Discover Materials

Research

Learning from machine learning: the case of band-gap directness in semiconductors

Elton Ogoshi[1] · Mário Popolin-Neto[2,3] · Carlos Mera Acosta[1] · Gabriel M. Nascimento[1] · João N. B. Rodrigues[1,4] · Osvaldo N. Oliveira Jr[5] · Fernando V. Paulovich[6] · Gustavo M. Dalpian[1,7]

Received: 30 October 2023 / Accepted: 8 February 2024
Published online: 29 February 2024

Abstract
Having a direct or indirect band gap can influence the potential applications of a semiconductor, for indirect band gap materials are usually not suitable for optoelectronic devices. Even though this is a fundamental property of semiconducting materials, discussed in textbooks, no unified theory exists to explain why a material has a direct or indirect band gap. Here we used an interpretable machine learning model, the multiVariate dAta eXplanation (VAX) method, to gather information from a dataset of materials extracted from the Materials Project. The dataset contains more than 10000 entries, and atomic properties such as the number of electrons, electronic affinity and orbital energies were used as features to build random forest models that successfully explain the directness of the band gaps. Our results indicate that symmetry is an important feature that dictates the target property, which is the reason why our analysis is made based on sub-groups with similar structures. These sub-groups include materials with zincblende, rocksalt, wurtzite, and perovskite structures. Besides the symmetry of the materials, the existence or not of *d* bands and the relative energy of atomic orbitals were found to be important in defining whether a material's band gap is direct or indirect. In conclusion, interpretable machine learning methods such as VAX can be useful in obtaining physical interpretation from materials databases.

1 Introduction

Machine learning methods are now an integral part of materials research [1, 2] thanks to the increasing availability of curated datasets containing structural and electronic data derived from quantum-mechanical calculations. Also, the accessibility and user-friendliness of Machine Learning algorithms and software have extended their utility to a broader scientific community. Machine Learning has been mostly used for predictive purposes, where algorithms are trained on data concerning a specific material property P_i with a set of features that can be related to P_i (i.e.,

Elton Ogoshi and Mário Popolin Neto equally contributed to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43939-024-00073-x.

✉ Gustavo M. Dalpian, dalpian@if.usp.br | [1]Center for Natural and Human Sciences, UFABC, Santo Andre, SP, Brazil. [2]Federal Institute of São Paulo, IFSP, Araraquara, SP, Brazil. [3]Institute of Mathematics and Computer Sciences, USP, São Carlos, SP, Brazil. [4]Physics Departament, Federal University of Pernambuco, Recife, PE, Brazil. [5]São Carlos Institute of Physics, USP, Sao Carlos, SP, Brazil. [6]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands. [7]Institute of Physics, University of Sao Paulo, Sao Paulo, SP, Brazil.

Discover

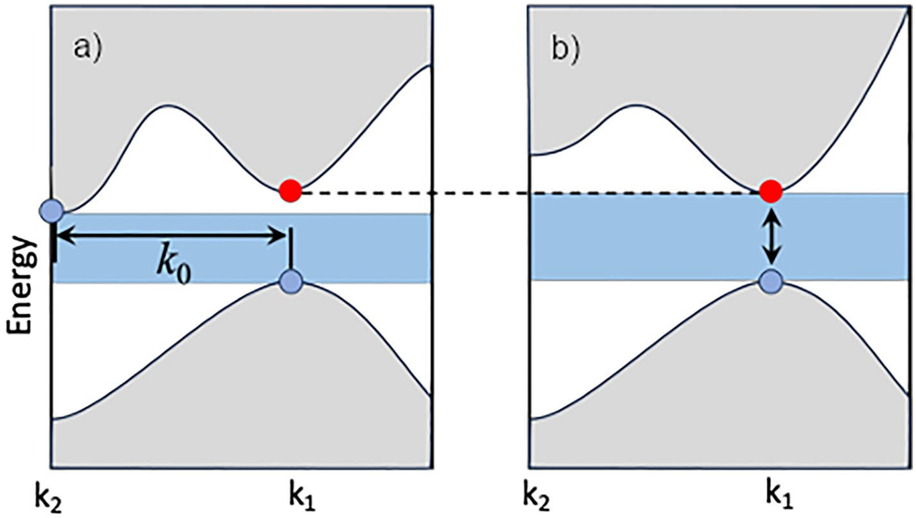

Fig. 1 Schematic representation of a band structure showing indirect and direct bandgaps (respectively, left and right panels). The blue regions stand for the bandgap and the blue dot at k_1 is the top of the valence band. The bottom of the conduction band changes from the k_2-wavevector (blue point) for indirect bandgap to the k_1-wavevector (red dots) for direct bandgap

the feature space $\{f_n\}$). Information such as numerical patterns and tendencies of P_i in the feature space $\{f_n\}$ is then inferred for new materials that are not in the initial dataset. An ultimate goal for Machine Learning, however, is to build models that can go beyond prediction, with scientific concepts being discovered via descriptive models. This is a formidable task, either because the feature space is too large, or because the physical mechanisms behind a specific phenomenon are unknown, barely understood or complex. Some cases of success have been reported [3–8], including one that looks towards the discovery of semiconductors [9].

In this article, we use interpretable Machine Learning methods to analyze a fundamental property of semiconductors which lacks a detailed understanding: the reason why some semiconductors have a direct band gap while others have an indirect one. In semiconductors, the top of the valence band (representing occupied states) and the bottom of the conduction band (representing empty states) are separated by an energy gap, as shown in the blue region in Fig. 1. The top of the valence band and the bottom of the conduction band can be located at different momentum *k*-wavevectors (i.e., indirect band gap, as shown in Fig. 1a) or at the same *k*-wavevector (i.e., direct band gap, as shown in Fig. 1b). Determining the type of band gap is relevant for semiconductor applications. For instance, indirect band gap materials are usually not suitable for applications in optoelectronic devices because the absorption or emission of a photon requires an electron-lattice momentum exchanged, and such two-step process is less likely to occur. The most used material in the microelectronics industry, i.e. silicon, has an indirect band gap and thus is not suitable for optoelectronic devices such as light emitting diodes (LEDs). Transforming indirect band gap materials into direct band gap ones is still challenging. Common strategies for tuning indirect-direct band gaps include alloying [10, 11], exploring strains [12, 13], and quantum confinement [14]. The directness of the band gap is not difficult to determine, either theoretically or experimentally, as discussed in fundamental physics textbooks [15]. However, no unified theory exists to explain why one material has a direct or an indirect band gap. Semicondutors encompass a variety of materials and crystal structures, from simple diamond structures as Si to perovskites containing at least three different types of atoms and many structural distortions that can change their properties [16]. To the best of our knowledge, only a couple of works have focused on the explanation of the band gap directness in semiconductors. Yuan and collaborators [17] focused on Zincblende semiconductors, and have shown that materials with cations with occupied *d*-levels tend to have direct band gaps owing to the symmetry of the zincblende lattice. In the zincblende point group, the deep *d*-band interacts with the VB/CB at the Γ, L and X high- symmetry points in the reciprocal space according to wavefunction symmetry at this point which is dictated by their symmetry at the *k*-point and the band symmetry representation. The *d*-bands do not interact with the bottom of the conduction band at the Gamma point (red dots in Fig. 1). This means that in the presence of occupied *d*-orbitals (Fig. 1b), the band repulsion at other *k*-points leads the CB higher in energy with respect to the Gamma point, which is kept fixed regardless the presence of *d*-orbitals due to the lack of interaction. This mechanism controls the direct/indirect gap for this specific crystal structure. In another paper, Choubisa and collaborators [9] proposed general rules to explain when a material has a direct or indirect band. The rules include the occupation of *p* orbitals, the position of the LUMO and the electronegativity of the constituent atoms. As we will show below, our rules are considerably different from these, providing new ways to design direct or indirect band gap materials.

Herein, we conduct a broad analysis using data science and interpretable Machine Learning methods such as Decision Trees (DT) [18] and Random Forests (RF) [19]. In particular, we perform a descriptive analysis using the VAX method [20], extracting Jumping Emerging Patterns (JEPs, descriptive logic rules) [21, 22] from Machine Learning models, and then look for causal relations or insights to explain why a material has a direct or indirect band gap. Detailed information of the VAX method is provided in the Methodology. Our findings indicate that the directness of the band gap depends on the symmetry of the different materials. No general pattern (rule) was encountered which would apply to all structural families of materials. However, if we divide the materials in smaller groups of compounds with similar symmetries (e.g, Zincblende, Wurtzite, Rock Salt, and Perovskite structures), VAX generates specific patterns that explain their bandgap directness. Specifically, we recovered the known result for Zincblende structures, i.e., the existence of *d* orbitals. Furthermore, we have also found that relative energies of highest occupied states determine the direct–indirect bandgap transitions in other structures. We believe that these results demonstrate that explainable machine learning is promising to advance the understanding of physics problems.

