communications materials

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43246-025-00928-7

Materials Expert-Artificial Intelligence for materials discovery

Check for updates

Yanjun Liu[1], Milena Jovanovic[2,3], Krishnanand Mallayya[1], Wesley J. Maddox[1,4], Andrew Gordon Wilson[1,4], Sebastian Klemenz[5], Leslie M. Schoop[2] & Eun-Ah Kim[1,6]

Advances in materials databases create an opportunity to uncover descriptors that predict emergent properties, yet most studies rely on high-throughput ab initio calculations that can diverge from experiment. Experimentalists instead depend on intuition honed by hands-on work. We present "Materials Expert-Artificial Intelligence" (ME-AI), a machine-learning framework that translates this intuition into quantitative descriptors extracted from curated, measurement-based data. Using a set of 879 square-net compounds described using 12 experimental features, we train a Dirichlet-based Gaussian-process model with a chemistry-aware kernel. ME-AI reproduces established expert rules for spotting topological semimetals (TSMs) and reveals hypervalency as a decisive chemical lever in these systems. Remarkably, a model trained only on square-net TSM data correctly classifies topological insulators in rocksalt structures, demonstrating transferability. Complementing electronic-structure theory, our framework scales with growing databases, embeds expert knowledge, offers interpretable criteria, and guides targeted synthesis, accelerating materials discovery and rapid experimental validation across diverse chemical families.

Global efforts of the last decade established vast databases of materials[1–11], allowing researchers to query data associated with individual entries. This coming age of materials data presents an unprecedented opportunity to discover descriptors of physical properties by inferring and articulating correlations and trends from the collection of entries. If successful, such efforts can lead to guided discovery of new materials. Historically, the most successful of such efforts was that by Mendeleev, who predicted properties of elemental materials from the empirical trends articulated in the periodic table. However, discoveries of atomistic features that capture the physical properties of non-elemental materials remain challenging. The combinatorial phase space of non-elemental materials and the large dimensionality of the feature space render the quest out of manual reach.

The increasing availability of machine-learning (ML) tools has enabled recent developments in finding stable materials by scaling up ab initio calculations[4,12] and synthesis[13,14]. However, much of the literature focused on leveraging ML for computational predictions in the tradition of ab initio calculations[15–21]. Left untapped are the intuition and insights that expert materials growers have honed through many years of experience. We propose to "bottle" the insights latent in the expert growers' human intellect to arrive at descriptors that articulate the insights. Our strategy will be to supply information and data curated by a materials expert (ME) following his/her intuition to an ML model, together with expert annotation of the desired material properties. The mission for the model we dubbed ME-AI will be to learn the descriptor that can predict the desired properties from readily available information (see Fig. 1a). By letting ME-AI investigate expert-curated experimental information, we hope to accelerate discoveries of empirical laws and allow for the research community to expand upon explicit articulation of the insight latent to the curated information.

Given the challenging nature of this problem, the likelihood of success will depend on the reliability of the data and the scope of the search. A reliable database starts with the ME curating a refined dataset with experimentally accessible primary features chosen based on intuition from literature, ab initio calculations, or chemical logic. Regarding the scope of the search, setting a scope that makes ample use of chemical understanding can increase the likelihood of success. Based on these criteria, we focus on learning the descriptor of topological nodal-line semimetals (TSMs) derived from basic atomistic and structural features, limiting the pool of candidate materials to square-net structures. Remarkably, we find that ME-AI trained on square-net structures to predict TSMs can also successfully predict topological insulators among rocksalt structures, exhibiting surprising generalization ability.

TSMs are of interest due to their novel physical properties and potential applications in energy conversion[22], electrocatalysts[23], and sensing[24]. A TSM is characterized by two key features: (1) the presence of nodal lines where

[1]Department of Physics, Cornell University, Ithaca, NY, USA. [2]Department of Chemistry, Princeton University, Princeton, NJ, USA. [3]Department of Chemistry, North Carolina State University, Raleigh, NC, USA. [4]New York University, New York, NY, USA. [5]Fraunhofer Research Institution for Materials Recycling and Resource Strategies IWKS, Hanau, Germany. [6]Department of Physics, Ewha Womans University, Seoul, South Korea. e-mail: ek436@cornell.edu

