chain lattice (Tl_2Te_5[2dm-1601]), linear chain lattice (NbS_3[2dm-1482]), isolated quadrilaterals ($MoSe_2$[2dm-2007]), and centered monoclinic lattice (Ta_2I_5 [2dm-5407]). More examples are given in Supplementary Tables I and II.

Bilayer flat band sublattices, in many cases, consist of stacked monolayer lattices with flat bands. For example, stacked kagome, square, centered orthorhombic, AB and AA hexagonal, and orthorhombic lattices belong to this category. Among bilayer sublattices, stacked centered-orthorhombic-square chains (SmF_3 [2dm-875]) and isolated tetrahedra lattice ($LiNiPO_4$ [2dm-5215], $Si(HgO_2)_2$ [2dm-4520]), are flat band lattices which are not reported in literature.

Meanwhile, several few-layer sublattices made of stacked monolayer flat band sublattices were also identified, including stacked hexagonal and orthorhombic, kagome-hexagonal, and coloring-triangle-hexagonal lattices. Apart from this, we observed many interesting multilayer flat band sublattices. For example, sublattices that consist of in-plane octahedras, where different octahedra connections generate seven new sublattices. These are vertex sharing octahedra chain ($YFeF_5$ [2dm-3862]), edge sharing octahedra-parallel chain (Sc_5Cl_8 [2dm-4448]), vertex sharing planar octahedra (ZrF_4 [2dm-5172]), vertex-edge sharing planar octahedra ($CaTlCl_3$ [2dm-4337]), isolated twisted octahedra (ReSeCl [2dm-5461]), edge-sharing zigzag octahedra chain ($ZnMoO_4$[2dm-3090]) and edge sharing octahedra chain (Y_2Cl_3 [2dm-3786]). Furthermore, the stacked $(3^2, 4, 3, 4)$ Archimedean sublattice ($CoBr_4$ [2dm-1906], MnF_4 [2dm-5281])with out-of-plane distortion, is also identified to host flat bands. Other multilayer sublattices, e.g. vertex sharing triangular pyramid chain (V_2CuO_6 [2dm-4387]), triangular bipyramid lattice (MoO_3 [2dm-5495]), distorted Kagome lattice ($MgCl_2$ [2dm-4672]), planar hexahedra lattice (TaF_5 [2dm-4011]), and stacked orthorhombic-square (HgI_2 [2dm-3966]) are also found to show flat bands.

A statistical analysis of the elements responsible for flat bands, presented in Supplementary Note V and Supplementary Figs. 4 and 5, gives an overall chemical intuition behind some of the nontrivial flat bands, for instance, those which are not generated due to localization of *f*-orbitals or quenched group velocity of the insulators. Following that intuition, Supplementary Table I focuses on the nontrivial flat band examples rather than *f*-orbital-localized actinides and lanthanides. Future research is expected to reveal the origin and the rich physics behind the majority of the potential flat-band lattices found in our work.

Our deep learning approach has identified many new flat-band sublattices by analyzing electronic structures of more than 5000 2D materials from 2Dmatpedia. Further improvements in the accuracy of DFT calculations used to generate 2Dmatpedia and other databases will make our approach even more robust. The 2Dmatpedia database uses a dispersion-corrected version of GGA(PBE) functional (vdW-optB88) to model the exchange-correlation interactions, resulting in accurate prediction of van der Waals forces, making the results more accurate for 2D materials. DFT+U corrections are implemented to model the onsite interactions. The magnetic materials are treated with spin-polarized charge densities and collinear magnetic structures[19]. However, this database does not include spin-orbit interactions in the calculations. Thus, the results may have the usual limitations of semilocal exchange-correlation functionals and lack spin-orbit coupling effects. Furthermore, strongly correlated systems are not the strongest feature of the Kohn-Sham DFT because most standard local and semilocal functionals fail to predict exchange correlations accurately in these systems. To tackle such cases, several methods have been formulated to date, e.g., DFT+DMFT, DFT+U, hybrid functionals, and Meta-GGA functionals, and the search for more accurate functionals are still underway. Higher order Meta-GGA functionals like MBJ, TPSS, and SCAN include higher order derivatives of the wavefunctions, thus more accurately predicting the exchange-correlation interactions compared to standard local and semilocal functionals, making the predicted electronic structure (e.g., band gap) more accurate.

