npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01446-9

Spectral operator representations

Check for updates

Austin Zadoks[1] ✉, Antimo Marrazzo[2,3] & Nicola Marzari[1,4,5]

Machine learning in atomistic materials science has grown to become a powerful tool, with most approaches focusing on atomic geometry, typically decomposed into local atomic environments. This approach, while well-suited for machine-learned interatomic potentials, is conceptually at odds with learning complex intrinsic properties of materials, often driven by spectral properties commonly represented in reciprocal space (e.g., band gaps or mobilities) which cannot be readily partitioned in real space. For such applications, methods that represent the electronic rather than the atomic structure could be more promising. In this work, we present a general framework focused on electronic-structure descriptors that take advantage of the natural symmetries and inherent interpretability of physical models. We apply this framework first to material similarity and then to accelerated screening, where a model trained on 217 materials correctly labels 75% of entries in the Materials Cloud 3D database, which meet common screening criteria for promising transparent-conducting materials.

The past two decades have seen an explosion in the amount and availability of materials structures properties database[1–14]. Simultaneously, the infrastructure and protocols for performing high-throughput studies have matured and now allow for producing large volumes of high-quality data with ease[15–22]. As in computer vision, natural language processing, and other fields where the combination of data availability and machine learning (ML) techniques have enabled powerful technologies from autonomous driving to machine translation, data-driven materials science is a promising new approach for accelerated materials discovery, property prediction, and inverse design.

This promise is somewhat tempered by the unique problem of representation: traditional "xyz" descriptions of atomic configurations as a set of atomic positions and species $\{\mathbf{R}_I, \alpha_I\}$ cannot be used directly to efficiently drive traditional statistical models. Properties of atomic systems, like total energy and forces, are either invariant or equivariant to rotations and translations of atoms and to permutations in the order in which they are listed, while the atomic coordinates and species are not. Therefore, one of the foremost problems in data-driven materials science is how to efficiently and compactly describe relevant information about an atomic system in a framework suited to ML applications. There exist a few broad categories of approaches to solving this problem[23,24], including atomic-density field features like the smooth overlap of atomic positions (SOAP)[25], internal coordinates representations like Behler-Parrinello symmetry functions[26,27], atomic cluster expansions[28], and end-to-end neural network models, often based on atom graphs, like the CGNN[29], NequIP[30], or MACE[31] models. All of these approaches take atomic structures as the fundamental objects to process into inputs for ML models, and most decompose them into atom-centered motifs for the purpose of imposing translational invariance and aiding transferability. These approximations are well-founded for learning problems where the target property or properties are extensive or conceptually decomposable into atomic contributions. However, these approaches can be limited by their strong scaling with compositional complexity, degeneracies in the local atomic expansion at low body orders[32], and by the fundamental concept of atomic decomposition, which struggles to capture important electronic quantities like band gap, quasi-particle energies, electrical conductivity, or optical properties, to name a few.

For these applications, a class of descriptors that can capture the physics and interactions pertaining to these complex properties would be beneficial. A promising approach can be seen in methods that leverage electronic structure for featurization. Methods in this class include the spectrum of approximated Hamiltonian matrices (SPA$^{\mathrm{H}}$M)[33], the D-fingerprint[34], and moments of the density of states (DOS)[35,36], among many others[37–39], with successful applications to structure similarity[34,39], regression of various quantum-chemical properties[33], and delta learning of G_0W_0 quasi-particle energies[38]. These algorithms follow the general approach of computing the spectrum of a physical operator applied to a model electronic structure followed by a transformation into an ML descriptor. We formalize and generalize this process in a framework for designing electronic-structure features, which we call spectral operator representations (SOREPs).

Results

SOREP framework

SOREPs aim to describe materials using targeted features of their electronic structure. However, neither experimental nor high-throughput ab initio

[1]Theory and Simulation of Materials (THEOS), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland. [2]Dipartimento di Fisica, Università di Trieste, I-34151 Trieste, Italy. [3]Scuola Internazionale Superiore di Studi Avanzati (SISSA), I-34136 Trieste, Italy. [4]National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland. [5]Laboratory for Materials Simulations (LMS), Paul Scherrer Institut, CH-5232 Villigen, Switzerland. ✉e-mail: austin.zadoks@epfl.ch

materials databases generally provide detailed electronic-structure objects (e.g., Kohn–Sham orbitals), so the process begins with knowledge of only atomic structure. The first step in featurizing a material with a SOREP is to build a model electronic structure from atomic positions. The quality of this electronic-structure calculation determines the quality of the information content of the representation – a consideration that must be carefully balanced in terms of the corresponding computational cost of directly determining the quantity of interest. Predicting more complex physical phenomena may require more expensive but more faithful electronic-structure representations, while applications involving millions of systems might necessitate more cost-effective approximations. In general, this first step entails applying some map f of the atomic structure (positions, species, etc.), which yields, in principle, a many-body wavefunction or another representation of the electronic structure (e.g., a density matrix or Green's function)

$$f : \{\mathbf{R}_I, \alpha_I, \ldots\} \rightarrow \Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N). \tag{1}$$

However, for ML applications, many-body electronic structure calculations are impractical. A more pragmatic approach, and the one we will consider in moving forward, is to generate a set of single-particle orbitals from the atomic structure:

$$f : \{\mathbf{R}_I, \alpha_I, \ldots\} \rightarrow \{|\phi_i\rangle\} \tag{2}$$

This electronic system, however, it may be represented, exists in a much higher-dimension space than its originating atomic configuration (an atomic structure of N atoms can be considered to exist in a $3N$-dimensional Cartesian space, while in principle, its electronic wavefunction exists in the Hilbert space of the problem considered). In order to extract compact and useful information from this raw electronic structure, a Hermitian operator $\hat{A}$ selected from physical intuition or constructed through careful engineering can be projected onto the set of orbitals to compute the operator matrix elements

$$A_{ij} = \langle \phi_i | \hat{A} | \phi_j \rangle. \tag{3}$$

$\hat{A}$ may be simple and efficient to evaluate, like the identity or kinetic energy operators, more expressive yet expensive like the Kohn–Sham Hamiltonian, or somewhere in between like the various guess Hamiltonians explored in ref. 33. Here, we consider only scalar operators (in the physical sense, i.e. independent of changes in frame of reference) in order to achieve rotation and translation invariance of the matrix elements. A further generalization can be made to higher-order tensor operators, like position, if the features are to be used in an equivariant model or if further consideration is taken in the following steps to enforce reference-frame invariance. Regardless, the resulting operator matrix A represents a distillation of the electronic structure, filtered through the lens of the operator and expressed in the chosen basis.

To make use of all the information contained within the operator matrix, one could consider leveraging the matrix elements A_{ij} as ML features, as explored in ref. 38. Although invariant to translation, rotation, and other physically relevant symmetries (because $\hat{A}$ is scalar), the matrix elements are sensitive the choice of the basis functions $|\phi_i\rangle$. So, a key step in formulating a SOREP, mirroring the standard procedure for electronic-structure calculations, is to diagonalize the operator matrix

$$A|\phi_i\rangle = \lambda_i S|\phi_i\rangle, \tag{4}$$

using the overlap matrix S, to retrieve its set of eigenvalues $\{\lambda_i\}$. This procedure removes explicit dependence on the choice of basis (for complete bases), and, significantly, it also mixes the information contained in the operator matrix in a non-trivial and physically meaningful manner[40,41]. In order to bring the eigenspectrum into a system-independent constant-dimensional space, as is required by all ML models, and to enforce invariance to permutations of the eigenvalue indices, the final step of the SOREP procedure is to apply a map g from the set of eigenvalues λ_i to a feature vector $\mathbf{x}$:

$$g : \{\lambda_i\} \rightarrow \mathbf{x}. \tag{5}$$

One simple and compact method for systems with few eigenvalues is to sort the spectrum and pad it with zeros up to a common constant dimension, as done in the SPA$^{\mathrm{H}}$M method[33]. However, the resulting features are discontinuous w.r.t. level crossings and are high-dimensional for systems where many eigenstates are considered (e.g., periodic systems sampled at many k-points and/or with many bands). To remedy these shortcomings, a DOS computed on a basis, e.g., as a sum-over-poles[42] or using polynomials[23], and sampled on a fixed domain can be more compact and is smooth w.r.t. level crossings. Other maps used in the literature are spectral histograms[34,37,39,43], moments of the DOS[35,40], and radially-decomposed projected densities of states[38].

In general, SOREPs exhibit many desirable properties for atomistic descriptors "for free" due to the properties of spectra[23]. Key symmetry invariances, such as rigid translation and rotation, are ensured by construction through the utilization of scalar operators. Beyond respecting physical symmetries, atomic descriptors should be complete; i.e. they should always distinguish (symmetry-)inequivalent structures. It has been shown[32] that low body-order local atomic descriptors can suffer to an extent from incompleteness, mapping distinct configurations to (nearly) identical descriptors. Overlap matrix (OM) fingerprints, as atom-centered spectral representations that leverage the overlap (identity) operator, have been seen in practice to lift these degeneracies[23,44]. However, the limits of the completeness of spectral representations, in particular of global (i.e., not atomically-decomposed) fingerprints, have not been rigorously bounded[45]. Additionally, because many properties of interest (e.g., total energies) vary smoothly with continuous deformations of the atomic structure, feature maps are often constructed to be similarly smooth. It is key to note that this criterion is intended to ensure that no *nonphysical* discontinuities can be found in the feature map, which could lead to, e.g., spurious discontinuities in a learned potential energy surface. Some electronic properties, like band gaps and Van Hove singularities, may not be smooth w.r.t. structural deformations. Unlike local atomic descriptors, spectral representations can capture these physical discontinuities and should be better suited for learning similarly non-smooth properties.

