Featureless adaptive optimization accelerates functional electronic materials design

Yiqun Wang, Akshay Iyer, Wei Chen and James M. Rondinelli

Department of Materials Science and Engineering, Northwestern University, Illinois 60208, USA
Department of Mechanical Engineering, Northwestern University, Illinois 60208, USA

Forward and inverse design\textsuperscript{1,2} of equilibrium materials properties have been accelerated with data-driven methods.\textsuperscript{3} Electronic materials exhibiting phase transitions\textsuperscript{4,5} between metastable states, such as metal-insulator transition (MIT) materials\textsuperscript{6} with abrupt electrical resistivity transformations, however, are challenging to decode. Conventional machine learning methods display limited predictive capability for MIT materials, because data scarcity and the absence of features\textsuperscript{7,8} impede model training. Here we demonstrate a discovery strategy based on multi-objective Bayesian optimization to directly circumvent these bottlenecks by utilizing latent variable Gaussian processes combined with high-fidelity electronic structure calculations for validation in the chalcogenide lacunar spinel family. We directly and simultaneously learn phase stability and band gap tunability from chemical composition alone to efficiently discover 12 superior compositions on the Pareto front. Previously unidentified electronic transitions also emerge from our featureless adaptive optimization engine (AOE). Our methodology readily generalizes to optimization of multiple properties, enabling co-design of complex multifunctional materials especially where prior data is sparse.

Upon traversing a critical temperature, the electrical resistivity of a metal-insulator transition (MIT) material can change by orders of magnitude. Athermal approaches may also trigger the electronic transitions, including (chemical) pressure, variable carrier-densities, and applied electromagnetic fields. The transformations can be used to encode, store, and process information for beyond von-Neumann microelectronics and overcome performance limits of conventional field-effect transistors\textsuperscript{9} for advanced logic/memory technologies.\textsuperscript{10} Because macroscopic MITs occur in materials with diverse chemistries and structures (Fig. 1a), various microscopic mechanisms – electron-lattice interactions, electron-electron interactions, or a combination thereof – lead to large variations in critical temperatures and accessible resistivity changes. This diversity exacerbates the efficient discovery and optimization challenge of achieving multiple property requirements to outperform silicon-based devices,\textsuperscript{11} including stability, large reversible resistivity changes ($\approx 10^5$), and above room-temperature operation.

The complex lacunar spinel family $AM^aM^bQ_8$ with trivalent main group $A$, transition metal $M$, and chalcogenide $Q$ ions demonstrate the complexity active in MIT materials design. The structure comprises transition-metal clusters (TMC) of $M^a$ and $M^b$ cations at the apical and basal positions of the tetrahedra (Fig. 1b inset). Although there are hundreds of possible elemental combinations on the four lattice sites in the crystal structure (Fig. 1a), only tens of the lacunar spinels have been experimentally reported.\textsuperscript{12,13} For example, Ga$V_2S_8$ ($M^a = M^b = V$) exhibits a MIT,\textsuperscript{14} exotic spin textures,\textsuperscript{15} and multiferroism\textsuperscript{16} while Ga$V$Ti$_2S_8$ shows negative magnetoresistance and half-metallic ferromagnetism.\textsuperscript{17} Most lacunar spinels are narrow-bandwidth semiconductors in their ground states;\textsuperscript{12,18} these electronic properties are governed by distortions of the local TMC from the ideal $T_d$ geometry,\textsuperscript{19} which manifest as low-frequency phonons as shown for Ga$Mo_3S_8$ (Fig. 1b, blue curve). Jahn-Teller-type distortions, which correspond to elongation along the [111] direction alter the TMC geometry, are particularly important; they transform the insulating Ga$Mo_3S_8$ ground state into a metastable metallic phase (Fig. 1c). The MIT arises from a redistribution of electrons among the structure-driven orbital hierarchy (Fig. 1c insets). Furthermore, these phases host low energy electronic structures, discernible from the projected density of states (pDOS) in Fig. 1c, that arise from the different $M^a$ and $M^b$ sites. This capability to exhibit distinct and tunable electronic phases poses a challenge in the design of lacunar spinels from physics-based models while also making them an ideal system for MIT performance optimization.