2 Materials and methods

2.1 Materials and features

The data used in the Machine Learning algorithms was derived from online materials databases containing detailed band structure for thousands of materials. These include synthesized inorganic materials taken from the Inorganic Crystal Structure Database (ICSD), [23] and thousands of hypothetical materials whose properties were obtained in calculations using an extrapolating approach where chemical elements are replaced by similar ones within the same crystal structure. We used the Materials Project [24] database, version 2022, that contains 76022 gapped materials (from a total of 126335 entries) with their band structures, including information on whether their band gap is direct or indirect. All these calculations were performed with a GGA exchange correlation functional. To make our database more suitable for the Machine Learning analysis, a data curation process was conducted where we screened out some problematic materials. Below we list the main filters used in this process, together with the number of available materials after each filter:

- We removed all materials with either a ferromagnetic or antiferromagnetic ground state configuration, leaving 41241 materials that have no spin polarization. This is done in order to simplify the analysis, and remove the complexity of magnetic systems;
- Compounds that have more than 18 atoms in the unit cell were removed, leaving our dataset with a total of 10477 materials. Owing to the reduction of the size of the Brillouin Zone in very large unit cells, it becomes harder to differentiate among direct and indirect band gap materials.

To make our dataset more suitable for an interpretable analysis, as discussed in detail in the results section, we split our data into smaller datasets of materials with the same crystal structure. We chose to concentrate on the following structures: Zincblende (ZB) structure (61 materials), Wurtzite (WZ) structure (38 materials), and Rock salt (RS) structure (61 materials). The choice of these groups facilitates the analysis, since all these compounds are binaries (A is the cation and B is the anion). Table 1 lists the features used together with the labels employed in other figures for reference. We've used the *Mendeleev* package for retrieving the atomic features [25]. We have also analyzed perovskites, with chemical formula ABX_3 (ABC_3 in our nomenclature). The features used for the ternary perovskites are basically the same as those for binary materials (Table 1), however we needed to add another C element. Also, when the features are related to differences among two values (e.g. Δ_X for the difference in electronegativity), we included the differences among atoms A and B, B and C, and A and C.

To select the materials from a specific group on the Materials Project, we restrict our search for a specific space group and number of sites in the primitive cell. For ZB we chose only the entries with space group F-43 m (216) and having only two sites in the primitive cell; for Rocksalt we selected space group Fm-3 m (225) with only 2 sites in the primitive cell and composition AB; for wurtzite we selected materials with space group P63mc (186) with only 4 sites and composition AB; for cubic perovskites we used the Pm-3 m symmetry with 5 sites in the primitive cell and composition ABC_3; finally, for non-cubic perovskites we manually selected the compounds with compositions ABC_3 with the characteristic topology of interlinked octahedra. The space groups were defined by the Materials Project database, using the package spglib with the default tolerance metrics.

Table 1 Set of features used in the zincblende, wurtzite and rock salt dataset; A: cation features; B: anion features. Atomic features were extracted from Ref. [25]

Features	Description
A_N, *B_N*	Number of electrons
A_group, *B_group*	Group in periodic table
A_row, *B_row*	Row in periodic table
A_NValence, *B_NValence*	Number of valence electrons
A_EA, *B_EA*	Electron Affinity
$Delta_X$	Electronegativity difference among atoms A and B
AB_bond_len	A–B distance
A_s, *A_p*, *A_d*, *B_s*, *B_p*, *B_d*	Delta energy of valence orbital to the closest occupied s, p, d, f orbitals (atomic)
A_top − energy, *B_top − energy*	Energy of the highest occupied atomic level
A_top − orbital, *B_top − orbital*	Orbital-type of the highest occupied atomic level

2.2 The VAX method

The *multiVariate dAta eXplanation* (VAX) method [20] provides multivariate data analysis leveraging Machine Learning models' descriptive power [26]. The main idea is to extract, select, and aggregate the so-called Jumping Emerging Patterns (JEPs) [22, 27] from Random Forest (RF) models [19, 28]. JEPs are descriptive logic rules [21, 22] where the variable selectors describe inter, and intra-class relationships. Once the JEPs are extracted, selected, and aggregated maximizing confidence and support, they are displayed into a matrix-like visual metaphor in VAX. The latter is shown in Fig. 2 (exploded view) for the 61 materials of the Zincblende dataset, 38 of which have direct gap. This example will be further discussed in another figure, without the notes, in the results section.

2.2.1 Patterns and variables

Patterns are arranged as rows (①) and variables appear as columns (②). In Fig. 2, patterns p_9 and p_5 are placed at the first and second rows, involving variables 'A_NValence' and 'total_N' at the first and second columns. Classes are mapped to categorical colors, with blue assigned to class "indirect gap" and orange assigned to the class "direct gap".

2.2.2 Local histograms and selectors

The matrix cells (③) present, for a particular pattern, local histograms showing the supported data instances distributed along the available variables. Moreover, a continuous line frame into the matrix cell represents a variable selector, that is a distinct range of values defined by upper and lower bounds. For example, in the histograms using 5 bins in Fig. 2, the rule p_9 (first row) contains a selector of 7.5 to 29.0 for variable 'A_NValence' (first column). The values in this cell can range from 2.0 to 29.0 (min. and max. values for 'A_NValence' considering all 61 materials). The pattern p_9 also contains a selector of 51.0 to 132.0 for variable 'total_N' (second column) from the possible range from 12.0 to 132.0 (min. and max. values). Thus, pattern p_9 supports (explains) direct gap materials (orange) with values between 7.5 and 29.0 (selector) for variable 'A_NValence' and 51.0 and 132.0 for variable 'total_N'. In contrast, pattern p_5 (second row) has a selector of 2.0 and 9.5 for variable 'A_NValence' (first column). Despite not having a selector for variable 'total_N' (second column), the local histogram is also placed into the matrix cell. Hence, pattern p_5 supports (explains) indirect gap materials (blue) with values between 2.0 and 9.5 for variable 'A_NValence'. Comparing patterns p_9 and p_5, indirect materials tend to have lower values than direct materials for both variables 'A_NValence' and 'total_N'.

2.2.3 Global Histograms, Support, and Cumulative Coverage

Global histograms for each class are shown on the top of the matrix (④) and the patterns' support (class percentage) is mapped onto a column on the matrix's left side (⑤). The cumulative dataset coverage is also mapped onto a column on the matrix left side (⑥), representing the cumulative percentage of data instances in a dataset covered (explained) by the patterns in that row and those in the rows above it. In this context, pattern p_9 in Fig. 2 has a support value of