To provide a more direct understanding of what this procedure entails and to show how specific constraints influence choices in each of the steps, we next consider the case where one aims to minimize as much as possible the computational cost of featurization while maintaining a spectral representation.

Kinetic energy SOREP

Often, the featurization of millions of structures may be required in order to apply ML to a given problem, for example, in testing structure uniqueness or learning from frames of molecular dynamics trajectories. In these situations low-cost featurizations are essential, so here we discuss how to design a SOREP with this constraint in mind. Generally, the diagonalization of operator matrices is the most computationally demanding step in producing SOREP features, but, as discussed above, it is essential in enforcing various symmetry invariances and in capturing non-local properties. Therefore, the important ingredients to consider for optimization are the choices of how to determine and represent the electronic structure, which operator to apply, and how to map the operator spectrum onto features after diagonalization.

An appealing electronic-structure model for these purposes is a linear combination of contracted Gaussian-type orbitals (cGTOs), for which many basis sets have been constructed alongside efficient libraries like `libcint`[46]

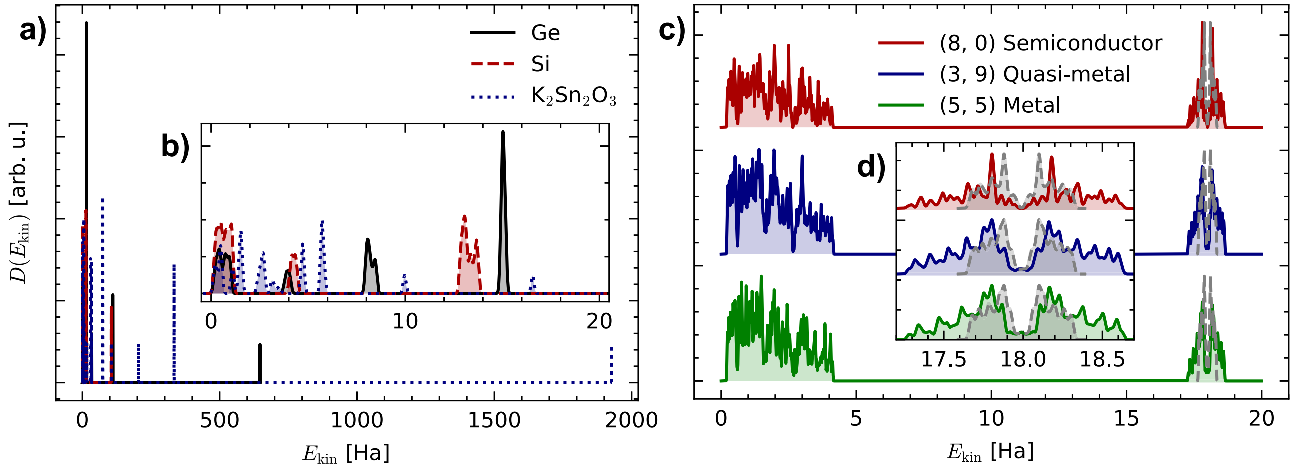


Fig. 1 | Illustrative examples of kinetic energy SOREP features. **a** Kinetic energy SOREPs of diamond-structure Ge and Si along with the transparent-conducting oxide $K_2Sn_2O_3$. The width of the spectral range (≈2000 Ha) is dominated by tightly bound semi-core orbitals of $K_2Sn_2O_3$. **b** The two elemental solids display qualitatively similar features at low energies ($\lesssim$20 Ha). **c** Kinetic energy SOREPs for three carbon nanotubes of different chiralities, overlaid with p_z tight-binding densities of state (dashed lines, shifted by +18 Ha to be similarly centered). **d** The transition from metallic to insulating configurations is correlated with a disappearance of the kinetic DOS in the (3,9) and (0,8) configurations around ~18 Ha.

for applying operators and computing integrals analytically. A cGTO with quantum numbers n, l, m for atom of species α is constructed as the product of a spherical harmonic $Y_l^m(\theta, \phi)$ and a radial function

$$R_{nl}^{\alpha}(r) = r^l \sum_p c_p^{\alpha} B(l, a_p^{\alpha}) e^{-a_p^{\alpha} r^2} \tag{6}$$

where c_p^{α} and a_p^{α} are the contraction coefficient and exponent for species α in the primitive Gaussian p, and B is a normalization constant. The cGTO for atom I of species α_I at position R_I is therefore

$$\phi_{nlm}^I(\mathbf{r}) = R_{nl}^{\alpha_I}(|\mathbf{r} - \mathbf{R}_I|) Y_l^m(\theta, \phi). \tag{7}$$

For periodic systems, we can write approximate Bloch states as Bloch sums of cGTOs

$$\phi_{\nu\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} \phi_{nlm}^I(\mathbf{r} - \mathbf{R}) \tag{8}$$

with band index ν capturing the cGTO indices n, l, m, and I, and crystal wave vector $\mathbf{k}$. We define the electronic-structure map as a simple decoration of the atomic positions with the cGTOs of the corresponding species, which requires no significant computational effort.

A simple yet descriptive one-electron integral is the kinetic energy, which can be applied to the Bloch sums

$$T_{\nu\mu\mathbf{k}} = \langle \phi_{\nu\mathbf{k}} | -\frac{\hbar^2}{2} \nabla^2 | \phi_{\mu\mathbf{k}} \rangle \tag{9}$$

and subsequently diagonalized. The density of kinetic energy eigenvalues per unit volume can then be calculated using Gaussian smearing. Volume normalization ensures that the spectra for unit cells and supercells are identical, which is the desired behavior in solids when predicting intrinsic properties such as structural similarity or electronic band gap. The kinetic energy SOREP features are therefore

$$x_i = \frac{1}{V} \frac{1}{N_\nu N_{\mathbf{k}}} \sum_{\nu\mathbf{k}} \exp\left[\frac{-[(E_i - \lambda_{\nu\mathbf{k}})/(k_B T_s)]^2}{\sqrt{\pi}}\right] \tag{10}$$

where T_s is a smearing temperature, E_i are uniformly-spaced energies running from $E_{\rm min}$ to $E_{\rm max}$, and $\lambda_{\nu\mathbf{k}}$ are the kinetic energy eigenvalues.

Figure 1a shows the kinetic energy SOREPs for silicon, germanium, and $K_2Sn_2O_3$ calculated using a customized version of the atomic natural orbital (ANO) cGTO basis set[47–55] which covers the full energy spectrum. This custom basis set, which we call ANO-ML-OS, is obtained from the relativistic ANO-type orbitals[50–55] known as ANO-RCC, as available on the Basis Set Exchange library[47–49], where we keep only the orbitals corresponding to the smallest closed-shell configuration. As one might expect, the SOREPs for the two elemental solids look qualitatively quite similar, mirroring their similar structural electronic properties. When comparing these materials to a more complex system with heavier elements, like $K_2Sn_2O_3$ as shown, we observe a rapidly growing spectral range due to the increasingly highly localized nature of the additional tightly bound (semi-) core orbitals. In order to improve the features for this purpose, one might consider modifying the operator by adding a nuclear potential or changing the final representation to one that compresses the spectrum more than the DOS. For example, the DOS could be truncated at the energy where the cumulative DOS is some fraction of its maximum, following the observation that only a small amount of the DOS is at high energy. Applying such a cutoff would yield features similar to those pictured in Fig. 1b. The underlying complication is not only that the spectral range is large but that it differs between chemistries in a way that makes it difficult to select the most important region(s) of the DOS beyond the intuition that low-curvature, relatively delocalized, and thus low-kinetic-energy states are more chemically meaningful than high-curvature, highly-localized, high-kinetic-energy ones. However, for systems of the same or similar compositions, such as the carbon nanotubes (CNTs) shown in Fig. 1c, this filtering is more intuitive. Here, we compare the kinetic SOREPs as above for CNTs of varying electronic character and find that the features around 18 Hartree, as seen in Fig. 1d, are quite similar to the p_z tight-binding DOSs[56,57] (shown in gray) and exhibit a gap forming for the semiconducting (0, 8) configuration. We conclude that the kinetic features are well-suited to comparing such compositionally similar materials, like in molecular dynamics, metallic alloys, or elemental systems with many allotropes.

One application in this regard is identifying unique structures of fixed stoichiometry in a large database. For illustration, we have selected the relaxed geometries for 127 $BaTiO_3$ entries in the Materials Cloud 3D database (MC3D)[58] feedstock and compare against the uniqueness analysis conducted by the MC3D developers using the pymatgen structure matcher[59]. To find an appropriate energy range for featurization, we first analyze the inexpensive-to-compute cumulative distribution of kinetic eigenvalues across all structures, shown in Fig. 2a. This cumulative