We seek lacunar spinels that exhibit high stability and large resistivity-switching ratios, which we formulate as design objectives for MIT-materials discovery. We reduce the approximately $O(10^3)$ compositional space to 270 candidates that maintain a $1 M^a$ to $3 M^b$ ratio. ($AM^aM^bQ_8$ compositions are excluded as they remove the $C_{3v}$ symmetry fundamental to the MIT; Cr is also excluded from occupying the $M^b$ site, because it destabilizes the cluster.) We define the first design objective as the decomposition enthalpy change ($\Delta H_d$, Fig. 1d), and use density functional theory (DFT) simulations to evaluate formation energies. Materials with larger $\Delta H_d$ are expected to be more synthesizable\textsuperscript{20} and stable during operation, making it a useful filter to prioritize compounds for subsequent theoretical analysis and synthetic processing. The second design objective is the ground state band gap ($E_g$). We use it as a proxy for the resistivity-switching ratio since $E_g$ is positively correlated with the resistivity change between different electronic states (Fig. 1d). A larger $E_g$ also allows for greater band-gap tunability through control over the $C_{3v}$ distortion, which is a desirable feature for programmable electronics. Importantly, because $E_g$ is small for most MIT materials, stability is expected to be lower and more difficult to achieve than that of nonpolymorphous compounds with majority ionic or covalent bonding.\textsuperscript{21}

The nonlinear responses of both design objectives bring severe challenges to compound optimization beyond those amplified by chemical combinatorics using data-driven models. Machine learning (ML) methods for materials design show promising results when sufficient training data is available for learning.\textsuperscript{8,23} The predictive performance (error and efficiency) is limited by the quality and quantity of the data, typically $> O(10^2)$ (Fig. 2), and features whose selection may be informed by domain knowledge or trial-and-error approaches. Thus, conventional ML models cannot accommodate the data.
Figure 1 | Metal-insulator transition materials and design objectives for the lacunar spinel family. a, The range in resistivity accessible (length of bar) across the MIT and transition temperature for a variety of MIT materials. (left inset) The crystal structure of GaTa$_4$Se$_8$. (right inset) Candidate elements on each site of the lacunar spinel structure. b, DFT-simulated phonon dispersion curves of GaMo$_4$S$_8$ in the rhombohedral ground state, the blue curve corresponds to the Jahn-Teller active cluster distortion mode. (inset) The transition-metal cluster with a single apical M$^a$ atom and three basal M$^b$ atoms. The arrows indicate displacements characterizing the Jahn-Teller active phonon mode. The intratetrahedral cluster angle $\theta_m$ formed by M$^b_1$-M$^a$-M$^b_2$. c, Electronic band structures and projected density of states of GaMo$_4$S$_8$ in its (right) semiconducting ground state and (left) metallic metastable phase shown with increasing $\theta_m$. The two R$3m$ phases are connected by the Jahn-Teller-type structural distortion with a F$43m$ intermediate state. (insets) Molecular orbital diagrams of the Mo$_4$ cluster with different local geometries. d, Design objective 1 with the definition of decomposition enthalpy change and the decomposition pathways of two lacunar spinels. The DFT-simulated temperature-dependent log ratio of the resistivity in the insulating and metallic phases of lacunar spinels, including experimentally known compounds and newly discovered compositions, serves as design objective 2. DFT band gaps specified in parentheses.

dearth posed by the limited number of lacunar spinels undergoing thermal MITs. Furthermore, the lack of microscopic understanding in how different compositions influence the MIT lead to ambiguity in whether any features can be suitably formulated for discovery of MIT materials from structure and composition alone rather than through effective Hamiltonians.$^{24}$ We overcome these obstacles by implementing a cyclic adaptive optimization engine (AOE) shown in Fig. 2, which consists of four iterative tasks (vide infra): property evaluation, aggregation of data (in a repository), featureless learning, and composition optimization. Beyond returning a predictive model capable of predicting properties from compositions alone, our iterative AOE leverages earlier approaches$^{25-27}$ to deliver materials with superior performance by design of composition-based solutions on the Pareto front spanned by the multiple objectives. The AOE has the important advantage of bypassing the feature engineering procedure as in conventional ML methods; it learns properties directly from the chemical composition at each site (i.e., A, M$^a$, M$^b$, Q). We achieve featureless learning by transforming discrete categorical variables (i.e., elemental compositions) into a continuous numerical space and then perform composition optimization under the multiple objectives through latent variable Gaussian processes (LVGP, see Methods).

We start the MIT-materials AOE for the lacunar spinel family through an initial design of experiment (DoE) consisting of four experimentally known compounds within the family and eight new compositions. The eight new compositions are obtained using Latin Hypercube Design to ensure each element is present in the dataset (see Methods and Supplementary Fig. 1). Next, we use high-fidelity DFT simulations to evaluate $\Delta H_d$ and $E_g$ (see Methods). This is the most resource-intensive step among the four tasks; therefore, it is desirable to iterate through the AOE (property evaluation) loop as few times as possible. Then, we create a data repository that contains entries for both composition and the evaluated properties. Unlike other ML methods, we do not rely on a large number of existing data at either the onset or later in the learning process.

