Multimodal Learning for Materials

Viggo Moro *,[1] Charlotte Loh *,†,[2] Rumen Dangovski *,[2] Ali Ghorashi,[1] Andrew Ma,[2]
Zhuo Chen,[1] Samuel Kim,[3] Peter Y. Lu,[4] Thomas Christensen,[5] and Marin Soljačić †[1]

[1]*Department of Physics, Massachusetts Institute of Technology, USA*
[2]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA*
[3]*Research and Exploratory Development, John Hopkins University Applied Physics Laboratory, USA*
[4]*Data Science Institute, University of Chicago, USA*
[5]*Department of Electrical and Photonics Engineering, Technical University of Denmark, Denmark*

Artificial intelligence is transforming computational materials science, improving the prediction of material properties, and accelerating the discovery of novel materials. Recently, publicly available material data repositories have grown rapidly. This growth encompasses not only more materials, but also a greater variety and quantity of their associated properties. Existing machine learning efforts in materials science focus primarily on single-modality tasks, i.e., relationships between materials and a single physical property, thus not taking advantage of the rich and multimodal set of material properties. Here, we introduce Multimodal Learning for Materials (MultiMat), which enables self-supervised multi-modality training of foundation models for materials. We demonstrate our framework's potential using data from the Materials Project database on multiple axes: (i) MultiMat achieves state-of-the-art performance for challenging material property prediction tasks; (ii) MultiMat enables novel and accurate material discovery via latent space similarity, enabling screening for stable materials with desired properties; and (iii) MultiMat encodes interpretable emergent features that may provide novel scientific insights.

I. INTRODUCTION

Data-based approaches have become increasingly prevalent in computational materials science [1–7], due to the rapid algorithmic innovations in the field of machine learning (ML) [8] as well as by the growing amount of data available in materials science databases [9–14]. An exciting aspect of ML in materials science lies in its potential to greatly accelerate calculations. Although training an ML model requires an up-front computational cost, predicting a material property using a trained ML model is substantially faster than running an ab initio calculation [15–17]. The discovery of new materials relies on that speedup since the vast combinatorial space of possible materials makes exhaustive ab initio calculations computationally infeasible. There have been a number of works that demonstrate the use of ML models to rapidly screen large amounts of materials with the aim of accelerating materials discovery [18–21]. Beyond these screening-based approaches—which rely on predictive models—there is also an emerging interest in the use of generative models for materials discovery [22–24]. Developing better graph neural networks (GNNs) [25–30] has represented the research frontier for achieving state-of-the-art predictive performance of materials. However, while interpretability has been a focus of ML for science, including in the domain of materials [21, 31–36], GNNs, as any other deep neural network, usually fall short when it comes to interpretability.

An increasingly important paradigm in ML is foundation models, which are general-purpose ML models that are pre-trained on large amounts of data and then fine-tuned for a variety of applications [37]. Notable examples include GPT-4 [38] and Gemini [39]. Because pre-training is performed using unsupervised methods, these foundation models are able to take advantage of extremely large amounts of data that would normally be difficult to utilize when directly training models for specific downstream tasks. A seminal work in multimodal learning is Contrastive Language Image Pre-training (CLIP) [40], which can be used to train multimodal foundation models. CLIP aligns an image encoder with a text encoder, encouraging the embeddings of the image and captions to be similar. Subsequent efforts [41–45], have predominantly focused on multimodal learning with just two modalities (usually images and text) [46–49]. How to best incorporate more than two modalities remains an open problem [50–52].

Here, we adapt CLIP to the materials domain and also extend it to multimodal pre-training with an arbitrary number of modalities. We leverage the fact that materials databases are inherently multimodal: e.g., besides the crystal structure, the density of states (DOS) [53, 54] and charge density [55] convey rich and complementary information about materials. Textual descriptions of the crystal, which can be machine-generated [56], offer a fourth, complementary modality that is additionally computationally cheap to acquire. Motivated by these opportunities, we introduce *Multimodal Learning for Materials* (MultiMat), a novel framework for training a foundation model for crystalline materials that allows for the incorporation of several modalities. The basis for MultiMat is a multimodal pre-training method that connects high-dimensional material properties (i.e., modalities) in a shared latent space to produce highly effective material representations that can then be transferred to various downstream tasks. Using MultiMat, we pre-train a state-of-the-art GNN on the Materials Project [10] database to demonstrate its ability to produce state-of-the-art foundation models for materials. Very recently, preliminary work has explored related ideas for molecules [57] and a structure-agnostic multi-task learning approach for crystals [58].

The MultiMat framework trains a foundation model for materials by aligning the latent spaces of encoders of different information-rich modalities, such as the crystal structure, DOS, charge density, and textual description, as shown in Fig. 1a.