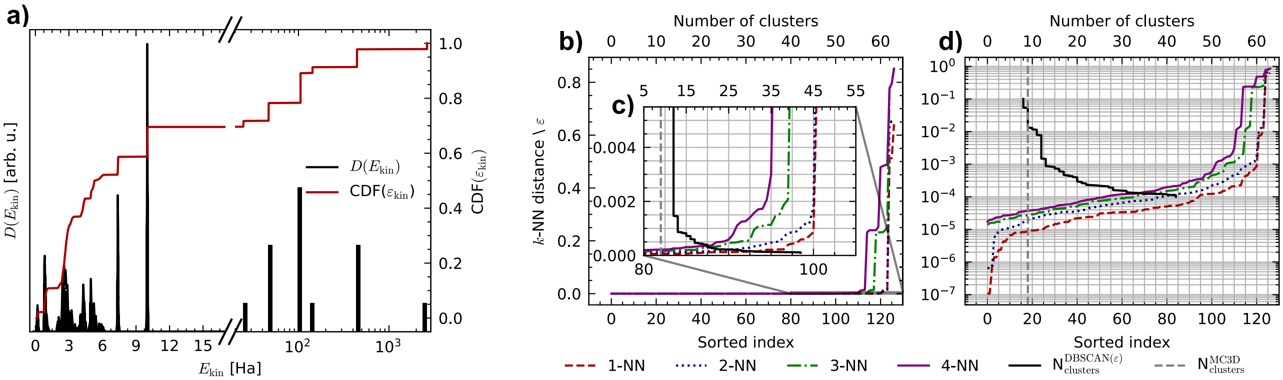
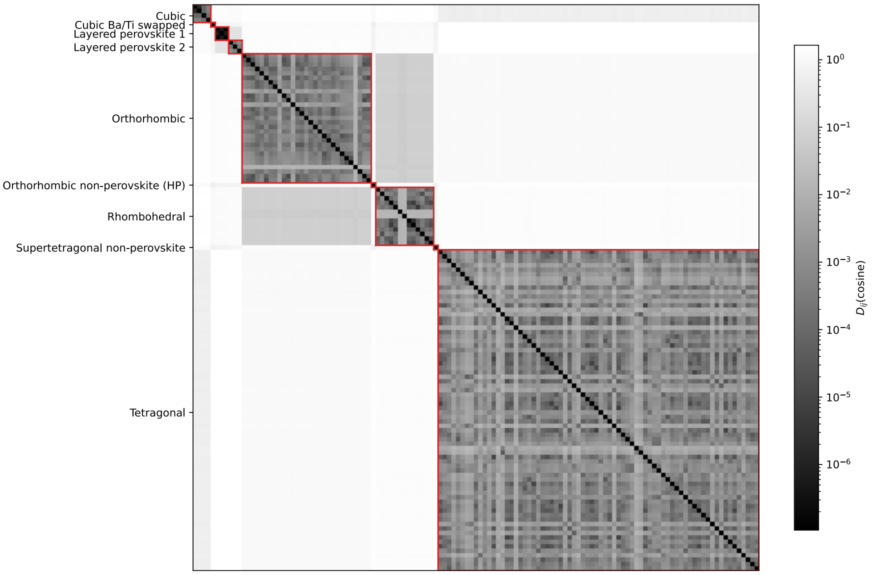


Fig. 2 | Feature selection and DBSCAN parameterization for $BaTiO_3$ uniqueness analysis. a Density and cumulative distribution function of kinetic eigenvalues for all $BaTiO_3$ structures considered. Below 15 Hartree, the DOS is computed with Gaussian smearing; above, a logarithmically binned histogram is shown. **b–d** Sorted k-nearest neighbor distance curves on linear and semi-log scales display an "elbow" at optimal values of the DBSCAN neighborhood radius parameter ε. DBSCAN models are fit for ε values within the elbow region (~1×10^{-4} to 1×10^{-1}), and the corresponding number of clusters is shown in solid black. The number of structure groups determined by the MC3D procedure is overlaid in dashed gray for reference. An optimal choice of ε exists in the region of relatively stable clustering around 1×10^{-3} to 1×10^{-1}.

Fig. 3 | Cosine distance matrix relating kinetic SOREPs for 127 $BaTiO_3$ structures with columns sorted by MC3D classification. Some outliers stand out visually in the orthorhombic and rhombohedral groups, alongside possible outliers in the tetragonal block.



distribution function increases in quasi-discrete steps at energies higher than approximately 7 Hartree. Past this point, the density of kinetic states is likely dominated by tightly bound (semi-)core states, which are likely uninformative. The DOS computed up to 15 Hartree, also in Fig. 2a, is indeed sparse, highly peaked, and therefore likely related to highly localized Ba and Ti cGTOs above ~6.5 Hartree. The final SOREP features are therefore computed from 0 to 6.5 Hartree using Gaussian-type smearing with a width of 0.03 Hartree sampled at 1024 equally-spaced energies.

To determine a set of unique prototype structures, the SOREP features are clustered using the density-based spatial clustering of applications with noise (DBSCAN)[60] algorithm, which has two parameters: the minimum number of samples required to create a cluster N_{min}, and a neighborhood radius ε. DBSCAN performs its clustering based on the distances between data points, not the data themselves; here, we use the cosine distance $d_{cos}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{||\mathbf{x}_i|| \, ||\mathbf{x}_j||}$. The distance matrix, sorted by the MC3D grouping, is shown in Fig. 3. Here, there are three groups containing only one structure each (an orthorhombic non-perovskite, a super-tetragonal non-perovskite, and an erroneous structure with the positions of barium and titanium swapped) along with six groups with multiple members (two-layered perovskites, and the four standard polymorphs). However, it can be seen that some groups (notably those labeled as orthorhombic and rhombohedral) contain structures that are relatively unlike the rest of their respective clusters.

As we expect to find some completely unique structures with no duplicates in the database, we set $N_{min} = 1$ for the following investigation. To determine ε, we inspect the sorted k-nearest neighbors (k-NN) distances (Fig. 2b–d) and find an "elbow" in the curve, which indicates a domain of reasonable values, from approximately 1×10^{-4} to 1×10^{-1}. Correlating these values of ε with the number of clusters found by DBSCAN, it can be seen that within this region of the k-NN distance curve, the number of unique structure groups ranges from >50–8 with increasing ε, compared to the 9 groups predicted by the MC3D procedure. However, the clustering is highly sensitive to small changes in the neighborhood radius for $\varepsilon < 9 \times 10^{-4}$, so we focus on values in the range [9×10^{-4}, 1×10^{-1}], which yield between 12 and 8 groups. The 9 clusters produced for ε in the range [0.014, 0.05] are identical to those from the MC3D as determined using `pymatgen`'s structure matcher. To understand how the clustering at lower and higher values of ϵ differs within the focus region, we describe the process of cluster merging as ε increases in Fig. 4. Of the 13 groups generated at low ε, most are in agreement with the reference groups aside from a splitting of the tetragonal, orthorhombic, and rhombohedral clusters, where the visual outliers seen in Fig. 3 are separated.

We conclude from this analysis that the kinetic SOREP features have the ability to capture and describe subtle structural differences in polymorphs of complex materials, similar to structure-based methods like the `pymatgen` structure matcher. However, significant weight is given specifically to structural changes that strongly affect the electronic structure, as seen in the CNTs. This simple and efficient example serves as a good case study for how one might approach crafting and electronic-structure featurization under the SOREP framework with a quite restrictive efficiency constraint. As mentioned above, more complex learning problems often require more expressive descriptors, so next, we consider constructing a featurization for such a situation.

Single-shot DFT SOREP

For applications such as screening large and diverse databases of materials, a representation is required that is rich enough to describe and compare any chemical composition but computationally efficient enough to be applied to tens of thousands of systems (containing up to ~100 atoms each). To guide the development of a SOREP fit for this application, we consider as a use case an ML-accelerated screening for transparent-conducting materials (TCMs) in the MC3D. TCMs are characterized by band gaps wide enough to allow for transparency across the visible spectrum, high mobility of charge carriers, and the ability to inject these carriers via n- or p-type doping. Most screening studies for these materials focus initially on approximating the first two properties via high-throughput density-functional theory (DFT) band-structure calculations[61–63]. From these calculations, the DFT-PBE[64] band gap and approximate electron and hole effective masses are used as figures of merit. Our aim is to define a featurization method descriptive enough to reproduce a classification based on effective masses and band gaps at a fraction of the cost. More concretely, we target an order of magnitude speedup compared to self-consistent DFT calculations; otherwise, it would be more efficient and practical to perform a traditional screening. Using these guiding principles, we propose a SOREP method based on a single-shot (i.e., non-self-consistent) DFT calculation of a superposition of pseudo-atomic valence charge densities $\tilde{\rho}$ and a linear combination of pseudo-atomic orbitals (PAOs) χ_{nl} taken from pseudopotentials from the standard solid-state pseudopotential (SSSP) library[65,66]. These pseudo-atomic quantities are exact matches to the all-electron quantities of an isolated atom outside a small pseudization radius and as such represent a reasonable guess of the true ground-state wavefunction and charge density of the chemically-active electrons of each element. Using a locally modified copy of the QUANTUM ESPRESSO[67] `pw.x` code, the PAOs (provided on a real-space radial grid) are transformed into Bloch orbitals. The Kohn–Sham DFT Hamiltonian and orbital overlap matrices are then calculated non-self-consistently from the potential derived from the superposition of atomic densities

$$H_{\nu\nu'\mathbf{k}} = \langle \chi_{\nu\mathbf{k}} | \hat{H} | \chi_{\nu'\mathbf{k}} \rangle \tag{11}$$

$$S_{\nu\nu'\mathbf{k}} = \langle \chi_{\nu\mathbf{k}} | \chi_{\nu'\mathbf{k}} \rangle \tag{12}$$

where

$$\hat{H} = \hat{T} + \hat{V}_{H[\tilde{\rho}]} + \hat{V}_{xc[\tilde{\rho}]} + \hat{V}_{ext} + \hat{V}_{PS}. \tag{13}$$

The Hamiltonian matrix then is diagonalized exactly (i.e., non-iteratively) on the basis of the PAOs to find the eigenvalues $\varepsilon_{n\mathbf{k}}$ and eigenstates $\psi_{n\mathbf{k}}$ at each $\mathbf{k}$-point

$$H_{\mathbf{k}} \psi_{n\mathbf{k}} = \varepsilon_{n\mathbf{k}} S_{\mathbf{k}} \psi_{n\mathbf{k}}, \tag{14}$$

yielding a $\mathbf{k}$-resolved eigenspectrum, i.e. a band structure. With respect to the kinetic energy operator used above, the Kohn–Sham Hamiltonian is well-behaved due to the inclusion of potential terms, with a meaningful Fermi energy and band extrema that can be leveraged as anchoring points. Finally, the features are the DOS calculated with Gaussian smearing as in Eq. (10) where the discretization over energies E_i, and smearing temperature T_s are taken as parameters of the featurization.