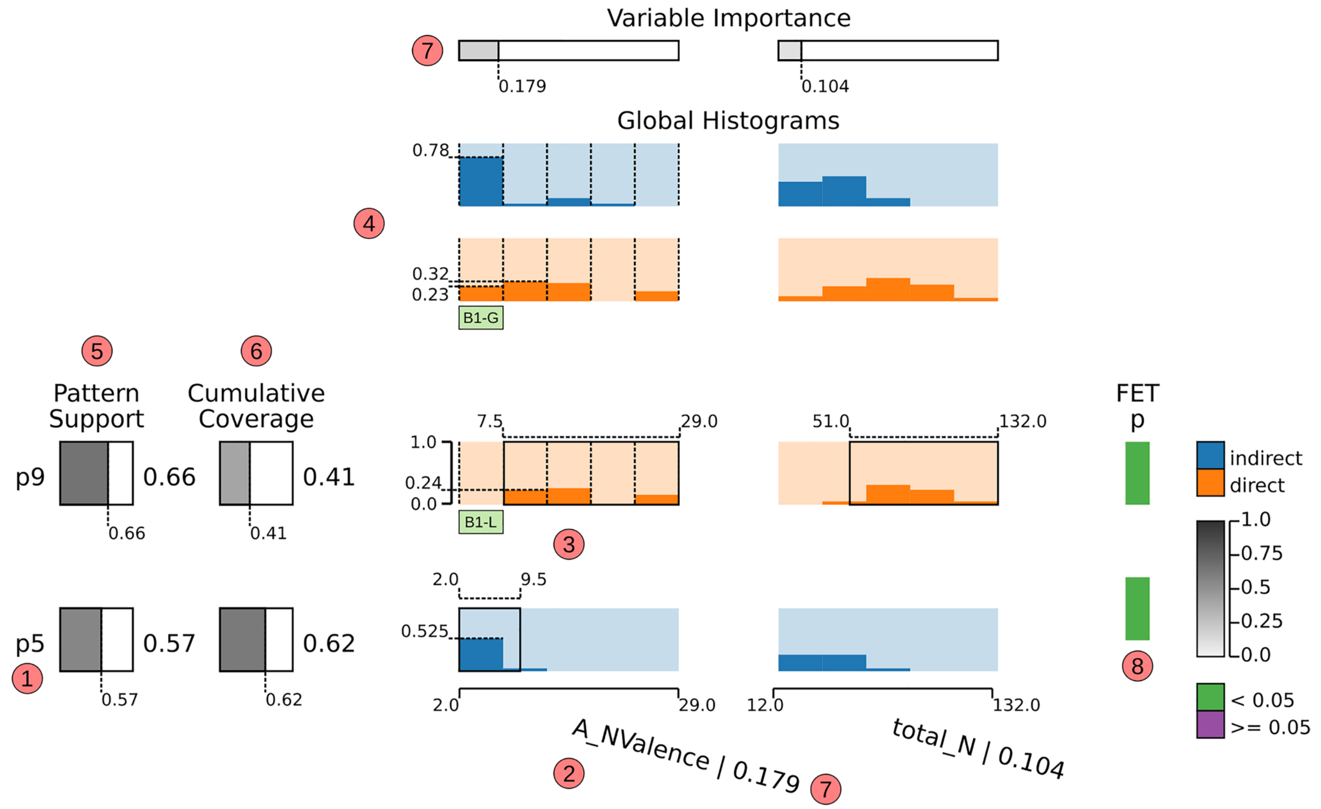

Fig. 2 The matrix-like visual metaphor employed by VAX using the Zincblende dataset. ① JEPs are displayed as rows (class color: indirect as blue and direct as orange). ② Variables are arranged as columns. ③ Local histograms are shown as cells. ④ Global histograms (one row per class) are placed on the top. ⑤ Pattern support. ⑥ Cumulative coverage assuming the matrix order (top to bottom). ⑦ Variable importance, pattern support, cumulative coverage, and variable importance are mapped to size and brightness (gray-scale). ⑧ FET (Fisher Exact Test) significance value (*p*) colored as green (statistically significant, $p \leq 0.05$) or purple (not significant, $p > 0.05$)

0.66, which means that 25 of the 38 (66%) direct gap materials are supported (explained) by p_9. The local histograms for pattern p_9 (matrix cells at first row) arrange the 25 supported materials distribution along variables 'A_NValence' and 'total_N'. The global histograms for the "direct gap" class (orange) display the distribution for each variable on the set of all 38 Zincblende materials. Comparing local and global histograms bins for variable 'A_NValence' (first column), 23% of the 38 direct gap materials at the first bin-global (B1-G – first bucket in the global histogram for direct gap class) are completely absent in pattern p_9, since it lacks (0%) the first bin-local (B1-L – first bucket in the local histogram in the first row). However, from the 32% in the second bin-global, 24% are enclosed by the second bin-local for pattern p_9. The pattern p_5 has a support value of 0.57, i.e it explains 13 of 23 (57%) indirect gap materials. For variable 'A_NValence' (first column), 78% of indirect materials are at the first bin-global, from which 52.5% are retained by the first bin-local for pattern p_5. The cumulative coverage for patterns p_9 and p_5 is 0.62, i.e. 38 (25 + 13) of 61 materials (62%) are covered by the two patterns. Therefore, more than half of the materials are explained by only two patterns (p_9 and p_5).

2.2.4 Variable Importance and FET p-value

The variable importance is shown visually above the global histograms and textually, next to the variables' name at the bottom (⑦). The FET $p - value$ (statistical significance test) for each pattern is displayed in a column to the matrix right side (⑧), where green means statistically significant (*p* values below 0.05) and purple otherwise. The variable 'A_NValence' in Fig. 2 is more important than 'total_N' (0.179 against 0.104). Moreover, both patterns p_9 and p_5 are statistically significant (green for FET p).

In addition to the visualization of JEPs, VAX provides similarity maps for data instances using Dimensional Reduction (DR) layouts [29] and patterns perspectives (dataset extension [30]). Clusters (groups of close data instances) and outliers (data instances apart) are often revealed in such maps, and more importantly, are interpreted by JEPs visualization analysis. Figure 3 presents a map for the Zincblende dataset, where clusters and outliers can be

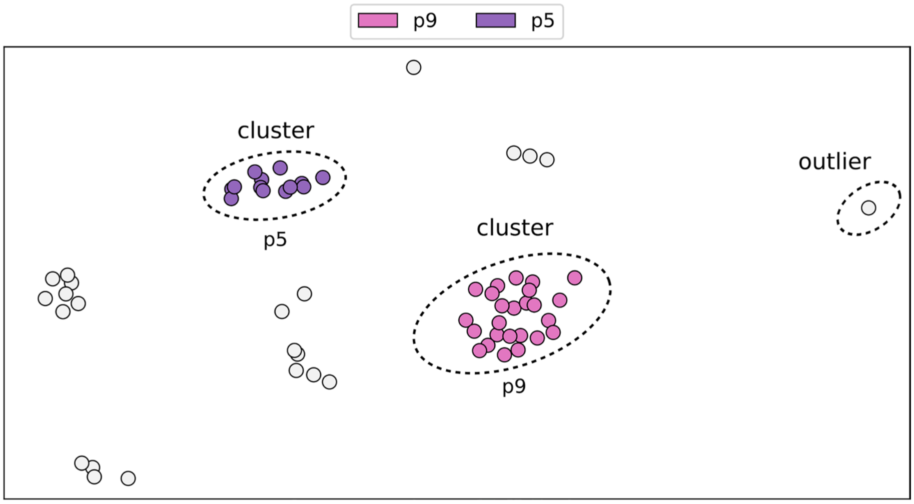

Fig. 3 Similarity Map for the Zincblende dataset within the JEPs perspectives. Materials whose band gap direct–indirectness are explained by patterns p_9 and p_5 cluster together and are identified as pink points (pattern p_9) and purple points (pattern p_5). An outlier is also highlighted (at the most right position)

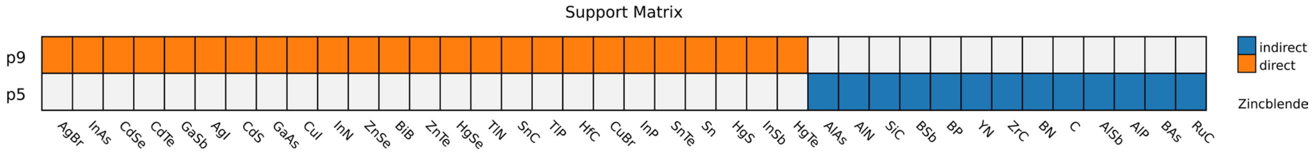

Fig. 4 Support Matrix for the Zincblende dataset. Materials explained by patterns p_9 and p_5 (class color: indirect gap materials as blue and direct gap materials as orange). Pattern p_9 supports 25 direct materials (AgBr, InAs, CdSe, CdTe, GaSb, AgI, CdS, GaAs, CuI, InN, ZnSe, BiB, ZnTe, HgSe, TIN, SnC, TIP, HfC, CuBr, InP, SnTe, Sn, HgS, InSb, and HgTe), whereas pattern p_5 supports 13 indirect (AlAs, AlN, SiC, BSb, BP, YN, ZrC, BN, C, AlSb, AlP, BAs, and RuC)

identified. Each point represents a zincblende material and the color refers to patterns support. Points colored in pink identify the materials supported (explained) by pattern p_9, whereas those supported by p_5 are in purple. The materials colored in the map compose two clusters, and the difference between such clusters is the difference between the patterns that originate them (p_9 and p_5). So, as noticed from the JEPs visualization in Fig. 2, the distinction between the cluster originated by pattern p_9 and the one formed by pattern p_5 is that materials in the latter trend to have lower values for variables 'A_NValence' and 'total_N'. Moreover, an outlier is highlighted in Fig. 3 (isolated point on the right), representing a material that does not belong to any of the clusters and is far away from other materials as well. A support matrix (binary heat map) can be shown relating patterns (rows) and the supported data instances (columns). For the Zincblende dataset, Fig. 4 presents the support matrix for patterns p_9 and p_5 of Fig. 2. Pattern p_9 supports AgBr, InAs, CdSe, CdTe, GaSb, AgI, CdS, GaAs, CuI, InN, ZnSe, BiB, ZnTe, HgSe, TIN, SnC, TIP, HfC, CuBr, InP, SnTe, Sn, HgS, InSb, and HgTe; and pattern p_5 supports AlAs, AlN, SiC, BSb, BP, YN, ZrC, BN, C, AlSb, AlP, BAs, and RuC.