Our hybrid approach combining supervised and unsupervised machine learning has demonstrated its accuracy and efficiency in navigating through enormous information space to hunt for flat band materials. This approach can be conveniently adapted in searching for materials with intended properties, and to mitigate issues arising from the quality of high-throughput density functional theory calculations in existing databases. An interesting direction is to identify topologically nontrivial flat bands, which show crossings with other dispersive bands at high symmetry points in the BZ[41]. Features like band crossings, which are both common and small, may require developing automatic graphical pattern search tools[42]. Moreover, in addition to using structural fingerprints, such as lattice coordination patterns as described in the current work, to distinguish different flat band materials, electronic fingerprints, e.g., the similarity of electronic states[25], may also be attempted in the future to offer new perspectives towards the discovery of new materials.

METHODS

Convolutional neural network

The CNN trained for the identification of flat bands consists of 6 layers (as shown in Fig. 1) and 6.9 million learnable parameters. It is a regression network which predicts the existence of a flat band in the image segment and outputs as a single value – the flatness score. The resolution of each segment is set to 96×96 pixels in consideration of both image quality and computing time. The first convolutional layer consists of 30 channels of 10×10 filters with [1,1] stride. The second convolutional layer has 12 channels of 3×3 filters with [1,1] stride. Sigmoid activation was used and found to give better performance than softmax and ReLU activation as the output of the network lies within [0,1]. The first fully-connected layer has 80 output nodes, whereas the second has a single output. For training purposes, Adam[43] optimization algorithm was used with learning rate 5×10^{-5} and L2-regularization parameter 0.003.

To train the CNN, band structure images of the first 5000 materials (based on material ID) from the Materials Project database were downloaded. After eliminating the blank entries of the database, these made a dataset of 3228 images. Our method involves subdivision of each band structure image horizontally and vertically. We found that only ≈12% of band structures have one of these subdivisions containing a flat band among them, whereas the rest ≈88% do not have any subdivisions with a flat band. Since each of these subdivisions is used as one training example, this could lead to a poorly trained network because of insufficient positive training examples. To avoid this, we increased the fraction of positive examples (energy bandwidth with flat band segments) in the training set by manually selecting 1900 band structure images that have at least one flat band containing segments. These band structures were then fed to train the network. This creates a training set with ≈ 46,500 training examples, with ≈25% positive examples. There are several reasons for selecting the Materials Project database for training. First, it offers more training flexibility compared to the relatively small database of 2Dmatpedia (≈5300 materials), avoiding applying the trained algorithm back to the same database. Second, training the algorithm on 3D compounds that have inherently more complicated band structures could improve its accuracy when applied to the 2Dmatpedia. Finally, training the model on a general 3D database allows future application to many other available materials databases.

The developed CNN model was then tested on 7000 image segments from Materials Project. The test set predictions match 92.1% with manual selection, whereas false negative and false positive predictions were seen in 3.3% and 4.6% of all cases, respectively. This is quite high precision since even manual selection of flat segments could not guarantee 100% consistency, as partially flat segments can often lead to confused judgment from a human selector.

Before applying the trained CNN to 2Dmatpedia, band structures from 2Dmatpedia in the energy range [-1,1] eV were segmented at high symmetry points of k-space. Each resulting segment has 96 × 96 resolution. For example, the band in Fig. 2b is divided horizontally into four 0.5 eV energy bandwidths, and vertically along k-path into three segments (as denoted in green, blue and yellow).

A threshold flatness score predicted from the trained CNN is used to determine whether a band-structure segment is flat (score ≥0.5) or non-flat (score < 0.5). We find a significantly higher fraction of flat band segments in 2Dmatpedia, approximately 25%, than that in 3D materials from Materials Project (approximately 13%). Dimensionality reduction and accompanied restriction in degrees of freedom of electrons in 2D materials are likely to be the reason. To identify plane-flat band 2D materials, which have flat bands throughout high symmetry lines and are of particular interest, we select the energy strip in which all the horizontal segments are recognized as flat. In the rare cases where multiple energy strips contain all-flat segments, priority is given to the energy bandwidth where electrons are more easily accessible, i.e. closer to the Fermi level.