Figure 5 shows the evolution of the DFT DOS with respect to the electronic self-consistent step for a reference semiconductor, elemental silicon, and the transparent-conducting oxide $K_2Sn_2O_3$. Remarkably, the qualitative shape of the DOS is quite similar to the fully converged calculation, especially around the Fermi level. Importantly for the application of

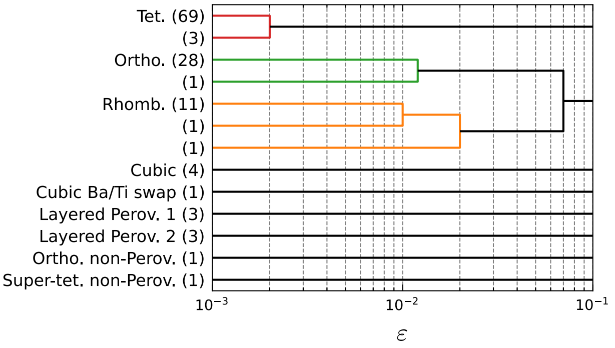


Fig. 4 | Progression of DBSCAN clusters with varying neighborhood radius parameters ε. The number of materials in each initial cluster is shown in parentheses. As the neighborhood radius is increased, clusters merge, agreeing with the MC3D's structure-based grouping for values from 2×10^{-2} to ~5×10^{-2}, after which meaningfully distinct clusters combine.

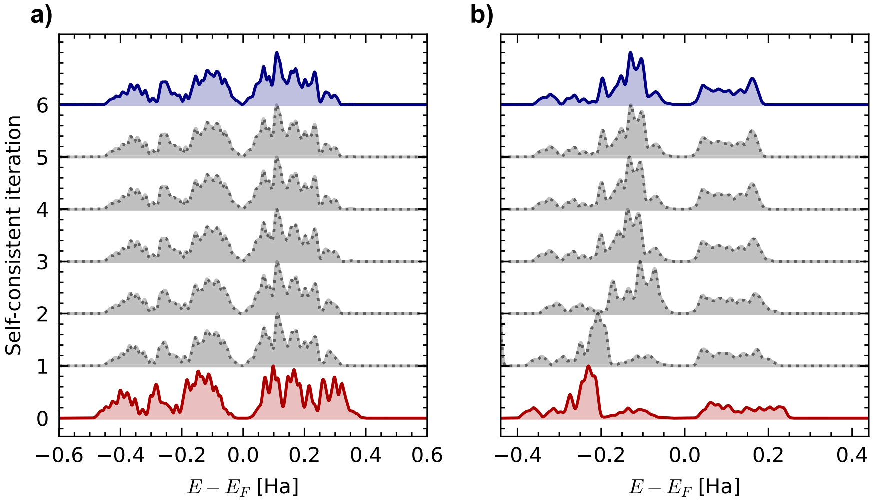


Fig. 5 | Densities of states for silicon and $K_2Sn_2O_3$ across self-consistent iterations. a shows elemental silicon, while **b** shows the transparent-conducting oxide $K_2Sn_2O_3$. "0" iterations correspond to a single-shot SOREP, while in both cases, the uppermost DOS corresponds to the self-consistent DFT ground state. The SOREP for silicon is remarkably similar to the converged DOS, while the features for the more complex $K_2Sn_2O_3$ are further from the ground-state solution.

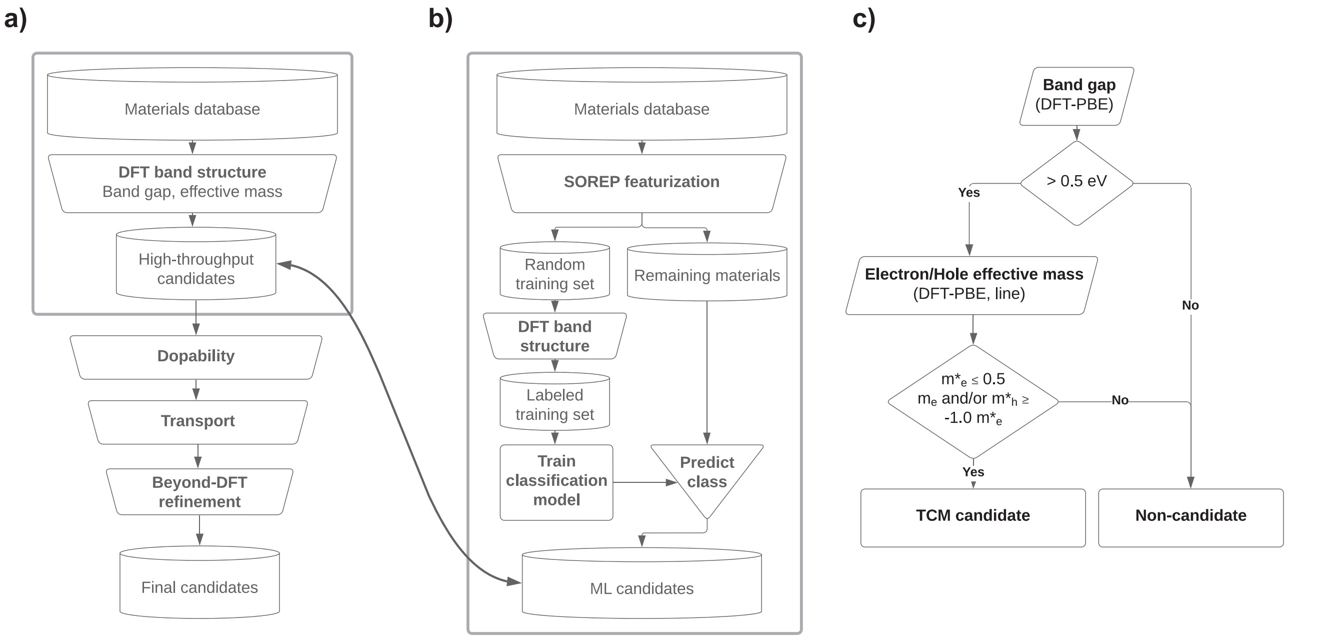


Fig. 6 | Flow diagrams illustrating the ML-accelerated screening procedure. **a** Standard high-throughput screening for TCMs, where the boxed section corresponds to the steps accelerated by SOREP-based ML. **b** ML-accelerated procedure: random materials from the database are used to train a classification model that labels the remaining materials. With a perfect classification model, the "ML candidates" in **a** would correspond exactly to the "high-throughput candidates" in **b**. **c** TCM screening criteria based on band gaps and hole and electron effective masses.

screening TCMs, the shape of the DOS at the conduction and valence band edges is well-reproduced in the single-shot SOREPs, so the features contain, to some extent, reliable information related to the electron and hole effective masses. A small-scale timing study is reported in Supplementary Information (SI) Sec. S3, where it is shown that single-shot DFT calculations do, in fact, meet the order-of-magnitude performance increase over self-consistent DFT targeted above.

Accelerated TCM screening

A generalized procedure for screening TCMs is shown in Fig. 6a, and consists of two broad steps: high-throughput DFT calculations (boxed) and low-throughput refined calculations. The most computationally expensive step is running DFT band-structure workflows for every material in the database, numbering often in the tens of thousands. These calculations are used to find the band gap and effective mass (proxies for transparency and conductivity, respectively) following the criteria shown in Fig. 6c. To accelerate the procedure, we featurize the database using the single-shot DFT SOREPs and construct a classification model which predicts which materials are likely to meet the DFT-based screening criteria (Fig. 6b). This approach significantly reduces the computational cost by performing DFT band-structure calculations only on a subset of the entire database, including, if available, some known TCMs in order to construct an ML model used to screen the rest of the database. Note that this approach is more general than the case of TCMs; screening studies for many materials classes and properties could follow a similar procedure.

As this is a validation study and not a true screening, we take the MC3D, a curated database of relaxed three-dimensional crystal structures, for ground-truth DFT simulation results. Using band structures from the MC3D, all materials are classified as TCM candidates or non-candidates based on the criteria shown in Fig. 6c. The filter, adapted from general guidelines outlined by Woods-Robinson et al.[63], selects candidate materials first by having a generalized gradient approximation (GGA) electronic band-gap wider than 0.5 electron-volts. Then, if the material meets either the electron ($m_e^* \le 0.5 m_e$) or hole ($m_h^* \ge -1.0 m_e$) effective mass condition (or both), it is considered a candidate material; otherwise, it is labeled as a non-candidate.

The band gap is simply calculated as the difference between the conduction band minimum (CBM) and valence band maximum (VBM) on a Monkhorst-Pack grid $E_g = E_{\mathrm{CBM}} - E_{\mathrm{VBM}}$. The electron effective masses are approximated from band structures computed along high-symmetry lines provided by SeeKPath[68] using the so-called "line effective mass" of Hautier et al.[62]:

$$\frac{1}{m^*_{e,\mathrm{line}}} = \max_{\alpha}\left[\frac{\sum_{n\in\mathrm{CB}}\int_{k_{\alpha a}}^{k_{\alpha b}} -\frac{\partial^2}{\partial k_\alpha^2}\varepsilon_n(k_\alpha)\theta_e(\varepsilon_n(k_\alpha))dk_\alpha}{\sum_{n\in\mathrm{CB}}\int_{k_{\alpha a}}^{k_{\alpha b}}\theta_e(\varepsilon_n(k_\alpha))dk_\alpha}\right] \qquad (15)$$

where the maximum is taken over high-symmetry lines α, the sum is over conduction bands, and θ is the Fermi-Dirac distribution at 300 K:

$$\theta_e(E) = \left[\exp\left(\frac{E - E_{\mathrm{CBM}}}{k_B T_{300K}}\right) + 1\right]^{-1}. \qquad (16)$$

Hole effective masses are approximated similarly by exchanging $E - E_{\mathrm{CBM}}$ in the Fermi-Dirac distribution for $E_{\mathrm{VBM}} - E$ and summing over valence, rather than conduction, bands in Eq. (15).