In summary, VAX employs JEPs and visualization of similarity maps to examine multivariate datasets aiming at knowledge discovery and phenomena analysis. By displaying JEPs using a matrix-like visual metaphor, meaningful visual explanations can be reached by filtering and ordering patterns (rows) and variables (columns). In turn, similarity maps convey an overview of the dataset (e.g., clusters and outliers) using JEPs context. It is worth mentioning that VAX was developed for descriptive analyses [20]. Unlike predictive models, which focus on creating generic models (splitting datasets into training and testing subsets for estimating performance [31]), descriptive tasks are not intended to be generic outside the scope of the data under analysis. Instead, they should offer explanations of the phenomena observed in a single dataset [32, 33]. Hence, VAX provides JEPs with "100% accuracy" for the dataset under examination. This is needed in order for us to be able to obtain patterns that work for all data instances. This is why we are able to find patterns even for all outliers. These are some of the advantages of VAX when compared to other methodologies.

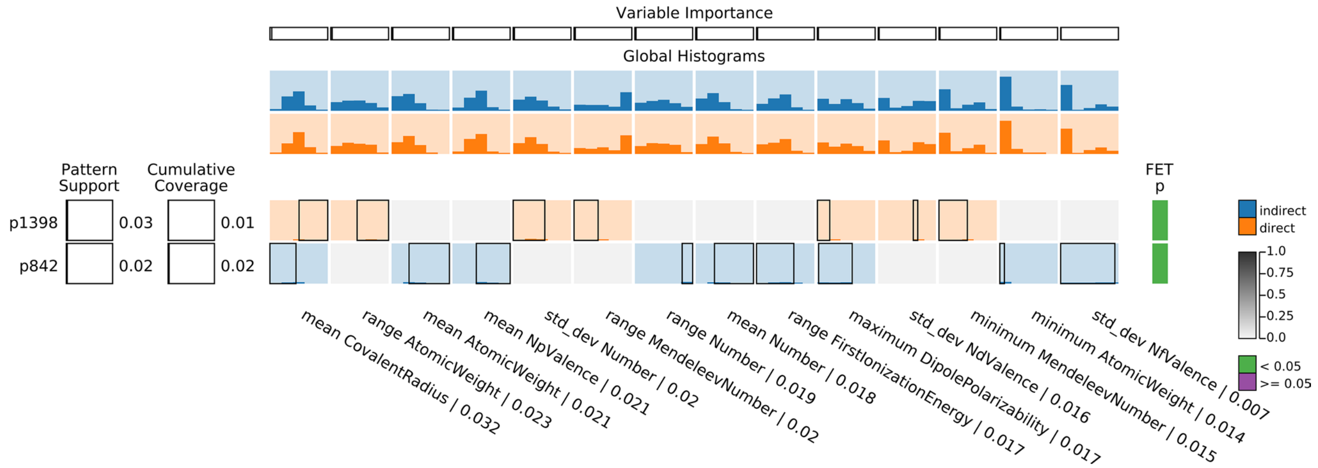

Fig. 5 The two patterns (p_{1398} and p_{842}) with the highest support filtered out the 1495 JEPs (extracted, selected, and aggregated) from the dataset with all materials (10477 – 7432 indirect gap and 3045 direct gap). Patterns (rows) are ordered by support and the variables (columns) are organized by importance. Pattern p_{1398} holds the highest support among the 1495 patterns; yet it supports only ∼ 3% of direct gap materials (80 of 3045). Since pattern p_{842} has ∼ 2% of support, the 10493 remaining patterns (not shown) have equal or less than 2%

3 Results and discussion

This section presents the descriptive analyses of the materials datasets using the VAX method [20]. The latter is available as a code package [1] for Python programming language. The source code for the results presented here are accessible as Python Jupyter notebooks, one for each material family: All Materials ,[2] Zincblende ,[3] Rock Salt ,[4] Wurtzite ,[5] and Perovskites .[6]

3.1 All materials

The VAX method was applied to the whole dataset of 10477 (post filtering) materials, 7432 of which have indirect gap while 3045 have direct gap. A total of 1495 patterns (JEPs) were obtained, and the patterns with highest support, p_{1398} and p_{842}, are shown in Fig. 5. Pattern p_{1398} explains ∼ 3% of direct gap materials (80 of 3045), whereas p_{842} supports ∼ 2% of indirect gap materials (148 of 7432). These values are statistically significant (green for FET p), but they together cover only 2% (228 of 10477) of the dataset (cumulative coverage). Interestingly, the remaining 1493 patterns (not presented in Fig. 5) also can have supports around to 2%. This decreased representativity of these patterns comes from the complexity of this dataset, with a huge variety of crystal structures and the lack of good features encoding symmetry.

By inspecting the materials supported by p_{1398} and p_{842} (in Figures S1 and S2 of the supplementary information), we noted that the Machine Learning model (descriptive RF post-processed by VAX) put together materials with similar spatial atomic arrangements. Pattern p_{1398} supports (explains) only two-dimensional heterostructures in the database. There were different types of stacking of 2D materials such as $MoS_2/MoSe_2$, WS_2/WSe_2, all of which have direct band gaps. For pattern p_{842}, which works for indirect band gaps, all the materials explained are oxides, mostly containing Bi. This motivated us to split the database into different families of materials with similar crystal structures. By doing so, we avoid the need to specify the crystal structure and symmetry features. Conversely, we infer that crystal structure and symmetry should be important to determine whether a material has a direct band gap. In the following subsections, we analyze materials grouped according to their crystal structures.

[1] https://pypi.org/project/vaxm/.
[2] https://popolinneto.gitlab.io/vaxm/papers/2023/dm/allmaterials/.
[3] https://popolinneto.gitlab.io/vaxm/papers/2023/dm/zincblende/.
[4] https://popolinneto.gitlab.io/vaxm/papers/2023/dm/salt/.
[5] https://popolinneto.gitlab.io/vaxm/papers/2023/dm/wurtzite/.
[6] https://popolinneto.gitlab.io/vaxm/papers/2023/dm/perovskites/.

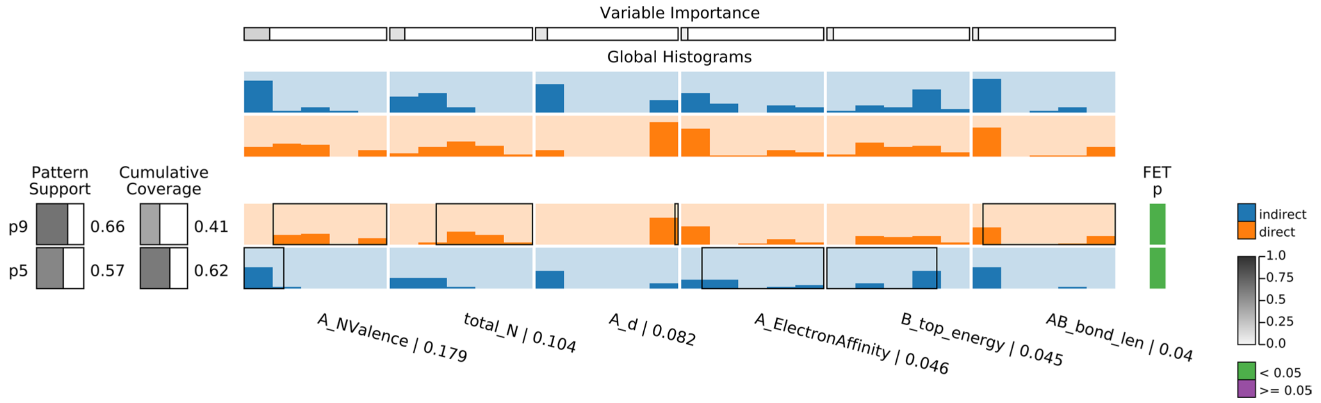

Fig. 6 The two patterns (p_9 and p_5 – first and second rows) with the highest support filtered out the 9 JEPs resulted (extracted, selected, and aggregated) from the Zincblende dataset (61 materials – 23 indirect and 38 direct). Patterns (rows) are ordered by support and the variables (columns) are ordered by importance. The variable 'A_NValence' (first column) is the most important, where low values are related to indirect materials (blue) and high (also medium) values to those direct (orange). Together, patterns p_9 and p_5 explain 62% (38 of 61 – more than half) of the Zincblende materials