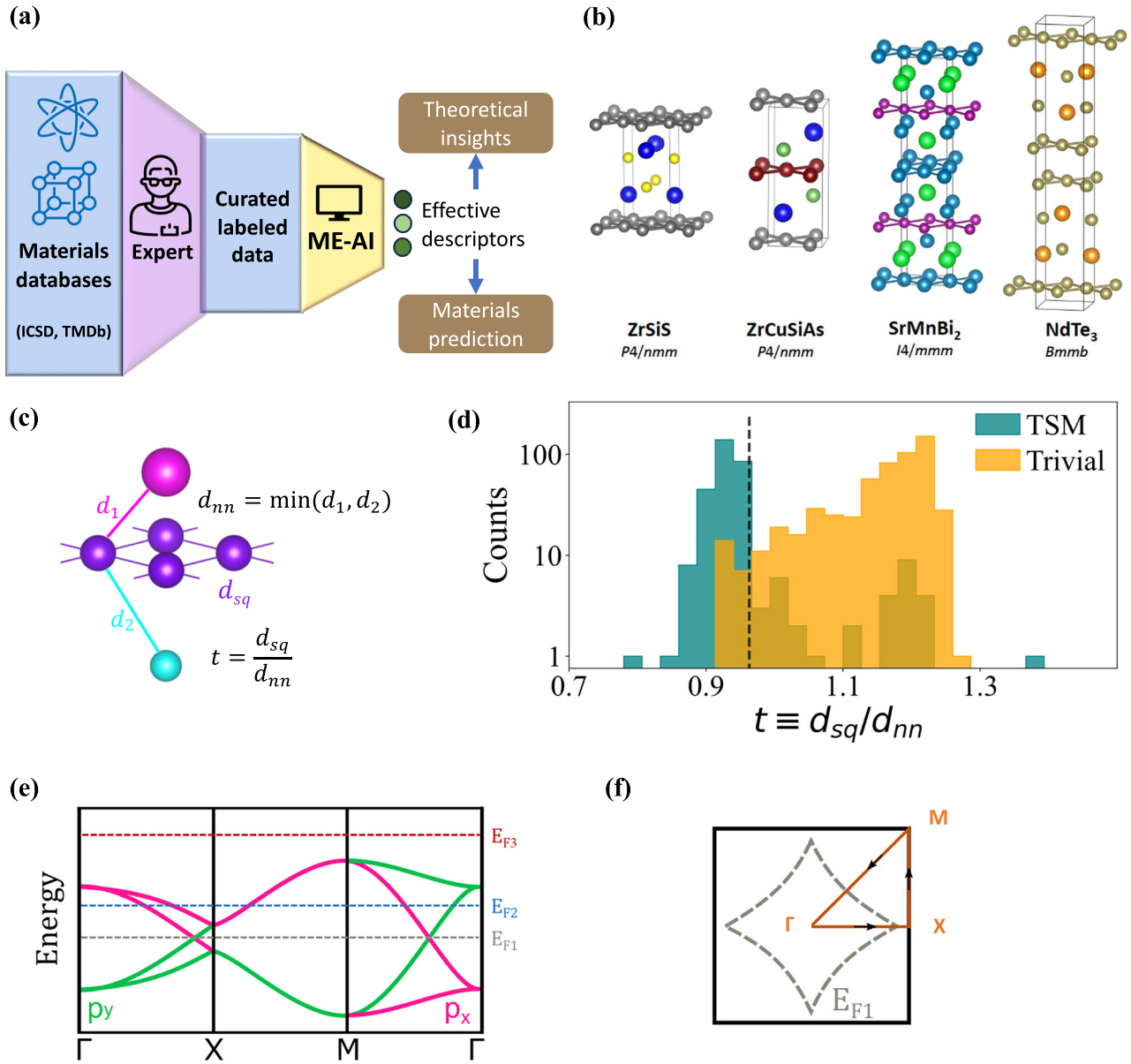


Fig. 1 | "Materials Expert-Artificial Intelligence" (ME-AI) for topological semimetals in square-net materials. a The conceptual framework of ME-AI. **b** Crystal structure of some 2D square-net materials. **c** The 3D square-net structure where the out-of-plane magenta and cyan atoms are separated by distances d_1 and d_2, respectively, from the square-net plane (purple). The tolerance factor $t \equiv d_{sq}/d_{nn}$, where $d_{nn} \equiv \min(d_1, d_2)$ is the out-of-plane nearest neighbor distance. **d** The distribution of tolerance factor (t) for the compounds in our dataset labeled as TSMs (teal) and trivial (orange) materials, where $t \equiv d_{sq}/d_{nn}$ and d_{nn} is the out-of-plane nearest neighbor distance. We find $t \approx 0.96$ (dashed line) as the value that separates TSMs and trivial materials with maximum accuracy. **e** The tight-binding band structure for hypothetical p_x (pink) and p_y (green) orbitals on the formation of the squarenet, plotted along the path shown in panel (**f**). Two symmetry-protected crossings with the band inversions are shown. The location of the Fermi level determines whether the material is a band insulator (E_{F3}), metal (E_{F2}) or TSM (E_{F1}). **f** BZ with the nodal line as the Fermi surface when the Fermi level is tuned through the nodal lines.

two spin-degenerate bands cross, and (2) the Fermi level tuned to the nodal lines. Conventional identification of TSMs requires a detailed analysis of materials' symmetry through the derivation of band representations and the number of electron filling[25]. However, some of us observed that chemical intuition can be instructive in predicting TSMs among 2D "square-net" materials- crystalline solids with a 2D-centered square-net motif (Fig. 1b, c). Following this intuition, some of us introduced a structural descriptor called the "tolerance factor (t-factor)" in refs. 26,27, defined as the ratio of square lattice distance to the out-of-plane nearest neighbor distance (Fig. 1c). The t-factor has proven effective in identifying target materials where specific physical properties relate to lower-dimensional bonding motifs in a 3D unit cell, such as materials with Kagome lattice signatures in their band structure[2,8]). The t-factor notably distinguishes TSMs, which have smaller t-values, from trivial materials (see Fig. 1d). However, at intermediate values ($t \sim 1$), it is clear that additional descriptors are necessary. The simplicity of the t-factor and strong domain knowledge incorporated into the choice of 2D square-net compounds make this problem an ideal starting point for developing and testing ME-AI.

For completeness, we first review the structural and chemical insight that led to the t-factor. Square-net materials have 2D square-net motifs stacked with other atoms in between. When the effects of out-of-plane atoms are minimally accounted for through the symmetry of a 2D tight-binding model with the out-of-plane atoms projected to the center of the squares and the square-net atoms have p_x and p_y orbitals partially filled, such a tight-binding band structure predicts symmetry-protected topological nodal lines (see Fig. 1e, f). If this simple picture holds in the full band

structure and the Fermi level is tuned to thread the nodal line, such a square-net material will be a TSM[29]. This inspired Klemenz et al.[27] to propose a ratio between two structural parameters to quantify the deviation from a 2D square-net plane structure, given by $t\text{-}factor \equiv d_{sq}/d_{nn}$, where d_{sq} is the square-net distance and d_{nn} is the out-of-plane nearest neighbor distance (see Fig. 1c). While the tight-binding model of pure lattice motifs can be informative, whether such insight can dominate the low-energy electronic property of the 3D crystal crucially depends on the filling, the existence of other interfering bands, and the location of the Fermi level[30].

In this work, we introduce ME-AI, a machine-learning framework that leverages experimentally curated expert intuition to uncover quantitative descriptors predictive of TSMs. Applying this approach to a dataset of 879 square-net compounds characterized by 12 primary features, ME-AI not only recovers the known structural descriptor, the "tolerance factor," but also identifies four new emergent descriptors. Remarkably, one purely atomistic descriptor aligns with classical chemical concepts of hypervalency and the Zintl line. ME-AI demonstrates excellent predictive capabilities and generalizes beyond square-net compounds to accurately identify topological insulators within the rocksalt structure family. Our findings highlight the potential of combining human expertise with artificial intelligence to facilitate targeted materials discovery.