These labeled data are then used to train a random forest classifier (RFC) to predict the DFT-derived binary classification based on the SOREP features; the RFC model is chosen for its simplicity and interpretability. Unlike neural network models, which are often quite opaque, the binary decisions in each of the trees of the random forest are easily understood as conditions on the DOS at particular energies, and the model can provide a measure of the importance of each of the input features. Because the input features have clear physical meanings, the results of the model training not only provide classification predictions but also useful feedback for improving the SOREP features if necessary.

Featurization

SOREP Featurization begins with performing single-shot calculations on a subset of the MC3D, followed by DOS sampling. Statistics describing the materials considered, various computed properties, and the time cost of these calculations are reported in the SI Sections S1.1 and S3. We investigate

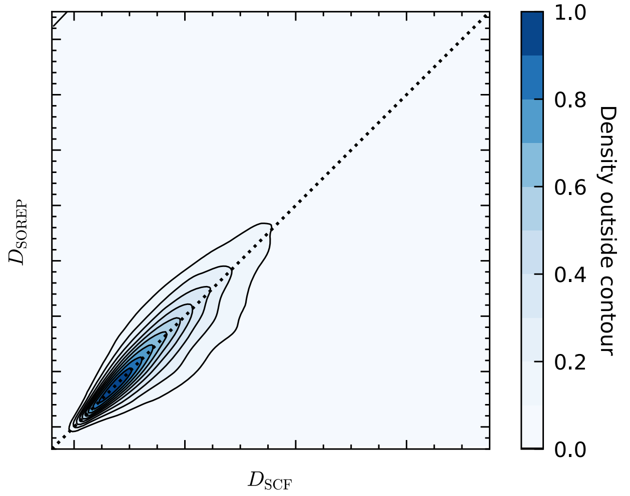


Fig. 7 | Parity plot comparing Euclidean distances among Fermi-centered SCF DOSs and corresponding distances among single-shot DFT SOREPs. The distances are highly correlated between the two sets of features, suggesting that the space of single-shot densities of states is not too different from that of fully self-consistent DOSs, which have been previously employed for materials cartography.

four different SOREP parameterizations derived from the single-shot band structures, focusing on different aspects of the DOS in a similar spirit to a weighted DOS fingerprint[43]. In the first, the DOS is sampled at 513 evenly-spaced energies 2 eV below and 6 eV above the VBM. These features are designed with selecting hole-conducing TCMs in mind; by fixing the DOS at the VBM to a specific feature, the model has a higher likelihood of learning to distinguish materials with high DOS and likely low effective mass at that point. Including 6 eV above the VBM should also provide enough information about the band gap so that insulators may be distinguished from conductors. For most insulators, this energy range should also provide a glimpse at the bottom of the conduction bands. The second set of parameters yields a standard Fermi level-centered DOS, sampled on a range of ±5 eV, allowing the valence and conduction bands to be captured in any material with a band gap less than 10 eV. For insulating materials, the Fermi level is taken as the mid-gap energy so that the valence and conduction are equally well represented. A fourth parameterization mirrors the VBM-centered features but with the CBM as the anchoring point, providing a feature set that is likely to be more useful for identifying n-type TCMs. Finally, a concatenated SOREP is constructed by combining features centered at the VBM, Fermi level, and CBM, each sampled on a range of ± 1 eV at 171 evenly-spaced energies to yield 513 total features.

In addition to these SOREP features, we also consider the SOAP[25] representation, which is a widely used and well-established local atomic environment descriptor. As an atomically decomposed representation, SOAP provides a vector of features for each atom in the unit cell whose dimension is determined by various parameters, notably the maximum n and l values in a spherical harmonic expansion of the local atomic density. In order to produce feature vectors that are comparable in dimension to the SOREPs, we fix $n_{\text{max}} = 10$, $l_{\text{max}} = 9$, and average over atoms using the "inner" approach implemented in DScribe[69,70], yielding 550 features.

As shown in refs. 34 and [39], distance metrics based on the self-consistent DFT DOS can be used in practice to both map out the space of electronic structures and to search for materials with complex electronic properties. To more rigorously investigate the information content of the SOREPs described above, we observe the correlation between the properties of the Fermi level-centered single-shot DFT SOREP features and identically sampled SCF densities of state from the MC3D. Figure 7 shows a strong correlation between self-consistent DFT and SOREP Euclidean distances across the database, confirming that this featurization not only contains physically relevant information but that it yields a topologically similar space to true self-consistent densities of state. This confirmation opens the door to applying SOREPs to materials cartography and other further investigations into electronic-structure space and its dimensionality using tools like DADApy[71].

Model training and performance

Using the three SOREP parameterizations described above, balanced RFC models, which correct for heavily imbalanced training data, as we have here, are trained using a gridded parameter cross-validation with $k = 5$ folds on a range of train-test splits. On a 14-core workstation, training all the models for all train-test splits and all SOREP features can be done in less than an hour.

A first question to ask after training the models is: what features do the models find to be most useful in their predictions? Figure 8 shows the mean mean decrease in impurity (MDI) feature importances for the four types of SOREP over the model ensemble trained on 50% of the database. The VBM-centered model gives significant importance to the valence bands and exhibits a strong peak just above the VBM and into the gap (if present), likely to probe the shape of the decay of the DOS at the band edge and the width of the gap. This follows the physical intuition used in constructing these features: the shape of the bands around the VBM and information about the band gap are both important for classifying materials as TCMs given the criteria imposed. This same behavior is seen mirrored in the CBM-centered features, with a strong peak just below the conduction band edge and a similar decay into the potential gap. The Fermi-centered SOREP has its importance peak just above and below the Fermi level, with smaller subpeaks just below and above the positive and negative limits, which define a 0.5 eV band gap, respectively. Because no anchoring point for the band edges is present in these features, the model expectedly focuses on distinguishing between conductors and insulators as defined by the screening criteria. Finally, the concatenated SOREP's importance show all the same major features as their component parts, as discussed above. As informative and affirming as this analysis seems to be with respect to the physical meaning of the features making up each SOREP, it is important to note that MDI feature importances are solely derived from training data and suffer from the same biases as the model itself (e.g., overfitting to noise, class imbalance, etc.). Despite these limitations, the feature importances provide an efficient and useful, if heuristic, guide for understanding in a broad sense how well the model agrees with an investigator's physical intuition and can point to possible improvements in SOREP engineering and parameterization. For example, importance seems to be much more highly localized around the anchoring points than we expected in all the parameterizations presented, suggesting that lower-dimensional features with a more targeted sampling of the DOS might be equally effective and more computationally efficient.

Although the feature importances tell quite different stories for each of the different SOREPs, classification metrics evaluated on a hold-out set of 10% of the database shown in Table 1 confirm that the relative performance of the features is comparable within the variance due to training data selection. In broad terms, the models trained using 1% of training data each achieve true positive rates of above 60% and up to 75% with corresponding false negative rates of between 25% and 38%. These quantities are important measures of how well a given classification model can accelerate a screening: a high false positive rate would mean that significant calculation time is wasted on false leads, while a low true positive rate would signal that many good candidates go overlooked. The effects of these factors are shown in Fig. 9. Figure 9a shows the relative speedup achieved by RFCs trained on each type of feature which is calculated as the ratio of TCMs found per calculation by the ML-accelerated screening to that of a high-throughput search:

$$\text{speedup} = \frac{\text{yield}_{\text{ML}}}{\text{yield}_{HT}} = \frac{\frac{N_{\text{TP}}+N_{\text{train}}*f_{\text{TCM}}}{N_{\text{TP}}+N_{\text{FP}}+N_{\text{train}}}}{f_{\text{TCM}}}. \quad (17)$$

N_{TP} is the number of true positives (TCMs labeled as such), N_{FP} the number of false positives (non-TCMs labeled as TCMs), N_{train} the number of