3.2 Zincblende

The zincblende (ZB) dataset is a good starting point for the analysis, since we can compare our results with those from Yuan et al. [17], and evaluate their *d-band model*. The ZB dataset contains 61 materials, with 38 direct gap materials. This dataset is larger than the one used in reference [17], thus permitting a broader analysis. In this dataset were also included elements that share the same crystal structure, but are composed by only one type of atom (C, Si, Ge and Sn). In this case, the materials actually have the diamond crystal structure. As shown in Table 1, our feature set includes those necessary to gather the information available in the work of Yuan [17], including the presence of *d* electrons and bond lengths. It also includes features such as electronegativity, electronaffinity, and the rows on the periodic table where the constituent atoms are located.

A total of 9 patterns (JEPs) were obtained for the Zincblende dataset. Figure 6 shows the two highest support patterns: p_9 and p_5 (first and second rows). Pattern p_9 supports 66% of direct materials (25 of 38), whereas p_5 supports 57% of those indirect (13 of 23). Together these patterns explain the majority of the compounds, that is 62% of the zincblende materials (38 of 61). The most important variable was the number of valence electrons of the cation, represented by 'A_NValence' (first column). The 25 direct gap materials from pattern p_9 have values between 7.5 to 29.0 (selector) for variable 'A_NValence'. On the other hand, the 13 indirect gap materials from p_5 have values between 2.0 and 9.5. It is worth mentioning that pattern selectors min. and max. values (ranges learned by the Machine Learning model) may not have feasible numbers, e.g. 7.5 for variable 'A_NValence'. In this case, these values must be interpreted by rounding up. Linking the patterns p_9 and p_5 with the limit of 6 rule found by Yuan et al. [17], if variable 'A_NValence' is larger than 6, there will be a high probability for the materials to have a direct band gap. If the number is smaller than 6, indirect band gaps are more frequent. This limit of A_NValence = 6 is correlated with the existence of *d* electrons in the cations: within the pseudopotential approach of our database, A cations can only have more than six valence electrons if they have filled d orbitals. This conclusion is the same made by Yuan et al. [17] for zincblende compounds. Therefore, we successfully reproduced the main result obtained in that paper. A similar analysis holds for the second most important variable (total_N).

Figure S3 in the Supplementary Information presents all 9 patterns for the ZB case, while Figure S4 shows the similarity map and Figure S5 shows the support matrix. Most of the 9 patterns returned by VAX are used to explain materials that do not fit into the 'd-electron' rule. Some of the patterns were actually used to explain only a few materials, which means that a different physical mechanism is needed to explain the directness of their band gap. These include Ge, GaP, CaSe and others. Moreover, pattern p_5 explains materials such as C, Si, SiC, BN, BP, AlN, AlP, YN, and ZrC as indirect band gap because they do not have *d* bands.

3.3 Rock salts

The Rock Salt (RS) dataset has the same number of materials as the one for ZB, i.e. 61 in total (32 with direct and 29 with indirect band gaps). We used the same variables employed for ZB, and found 9 patterns (JEPs) for the RS dataset. The

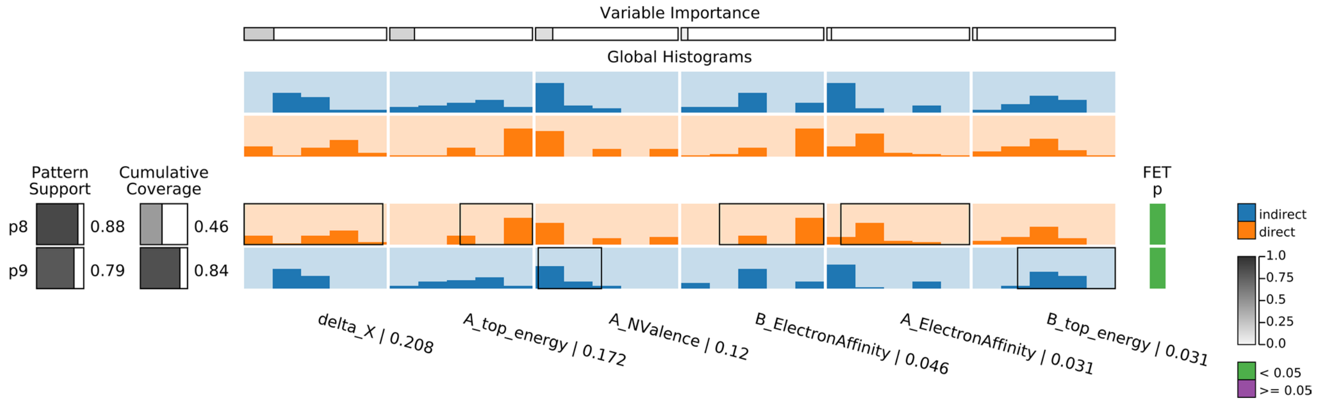

Fig. 7 The two patterns (p_8 and p_9 – first and second rows) with the highest support filtered out the 9 JEPs (extracted, selected, and aggregated) for the Rock Salt dataset (61 materials – 29 with indirect and 32 with direct band gaps). Patterns (rows) are ordered by support and the variables (columns) are ordered by importance. The variables 'delta_X' and 'A_top_energy' (first and second columns) are the most important. For 'A_top_energy' (second column), direct materials (orange) concentrate (66% – 21 of 32) in the last histogram bin-local for pattern p_8, whereas pattern p_9 reveals only a few (7% – 2 of 29) indirect materials (blue) for such a bin. Together, patterns p_8 and p_9 explain 84% (51 of 61 – strong generic behaviour) of the rock salt materials

highest support patterns p_8 and p_9 are shown in Fig. 7. It is important to note that pattern p_9 for the RS dataset is different from pattern p_9 of the ZB dataset since they involve different Machine Learning models. Pattern p_8 supports 88% of direct gap materials (28 of 32), whereas p_9 supports 79% of those with indirect band gaps (23 of 29). Together these patterns explain 84% of the rock salt materials (51 of 61). Even with the same number of patterns (9 JEPs), the coverage for the two highest support patterns (p_8 and p_9) for RS is larger than for ZB (with patterns p_9 and p_5), i.e. 84% against 62%. Thus, determining the directness of the band gap is easier for RS than for ZB. The remaining results for the RS dataset are shown in the Supplementary Information with all the patterns found, the similarity map and the support matrix shown in Figure S6, Figure S7 and Figure S8, respectively. By inspecting the second most important variable (second column) for patterns p_8 and p_9 in Fig. 7, which is 'A_top_energy', one notices distinct histogram distributions for direct and indirect band gaps. 'A_top_energy' indicates the energy of the highest occupied atomic level of atom A. 66% of direct gap materials (21 of 32) are found in the last bin-local for variable 'A_top_energy' in pattern p_8 in contrast to 7% of indirect gap materials (2 of 29) in pattern p_9. Hence, if 'A_top_energy' is high, there is a strong tendency for the band gap to be direct. This trend is used here to explain the indirect gap materials supported by pattern p_9, and explicitly listed in Figure S9. By analyzing this list we realize that these indirect band gap materials have as cations Be, Mg, Ca, Sr or Ba (smaller 'A_top_energy'). These are compounds where the cation has a completely filled valence *s* orbital, which can be related to a smaller 'A_top_energy'.

To further evaluate the explanations provided in patterns p_8 and p_9, we selected two compounds with similar chemical structures, but with different band gap behaviors: CaSe has an indirect band gap, and KBr has a direct band gap. All these atoms belong to the same period of the periodic table. Figure 8 shows the band structure of these compounds. The conduction band minimum of KBr (direct band gap) is composed of K *s* orbitals and the band gap is at the Γ point. For CaSe, CBM is composed of Se *p* orbitals and is located at the X point. The Ca *s* orbitals are higher in energy, turning the band gap indirect. The relative position of the K (Ca) *s* orbitals, related to the variable 'A_top_energy', is key in defining the directness of the band gap.