Results

ME-AI Workflow

ME-AI aims to reveal the expert intuition (t-factor) and further insights. This starts with the ME curating a dataset of square-net materials (see Supplementary Methods A). With this curated data, the AI part of ME-AI reveals the correlations between different primary features (PFs) and discovers the emergent descriptors composed of primary features using a Gaussian process (GP). We choose the PFs to be atomistic or structural, such that we can interpret the ML results from a chemical perspective. For atomistic features, we focus on (a): electron affinity, the change in energy after a neutral atom gains an electron; (b): (Pauling) electronegativity, the relative ability of an atom to attract a shared electron pair; and (c): valence electron count, corresponding to the number of electrons in a neutral atom. Since each compound consists of different numbers of elements, we built primary atomistic features in a uniform structure by choosing maximum and minimum values of these features and the square-net element features (see Fig. 2a). The last atomistic feature is (d) the estimated face-centered cubic lattice parameter fcc of the square-net element, which reflects the radius of the square-net atom. Given the prior choice of square-net compounds, two lengths are sufficient to represent the structural feature variations: (e) the crystallographic characteristic distances d_{sq} and d_{nn}. In total, we chose 12 PFs (see Fig. 2a).

Having determined the PFs, we curated an experimentally measured database of PFs for 879 compounds belonging to the 2D-centered square-net class in the inorganic crystal structure database (ICSD)[31]. The structure types of these compounds include PbFCl, ZrSiS (ZrSiTe, $AmTe_{2-x}$, and UP_2), PrOI, Cu_2Sb, ZrCuSiAs-$HfCuSi_2$, $LaZn_{0.5}Sb_2$, $PrNiSb_2$ and $CaBe_2Ge_2$. While we focus on the space group 129, the methodology and features we have found would work in other space groups, as long as one can define a centrosymmetric square net. The expert labeling of the materials is a critical part of the data curation, where the expert insight is passed on to the model (see Fig. 1a). The ambitious goal of ME-AI is to go beyond mechanical scraping of data, but to use the logic routinely employed by materials experimentalists to give the ML model the opportunity to articulate the expert's insight. Our priority in the labeling process was to capture as comprehensive information as possible. When experimental or computational band structure is available for a given entry, we label the material through visual comparison of the band structure to the square-net tight-binding model band structure in Fig. 1e. This applies to 56% of the entire database. For alloys, we use the expert chemical logic to label the material based on the labels of the parent materials. This applies to 38% of the database. For instance, since HfSiS, HfSiSe, ZrSiS, and ZrSiSe are all TSM, we expect (Hf,Zr)Si(S,Se) to also be TSM. 6% of the database consists of stochiometric compounds without available band structure information, but that are closely related to materials with band structure, for instance, by cation substitution. We again used chemical logic in such cases. Once our GP model is trained, we inspect the distribution of different labeling mechanisms in the space of learned emergent features to confirm that consistent and uniform feature learning occurred across different branches of the labeling process.

See Supplementary Methods A for details.

From the machine-learning perspective, the task at hand is to use this relatively small labeled dataset of 12-dimensional PFs to discover effective descriptors that predict TSMs automatically. The simplest dimensional reduction approach of principal component analysis (PCA) fails (see Supplementary Discussion A) since PCA cannot incorporate prior knowledge of the labels. On the other hand, the go-to supervised learning architectures using neural networks are also inadequate due to the small dataset, lack of interpretability, and the danger of overfitting. ME-AI treats this problem as a Gaussian Process (GP) classification task[32] and obtains the descriptors by learning a purposefully chosen factor analysis matrix parameterizing the kernel for the GP[33]. ME-AI achieves a classification accuracy above 0.98 (see Supplementary Methods C for hyperparameters for training and stability testing). While random forest offers a comparable accuracy, it fails to reveal scientific insight due to limited interpretability (see Supplementary Discussion B).

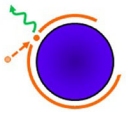
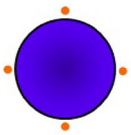
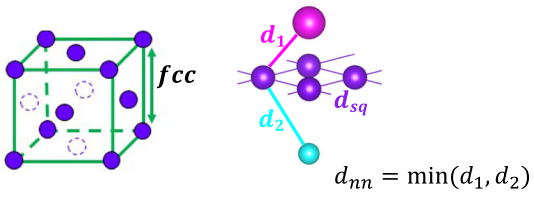
The emergent descriptors

The correlation matrix learned from the GP classifier (see Methods) reveals the emergent descriptors that identify TSMs. The elements of the normalized-correlation matrix, $\{\mathbf{C}_{mn}\}$ [eq: (4)] shown in Fig. 3a reveal the relative strengths and signs of interaction among all the 72 pairs of PFs. Five elements of $\{\mathbf{C}_{mn}\}$ clearly stand out in their magnitudes with varying signs. We hypothesize that positive (negative) and strong $\{\mathbf{C}_{mn}\}$ can be effectively captured by emergent descriptors formed from products (ratios) of PF m and PF n. To test this hypothesis, we plot the distribution of the products and ratios, i.e., d_{sq}/d_{nn}, d_{sq}/fcc, χ_{sq}/d_{sq}, $\chi_{sq}{}^*d_{nn}$, and $\chi_{sq}{}^*fcc$, as shown in Fig. 3b–f. As shown in Supplementary Discussion C, the distributions confirm small values of d_{sq}/d_{nn}, d_{sq}/fcc and intermediate values of χ_{sq}/d_{sq}, $\chi_{sq}{}^*fcc$, and $\chi_{sq}{}^*d_{nn}$ as successful descriptors of TSM. These descriptors synthesize atomistic information of elements forming the material and the structural information of the compound in three different ways: (1) d_{sq}/d_{nn} relies solely on structural features, (2) d_{sq}/fcc, $\chi_{sq}{}^*d_{sq}$, and $\chi_{sq}{}^*d_{nn}$ combine structural and atomistic features, and (3) $\chi_{sq}{}^*fcc$ is a descriptor solely based on atomistic features. It is noteworthy that ME-AI reproduces $t = d_{sq}/d_{nn}$ as one of the emergent descriptors without prior knowledge. This bolsters the insight underlying the t-factor and establishes ME-AI's capability to "bottle" expert insight latent to the expert curation of the data.

We now turn to our four newly discovered descriptors. Among these, the distribution of d_{sq}/fcc (Fig. 3c) is similar to the distribution of the t-factor (Fig. 3b), although d_{sq}/fcc does a better job in separating TSM from trivial materials (see Supplementary Discussion C for quantitative analysis of the distributions). Since fcc estimates the size of the unbonded square-net atoms, small values of d_{sq}/fcc and that of the t-factor both point to a chemically and electronically significant role of the square-net planes.