Next, we construct a LVGP model by mapping the elemental compositions (e.g., Al, Ga, In) into a two-dimensional (2D) latent space (Fig. 2, lower right inset) such that relative positions in this space are obtained using maximum likelihood estimation (MLE). This latent space representation enables us to construct a Gaussian process surrogate model of the unknown underlying
Conventional materials informatics with features

Data acquisition
- Materials Project
- OQMD
- AFLOW ...
- (size > 10^9)

Feature engineering
- valence electrons electron negativity atomic radii ...

Model construction
- Preprocessing
- Model training
- Validation

Output
- Predictive model on new data

Adaptive materials discovery with featureless Bayesian optimization

Initial DoE
- Problem definition
- Starting candidate materials (~10^3)

Property evaluation
- Conventional machine learning models typically require a large amount of raw data > O(10^2) from existing databases. Feature engineering is an essential step to extract effective materials descriptors based on chemical intuition. Then, machine learning models are constructed after some preprocessing procedure (e.g. train-test-split, data standardization). The validation step checks the model performance and helps with hyper-parameter tuning. Finally, the predictive model is ready for production and can be applied to unobserved input data.

Lower panel, The adaptive materials discovery scheme starts from an initial set of design of experiments (DoE), where system variables, design objectives, and design space are first defined for the problem, providing a few O(10^1) candidate materials to initialize the discovery procedure. Property evaluation, The target material properties (design objectives) are evaluated either by experimental measurement or theoretical simulations. Candidate composition and its evaluated properties are then added to a Data repository, which initially may either be empty or only contains entries for existing materials within the design space. Its size grows as more candidate materials are evaluated during the adaptive optimization process. Featureless learning involves directly learning from the chemical composition of materials comprising the data repository by mapping each compositional variable into a two-dimensional latent space (spanned by z_1 and z_2) using maximum likelihood estimation, which enables the construction of a latent variable Gaussian process (LVGP) surrogate model. One surrogate model is constructed for each design objective using all currently available knowledge within the data repository. Composition optimization, Multi-objective Bayesian optimization is then performed with the LVGP models to obtain the next candidate material composition with the highest expected maximin improvement (EMI) value. The model accounts for uncertainty with the 95% confidence interval shown as the shadowed area around the new compositions (the green symbols). In the lower left inset, the green star composition outperforms the green circle composition, and will be passed to the next property evaluation procedure. The iterative optimization step continues until all compounds satisfying the objectives are discovered, forming the Pareto front, or computational resources expire.

Latent-variable Gaussian processes

Termination Criteria
- Satisfying compounds identified
- Research budget exhausted

Output
- Functional materials with superior properties by design

Figure 2 | Comparison of conventional (feature-required) machine learning with the featureless adaptive optimization engine. Upper panel, Conventional machine learning models typically require a large amount of raw data > O(10^2) from existing databases. Feature engineering is an essential step to extract effective materials descriptors based on chemical intuition. Then, machine learning models are constructed after some preprocessing procedure (e.g. train-test-split, data standardization). The validation step checks the model performance and helps with hyper-parameter tuning. Finally, the predictive model is ready for production and can be applied to unobserved input data. Lower panel, The adaptive materials discovery scheme starts from an initial set of design of experiments (DoE), where system variables, design objectives, and design space are first defined for the problem, providing a few O(10^1) candidate materials to initialize the discovery procedure. Property evaluation, The target material properties (design objectives) are evaluated either by experimental measurement or theoretical simulations. Candidate composition and its evaluated properties are then added to a Data repository, which initially may either be empty or only contains entries for existing materials within the design space. Its size grows as more candidate materials are evaluated during the adaptive optimization process. Featureless learning involves directly learning from the chemical composition of materials comprising the data repository by mapping each compositional variable into a two-dimensional latent space (spanned by z_1 and z_2) using maximum likelihood estimation, which enables the construction of a latent variable Gaussian process (LVGP) surrogate model. One surrogate model is constructed for each design objective using all currently available knowledge within the data repository. Composition optimization, Multi-objective Bayesian optimization is then performed with the LVGP models to obtain the next candidate material composition with the highest expected maximin improvement (EMI) value. The model accounts for uncertainty with the 95% confidence interval shown as the shadowed area around the new compositions (the green symbols). In the lower left inset, the green star composition outperforms the green circle composition, and will be passed to the next property evaluation procedure. The iterative optimization step continues until all compounds satisfying the objectives are discovered, forming the Pareto front, or computational resources expire.