3.4 Wurtzites

The Wurtzite dataset has 38 materials, 29 of which have direct band gaps. Wurtzite (WZ) compounds have a crystal structure similar to that of Zincblende materials. Each cation is bound to four anions, but differently from zincblende, one of these bonds has a different length than the others. Also, the stacking along the (111) direction is different in these two crystal structures, with ZB showing an ABCABC... stacking (three different layers) and WZ having an ABAB... configuration (two different layers) [34]. A total of 4 patterns (JEPs) were obtained for the WZ dataset, whose patterns are in descending order of support in Fig. 9. Patterns p_3 and p_2 (first and second rows) yield the highest support, covering 82% of the wurtzite materials (31 of 38). This coverage with the two highest support patterns is almost the same as that in the RS database. Pattern p_3 supports 89% of indirect gap materials (8 of 9), whereas p_2 supports 79% of the direct gap ones (23 of 29). A comparison of the two patterns indicates that the direct gap materials (orange) tend to have higher

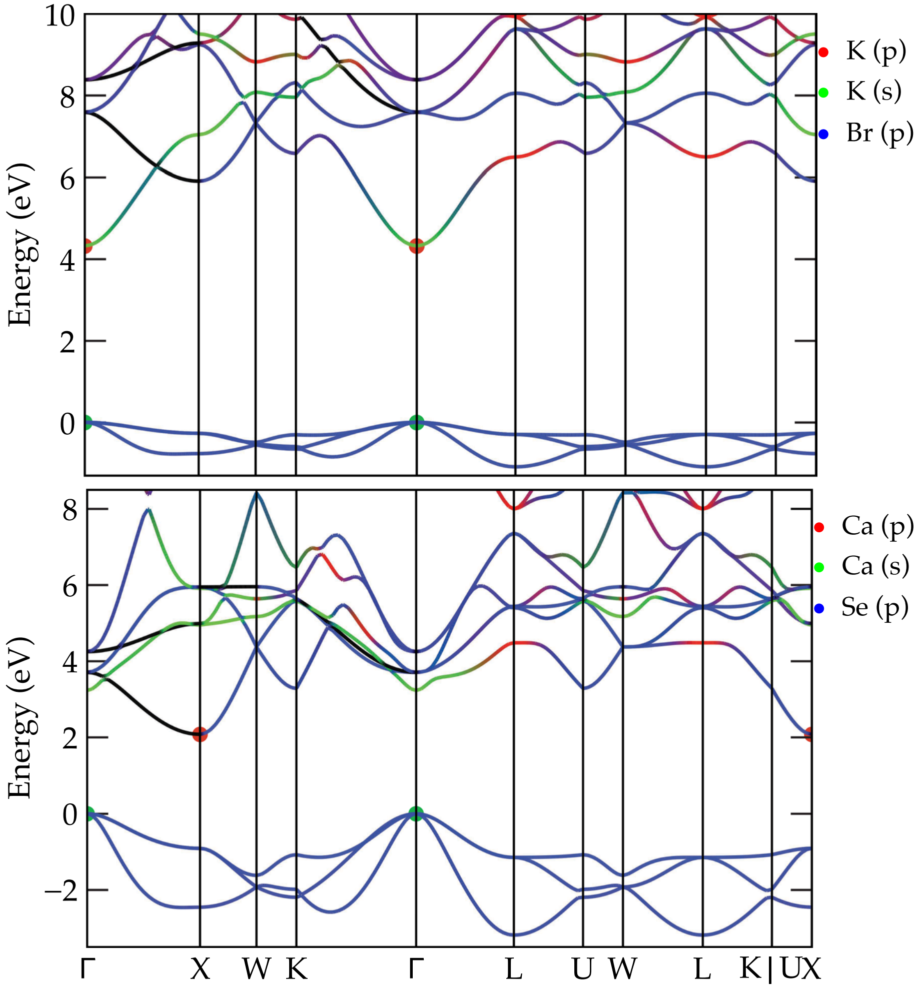

Fig. 8 Band structures of KBr (above) and CaSe (below). The colors in the graphs indicate the atomic contribution for each eigenvalue. These band structures were calculated using DFT with a GGA exchange correlation functional

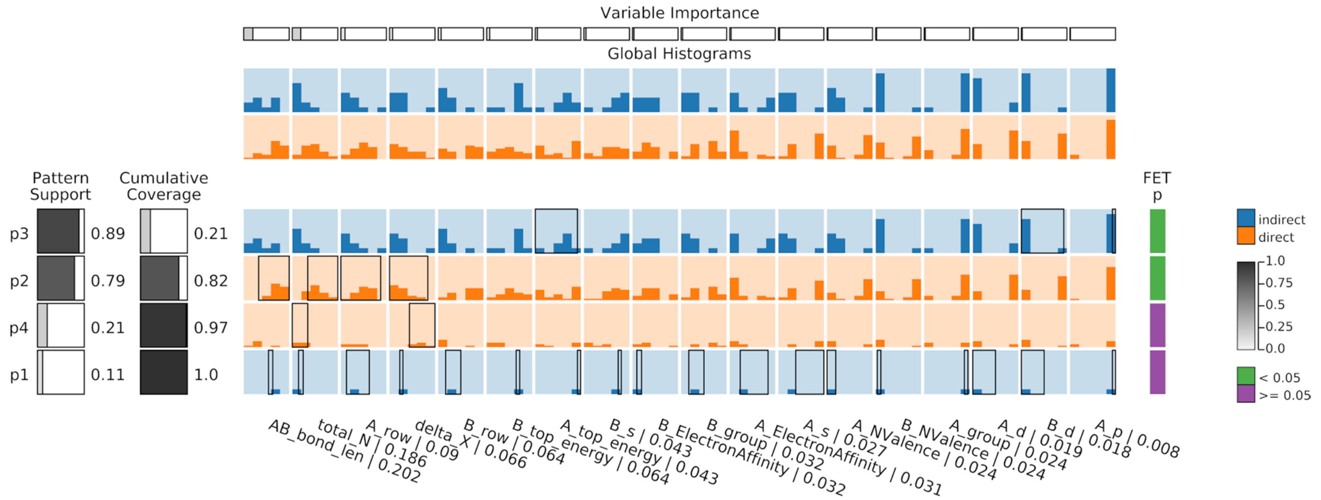

Fig. 9 All 4 patterns resulted (extracted, selected, and aggregated) from the Wurtzite dataset (38 materials – 9 with indirect and 29 with direct gaps). Patterns (rows) are ordered by support and the variables (columns) are ordered by importance. The patterns p_3 and p_2 (first and second rows) have the highest support. For variables 'total_N' and 'AB_bond_len' (second and first columns – the most important) direct gap materials (orange) tend to have higher values than indirect gap compounds (blue). Together, patterns p_3 and p_2 explain 82% (31 of 38 – strong generic behaviour) of the wurtzite materials