The distributions of the remaining three descriptors, $\chi_{sq}{}^*d_{nn}$, χ_{sq}/d_{sq}, and $\chi_{sq}{}^*fcc$, are similar in that TSM favors intermediate values of these descriptors (Fig. 3d, f). This result reveals the underappreciated significance of hypervalency for TSM. The concept of hypervalency accounts for the seeming violation of the octet rule in molecules and solids, a classic chemical rule, the main group elements want to reach noble gas configuration with 8 valence electrons to fill the s and p shells[34]. Hypervalent bonds connect more than two atoms and allow cases with the seeming violation of the octet rule to satisfy the octet rule in an extended sense[35,36]. Hypervalent bonds make the desired electron count for placing the Fermi level at the right position possible: it stabilizes half-filled p_x and p_y orbitals by creating covalent bonds with only one electron per bond. While hypervalent bonds were known to be possible in square-net materials, a descriptor quantifying hypervalency was not known. The three descriptors address complementary but distinct

(a)

Features	Symbols	Illustration
Electron Affinity	◦ EA_{max} ◦ EA_{min} ◦ EA_{sq}	
Electronegativity	◦ χ_{min} ◦ χ_{sq}	
Valence electrons	◦ NE_{max} ◦ NE_{min} ◦ NE_{sq} ◦ NE_{tot}	
Length scales	◦ r_w ◦ d_{sq} ◦ d_{nn}	

(b)

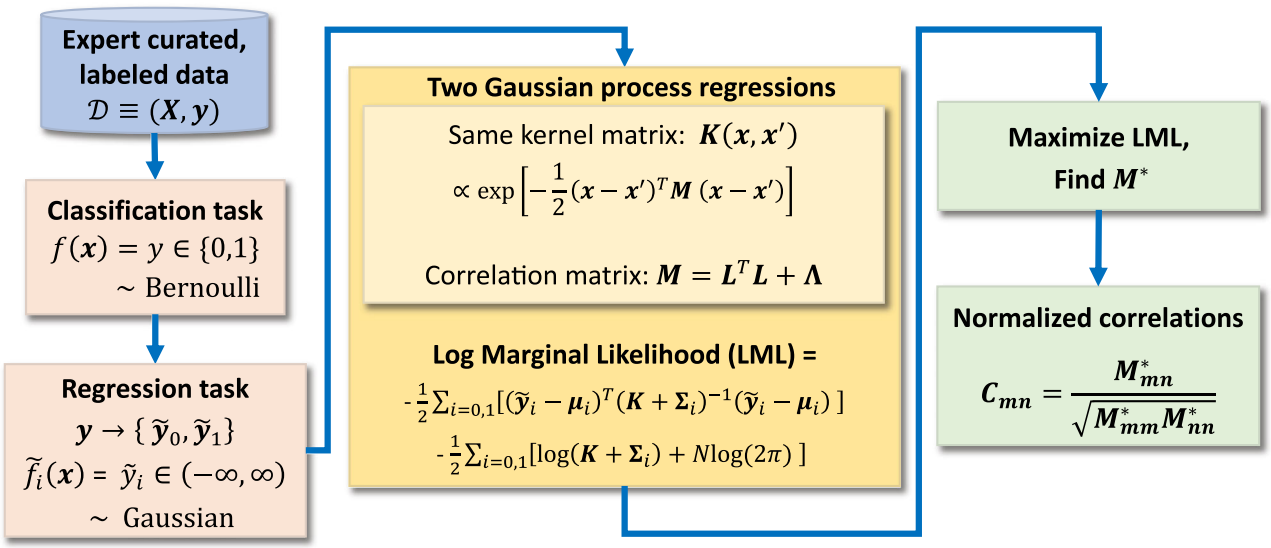


Fig. 2 | ME-AI machine-learning scheme. a Primary features. Subscripts *max*, *min*, and *sq* refer to the maximum, minimum, and square-net element values. The subscript *tot* refers to the total number of electrons based on the stoichiometric information. We omit χ_{max} because of its high covariance with χ_{sq}. Graphical illustrations of the atomic features are shown in the third column. **b** A flowchart illustration of ME-AI. We first create a preprocessed and labeled dataset $\mathcal{D} \equiv (\mathbf{X}, \mathbf{y})$ composed of the 12 primary features, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ with $\mathbf{x}_n \in R^D$ where $D = 12$, and their class labels $\mathbf{y} = (y_1, \ldots, y_N)$ with $y_n \in \{0, 1\}$. The subscript n denotes different materials in the total N entries database. After converting the classification task into two independent regression tasks, we set up two GPRs with the shared kernel $\mathbf{K}(\mathbf{x}, \mathbf{x}')$. The correlation matrix $\mathbf{M}$ inside the kernel couples the different primary features and is designed to have a factor analysis structure $\mathbf{M} = \mathbf{L}^T\mathbf{L} +, \mathbf{\Lambda}$, which allows us to learn interactions between different PFs with a limited number of hyperparameters. We train the model with the whole dataset by maximizing the log marginal likelihood. This automatically includes regularization. We get the correlations $\mathbf{M}^*$ from the trained model parameters between the primary features. The significance of the coupling between the m^{th} and n^{th} primary features is measured by the normalized correlation $\mathbf{C}_{mn}$. This reveals the pairs of significant features that identify the square-net TSMs.