design objectives, $\Delta H_d$ and $E_g$, as a function of composition. The multi-objective Bayesian optimization (MOBO) step then begins and we use the LVGP models to predict $\Delta H_d$ and $E_g$ of the unexplored compositions in our design space; we choose the next candidate composition for evaluation using the expected maximin improvement (EMI, see Methods) as the acquisition function, which quantitatively describes the performance gain compared against the compositions at the current Pareto front. This acquisition function considers both exploration of compositions with high uncertainty (Fig. 2, shaded ellipses, lower left inset) as well as exploitation of candidates with high performance gain. The composition with highest EMI is then selected for DFT simulation (property evaluation), at which point another AOE cycle commences.

The aforementioned iterative optimization procedure progresses and explores the available design space. One new lacunar spinel composition is evaluated and added to repository after each AOE iteration. The LVGP models are also updated in each iteration as more knowledge becomes available. Owing to the high computational cost of the property evaluation process, we terminate the optimization process after searching through 1/3 of the entire design space. In order to validate the effectiveness of this method, we ultimately evaluated $\Delta H_d$ and $E_g$ with DFT calculations of all 270 compositions within the
Figure 3 | The results of adaptive optimization on the lacunar spinel family. 

**a.** Upper panel: Evolution of the highest expected maximin improvement (EMI, blue line) and percentage of true Pareto front compounds identified (green line) as a function of iteration number. Results of the first 60 iterations are shown here. The red stars represent sampling points where a true Pareto front design is successfully identified. Lower panel: The moving average of absolute error in the predicted $E_g$ and $\Delta H_d$ values for a compound selected by the acquisition function for property evaluation.

**b.** The distribution of initial design of experiment and the first 60 evaluated compounds. Compounds evaluated in earlier stages have darker colors. True Pareto front designs are marked with red stars.

**c.** Distribution of Bayesian optimization-sampled elemental compositions for the first 60 iterations.

**d, e.** Latent space representation of elemental composition at different crystal structure sites in the $\Delta H_d$ and $E_g$ surrogate model, respectively. Results obtained after 60 iterations.

Design space by expending approximately $3 \times 10^6$ CPU hours.

**Fig. 3a** displays the results of the AOE. We successfully identify all 12 materials at the true Pareto front within 53 iterations (red asterisks, upper panel)—compositions and objective-related properties are enumerated in Table 1. Combined with the 12 compounds from our initial DoE, we explored less than 25% of the entire design space before identifying all lacunar spinels on the Pareto front. Interestingly, Pareto-front compositions are mostly found with high EMI values, showing that our model makes beneficial recommendations on which composition to evaluate next. High prediction uncertainty likely explains why a Pareto-front composition is not identified for some iterations with a large EMI. The EMI values reduce to nearly zero after all Pareto front compositions are identified (blue, upper panel) since all candidates not sampled are dominated by the Pareto front compositions. We also show the absolute error in the LVGP-predicted $\Delta H_d$ (pink) and $E_g$ (orange) values of the evaluated composition at each iteration to further demonstrate the effectiveness of our model (Fig. 3a).

We find a general decreasing trend in error and therefore better model predictability as it becomes aware of more composition-property knowledge. The aforementioned performance is robust as revealed by our multi-trial results (Supplementary Fig. 2), where we find the AOE successfully identifies 90% of the true Pareto-front compositions by exploring 30% of design space with different initial DoE sets.

**Fig. 3b** shows the history of composition explored by the AOE for the first 60 iterations. The detailed optimization history is shown in Supplementary Fig. 3. The initial DoE sets are relatively scarcely distributed away from the Pareto front, yet the model explores regions far from that covered by the DoE sets and is able to identify 75% of Pareto front compositions within the first 40 iterations. First, we begin to understand this performance by examining the distribution of elements sampled by the MOBO (Fig. 3c). Our model does not exhibit much compositional bias upon sampling elements for the A site; however, it shows clear preferences for choosing certain elements on other sites. V and Mo are sampled more frequently on the basal $M_b$ site, while Nb and Ta are less favored on the apical $M_a$ site. Se is also preferred over S and Te for the $Q$ site.