Sublattice extraction and vectorization

We use the element projected DOS, as shown in Fig. 3a, b, to identify the element corresponding to the flat band sublattice. Because flat electronic bands are accompanied by corresponding peaks in DOS, an element with maximum density contribution in the flat band is most likely to be responsible for the flat dispersion, although orbital mixing among different species is also a contributing factor. Here we conjecture that the symmetry operations pertaining to this elemental sublattice lead to the flat band, thus only the sublattice with the chosen element for each compound is kept for further analysis. After the energy segment containing flat band is identified, the corresponding bandwidth in the element projected DOS is analyzed to obtain the element and, subsequently, its sublattice. When a band structure contains multiple energy segments with flat bands, we use a priority scheme to identify the dominating flat band (see Supplementary Note VII and Supplementary Fig. 7).

A structure descriptor was then applied to a selected sublattice to create a vector representation for machine learning algorithms. Here we use a structure descriptor CrystalNNFingerprint[27], implemented in the Matminer package[44]. Other structure descriptors, like Coulomb matrix, Ewald sum matrix, Smooth Overlap of Atomic Positions (SOAP), Many-body Tensor Representation (MBTR), usually yield a constant length vector representing the structure[37,45,46], were found less suitable for our analysis. CrystalNNFingerprint retains more information when the number of atomic sites is high, thus particularly suitable for identifying local coordination patterns in crystals even in the presence of small lattice distortions. For example, popular descriptors like SOAP calculating spectral average of the local fingerprints, would fail to retain local structural information (see Supplementary Note VI and Supplementary Fig. 6 for details). CrystalNNFingerprint first determines the local environment for each atomic site using a neighbor-finding algorithm based on Voronoi decomposition. The resulting coordination pattern is then compared with Local Structure Order Parameters (LoStOPs) which are different coordination templates, and a 61-dimensional fingerprint vector is assigned to the atomic site. The global fingerprint for the whole structure is then represented by arranging the mean, standard deviation, maximum and minimum of each local (atomic site) CrystalNNFingerprint within a structure. The final structure fingerprint lies in a 244-dimensional vector space.

Clustering module

We then used unsupervised machine learning module consisting of density-based clustering algorithm HDBSCAN[28] and t-SNE[29] to further cluster 244-dimensional structure fingerprints. HDBSCAN creates clusters with fingerprints that are densely populated, acting as a strict identifier of similarity in coordination patterns with many fingerprints designated as outliers. Simultaneously, soft clustering feature of HDBSCAN yields a probability for each material to be included in a cluster. Consequently, members of clusters with very high inclusion probability (exemplars) facilitate straightforward identification of corresponding sublattice structures. The exemplars are the members of a cluster which persist in the cluster for the largest range of density. Hence, these may be identified as the central members even if the shape of the cluster is fairly complicated as they do not depend on any distance metric. However, HDBSCAN tends to assign fingerprints with even small discrepancies into different clusters, risking overlooking local neighborhood information. As a complement, we use t-SNE to visualize and identify similarities among HDBSCAN clusters, acting as a second layer of unsupervised classification. The nonlinear dimensionality reduction capability of t-SNE allows both local neighborhood information and global distance relations to be preserved, in comparison to other algorithms like Isomap[47], Locally Linear Embedding (LLE)[48], Hessian eigenmapping[49], etc. We also tested UMAP[50], no significant improvement over t-SNE is observed (results are discussed in the Supplementary Note III and Supplementary Fig. 2).