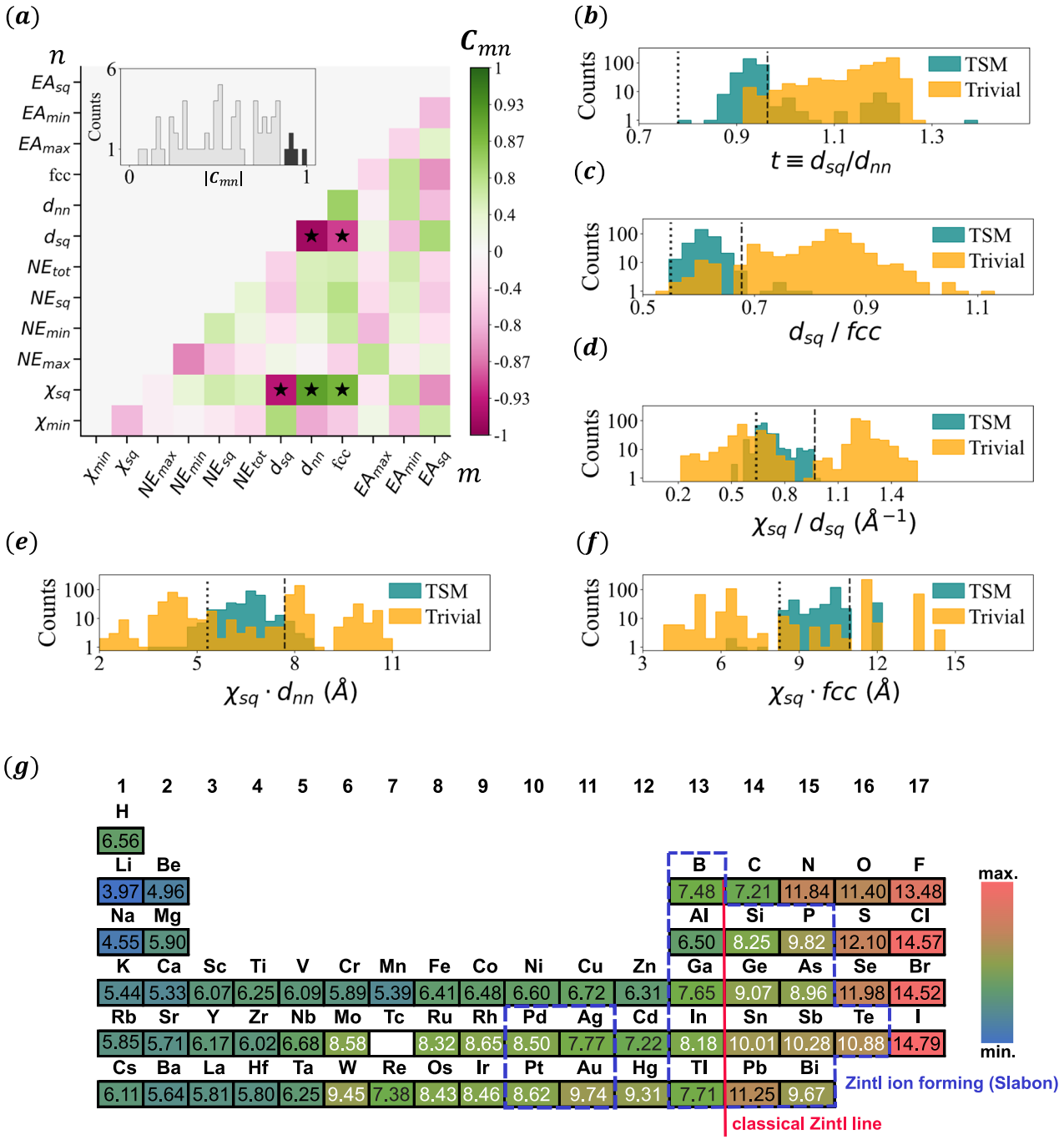


Fig. 3 | The main results from the ME-AI. a The normalized-correlation matrix C_{mn}, where m and n denote the primary features. Only the lower triangle is shown. The inset shows the distribution of magnitudes of non-diagonal elements, $|C_{mn}|$, with the strongest elements shown in black. The strongest elements: (d_{sq}, d_{nn}), (d_{sq}, fcc), (χ_{sq}, d_{sq}), (χ_{sq}, d_{nn}), and (χ_{sq}, fcc) are marked with a star. The values of **b** d_{sq}/d_{nn}, **c** d_{sq}/fcc, **d** χ_{sq}/d_{nn}, **e** $\chi_{sq} \cdot d_{nn}$, and **f** $\chi_{sq} \cdot$ fcc, for the TSM and trivial materials in our dataset. The vertical lines mark the range of the emergent descriptor supporting TSMs. **g** The periodic table with the new element-specific descriptor $\chi_{sq} \cdot$ fcc color-coded from lowest to highest values from blue to red, respectively. The entries in the white font have values between 8 and 11 (TSM window from panel (**f**)). The red line represents the classical Zintl line, while the blue-framed elements are known to form Zintl ions as described by Slabon[44].

aspects determining hypervalent bonds in square-net atoms. $\chi_{sq}{}^{*}d_{nn}$ and χ_{sq}/d_{sq} reveal the bonding tendency with out-of-plane atoms, and $\chi_{sq}{}^{*}$fcc reflects the atomistic features that facilitate hypervalent bonding. Specifically, only elements of intermediate electronegativity and size will be inclined to form electron-sharing hypervalent bonds.

Since $\chi_{sq}{}^{*}$fcc is made of purely atomistic features of square-net atoms, we can calculate the descriptor for each element of the periodic table, as shown in Fig. 3g (we exclude radioactive elements and rare earth elements). The positions of elements in the periodic table with $8 < \chi_{sq}{}^{*}\text{fcc} < 11$, the range identified to support TSM in Fig. 3f, reveal a remarkable connection between this ME-AI learned descriptor and the classic chemical concept of "Zintl line"[37]. The Zintl-line divides elements that form covalent bonds with each other to yield polyanions (or polycations) from elements that form regular monoatomic cations (or anions)[37,38]. Thus, it can also point to hypervalent bonds, as they can be viewed as a type of covalent bond. Classically, the Zintl line is drawn between group 13 and 14[39]; however, extensions of the Zintl-Klemm concept included the higher homologs of group 13 and heavy transition metals such as Au[38]. Thus, the ME-AI learned