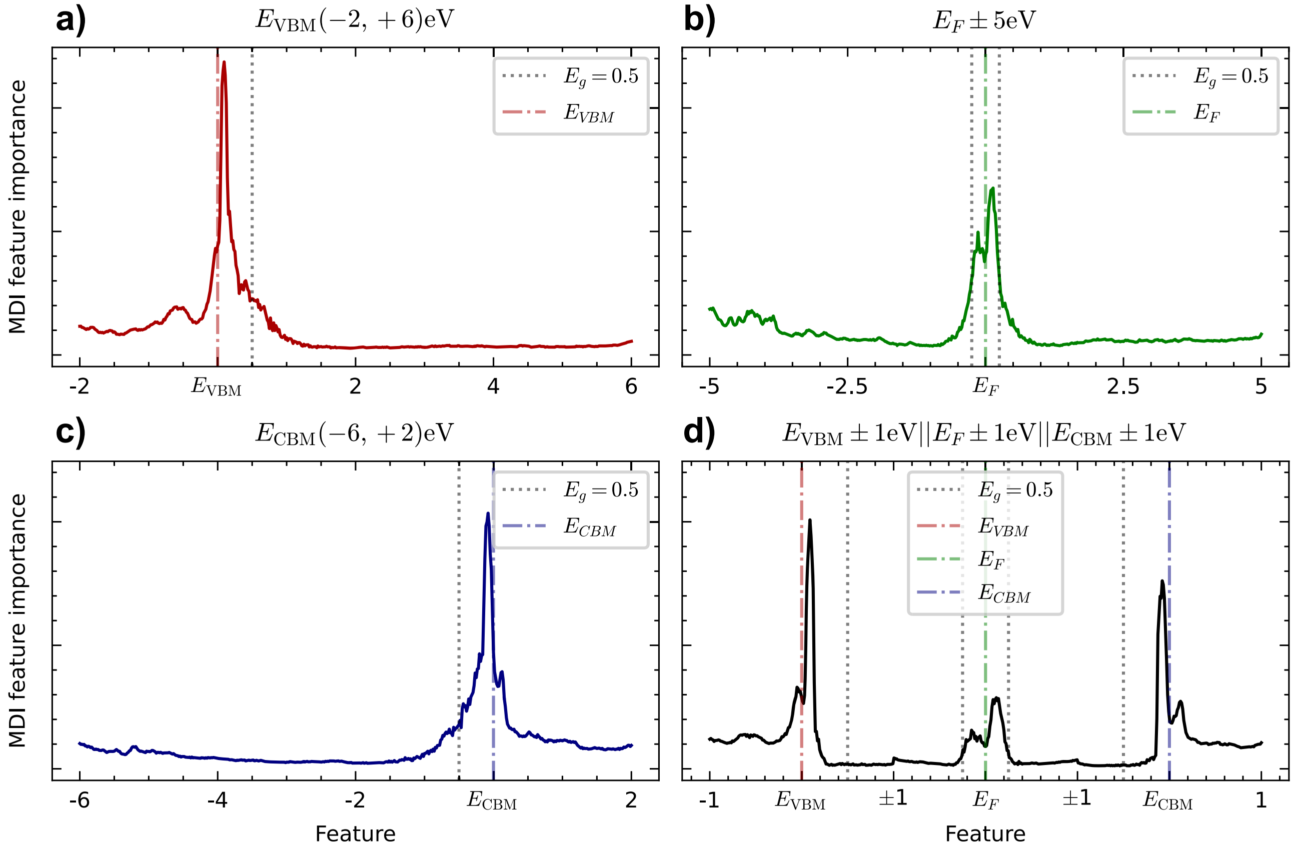


Fig. 8 | Average mean decrease in impurity feature importances over 30 random forest classification models trained on independent 50:50 splits of the training data. a Importance reaches a maximum at the VBM anchor point and decays far away, with a slight increase of importance ~0.5 eV above the VBM. **b** Fermi-centered SOREPs give large importance within a region of ±0.25 eV around the (gap-centered) Fermi level. **c** CBM-centered SOREPs show an importance distribution mirroring that of the VBM features. **d** Concatenated SOREPs give, as expected, a superposition of the importance distributions of their individual components. In all SOREPs, feature importances peak around the anchoring points (Fermi level, band edges) and decay strongly outside the 0.5 eV band gap range away from the anchoring points.

Table 1 | Mean values and standard deviations of classification metrics for models trained on 1% of the data computed over 30 independent train-test splits and evaluated on a hold-out set

Features	True negative (Rate)	True positive (Rate)	False negative (Rate)	False positive (Rate)	Balanced accuracy
VBM-centered	1704 ± 52	151 ± 14	62 ± 14	253 ± 52	0.58 ± 0.06
	(0.871 ± 0.026)	(0.710 ± 0.067)	(0.290 ± 0.067)	(0.129 ± 0.026)	
Fermi-centered	1684 ± 51	134 ± 18	79 ± 18	273 ± 51	0.49 ± 0.08
	(0.860 ± 0.026)	(0.628 ± 0.087)	(0.372 ± 0.087)	(0.140 ± 0.026)	
CBM-centered	1682 ± 72	134 ± 20	79 ± 20	275 ± 72	0.49 ± 0.07
	(0.859 ± 0.037)	(0.628 ± 0.095)	(0.372 ± 0.095)	(0.141 ± 0.037)	
Concatenated	1705 ± 48	**159** ± **13**	**54** ± **13**	252 ± 48	0.62 ± 0.05
	(0.871 ± 0.024)	(**0.746** ± **0.061**)	(**0.254** ± **0.061**)	(0.129 ± 0.024)	
SOAP	**1783** ± **71**	36 ± 13	177 ± 13	**174** ± **71**	0.08 ± 0.04
	(**0.911** ± **0.036**)	(0.168 ± 0.060)	(0.832 ± 0.060)	(**0.089** ± **0.036**)	

The best-performing features are shown in bold; concatenated SOREPs have the best true positive and false negative rates, while the SOAP features have the best true negative and false positive rates. SOAP's performance here can be attributed to predicting closer to the mean of the data, which, due to the heavy class imbalance, is a negative classification. This is reflected in the balanced accuracy, which for SOAP features is not significantly better than random.

training samples, and f_{TCM} the fraction of materials in the database, which meet the TCM criteria. This speedup factor is an informative metric for such a screening task because generating training data comes at a non-negligible cost and is actually less effective at discovering promising materials than a well-performing classifier. The turning point where generating more training data reduces the speedup can be seen in Fig. 9a beginning at ~1000 training samples. Above approximately 100 training samples, all SOREP features significantly outperform SOAPs in terms of speedup, with the concatenated features showing the highest mean performance followed by VBM-centered, CBM-centered, and Fermi-centered features. The relative performance of the various SOREP parameterizations is generally in line with expectations and the analysis of feature importances above. Notably,

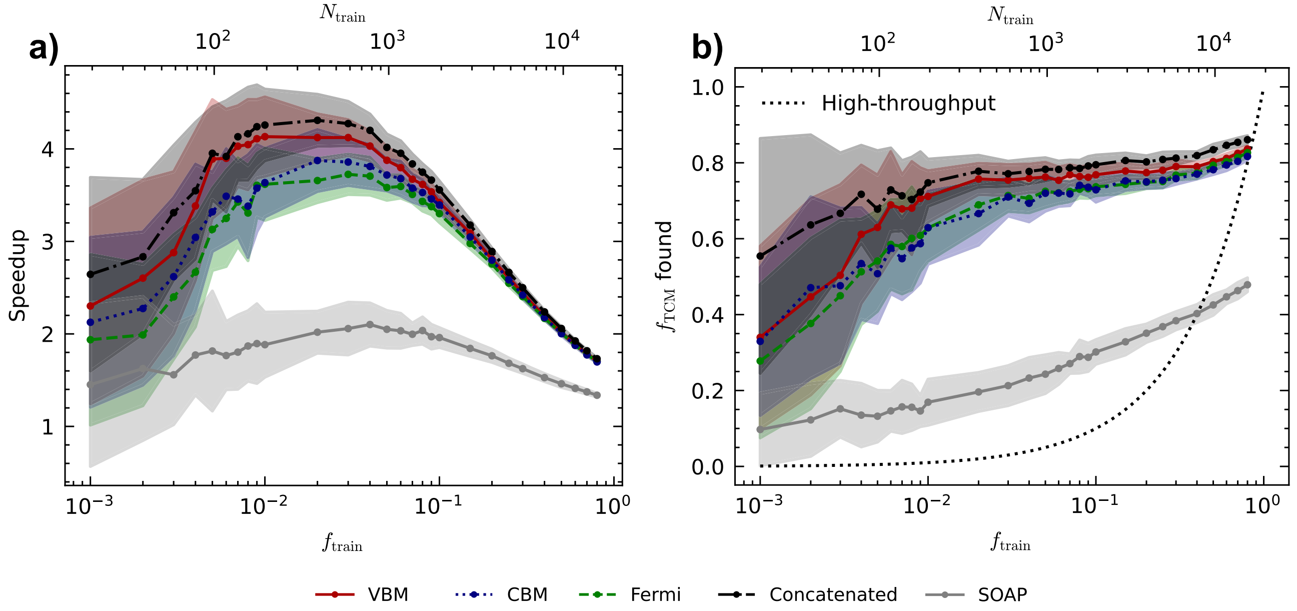


Fig. 9 | Performance metrics as a function of training fraction for the different SOREP features and SOAP features. a Speedups for the different descriptors versus an exhaustive high-throughput search. **b** Fraction of TCMs found by ML-accelerated screening for different amounts of training data. In both panels, the dotted lines plot the mean performance over 30 independent train-test splits, while the shaded regions show the mean plus and minus the standard deviation. All SOREPs significantly outperform the SOAP features in terms of speedup and fraction of TCMs found for this classification task. Of the SOREPs, the concatenated features have the highest mean performance, although the standard deviations are large.

the VBM-centered features do, to a minor extent, outperform the CBM-centered features, likely due to the valence bands being better reproduced in the single-shot DFT calculations than the conduction bands. An analysis of self-consistent field (SCF) DFT-based SOREP features is reported in SI Sec. S1.2, where the difference between the VBM-centered and CBM-centered features is indeed reduced.

The speedup can hide a non-negligible decrease in total yield, i.e., a high false negative rate. Therefore, in Fig. 9b, we show a fraction of TCMs found via the ML-accelerated screening versus what would be found in an exhaustive high-throughput search as a function of the number of training samples. All models strongly outperform a brute-force screening at low training fractions, in agreement with the yield results, but even the best model saturates at around an 80% discovery rate. Notably, SOAP features underperform a high-throughput search when trained on more than ~40% of the data, while all of the SOREPs continue to outperform the exhaustive search up to an 80:20 train-test split. As seen in the speedups, SOREPs consistently outperform SOAP features across the data regime by approximately a factor of 2, well outside the standard deviation of the metric.

The best-performing model, both in terms of yield and fraction of TCMs found, uses the concatenated features and is trained on 0.5–1.5% of the database, or between 100 and 300 materials. After training on 217 materials, the RFC predicts on average 1410 true positives and 2525 false positives, requiring in total 4152 SCF and band-structure calculations, or ~20% of the database. In turn, it finds ~75% of the TCM candidates present in the database. This is a significant improvement over SOAP features, which, even when trained on 80% of the database (~15,000 materials), only find 40% of the TCM candidates.