The choice of HDBSCAN as the first layer of unsupervised clustering algorithm is based on the fact that more commonly utilized clustering algorithms, like K-means and hierarchical clustering, would yield non-optimum segregation. K-means algorithm solely depends on the distance among inputs, generating accurate results only for inputs that form ellipsoid clusters. At the same time, hierarchical clustering is too sensitive to the presence of noise in the data, making it unsuitable in this context since lattice distortion is an avoidable source of noise. For our 244-dimensional data we used Manhattan distance metric, L_1 norm, since Euclidean distance metric (L_2 norm) becomes less effective in calculating distances among points in high-dimensional spaces[51].

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are available from the corresponding authors on reasonable request.

CODE AVAILABILITY

The codes generated during the current study are available in the Github repository, https://github.com/Anupam-Bh/ML_2D_flat_band.

Received: 25 November 2022; Accepted: 25 May 2023;
Published online: 08 June 2023

REFERENCES

1. Regnault, N. et al. Catalogue of flat-band stoichiometric materials. *Nature* **603**, 824–828 (2022).
2. Ma, D.-S. et al. Spin-orbit-induced topological flat bands in line and split graphs of bipartite lattices. *Phys. Rev. Lett.* **125**, 266403 (2020).
3. Bistritzer, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl. Acad. Sci. USA* **108**, 12233–12237 (2011).
4. Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
5. Li, T. et al. Quantum anomalous hall effect from intertwined moiré bands. *Nature* **600**, 641–646 (2021).
6. Rhim, J. W., Kim, K. & Yang, B. J. Quantum distance and anomalous Landau levels of flat bands. *Nature* **584**, 59–63 (2020).
7. Choi, Y. et al. Correlation-driven topological phases in magic-angle twisted bilayer graphene. *Nature* **589**, 536–541 (2021).
8. Li, H. et al. Imaging two-dimensional generalized Wigner crystals. *Nature* **597**, 650–654 (2021).
9. Wang, X. et al. Light-induced ferromagnetism in moiré superlattices. *Nature* **604**, 468–473 (2022).
10. Huang, T. et al. Observation of chiral and slow plasmons in twisted bilayer graphene. *Nature* **605**, 63–68 (2022).
11. Ma, C. et al. Intelligent infrared sensing enabled by tunable moiré quantum geometry. *Nature* **604**, 266–272 (2022).
12. Leykam, D., Andreanov, A. & Flach, S. Artificial flat band systems: from lattice models to experiments. *Adv. Phys. X* **3**, 1473052 (2018).
13. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
14. Curtarolo, S. et al. Aflow: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
15. Ashton, M., Paul, J., Sinnott, S. B. & Hennig, R. G. Topology-scaling identification of layered solids and stable exfoliated 2d materials. *Phys. Rev. Lett.* **118**, 106101 (2017).
16. Mounet, N. et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **13**, 246–252 (2018).
17. Boland, T. M. & Singh, A. K. Computational synthesis of 2d materials: A high-throughput approach to materials design. *Comput. Mater. Sci.* **207**, 111238 (2022).
18. Haastrup, S. et al. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Materials* **5**, 042002 (2018).
19. Zhou, J. et al. 2dmatpedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci. Data* **6**, 1–10 (2019).
20. Călugăru, D. et al. General construction and topological classification of all magnetic and non-magnetic flat bands. *Nat. Phys.* **18**, 185–189 (2022).
21. Liu, H., Meng, S. & Liu, F. Screening two-dimensional materials with topological flat bands. *Phys. Rev. Mater.* **5**, 084203 (2021).
22. Duan, J. et al. Inventory of high-quality flat-band van der waals materials. *Preprint at* https://doi.org/10.48550/arXiv.2204.00810 (2022).
23. Scheurer, M. S. & Slager, R. J. Unsupervised machine learning and band topology. *Phys. Rev. Lett.* **124**, 226401 (2020).
24. Nuñez, M. Exploring materials band structure space with unsupervised machine learning. *Comput. Mater. Sci.* **158**, 117–123 (2019).
25. Knøsgaard, N. R. & Thygesen, K. S. Representing individual electronic states for machine learning GW band structures of 2d materials. *Nat. Commun.* **13**, 1–10 (2022).