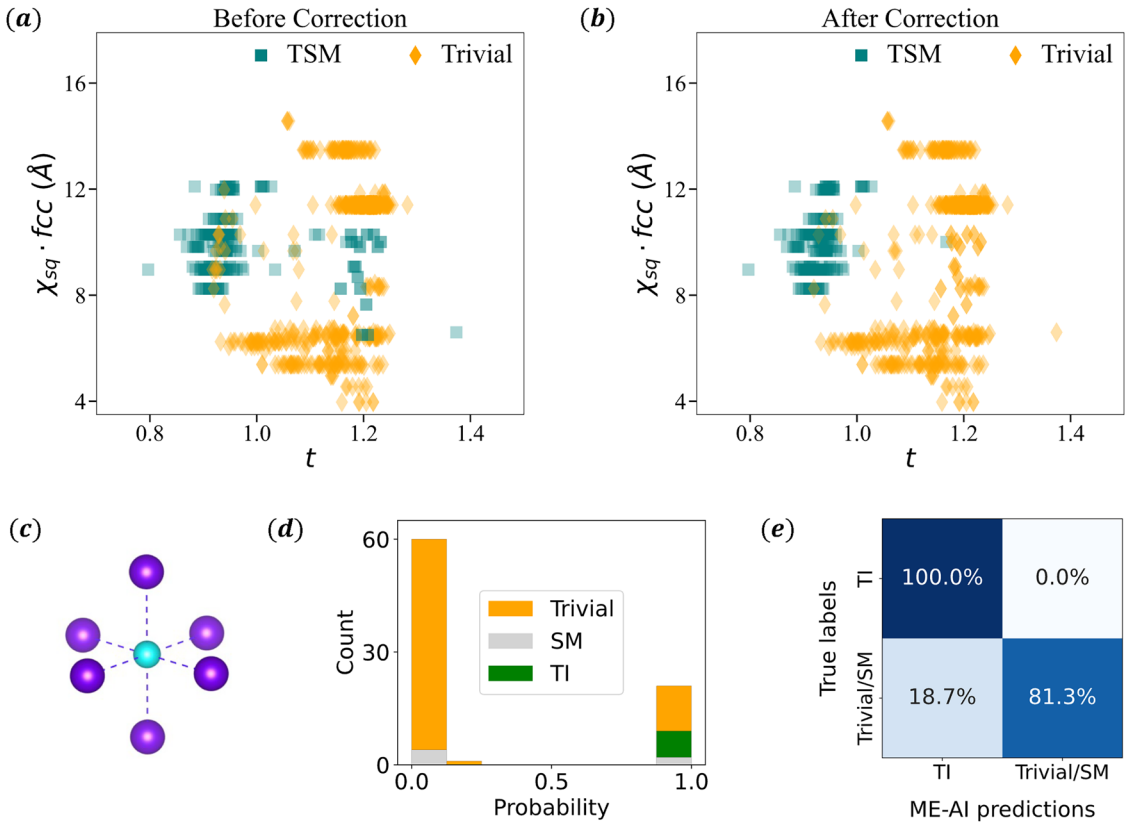


Fig. 4 | The predictive power of the ME-AI. a, **b** The scatter plot showing TSM and trivial materials in the 2D descriptor space spanned by the *t*-factor and $\chi_{sq} \cdot$ fcc. The TSM (teal) and trivial (orange) materials are separated sharply in this 2D space. **a** The originally labeled TSM (teal) and trivial (orange) materials show some outliers in the clustered regions of TSM and trivial materials. (**b** Further analysis enabled us to correct mislabeled materials in our original dataset. **c** 3D Rocksalt structure. **d** Stacked bar plot of the distribution of predicted probabilities $P(T|AB)$ for rocksalt materials classified by their expert-assigned labels: trivial, semimetal (SM), and topological insulator (TI) (see Supplementary Table 1). The GP classifier labels the material *AB* as topological if $P(T|AB) > 0.5$ and trivial/SM if $P(T|AB) < 0.5$. **e** Normalized confusion matrix to compare the expert-assigned labels (true labels) and the generalized ME-AI predictions for rocksalt materials. The matrix displays the proportion of correctly and incorrectly classified instances for each class, normalized by the total number of instances in the respective class. See Supplementary Fig. 7 for the unnormalized matrix and the accuracy metrics.

atomistic descriptor χ_{sq}*fcc agrees with the extended Zintl-Klemm concept and highlights the importance of covalent bonding for TSM formation.

Discussion

We now turn to new insights ME-AI enabled. Firstly, the excellent separation between TSM and trivial materials we achieved with the ME-AI learned descriptors allowed us to detect misplaced labels in the original curation of the material dataset, as shown in Fig. 4. With the *t*-factor alone, significant mixing between TSM and Trivial (see Fig. 1e) made it difficult to identify true outliers and potential mislabeling. However, using a two-dimensional descriptor space (t, χ_{sq}*fcc) as shown in Fig. 4a, the TSM and trivial materials separate sharply and draw attention to the 32 outliers. Upon inspection, we found the original labeling left some ambiguity for these outlier compounds. Those incorrect labels were a result of human error, due to the limitations of the intuition-based evaluation of band structures (Supplementary Discussion D). For a closer inspection, we performed in-house DFT calculations on these outliers and compared the orbital-projected band structure with the tight-binding model band structure in Fig. 1g (see Supplementary Discussion D). For 32 compounds, this close-up study resulted in label correction. The corrected distribution in Fig. 4b demonstrates the power of the ME-AI approach in bottling the essence of the expert intuition in a robust manner that can withstand small unintended errors inevitable in data curation.

Moreover, the significance of the new atomistic features allows us to extend our considerations beyond square-net compounds by predicting the probability for a given element *A* to form a material with a topological band structure, $P(T|A)$, in a different structure type. As we show in Supplementary Discussion E, such probability can be estimated by training a GP with just (χ_{sq}, fcc) and the TSM label, using the Bayes theorem. Given the structural prior in the training data, we choose to test the generalization possibility in a closely related structural family of rocksalts (see Fig. 4c). Since the rocksalts consist of two elements, we can estimate the probability for the rocksalt *AB* to have a topological band structure $P(T|AB) = P(T|A)P(T|B)$. Comparing the resulting probability with the band structure characterization given in Supplementary Table 1 shows a remarkable agreement: the generalized ME-AI prediction of $P(T|AB) \gg 0.5$ captures 100% of the topological insulators, while the majority of trivial materials have $P(T|AB) \ll 0.5$ (see Fig. 4d, e). The unreasonable success of the ME-AI approach in generalizing the predictions of the model trained only on the squareness compounds to the rocksalts hints at a broader overarching perspective. Most of these TIs form the so-called metavalent bonds[40], which are delocalized bonds between different atoms. Hence, the generalization can be understood as ME-AI uncovering the role of bonding through its training over atomistic and structural primary features, based on the band structure labels.