An identical study to the one described above has also been performed using SCF DFT SOREP features, and its results are reported in SI Sec. S1.2. The performance of SCF features is, on average, slightly better with a smaller variance but within two standard deviations of that of the single-shot features reported here. In order to confirm the improved performance w.r.t. SOAP observed in this classification study, an additional regression study into DFT band gap prediction is presented in SI Sec. S2, where it is shown that SOREP features do indeed outperform SOAP.

Discussion

In this work, we have presented a unified framework for constructing machine-learning features based on the electronic structure of molecules and materials, leveraging the symmetry preservation and conceptual clarity of physical approaches. By formalizing the process of featurization as a multistep algorithm, involving the selection of an electronic model, design and application of a spectral operator, and reduction of the spectrum into compact, invariant features, we were able to rapidly design and apply two sets of features to the problems of polymorph similarity and the discovery of TCMs.

We have described and investigated a kinetic-operator-based SOREP method and applied it successfully to distinguishing structurally similar, but electronically diverse, CNTs. Using the SOREPs, metallic and insulating polymorphs were clearly distinguishable, with particular features paralleling p_z-tight-binding densities of states. Applied to the uniqueness analysis of $BaTiO_3$ structures from the MC3D feedstock, these kinetic features also showed a remarkable ability to highlight configurations that are missed by the currently employed atomic-structure-based method. In combination with advanced clustering algorithms, dimensionality reduction schemes, and intrinsic dimension analyses such as those implemented in DADApy, the physical interpretability of SOREP features may also lead in future work to a better understanding of the important electronic collective variables in similar datasets.

A second SOREP featurization based on a single-shot evaluation of the Kohn–Sham Hamiltonian was then investigated for the more complex and compositionally diverse problem of TCM discovery within the MC3D. By leveraging the SOREP framework, minimal modifications to the kinetic SOREPs were identified and remedied, producing features remarkably similar to self-consistent DFT DOS features at a fraction of the computational effort. Used to train a random forest classifier, these features allowed for the “discovery” of 75% of materials in the MC3D that meet common TCM screening criteria while relying on only 1% of the database for reference data. Comparing these results against those of classifiers trained on SCF-based SOREP features, we find no significant ML performance improvement, particularly considering the associated increase in computational cost. The success of this approach is not only due to their inherently physical information content but also to their interpretability, allowing researchers to select the most meaningful features by leveraging their scientific knowledge and experience. The strong data efficiency of SOREP features, i.e., that they can be used to train accurate models with little training data, opens the door to their application in learning difficult-to-compute properties or predictions from levels of theory beyond DFT, which are much more computationally expensive.

Methods

Kinetic SOREP

Ge, Si, and $K_2Sn_2O_3$. The germainum, silicon, and $K_2Sn_2O_3$ unit cell structures are taken from the Materials Project[59] database entries mp-32, mp-149, and mp-7502, respectively. Using the ANO-ML-OS cGTO basis set, the kinetic energy and overlap matrices are computed with a $\mathbf{k}$-point distance of 0.1 $Å^{-1}$ and Gaussian smearing with a width of $\sigma = 0.05$ Hartree. PySCF's `safe_eigh` function is used to diagonalize the kinetic energy matrix, and the kinetic DOS is sampled on an evenly-spaced grid of points from -5σ to the maximum eigenvalue with a spacing of 0.01 Hartree.

Carbon nanotubes. The CNTs for kinetic SOREP featurization are generated using ASE's[15,16] `ase.build.nanotube` method with a bond length of 1.42 Å and vacuum of 30 Å. Using the ANO-ML-OS cGTO basis set, the kinetic energy and overlap matrices are computed with a $\mathbf{k}$-point distance of 0.05 $Å^{-1}$ and Gaussian smearing with a width of $\sigma = 0.015$ Hartree. SciPy's[72] `eigh` function is used to diagonalize the kinetic energy matrix, and the kinetic DOS is sampled on an evenly-spaced grid of points from 0 to 20 Hartree with a spacing of 0.001 Hartree. The nanotube structures for the p_z tight-binding DOS are generated by the CNTbands[57] tool hosted on nanoHUB with default parameters. The p_z densities of states are computed from the band dispersion provided by CNTbands using the smearing parameters reported above and are sampled from the minimum to the maximum tight-binding energy ± 5σ with a spacing of 0.001 Hartree.

$BaTiO_3$ uniqueness analysis

The $BaTiO_3$ dataset consists of 127 structures relaxed with the MC3D[73] protocol and PBE functional with unit cells containing either 5 or 30 atoms (1 or 6 formula units). As a baseline method, the structures are grouped using Pymatgen's[59] `StructureMatcher` with default parameters using the "first come, first serve" algorithm implemented in the mc3d-source package (https://github.com/mbercx/mc3d-source). Each structure is featurized using the kinetic energy SOREP via PySCF[74,75] v1.7.6 using the ANO-ML-OS cGTO basis set, a $\mathbf{k}$-point distance of 0.2 $Å^{-1}$, Gaussian smearing with a width of $\sigma = 0.03$ Hartree, and scipy's `eigh` function for diagonalization. The kinetic DOS is sampled on an evenly spaced grid of 512 points from -5σ to 6.5 Hartree.

TCM screening

DFT calculations. SCF calculations were performed using Quantum ESPRESSO[76–78] (QE) (v6.4.1 and v6.5) via AiiDA[17,18] and the AiiDA QE plugin (v3.0.0a3, v3.0.0a5, v3.0.0, v3.2.1, v3.1.0, and v3.4.2) with SSSP[66] PBE efficiency (v.1.1) pseudopotentials and associated plane-wave cutoffs, PBE[64] exchange-correlation functional, cold smearing[79], and a 0.15 $Å^{-1}$ $\mathbf{k}$-point spacing. High-symmetry line band structure calculations were performed via AiiDA and the AiiDA QE plugin (v3.4.2) using QE (v6.8) with SSSP PBEsol efficiency (v.1.2) pseudopotentials and associated plane-wave cutoffs, PBEsol[80] exchange-correlation functional, cold smearing, and a 0.025 $Å^{-1}$ $\mathbf{k}$-point spacing. Single-shot calculations are performed via AiiDA and a modified version of the AiiDA QE plugin (v3.4.2) using a modified version of QE (v6.7, v6.8) with identical input parameters to the respective SCF calculations excepting the number of allowed SCF iterations (set to 0).

Post-processing is performed on all calculations in order to determine the Fermi level consistently across different QE versions. Due to the modifications made to QE in order to enable the single-shot calculations, the Fermi level and occupations are not computed by the code. Additionally, QE versions up to and including 6.8 use a bisection algorithm with known limitations when cold smearing is employed[81]. To address these issues, the Fermi level determination algorithm described in ref. 81 is implemented in Python and applied while retrieving relevant calculations from AiiDA.

SOAP. SOAP features are computed without distinguishing atomic species (due to rapidly increasing feature dimension) using DScribe[69,70] (v1.1.0) with a radial cutoff of 6 Å, $n_{\max} = 10$, $l_{\max}$=9, σ = 0.3 Å, "gto" type radial basis, "inner" type averaging, and "poly" type weighting with r_0 = 5.0 Å, c = 1.0, m = 1.0.

Random forest classification. Random forest classification implementations in the scikit-learn[82] (v1.5.1) and imbalanced-learn[83] (v0.12.3) packages are used. Prior to any hyperparameter optimization or model training, the 21,691 materials in the MC3D subset are split into training and hold-out datasets, with 10% of the data held out. The training data are then used to optimize the RFC hyperparameters via fivefold grid search cross-validation with a 60:40 train-test split. The final models are trained with 500 estimators with `ccp_alpha=0` and "balanced-subsample" class weighting with replacement, bootstrapping, and a 50:50 class balance. Using these parameters, models are trained using each set of features on 30 independent train-test splits for training fractions ranging from 0.001 to 0.8. Each trained model is evaluated on the hold-out dataset, and the mean and standard deviation of the classification metrics are reported.

Data availability

The datasets are available on the Materials Cloud Archive at https://archive.materialscloud.org/record/2024.128.

Code availability

The code used to process and feature the data and train the models is available in static form with the data on the Materials Cloud in the entry listed above and also in a GitHub repository at https://github.com/azadoks/sorep.

Received: 29 April 2024; Accepted: 24 October 2024;
Published online: 02 December 2024

References

1. Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
2. Hellenbrandt, M. The inorganic crystal structure database (ICSD)—present and future. *Crystallogr. Rev.* **10**, 17–22 (2004).
3. Wishart, D. S. et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**, D901–D906 (2008).
4. Gražulis, S. et al. Crystallography Open Database – an open-access collection of crystal structures. *J. Appl. Crystallogr.* **42**, 726–729 (2009).
5. Curtarolo, S. et al. Aflowlib.org: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
6. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
7. Kim, S. et al. PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2016).
8. Puchala, B. et al. The materials commons: a collaboration platform and information repository for the global materials community. *Jom* **68**, 2035–2044 (2016).