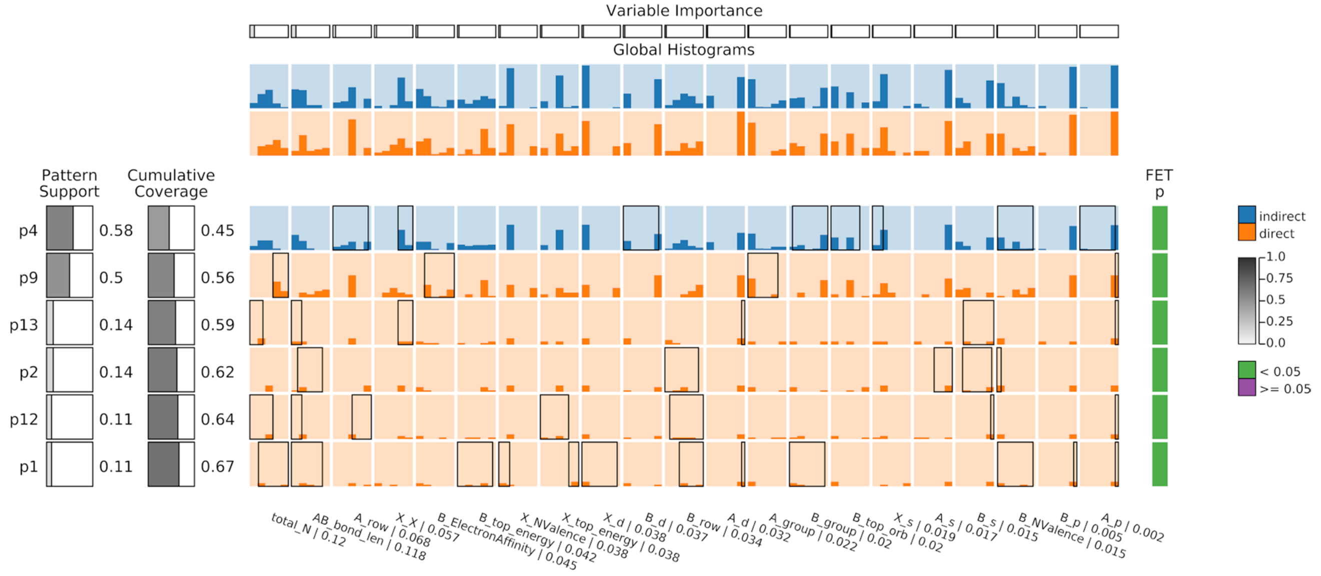

Fig. 10 The 6 patterns with the highest support filtered out from the 13 JEPs resulted (extracted, selected, and aggregated) for the Perovskites dataset (132 materials – 104 indirect and 28 direct). Patterns (rows) are ordered by support and the variables (columns) are ordered by importance. The variable 'total_N' (first column) is the most important; for patterns p_4 and p_9 (first and second rows) medium and low values are related to indirect gap materials (blue) while high values refer to direct gap materials (orange). For variable 'A_p' (last column), all patterns (p_9, p_{13}, p_2, p_{12}, and p_1) for direct gap materials (second to sixth row) contain higher values than for pattern p_4 (indirect gap materials). These six patterns explain 67% (89 of 132) of the Perovskites materials

values for variables 'total_N' and 'AB_bond_len' (second and first columns). Note that there is a correlation between these variables: compounds with a larger number of electrons tend to have larger bond lengths. This analysis also leads to a similar interpretation as that for the ZB dataset: a larger number of electrons appears when the constituting atoms have *d* orbitals. The similarity of the ZB and WZ patterns comes from their similar crystal structures.

The similarity map in Figure S10 in the Supplementary Information highlights two clusters formed by patterns p_3 and p_2 (pink and purple). A third cluster is formed with pattern p_4 (maroon) and there is an outlier supported by pattern p_1 (emerald). Figure S11 presents the materials supported by patterns p_3, p_2, p_4, and p_1.

The difference among clusters (purple and maroon – Figure S10) formed by patterns p_2 and p_4 is in variable 'total_N' (second column) and 'delta_X' (fourth column). This last variable indicates the difference in electronegativity between atoms A and B. The materials from pattern p_4 have large values of 'delta_X', indicating they are more ionic. This also leads to larger band gaps, making the coupling between *d* bands and the conduction band smaller, decreasing the importance of the *d* bands and making the band gap direct.

The indirect material AlP, an outlier in Figure S10, is supported by pattern p_1 (fourth row in Fig. 9) and it differs from the 8 indirect materials (C, Si, BP, BN, GeC, BaO, BAs, SiC) by pattern p_3 (first row in Fig. 9) in variable 'A_top_energy' (seventh column). The AlP compound has a higher value for variable 'A_top_energy' (between −0.11 to −0.1 – variable selector) than the 8 materials explained by pattern p_3 (between −0.22 to −0.11 – variable selector).

3.5 Perovskites

Perovskites are much more complex than the other materials discussed before. They comprise three types of atoms (ABX_3), with a variety of chemical formulas and many crystal structures. Although the general chemical formula has an ABX_3 stoichiometry, there are also double perovskites such as $A_2BB'X_6$ [35] and various magnetic configurations [36]. As for the crystal structure, perovskites can be cubic, tetragonal, orthorhombic and even polymorphous [37]. All these properties can influence the band gaps [38]. The Perovskite dataset has 132 materials, of which 104 have indirect band gaps. For perovskites we used a different set of variables compared to the binary compounds. The variables in Table 1 for binary materials are not able to fully encode their chemical space. Additionally, the dataset contains not only cubic perovskites, but also perovskites with rotated octahedra. Thus, we need variables that also discriminate this structural degree of freedom.

A total of 13 patterns (JEPs) were obtained for the Perovskite dataset, and Fig. 10 shows the 6 patterns yielding the highest support. These 6 patterns are the only ones with statistical significance. Pattern p_4 (first row) supports 58% of indirect gap materials (60 of 104), whereas p_9 supports 50% of those with direct gap (14 of 28). Together these patterns explain about half (56%) of the compounds, i.e. 74 of 132 materials. The main difference between patterns p_4 and p_9 is in variables 'total_N' and 'A_p' (first and last column), since direct gap materials (orange) tend to have higher values than the indirect gap materials (blue). These variables respectively indicate the total number of electrons and the relative energy of the highest occupied *p* orbital of atom A.

The 14 direct gap materials from pattern p_9 contain values between 121 and 190 (selector) for 'total_N' and from −0.54 to 0.0 (selector) for variable 'A_p'. This latter tendency for variable 'A_p' is also found in the remaining patterns p_{13}, p_2, p_{12}, and p_1 (third to sixth row) for direct gap materials. As observed for Zincblende and Wurtzites, the trend for the variable 'total_N' indicates that compounds with a larger number of electrons have direct band gaps, which is correlated with the existence of filled *d* bands.

Figure S12 in the Supplementary Information shows all the 13 patterns for the perovskites dataset, while Figures S13 and S14 display the similarity map and the support matrix, respectively.

4 Conclusions

Machine learning methods in materials science have been mostly used to predict a specific property of a material. Examples include models to predict properties such as the value of band gaps [39], stability of 2D compounds [40], or the magnetic order of compounds [5]. Recent works [6, 41], however, indicate that Machine Learning can be used in more complex tasks, including interpretation of physical phenomena. In this paper, we discovered through using VAX that to explain the band gap directness the semiconductors have to be separated into smaller groups with the same crystal structure. Indeed, we could determine the rules that explain why a semiconductor structural family has a direct or indirect gap. This has been done for zincblende materials, confirming the conclusions from an earlier work [42], and for other crystal structures. Overall we observed that two main features are essential to determine if a material has a direct or indirect band gap. *The first feature* is the symmetry of the crystal, and consequently the symmetry of each high-symmetry *k*-point in the reciprocal space. This conclusion was inferred from the fact that we were unable to get a good Machine Learning model using all materials. The creation of subsets with the same crystal structure was necessary to get good results. *The second feature* is the existence of specific orbitals (e.g., d-orbitals) and their respective energy and occupation. The interaction among these orbitals may push the valence band maximum/conduction band minimum (VBM/CBM) at specific *k*-points towards higher/lower energies, leading to changes in the band gap directness.

In addition to providing explanations about the directness of semiconductor band gaps, the use of the VAX in this paper showcases new ways to employ machine learning to study materials in an interpretable way. For the VAX method can be used in conjunction with any rule-based machine learning algorithm.

Acknowledgements This research was supported by Brazilian funding agencies FAPESP (grants number 17/02317-2, 18/11856-7, 18/11641-0, 2018/22214-6 and 19/04176-2) and CNPq - INCT (National Institute of Science and Technology on Materials Informatics, grant number 371610/2023-0). Computer simulations were performed at the Santos Dumont supercomputer.

Author contributions EO retrieved the data from Materials project and worked on data curation. MPN, ONOJ, and FVP performed the VAX analysis. EO, CMA, GMN, JN and GMD evaluated the results from VAX and substantiated them with physical principles. All authors participated in the discussion of the results and writing and revising the paper.

Data availability All codes and files were made available in the links provided in the manuscript.

Declarations

Competing interests Authors declare no competing interests.