Second, we examine the 2D latent space representations for both design objectives obtained after 60 iterations of AOE (Fig. 3d and e). The relative positioning of elements in the latent space reflects correlations in their influence on properties; elements in close proximity exhibit similar impact. Interestingly, different transition metals exhibit distinct correlation patterns across various sites and objective properties. This variation leads us to conclude that (i) the transition metals contribute to stability and band gap in different and unexpected ways, and (ii) the lack of any resemblance in element positioning in the site-dependent latent spaces, except for the $M^a$ site, to the periodic table indicates that chemical-intuition-based MIT design within the lacunar spinels is highly nontrivial. For example, chromium is located far from the other elements in
the M\(^a\) latent space, indicating that its influence on properties is distinct. Indeed, Cr-containing compounds have significantly lower \(E_g\) and higher \(\Delta H_d\) (Supplementary Fig. 4).

Third, we use DFT simulations to examine the properties of the identified Pareto-front compositions, focusing on \(\Delta H_d\), \(E_g\), and the Jahn-Teller active phonon \(\nu_{JT}\) involved in the MIT (Table 1). We find most Pareto-front compositions consist of two different cations on the M\(^a\) and M\(^b\) site, only three have M\(^a\) = M\(^b\), with 75% of the optimized materials being selenides. GaV\(_4\)Se\(_8\) is the only Pareto front compound previously synthesized. All compounds exhibit \(R\bar{3}m\) symmetry and are dynamically stable in their ground state (\(\nu_{JT} > 0\)). The phonon frequencies of the selenides, including \(\nu_{JT}\) are lower than those of the sulfides. All of the designed lacunar spinels also exhibit semiconducting gaps with semilocal exchange-correlation and static Coulomb interactions and exhibit nonzero electric polarizations. Compositions with larger band gaps tend to have lower stability as determined by \(\Delta H_d\); \(2/3\) are stable (\(\Delta H_d > 0\), indicating decomposition is endothermic), whereas four of the 12 compounds comprising Mo have small values of \(\Delta H_d < 0\), which could nonetheless be stable and synthesizable\(^{21,28}\). Chemical intuition-based interpretations of the observed \(\Delta H_d\)-\(E_g\) trade-offs among these compounds, however, are difficult to formulate and reinforces the effectiveness of the AOE.

Although the ground states of these materials are all semiconducting, we find two different electronic transitions upon traversing the ideal TMC geometry (\(\theta_m = 60^\circ\)): the expected (Type I) metal-to-insulator transition and an unexpected (Type II) semiconductor-to-insulator transition (SIT). Fig. 4a shows the changes to the electronic structure for the MIT lacunar spinels AlTaV\(_3\)Se\(_8\) and InWMo\(_3\)Se\(_8\) with the insulating state (lower panel) always lower in energy than the metastable metallic phase (upper panel) after the Jahn-Teller-type distortion (\(\theta_m \neq 60^\circ\). Table 1). The pDOS of these compounds show that the metallic state in the Type I transition arises from cluster distortion-triggered orbital ordering and occupancy changes, similar to the mechanism depicted in Fig. 1b. However, the metallic states are different owing to the chemistry of the metals comprising the TMCs. We also find that the basal M\(^b\) site plays a more decisive role near the Fermi level with
we successfully identify all 12 Pareto-front compositions by

The simulated DC resistivity of InWMo

The AOE is particularly useful when data availability and physical understanding of the target materials system is limited at either the atomic or microstructural scale.

**Online Content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at URL to be inserted by Publisher.

**References**

1. A. Zunger, *Nature Reviews Chemistry* **2** (2018), 10.1038/s41570-018-0121.
2. J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, and Y. Jung, *Matter* **1**, 1370 (2019).
3. A. Agrawal and A. Choudhary, *MPL. Materials* **4**, 053208 (2016).
4. Z. Yang, C. Ko, and S. Ramanathan, *Annual Review of Materials Research* **41**, 337 (2011).
5. W. Zhang, R. Mazzarello, M. Wuttig, and E. Ma, *Nature Reviews Materials* **4**, 150 (2019).
6. M. Imada, A. Fujimori, and Y. Tokura, *Reviews of Modern Physics* **70**, 1039 (1998).
7. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Nature* **559**, 547 (2018).
8. J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, *npj Computational Materials* **5**, 1013864 (2019).
9. N. Shukla, A. V. Thathachary, A. Agrawal, H. Paik, A. Aziz, D. G. Schlom, S. K. Gupta, R. Engel-Herbert, and S. Datte, *Nature Communications* **6** (2015), 10.1038/ncomms8812.
10. Y. Zhou and S. Ramanathan, *Proceedings of the IEEE* **103**, 1289 (2015).