9. Borysov, S. S., Geilhufe, R. M. & Balatsky, A. V. Organic materials database: an open-access online database for data mining. *PLoS One* **12**, e0171501 (2017).
10. Villars, P., Cenzual, K., Gladyshevskii, R., Franko, I. & Iwata, S. Pauling file - towards a holistic view. *Chem. Met. Alloy.* **11**, 43–76 (2018).
11. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2018).
12. Draxl, C. & Scheffler, M. The nomad laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).
13. Choudhary, K. et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).
14. Talirz, L. et al. Materials Cloud, a platform for open computational science. *Sci. Data* **7**, 299 (2020).
15. Bahn, S. R. & Jacobsen, K. W. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.* **4**, 56–66 (2002).
16. Larsen, A. H. et al. The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
17. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. Aiida: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).
18. Huber, S. P. et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data* **7**, 300 (2020).
19. Uhrin, M., Huber, S. P., Yu, J., Marzari, N. & Pizzi, G. Workflows in AiiDA: engineering a high-throughput, event-based engine for robust and modular computational workflows. *Comput. Mater. Sci.* **187**, 110086 (2021).
20. Jain, A. et al. FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurr. Comput. Pract. Exp.* **27**, 5037–5059 (2015).
21. Mathew, K. et al. Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. *Comput. Mater. Sci.* **139**, 140–152 (2017).
22. Huber, S. P. et al. Common workflows for computing material properties using different quantum engines. *npj Comput. Mater.* **7**, 136 (2021).
23. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
24. Marzari, N., Ferretti, A. & Wolverton, C. Electronic-structure methods for materials design. *Nat. Mater.* **20**, 736–749 (2021).
25. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
26. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
27. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
28. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
29. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
30. Musaelian, A. et al. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).
31. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process. Syst.* **35**, 11423–11436 (2022).
32. Pozdnyakov, S. N. et al. Incompleteness of atomic structure representations. *Phys. Rev. Lett.* **125**, 166001 (2020).
33. Fabrizio, A., Briling, K. R. & Corminboeuf, C. SPA^{H}M: the spectrum of approximated hamiltonian matrices representations. *Digit. Discov.* **1**, 286–294 (2022).
34. Isayev, O. et al. Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chem. Mater.* **27**, 735–743 (2015).
35. Hammerschmidt, T., Ladines, A., Koßmann, J. & Drautz, R. Crystal-structure analysis with moments of the density-of-states: Application to intermetallic topologically close-packed phases. *Crystals* **6**, 18 (2016).
36. Jenke, J. et al. Electronic structure based descriptor for characterizing local atomic environments. *Phys. Rev. B* **98**, 144102 (2018).
37. Fung, V., Hu, G., Ganesh, P. & Sumpter, B. G. Machine learned features from density of states for accurate adsorption energy prediction. *Nat. Commun.* **12**, 88 (2021).
38. Knøsgaard, N. R. & Thygesen, K. S. Representing individual electronic states for machine learning GW band structures of 2D materials. *Nat. Commun.* **13**, 468 (2022).
39. Geilhufe, R. M., Borysov, S. S., Kalpakchi, D. & Balatsky, A. V. Towards novel organic high-T_c superconductors: data mining using density of states similarity search. *Phys. Rev. Mater.* **2**, 024802 (2018).
40. Sadeghi, A. et al. Metrics for measuring distances in configuration spaces. *J. Chem. Phys.* **139**, 184118 (2013).
41. Zhu, L. et al. A fingerprint based metric for measuring similarities of crystalline structures. *J. Chem. Phys.* **144**, 034203 (2016).
42. Chiarotti, T., Marzari, N. & Ferretti, A. Unified green's function approach for spectral and thermodynamic properties from algorithmic inversion of dynamical potentials. *Phys. Rev. Res.* **4**, 013242 (2022).
43. Kuban, M., Rigamonti, S., Scheidgen, M. & Draxl, C. Density-of-states similarity descriptor for unsupervised learning from materials data. *Sci. Data* **9**, 646 (2022).
44. Parsaeifard, B. & Goedecker, S. Manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions. *J. Chem. Phys.* **156**, 034302 (2022).
45. Pozdnyakov, S. N. et al. Comment on "Manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions". *J. Chem. Phys.* **157**, 177101 (2022).
46. Sun, Q. Libcint: an efficient general integral library for gaussian basis functions. *J. Comput. Chem.* **36**, 1664–1671 (2015).
47. Pritchard, B. P., Altarawy, D., Didier, B., Gibsom, T. D. & Windus, T. L. A new basis set exchange: an open, up-to-date resource for the molecular sciences community. *J. Chem. Inf. Model.* **59**, 4814–4820 (2019).
48. Feller, D. The role of databases in support of computational chemistry calculations. *J. Comput. Chem.* **17**, 1571–1586 (1996).
49. Schuchardt, K. L. et al. Basis set exchange: a community database for computational sciences. *J. Chem. Inf. Model.* **47**, 1045–1052 (2007).
50. Veryazov, V., Widmark, P.-O. & Roos, B. O. Relativistic atomic natural orbital type basis sets for the alkaline and alkaline-earth atoms applied to the ground-state potentials for the corresponding dimers. *Theor. Chem. Acc.* **111**, 345–351 (2004).
51. Roos, B. O., Lindh, R., Malmqvist, P.-Å., Veryazov, V. & Widmark, P.O. Main group atoms and dimers studied with a new relativistic ANO basis set. *J. Phys. Chem. A* **108**, 2851–2858 (2004).
52. Roos, B. O., Lindh, R., Malmqvist, P.-Å., Veryazov, V. & Widmark, P.O. New relativistic ANO basis sets for transition metal atoms. *J. Phys. Chem. A* **109**, 6575–6579 (2005).
53. Roos, B. O., Lindh, R., Malmqvist, P.-Å., Veryazov, V. & Widmark, P.O. New relativistic ANO basis sets for actinide atoms. *Chem. Phys. Lett.* **409**, 295–299 (2005).
54. Roos, B. O. et al. New relativistic atomic natural orbital basis sets for lanthanide atoms with applications to the Ce diatom and LuF$_3$. *J. Phys. Chem. A* **112**, 11431–11435 (2008).

55. Widmark, P.-O., Malmqvist, P.-Å. & Roos, B.O. Density matrix averaged atomic natural orbital (ANO) basis sets for correlated molecular wave functions. *Theor. Chim. Acta* **77**, 291–306 (1990)..
56. Yang, L., Anantram, M., Han, J. & Lu, J. Band-gap change of carbon nanotubes: effect of small uniaxial and torsional strain. *Phys. Rev. B* **60**, 13874 (1999).
57. Seol, G. et al. Cntbands https://nanohub.org/resources/cntbands-ext (2006).
58. Huber, S. et al. Materials cloud three-dimensional crystals database (mc3d) https://doi.org/10.24435/materialscloud:rw-t0 (2022).
59. Ong, S. P. et al. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
60. Ester, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, **96**, 226–231 (1996).
61. Hautier, G., Miglio, A., Ceder, G., Rignanese, G.-M. & Gonze, X. Identification and design principles of low hole effective mass p-type transparent conducting oxides. *Nat. Commun.* **4**, 2292 (2013).
62. Hautier, G., Miglio, A., Waroquiers, D., Rignanese, G.-M. & Gonze, X. How does chemistry influence electron effective mass in oxides? a high-throughput computational analysis. *Chem. Mater.* **26**, 5447–5458 (2014).
63. Woods-Robinson, R. et al. Assessing high-throughput descriptors for prediction of transparent conductors. *Chem. Mater.* **30**, 8375–8389 (2018).
64. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
65. Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. *Science* **351**, aad3000 (2016).
66. Prandini, G., Marrazzo, A., Castelli, I. E., Mounet, N. & Marzari, N. Precision and efficiency in solid-state pseudopotential calculations. *npj Comput. Mater.* **4**, 72 (2018).
67. Giannozzi, P. et al. Quantum ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
68. Hinuma, Y., Pizzi, G., Kumagai, Y., Oba, F. & Tanaka, I. Band structure diagram paths based on crystallography. *Comput. Mater. Sci.* **128**, 140–184 (2017).
69. Himanen, L. et al. DScribe: library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
70. Laakso, J. et al. Updates to the dscribe library: new descriptors and derivatives. *J. Chem. Phys.* **158**, https://arxiv.org/abs/2303.14046 (2023).
71. Glielmo, A. et al. DADApy: distance-based analysis of data-manifolds in Python. *Patterns* 100589 https://www.sciencedirect.com/science/article/pii/S2666389922002070 (2022).
72. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
73. Huber, S. et al. Materials cloud three-dimensional crystals database (mc3d) https://archive.materialscloud.org/record/2022.38 (2022).
74. Sun, Q. et al. PySCF: the python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.* **8**, e1340 (2018).
75. Sun, Q. et al. Recent developments in the PySCF program package. *J. Chem. Phys.* **153**, 024109 (2020).
76. Giannozzi, P. et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (19pp) (2009).
77. Giannozzi, P. et al. Advanced capabilities for materials modelling with quantum espresso. *J. Phys. Condens. Matter* **29**, 465901 (2017).
78. Giannozzi, P. et al. Quantum espresso toward the exascale. *J. Chem. Phys.* **152**, 154105 (2020).
79. Marzari, N., Vanderbilt, D., De Vita, A. & Payne, M. C. Thermal contraction and disordering of the Al(110) surface. *Phys. Rev. Lett.* **82**, 3296–3299 (1999).
80. Perdew, J. P. et al. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **100**, 136406 (2008).
81. Dos Santos, F. J. & Marzari, N. Fermi energy determination for advanced smearing techniques. *Phys. Rev. B* **107**, 195122 (2023).
82. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
83. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).

Acknowledgements

A.M. acknowledges funding from the European Union - NextGenerationEU, through the ICSC-Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing-(Grant No. CN00000013, CUP J93C22000540006, PNRR Investimento M4.C2.1.4). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them. This research was supported by the NCCR MARVEL, a National Centre for Competence in Research, funded by the Swiss National Science Foundation (grant number 205602).

Author contributions

A.Z., A.M., and N.M. conceived of the work. A.Z. performed the calculations and model training and evaluation. A.Z. implemented the single-shot DFT calculations in Quantum ESPRESSO and the SOREP featurizations in the `sorep` package, the latter of which is based on earlier work by A.M. A.M. prepared the custom ANO basis sets. N.M. and A.M. supervised the work. All authors analyzed the results and contributed to the writing of 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/s41524-024-01446-9.

Correspondence and requests for materials should be addressed to Austin Zadoks.

Reprints and permissions 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.