References

1. Rodrigues JF, Florea L, de Oliveira MCF, Diamond D, Oliveira ON. Big data and machine learning for materials science. Discover Mater. 2021. https://doi.org/10.1007/s43939-021-00012-0.
2. Schleder GR, Padilha ACM, Reily Rocha A, Dalpian GM, Fazzio A. Ab initio simulations and materials chemistry in the age of big data. J Chem Inform Modeling. 2019;60:452.
3. Iten R, Metger T, Wilming H, Rio L, Renner R. Discovering physical concepts with neural networks. Phys Rev Lett. 2020;124:10508. https://doi.org/10.1103/PhysRevLett.124.010508.
4. Häse F, Roch LM, Friederich P, Aspuru-Guzik A. Designing and understanding light-harvesting devices with machine learning. Nat Commun. 2020. https://doi.org/10.1038/s41467-020-17995-8.
5. Acosta CM, Ogoshi E, Souza JA, Dalpian GM. Machine learning study of the magnetic ordering in 2d materials. ACS Appl Mater Interfaces. 2022;14:9418. https://doi.org/10.1021/acsami.1c21558.
6. Oviedo F, Ferres JL, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and chemistry. Accounts Mater Res. 2022;3:597–607. https://doi.org/10.1021/accountsmr.1c00244.
7. Pilania G. Machine learning in materials science: from explainable predictions to autonomous design. Comput Mater Sci. 2021;193: 110360. https://doi.org/10.1016/j.commatsci.2021.110360.
8. Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY-J. Explainable machine learning in materials science. NPJ Comput Mater. 2022. https://doi.org/10.1038/s41524-022-00884-7.
9. Choubisa H, Todorović P, Pina JM, Parmar DH, Li Z, Voznyy O, Tamblyn I, Sargent EH. Interpretable discovery of semiconductors with machine learning. NPJ Comput Mater. 2023. https://doi.org/10.1038/s41524-023-01066-9.
10. Fadaly EM, Dijkstra A, Suckert JR, Ziss D, van Tilburg MA, Mao C, Ren Y, van Lange VT, Korzun K, Kölling S, Verheijen MA, Busse D, Rödl C, Furthmüller J, Bechstedt F, Stangl J, Finley JJ, Botti S, Haverkort JE, Bakkers EP. Direct-bandgap emission from hexagonal Ge and SiGe alloys. Nature. 2020;580:205. https://doi.org/10.1038/s41586-020-2150-y.
11. Lee M-H, Liu P-L, Hong Y-A, Chou Y-T, Hong J-Y, Siao Y-J. Electronic band structures of Ge1- xSnx semiconductors: a first-principles density functional theory study. J Appl Phys. 2013;113: 063517.
12. Zhang S, Yan Z, Li Y, Chen Z, Zeng H. Atomically thin arsenene and antimonene: semimetal-semiconductor and indirect-direct band-gap transitions. Angewandte Chemie. 2015;127:3155.
13. Desai SB, Seol G, Kang JS, Fang H, Battaglia C, Kapadia R, Ager JW, Guo J, Javey A. Strain-induced indirect to direct bandgap transition in multilayer wse2. Nano lett. 2014;14:4592.
14. Mak KF, Lee C, Hone J, Shan J, Heinz TF. Atomically thin mos2: a new direct-gap semiconductor. Phys Rev Lett. 2010. https://doi.org/10.1103/PhysRevLett.105.136805.
15. Sze SM, Li Y, Ng KK. Physics of semiconductor devices. 4th ed. Hoboken, NJ: Wiley-Blackwell; 2021.
16. Yin W-J, Yang J-H, Kang J, Yan Y, Wei S-H. Halide perovskite materials for solar cells: a theoretical review. J Mater Chem A. 2015;3:8926. https://doi.org/10.1039/c4ta05033a.
17. Yuan LD, Deng HX, Li SS, Wei SH, Luo JW. Unified theory of direct or indirect band-gap nature of conventional semiconductors. Phys Rev B. 2018;98:1. https://doi.org/10.1103/PhysRevB.98.245203.
18. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Chapman and Hall/CRC; 1984.
19. Breiman L. Random forests. Mach Learning. 2001;45:5. https://doi.org/10.1023/A:1010933404324.
20. Neto MP, Paulovich FV. Multivariate data explanation by jumping emerging patterns visualization. IEEE Trans Visualizat Comput Graph. 2022. https://doi.org/10.1109/tvcg.2022.3223529.
21. Novak PK, Lavrac N, Webb GI. Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. J Mach Learn Res. 2009;10:377.
22. García-Vico A, Carmona C, Martín D, García-Borroto M, del Jesus M. An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects. WIREs Data Mining Knowl Discovery. 2018;8: e1231. https://doi.org/10.1002/widm.1231.
23. Zagorac D, Müller H, Ruehl S, Zagorac J, Rehme S. Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. J Appl Crystallogr. 2019;52:918–25. https://doi.org/10.1107/s160057671900997x.
24. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. APL Mater. 2013;1: 011002. https://doi.org/10.1063/1.4812323.
25. u. Mentel, mendeleev – a python resource for properties of chemical elements, ions and isotopes.https://github.com/lmmentel/mendeleev.
26. Tan P-N, Steinbach M, Kumar V. Introduction to data mining, 1st ed., Pearson, 2005;792.
27. Kane B, Cuissart B, Crémilleux B. Minimal jumping emerging patterns: computation and practical assessment. In: Cao T, Lim E-P, Zhou Z-H, Ho T-B, Cheung D, Motoda H, editors. Advances in knowledge discovery and data mining. Cham: Springer International Publishing; 2015. p. 722–33.
28. García-Borroto M, Martínez-Trinidad JF, Carrasco-Ochoa JA. Finding the best diversity generation procedures for mining contrast patterns. Expert Syst Appl. 2015;42:4859. https://doi.org/10.1016/j.eswa.2015.02.028.
29. Nonato LG, Aupetit M. Multidimensional projection for visual analytics: linking techniques with distortions, tasks, and layout enrichment. IEEE Trans Visualizat Comput Graphics. 2019;25:2650. https://doi.org/10.1109/TVCG.2018.2846735.
30. Pérez D, Zhang L, Schaefer M, Schreck T, Keim D, Díaz I. Interactive feature space extension for multidimensional data projection. Neurocomputing. 2015;150:611. https://doi.org/10.1016/j.neucom.2014.09.061.
31. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R. New York: Springer; 2013. https://doi.org/10.1007/978-1-4614-7138-7.
32. Gleicher M. Explainers: expert explorations with crafted projections. IEEE Trans Visualizat Comput Graphics. 2013;19:2042.

33. Knittel J, Lalama A, Koch S, Ertl T. Visual neural decomposition to explain multivariate data sets. IEEE Trans Visualizat Comput Graphics. 2020. https://doi.org/10.1109/TVCG.2020.3030420.
34. Dalpian GM, Wei S-H. Hole-mediated stabilization of cubic GaN. Phys Rev Lett. 2004. https://doi.org/10.1103/physrevlett.93.216401.
35. Bartel CJ, Sutton C, Goldsmith BR, Ouyang R, Musgrave CB, Ghiringhelli LM, Scheffler M. New tolerance factor to predict the stability of perovskite oxides and halides. Sci Adv. 2019. https://doi.org/10.1126/sciadv.aav0693.
36. Varignon J, Grisolia MN, Íñiguez J, Barthélémy A, Bibes M. Complete phase diagram of rare-earth nickelates from first-principles. npj Quantum Mater. 2017. https://doi.org/10.1038/s41535-017-0024-9.
37. Zhao X-G, Dalpian GM, Wang Z, Zunger A. Polymorphous nature of cubic halide perovskites. Phys Rev B. 2020. https://doi.org/10.1103/physrevb.101.155137.
38. Zhao X-G, Wang Z, Malyi OI, Zunger A. Effect of static local distortions vs. dynamic motions on the stability and band gaps of cubic oxide and halide perovskites. Mater Today. 2021;49:107. https://doi.org/10.1016/j.mattod.2021.05.021.
39. Zhuo Y, Mansouri Tehrani A, Brgoch J. enPredicting the band gaps of inorganic solids by machine learning. J Phys Chem Lett. 2018;9:1668.
40. Schleder GR, Acosta CM, Fazzio A. Exploring two-dimensional materials thermodynamic stability via machine learning. ACS Appl Mater Interfaces. 2019;12:20149. https://doi.org/10.1021/acsami.9b14530.
41. Schmidt J, Marques MRG, Botti S, Marques MAL. enRecent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5:83.
42. Yuan LD, Wang Z, Luo JW, Rashba EI, Zunger A. Giant momentum-dependent spin splitting in centrosymmetric low- z antiferromagnets. Phys Rev B. 2020. https://doi.org/10.1103/PhysRevB.102.014422.

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