**Table 1** DFT-evaluated ground state properties of the Pareto front compounds. NOI is the number of iterations taken to discover the compound during the adaptive optimization process. Values of $\Delta H_f > 0$ (units of eV u.$^{-1}$) indicate an endothermic reaction occurs and the stable compound disfavors decomposition. $E_g$ is the DFT band gap in eV. $\nu_{JT}$ is the frequency (THz) of the Jahn-Teller-type phonon involving the TMC. $P$ is the electric polarization in $\mu$C cm$^{-2}$. The value of $\theta_m$ in the insulating ground state and transition type, Type I (MIT) or Type II (SIT), are also specified.

| Compound           | NOI | $\Delta H_f$ | $E_g$ | $\nu_{JT}$ | $P$ | $\theta_m$ | Type |
|--------------------|-----|--------------|-------|------------|-----|------------|------|
| InWV$_3$S$_8$      | 4   | 0.09         | 0.58  | 5.83       | 0.56| 65.0       | II   |
| AlCrV$_3$Se$_8$    | 8   | 3.17         | 0.19  | 3.77       | 1.87| 56.4       | II   |
| InMo$_3$Se$_8$     | 14  | -0.69        | 0.62  | 4.55       | 1.08| 63.4       | I    |
| InWMo$_3$Se$_8$    | 19  | -0.99        | 0.63  | 4.43       | 0.24| 63.8       | I    |
| InCrV$_3$S$_8$     | 20  | 2.59         | 0.40  | 4.75       | 0.28| 56.6       | II   |
| AlCrV$_3$S$_8$     | 21  | 2.63         | 0.39  | 5.81       | 1.02| 57.0       | II   |
| InCrV$_3$Se$_8$    | 25  | 3.10         | 0.22  | 3.45       | 0.58| 56.0       | II   |
| InTaMo$_3$Se$_8$   | 28  | -0.88        | 0.62  | 4.25       | 1.38| 54.8       | II   |
| AlTaV$_3$Se$_8$    | 38  | 0.56         | 0.56  | 3.90       | 0.15| 57.3       | I    |
| AlV$_4$Se$_8$      | 47  | 1.06         | 0.46  | 4.08       | 2.80| 54.9       | I    |
| InNbMo$_3$Se$_8$   | 49  | -0.66        | 0.59  | 4.44       | 0.75| 55.2       | II   |
| GaV$_4$Se$_8$      | 53  | 1.18         | 0.44  | 4.09       | 2.37| 55.0       | I    |
We perform full lattice relaxations until the residual forces with Blöchl corrections. We initialize the relaxation with various relaxed crystal structures of the Pareto front compositions are obtained using the frozen-phonon method with pre- and post-processing performed with the Phonopy package. The decomposition pathways are automatically generated using Grand Canonical Linear Programming from the Open Quantum Materials Database. Resistivity simulations are performed using electronic structures computed with VASP as previously described, but with an increased \(24 \times 24 \times 24\) \(k\)-point mesh and the BoltzTrap2 package. We also assume that all M\(^{\text{III}}\) sites have the same orientation within the crystal. In order to validate this model, we simulated a \(2 \times 2 \times 2\) supercell of InNbMo\(_{3}\)Se\(_{8}\) with one Nb atom oriented in a different direction from the other seven. We find that the ground state \(E_g\) as well as \(\Delta H_d\) exhibit negligible changes from the homogeneous description. We also compared the change in properties with the anti-ferromagnetic spin configuration using a doubled simulation cell with the ferromagnetic ground state. As before, we find there are no significant changes in the aforementioned properties. These results are reasonable because the local structure of the TMC dictates the low-energy band structure near the Fermi level.