26. Kuroda, T., Mizoguchi, T., Araki, H. & Hatsugai, Y. Machine learning study on the flat-band states constructed by molecular-orbital representation with randomness. *J. Phys. Soc. Jpn* **91**, 044703 (2022).
27. Zimmermann, N. E. & Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **10**, 6063–6081 (2020).
28. Campello, R. J., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In *PAKDD*, p. 160–172 (Springer, 2013).
29. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
30. Kang, M. et al. Topological flat bands in frustrated kagome lattice CoSn. *Nat. Commun.* **11**, 1–9 (2020).
31. Zhang, S. et al. Kagome bands disguised in a coloring-triangle lattice. *Phys. Rev. B* **99**, 100404 (2019).
32. Nakai, H. & Hotta, C. Perfect flat band with chirality and charge ordering out of strong spin-orbit interaction. *Nat. Commun.* **13**, 1–9 (2022).
33. Moulavi, D., Jaskowiak, P. A., Campello, R. J., Zimek, A. & Sander, J. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, 839–847 (SIAM, 2014).
34. Halkidi, M. & Vazirgiannis, M. Clustering validity assessment: Finding the optimal partitioning of a data set. In *Proceedings 2001 IEEE International Conference on Data Mining*, 187–194 (2001).
35. Belkina, A. C. et al. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat. Commun.* **10**, 1–12 (2019).
36. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 1–14 (2019).
37. Himanen, L. et al. Dscribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
38. Essafi, K., Jaubert, L. & Udagawa, M. Flat bands and dirac cones in breathing lattices. *J. Phys. Condens. Matter* **29**, 315802 (2017).
39. Grémaud, B. & Batrouni, G. G. Haldane phase on the sawtooth lattice: Edge states, entanglement spectrum, and the flat band. *Phys. Rev. B* **95**, 165131 (2017).
40. Mondaini, R., Batrouni, G. G. & Grémaud, B. Pairing and superconductivity in the flat band: Creutz lattice. *Phys. Rev. B* **98**, 155142 (2018).
41. Rhim, J.-W. & Yang, B.-J. Classification of flat bands according to the band-crossing singularity of bloch wave functions. *Phys. Rev. B* **99**, 045107 (2019).
42. Borysov, S. S., Olsthoorn, B., Gedik, M. B., Geilhufe, R. M. & Balatsky, A. V. Online search tool for graphical patterns in electronic band structures. *NPJ Comput. Mater.* **4**, 1–8 (2018).
43. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* San Diego, CA, USA (2015).
44. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
45. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
46. Lee, J. et al. Descriptors of atoms and structure information for predicting properties of crystalline materials. *Mater. Res. Express.* **8**, 026302 (2021).
47. Tenenbaum, J. B., Silva, V. D. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
48. Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326 (2000).
49. Donoho, D. L. & Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA* **100**, 5591–5596 (2003).
50. McInnes, L., Healy, J. & Melville, J. Umap: uniform manifold approximation and projection for dimension reduction. *Preprint at* https://doi.org/10.48550/arXiv.1802.03426 (2018).
51. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. On the surprising behavior of distance metrics in high dimenional space. In *ICDT*, p. 420–434 (Springer, 2001).

ACKNOWLEDGEMENTS

This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 865590), and the Royal Society International Exchanges 2019 Cost Share Program (IEC\R2\192001). A.B. acknowledges the Commonwealth Scholarship Commission in the UK for financial assistance. Q.Y. acknowledges the funding from Leverhulme Early Career Fellowship ECF-2019-612, Dame Kathleen Ollerenshaw Fellowship from the University of Manchester, and Royal Society University Research Fellowship URF\R1\221096.

AUTHOR CONTRIBUTIONS

A.B. carried out the calculations, implemented the algorithms in codes, analyzed the data, and drafted the manuscript. A.M. and Q.Y. conceived the research plan, analyzed the data, drafted the manuscript, and supervised the research work. R.C. co-supervised the research work. I.T. analyzed the data and contributed to writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-023-01056-x.

Correspondence and requests for materials should be addressed to A. Bhattacharya, Q. Yang or A. Mishchenko.

Reprints and permission information is available at http://www.nature.com/reprints

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