Conclusion

In summary, we introduced a machine-learning strategy, ME-AI, that bottles expert intuition underlying expertly curated data and articulates the intuition into quantifiable descriptors for target features. The strategy starts with expertly curated and labeled data. Deployed to a 12-dimensional experimentally measured PF set for an expertly curated and labeled dataset of 879 square-net compounds, ME-AI rediscovered the *t*-factor that human experts gleaned from the same data independently. Moreover, ME-AI revealed four new emergent descriptors. While one of the new descriptors addresses the same chemical intuition as the *t*-factor, highlighting the critical

role of the square-net plane, the remaining three new descriptors all point to the new insight regarding hypervalency. In particular, one of the new descriptors χ_{sq}*fcc is entirely atomistic concerning the square-net atoms, allowing us to place the new descriptor on the periodic table and check the hypervalency interpretation against chemical knowledge. The new descriptors led to predictive probability that successfully predicted topological insulators among rocksalt compounds, demonstrating a remarkable generalization. The success demonstrates ME-AI's capability to encapsulate human chemical intuition and expertise from relatively small, expertly curated data to advance materials discovery. Such synergy between human experts and machine intelligence can accelerate materials discovery, building on precious expert experiences.

An exciting future direction to build on the initial success of ME-AI would be to target a database of materials with target properties such as superconductivity or magnetism. To apply ME-AI across a heterogeneous mix of structures, it will be critical to develop a principled approach to featurizing atomistic information together with the structural environment for each element in the material. With database built on published experimental results, such as SuperCon[41] it will be productive to engage a large language model.

Methods

ME-AI starts with an expert-curated database that carries and informs the expert's intuition (see Fig. 1a). Of the 879 materials, we select 861 materials with structure types other than $CaBe_2Ge_2$ as our training data. We exclude the $CaBe_2Ge_2$ structure type for training as these entries have errors in their labels (see Supplementary Methods A for details). To ensure a fair comparison of the PFs with different magnitudes and units, each PF is linearly normalized to lie in the range [0, 1] (min-max normalization). Thus, our preprocessed training dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ for the $N = 861$ materials consists of the input features (normalized PFs) $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and the output labels $\mathbf{y} = (y_1, \ldots, y_N)^T$. The features $\mathbf{x}_n \in \mathbb{R}^D$ is a $D = 12$-dimensional vector of the n^{th} material, whose output label $y_n \in \{0, 1\}$ represents the trivial class as 0 and the TSM class as 1. The ME-AI proceeded in three steps, as depicted in Fig. 2b and Supplementary Methods B.

Step 1: Assuming a Bernoulli model for the discrete class labels, we start with a Beta prior to the distribution of labels. Next, we transform these labels to a latent space where we can approximate the likelihood by a Gaussian, following the route of ref. 32 (see Supplementary Methods B). With this transformation, each label $y_n \in \{0, 1\}$ gives two new variables $\tilde{y}_0, \tilde{y}_1 \in (-\infty, \infty)$ (given by Eq.(15) in Supplementary Methods B), whose distributions are approximately Gaussian. We can now proceed with two Gaussian Process Regressions (GPRs) for the $\tilde{\mathbf{y}}_0$ and $\tilde{\mathbf{y}}_1$.

Step 2: As our goal is to learn the feature correlations that encompass both labels, we design a Mahalanabois kernel $\mathbf{K}$ to be shared between the two GPRs, as

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') := \sigma_p^2 \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')\right], \tag{1}$$

whose elements connect the features $\mathbf{x}$ and $\mathbf{x}'$. Assuming the existence of a smaller number of hidden features emerging from combinations of the primary features[33], we use a factor analysis model for the structure of the correlation matrix $\mathbf{M}$ as,

$$\mathbf{M} = \mathbf{L}^T \mathbf{L} + \mathbf{\Lambda}, \tag{2}$$

where $\mathbf{L}$ is a $q \times D$ matrix with $q < D$, and $\mathbf{\Lambda}$ is a positive definite diagonal matrix. The q sets the number of common factors, and we find $q = 6$ to be the optimal choice, as found through hyperparameter tuning (see Supplementary Methods C). The smaller value of q limits the number of hyperparameters to be learned from our relatively small volume of data.

Step 3: Now we carry out the standard GPR by maximizing the log marginal likelihood (LML) to learn the hyperparameters, including the correlation matrix $\mathbf{M}$, which will reveal the descriptors. Given the assumed Gaussian distribution of the transformed variables, $\tilde{\mathbf{y}}_i$ the LML has the usual closed-form expression[42] given by,

$$\begin{aligned} \mathrm{LML} = -\frac{1}{2}\sum_{i=0}^{1}\big[&(\tilde{\mathbf{y}}_i - \boldsymbol{\mu}_i)^T (\mathbf{K} + \mathbf{\Sigma}_i)^{-1} (\tilde{\mathbf{y}}_i - \boldsymbol{\mu}_i) \\ &+ \log|\mathbf{K} + \mathbf{\Sigma}_i| + N \log 2\pi\big], \end{aligned} \tag{3}$$

where $\mathbf{K}$ and $\boldsymbol{\mu}_i$ are the parameters to be learned, and $\mathbf{\Sigma}_i$ is a diagonal matrix set by the Beta prior (see Supplementary Methods B). In the above expression, the first term promotes the model fitness while the second term penalizes model complexity. The marginal likelihood produces principled and automatic regularization, satisfying the "Occam's razor" property[43].

With our database, training to maximize LML takes ~8 min using a Quadro GV100 GPU (see Supplementary Methods C for details). We interpret the LML maximizing matrix elements $\mathbf{M}^*$ to reflect how PFs interact to form the descriptors of TSM. For this, we construct a normalized correlation with elements given by

$$\mathbf{C}_{mn} = \frac{\mathbf{M}^*_{mn}}{\sqrt{\mathbf{M}^*_{mm}\mathbf{M}^*_{\mathrm{nn}}}}. \tag{4}$$

Now, the correlation matrix reveals the interaction between PFs in forming descriptors of TSM.

Data availability

All the data used for the analysis are available at github.com/KimGroup/MEAI_TSM.

Code availability

All the codes used for the analysis are available at github.com/KimGroup/MEAI_TSM.

Received: 25 September 2024; Accepted: 24 August 2025;
Published online: 29 September 2025

References

1. Horton, M. K., Dwaraknath, S. & Persson, K. A. Promises and perils of computational materials databases. *Nat. Comput. Sci.* **1**, 3–5 (2021).
2. Landis, D. D. et al. The computational materials repository. *Comput. Sci. Eng.* **14**, 51–57 (2012).
3. Jain, A. et al. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
4. 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).
5. De Jong, M. et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 150009 (2015).
6. Ashton, M., Paul, J., Sinnott, S. B. & Hennig, R. G. Topology-scaling identification of layered solids and stable exfoliated 2D materials. *Phys. Rev. Lett.* **118**, 106101 (2017).
7. Haastrup, S. et al. The computational 2D materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Mater.* **5**, 042002 (2018).
8. Zhou, J. et al. 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci. Data* **6**, 86 (2019).
9. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).
10. Talirz, L. et al. Materials Cloud, a platform for open computational science. *Sci. Data* **7**, 299 (2020).
11. Choudhary, K. et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).

12. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
13. Stanev, V., Choudhary, K., Kusne, A. G., Paglione, J. & Takeuchi, I. Artificial intelligence for search and discovery of quantum materials. *Commun. Mater.* **2**, 105 (2021).
14. Gregoire, J. M., Zhou, L. & Haber, J. A. Combinatorial synthesis for AI-driven materials discovery. *Nat. Synth.* **2**, 493–504 (2023).
15. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).
16. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
17. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
18. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
19. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
20. Isayev, O. et al. Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chem. Mater.* **27**, 735–743 (2015).
21. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
22. Fu, C. et al. Large Nernst power factor over a broad temperature range in polycrystalline Weyl semimetal NbP. *Energy Environ. Sci.* **11**, 2813–2820 (2018).
23. Xie, R., Zhang, T., Weng, H. & Chai, G.-L. Progress, advantages, and challenges of topological material catalysts. *Small Sci.* **2**, 2100106 (2022).
24. Liu, J., Xia, F., Xiao, D., García De Abajo, F. J. & Sun, D. Semimetals for high-performance photodetection. *Nat. Mater.* **19**, 830–837 (2020).
25. Bradlyn, B. et al. Topological quantum chemistry. *Nature* **547**, 298–305 (2017).
26. Klemenz, S., Lei, S. & Schoop, L. M. Topological semimetals in square-net materials. *Annu. Rev. Mater. Res.* **49**, 185–206 (2019).
27. Klemenz, S. et al. The role of delocalized chemical bonding in square-net-based topological semimetals. *J. Am. Chem. Soc.* **142**, 6350–6359 (2020).
28. Jovanovic, M. & Schoop, L. M. Simple chemical rules for predicting band structures of Kagome materials. *J. Am. Chem. Soc.* **144**, 10978–10991 (2022).
29. Klemenz, S., Schoop, L. & Cano, J. Systematic study of stacked square nets: from Dirac fermions to material realizations. *Phys. Rev. B* **101**, 165121 (2020).
30. Ye, L. et al. Hopping frustration-induced flat band and strange metallicity in a kagome metal. *Nat. Phys.* **20**, 610–614 (2024).
31. Levin, I. *Inorganic Crystal Structure Database (ICSD).* http://icsd.fiz-karlsruhe.de (2020).
32. Milios, D., Camoriano, R., Michiardi, P., Rosasco, L. & Filippone, M. Dirichlet-based Gaussian processes for large-scale calibrated classification (eds. Bengio, S. et al.) In: *Advances in Neural Information Processing Systems*, Vol. 31 https://proceedings.neurips.cc/paper_files/paper/2018/file/b6617980ce90f637e68c3ebe8b9be745-Paper.pdf (Curran Associates, Inc., 2018).
33. Vivarelli, F. & Williams, C. Discovering hidden features with Gaussian processes regression (eds. Kearns, M., Solla, S. & Cohn, D.) In: *Advances in Neural Information Processing Systems*, Vol. 11. https://proceedings.neurips.cc/paper_files/paper/1998/file/ed4227734ed75d343320b6a5fd16ce57-Paper.pdf (MIT Press, 1998).
34. Tro, N., Fridgen, T., Shaw, L. & Boikess, R. *Chemistry: A Molecular Approach* (Pearson Canada, 2013).
35. A. Papoian, G. & Hoffmann, R. Hypervalent bonding in one, two, and three dimensions: extending the Zintl-Klemm concept to nonclassical electron-rich networks. *Angew. Chem. Int. Ed.* **39**, 2408–2448 (2000).
36. Rogachev, A. Y. & Hoffmann, R. Hypervalent compounds as ligands: I_3 -anion adducts with transition metal pentacarbonyls. *Inorg. Chem.* **52**, 7161–7171 (2013).
37. Zintl, E. Intermetallische verbindungen. *Angew. Chem.* **52**, 1–6 (1939).
38. Janka, O. & Kauzlarich, S. M. Zintl compounds. *Encyclopedia of Inorganic and Bioinorganic Chemistry.* 1–19 (John Wiley & Sons, Ltd, 2011).
39. Wang, F. & Miller, G. J. Revisiting the Zintl–Klemm concept: alkali metal trielides. *Inorg. Chem.* **50**, 7625–7636 (2011).
40. Wuttig, M., Deringer, V. L., Gonze, X., Bichara, C. & Raty, J.-Y. Incipient metals: functional materials with a unique bonding mechanism. *Adv. Mater.* **30**, 1803777 (2018).
41. Center for Basic Research on Materials. MDR SuperCon Datasheet Ver.240322. https://doi.org/10.48505/nims.4487.
42. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning) (MIT Press, Cambridge, Mass, 2006).
43. MacKay, D. *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).
44. Slabon, A. *Structure-Property Relationships in Ternary Zintl Phases Based on Europium-metal-tetrelides and Lithium-transition Metal-silicides*. Ph.D. thesis, ETH Zurich (2013).

Acknowledgements

M.J., W.M., A.W., L.S., and E.A.K. were supported by NSF grant OAC-2118310. The computation was done using a high-powered computing cluster that was established through the support of the Gordon and Betty Moore Foundation's EPiQS Initiative, Grant GBMF10436 to E.A.K., and through the support of the MURI grant FA9550-21-1-0429. Y.L. was supported by the Gordon and Betty Moore Foundation's EPiQS Initiative, Grant GBMF10436 to E.A.K., and by the MURI grant FA9550-21-1-0429. K.M. was supported by Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship: a Schmidt Futures program. S.K. was supported by the Fraunhofer Internal Programs under Grant No. Attract 170-600006.

Author contributions

E.K., L.S., and A.W. conceived of the idea. Y.L., W.M. contributed to the coding. S.K., M.J. built the dataset. Y.L. performed the model training. Y.L. and K.M. analyzed the ML results. All authors contributed to the interpretation of the results and 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/s43246-025-00928-7.

Correspondence and requests for materials should be addressed to Eun-Ah Kim.

Peer review information *Communications Materials* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary handling editors: Pascal Friederich and Aldo Isidori. A peer review file is available.

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025