**Methods**

**Density functional calculations.** We perform DFT simulations as implemented in the Vienna Ab initio Simulation Package (VASP). The projector augmented-wave (PAW) potentials are used for all elements in our calculations with the following valence electron configurations: Al (3\(s^2\)3\(p^1\)), Ga (3\(d^4\)4\(s^2\)4\(p^1\)), In (4\(d^1\)5\(s^2\)5\(p^1\)), V (3\(d^2\)3\(p^6\)3\(d^4\)4\(s^2\)), Nb (4\(s^2\)4\(p^6\)4\(d^5\)5\(s^1\)), Ta (5\(p^6\)5\(d^4\)6\(s^1\)), Cr (3\(d^3\)3\(p^6\)3\(d^2\)4\(s^1\)), Mo (4\(s^2\)4\(p^6\)4\(d^5\)5\(s^1\)), W (5\(s^3\)5\(p^6\)5\(d^5\)6\(s^1\)), S (3\(d^3\)3\(p^1\)), Se (4\(s^2\)4\(p^3\)), and Te (5\(s^3\)5\(p^7\)). We use exchange-correlation potentials (\(V_{xc}\)) as implemented by Perdew-Burke-Ernzerhof (PBE). The effect of on-site Coulomb interactions (PBE+U) is considered with a \(U\) value of 2.0 eV for all 6 transition metals, as suggested by previous studies. Numerous spin configurations are evaluated to ensure the global ground state is achieved and that those states are consistent with available experimental magnetic data. Spin-orbit interactions (SOI) are not considered in our calculations. Although it has been shown that SOI leads to interesting molecular \(j\text{eff}\) states, this order does not strongly affect the size of the ground state electronic band gaps, even 5\(d\) transition metals lacunar spinels. A \(\Gamma\)-centered 6 \(\times\) 6 \(\times\) 6 \(k\)-point mesh with a 500 eV kinetic energy cutoff is used. We employ Gaussian smearing with a small 0.05 eV width. For density-of-state calculations, we use the tetrahedron method with Blöchl corrections. Electric polarizations along the [111] direction are simulated using the Berry phase method.

The crystal structures of the existing lacunar spinels are obtained from our previous DFT studies, structures of new compositions are obtained by replacing the elements on the corresponding crystallographic sites from existing structures. We perform full lattice relaxations until the residual forces on each individual atom are less than 1.0 meV\(\text{Å}^{-1}\). The DFT-relaxed crystal structures of the Pareto front compositions are available at Ref. 41. We initialize the relaxation with various magnetic moment configurations, the converged configuration with the lowest energy is reported as the DFT ground state. Zone center (\(k = 0\)) phonon frequencies and eigendisplacements are obtained using the frozen-phonon method with pre- and post-processing performed with the Phonopy package. The decomposition pathways are automatically generated using Grand Canonical Linear Programming from the Open Quantum Materials Database. Resistivity simulations are performed using electronic structures computed with VASP as previously described, but with an increased \(24 \times 24 \times 24\) \(k\)-point mesh and the BoltzTrap2 package. We also assume that all M\(^{\text{III}}\) sites have the same orientation within the crystal. In order to validate this model, we simulated a \(2 \times 2 \times 2\) supercell of InNbMo\(_{3}\)Se\(_{8}\) with one Nb atom oriented in a different direction from the other seven. We find that the ground state \(E_g\) as well as \(\Delta H_d\) exhibit negligible changes from the homogeneous description. We also compared the change in properties with the anti-ferromagnetic spin configuration using a doubled simulation cell with the ferromagnetic ground state. As before, we find there are no significant changes in the aforementioned properties. These results are reasonable because the local structure of the TMC dictates the low-energy band structure near the Fermi level.
evaluated compositions. The positioning of latent variables \( z^A, z^{M^a}, z^{M^b}, \) and \( z^{Q} \) in their corresponding latent space are estimated via MLE as described in Ref 47. The LVGP model for \( \Delta H_d \) also utilizes the 2D latent variable representation \( \kappa^A, \kappa^{M^a}, \kappa^{M^b}, \) and \( \kappa^Q \) as previously defined to evaluate the correlation \( \rho(\Delta H^d_{GdMoV:Se}, \Delta H^d_{AlNbW:Se}) \) in a similar manner.

**Multibjective Bayesian optimization.** Consider the lacunar spinel family \( A^{M^a}M^bQ_8 \) with \( A \in \{Al, Ga, In\}, \) \( M^a \in \{V, Nb, Ta, Cr, Mo, W\}, \) \( M^b \in \{V, Nb, Ta, Mo, W\} \) and \( Q \in \{S, Se, Te\} \). The design space (C) comprises 270 compounds, each compound is represented by four design variables \( A, M^a, M^b \) and \( Q \) with three, six, five, and three choices, respectively. Our objective is to maximize \( E_R \) and \( \Delta H_d \), which is represented in standard optimization formulation as:

$$\min_{c \in C} -E_R(c), -\Delta H_d(c). \quad (2)$$

Starting from the initial dataset, the AOE evaluates new candidate compounds by gauging their improvement in the design objectives. Here, we use the expected maximin improvement (EMI) metric\(^{49}\) to guide the adaptive sampling framework. The Maximin Improvement (\( I_M \)) for compound \( c \) is:

$$I_M(c) = \min_{c_i \in C_{PF}} \left\{ \max \left( E_R(c_i), \Delta H_d(c_i) - \Delta H_d(c) \right) \right\} \quad (3)$$

where \( C_{PF} \) is the current set of Pareto front compositions. To facilitate the comparison in Equation 3, we scale the value of each design objective \( P \) using the scheme \( \tilde{P}(\cdot) = P(\cdot)/P_{\min} \) where \( P_{\max} \) and \( P_{\min} \) are the maximum and minimum value of property observed so far. The EMI of compound \( c \) is defined as the expected value of \( I_M \):

$$EMI(c) = \mathbb{E}[I_M(c)]. \quad (4)$$

We evaluate the EMI through Monte Carlo sampling with 500 trials. An alternative approach to EMI calculations is based on its geometrical interpretation.\(^{50}\) At each AOE iteration, the EMI is calculated for all compositions that are not yet present in the data repository. The composition with largest EMI will be sampled next in property evaluation and then added to the data repository.

**Data Availability**

The simulation output files are available upon reasonable request. They are not publicly available due to the very large file sizes. Parameters of the input files are described in the computational methods. Atomic structures for the simulation cells are available upon reasonable request.

**Code Availability**

The Vienna \textit{Ab initio} Simulation Package (VASP) is a proprietary software available for purchase at https://www.vasp.at/. Data processing scripts written to process output files and create figures are available upon request. The R-package \texttt{LGPLAI} contains the code for fitting LVGP models to general mixed-variable datasets and is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.Rproject.org/package=LGPLAI. The MOBO codes used in the current study are available from the corresponding authors upon reasonable request.

**References**

31. G. Kresse and J. Furthmüller, Phys. Rev. B 54, 11169 (1996).
32. G. Kresse and D. Joubert, Phys. Rev. B 59, 1758 (1999).
33. P. E. Blöchl, Phys. Rev. B 50, 17953 (1994).
34. J. P. Perdew, Phys. Rev. Lett. 77, 3865 (1996).
35. Y. Wang, D. Puggioni, and J. M. Rondinelli, Physical Review B 100, 115149 (2019).
36. H.-S. Kim, J. Im, M. J. Han, and H. Jin, Nat. Commun. 5, 1 (2014).
37. A. K. Rastogi, A. Berton, J. Chaussy, R. Tournier, M. Potel, R. Chevrel, and M. Sergent, Journal of Low Temperature Physics 52, 539 (1983).
38. P. E. Blöchl, O. Jepsen, and O. K. Andersen, Phys. Rev. B 49, 16223 (1994).
39. R. Resta, Reviews of Modern Physics 66, 899 (1994).
40. Crystal structures available at our group GitHub Page.
41. Crystal structures available at our group GitHub Page.
42. A. Togo and I. Tanaka, Scr. Mater. 108, 1 (2015).
43. S. Kirklin, B. Meredig, and C. Wolverton, Advanced Energy Materials 3, 252 (2013).
44. S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aylol, S. Rühl, and C. Wolverton, npj Computational Materials 1, 1 (2015).
45. G. K. Madsen, J. Carrete, and M. J. Verstraete, Comput. Phys. Commun. 231, 140 (2018).
46. M. D. McKay, R. J. Beckman, and W. J. Conover, Technometrics 21, 239 (1979).
47. Y. Zhang, S. Tao, W. Chen, and D. W. Apley, Technometrics, 231 (2018).
48. Y. Zhang, D. W. Apley, and W. Chen, Scientific Reports 10, 1 (2020).
49. D. C. T. Bautista, A sequential design for approximating the pareto front using the expected pareto improvement function, Ph.D. thesis, The Ohio State University (2009).
50. A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, Scientific Reports 8, 3738 (2018).

**Acknowledgments**

This work was supported in part by the National Science Foundation (NSF) under award number DMR-1729303 and DMREF-1729473. The information, data, or work presented herein was also funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0001209. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. \textit{Ab initio} DFT simulations were performed on the DoD-HPC (Copper cluster) and Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF (ACI-1548562).

**Author Contribution**

Y.W. and A.I. contributed equally to this work. W.C. and J.M.R. conceived and administered the project. Y.W. performed the DFT simulations and automated the AOE together with A.I. A.I. developed the Design of Experiment procedure, implemented MOBO, and performed multiple trials of AOE.
Y.W. and A.I. wrote the first draft of the paper, which was revised based on input from all authors.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at URL.

Correspondence and requests for materials should be addressed to J.M.R.

Reprints and permissions information is available at www.nature.com/reprints.