---

* These authors contributed equally to this work.
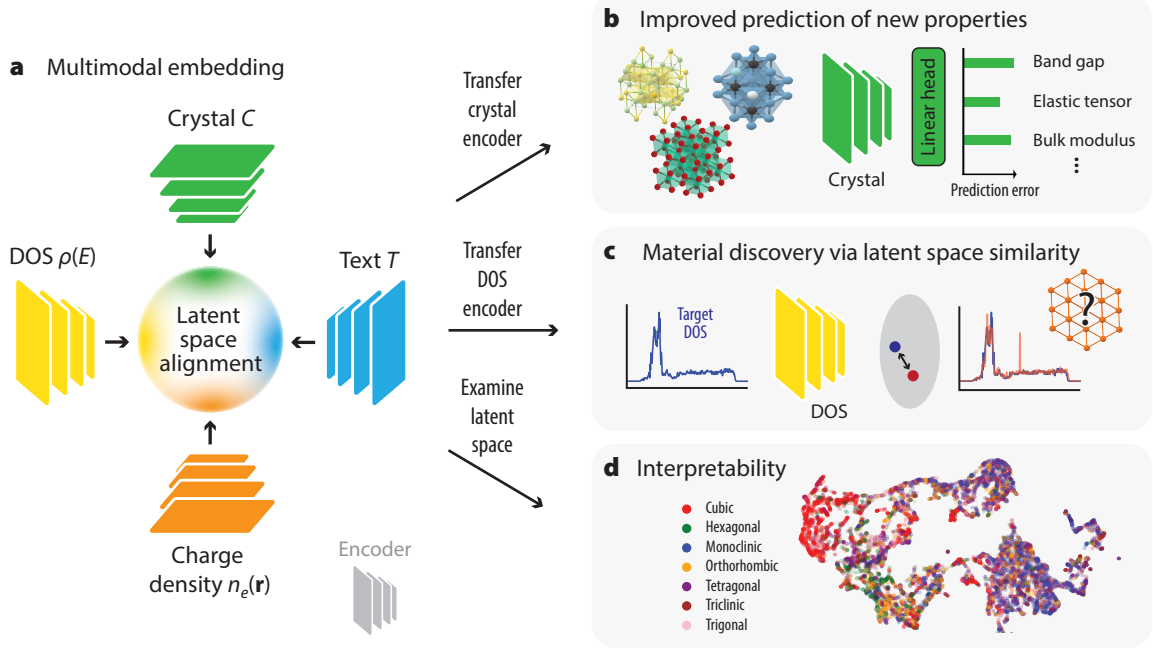† {cloh, soljacic}@mit.edu

Figure 1. The Multimodal Learning for Materials (MultiMat) approach. a, Crystal (C), DOS ($\rho(E)$), charge density ($n_e(\mathbf{r})$), and text (T) encoders map each modality to embeddings in a shared multimodal latent space (center). MultiMat's training objective aligns the embeddings of different modalities corresponding to the same material. **b**, Application of MultiMat in improved prediction of materials' properties. The C encoder from (a) is transferred, and a randomly initialized linear head is trained jointly with the transferred encoder to predict material properties. **c**, Application of MultiMat in material discovery. The DOS encoder embeds a target DOS (in blue). In the shared latent space, the closest crystal embedding (in red) from a large collection of crystal embeddings is selected. Since the embeddings of DOS and crystal are aligned during training, the crystal whose embedding is closest to the target DOS embedding is highly likely to have a DOS (in red) that closely resembles the target. Therefore, this crystal is identified as the best candidate. **d**, Application of MultiMat in enabling interpretability. We visualize the latent space of the crystal encoder using dimensionality reduction to reveal information about properties of materials that are implicitly encoded in the embeddings.

This alignment process produces shared latent spaces and effective material representations which can then be leveraged for a series of downstream tasks (Fig. 1b–d). For instance, the crystal encoder can be transferred and fine-tuned for material property prediction, enabling improved predictive performance compared to traditional training techniques. Since MultiMat aligns the latent spaces of different modalities, it can also be used in a novel material discovery strategy by screening large crystal-structure databases with comparisons between target properties and candidate crystals based on the latent-space similarity. Finally, we demonstrate the interpretability enabled by the MultiMat approach, by exploring the latent space from MultiMat using a dimensionality reduction approach.

II. RESULTS

A. Modalities and Architecture

To illustrate the MultiMat framework, we consider four modalities for each material, all from the Materials Project database (i) the crystal structure, which we denote by $C = (\{(\mathbf{r}_i, E_i)\}_i, \{\mathbf{R}_j\}_j)$, where $\{(\mathbf{r}_i, E_i)\}_i$ is a set containing the coordinates $\mathbf{r}_i$ and chemical element E_i of the i-th atom in the unit cell, and $\{\mathbf{R}_j\}_j$ is the set of unit cell lattice vectors; (ii) the DOS, $\rho(E)$, as a function of energy E; (iii) the charge density

$n_e(\mathbf{r})$ as a function of position $\mathbf{r}$; and (iv) a textual description T of the crystal obtained from Robocrystallographer [59]. For each material modality, we train a separate neural network encoder to learn a parameterized transformation from raw data to an embedding in a shared latent space. The C encoder uses PotNet, a state-of-the-art GNN [30]; the encoders of $\rho(E)$ and $n_e(\mathbf{r})$ are based on the Transformer [60] and 3D-CNN architectures [61]. The T encoder uses a frozen MatBERT [62] model, a Bidirectional Encoder Representations from Transformers (BERT) textual model [63] that has been pre-trained on material science literature. A key advantage of the T modality is that its data collection is relatively low cost, since Robocrystallographer can be used to generate a T modality for every C; thus T can be used to obtain a much larger pre-training dataset. Conversely, T may not contain as rich information as other "high-cost" modalities like $\rho(E)$ and $n_e(\mathbf{r})$, which are usually obtained from ab initio simulations. Additional modality and architecture details are provided in the Methods section.

B. Overview of Multimodal Pre-training Methods

MultiMat adapts CLIP [40] to the materials science domain through several extensions that allow for the integration of more than two modalities. Below, we give a brief summary of CLIP and these extensions (see Methods for additional
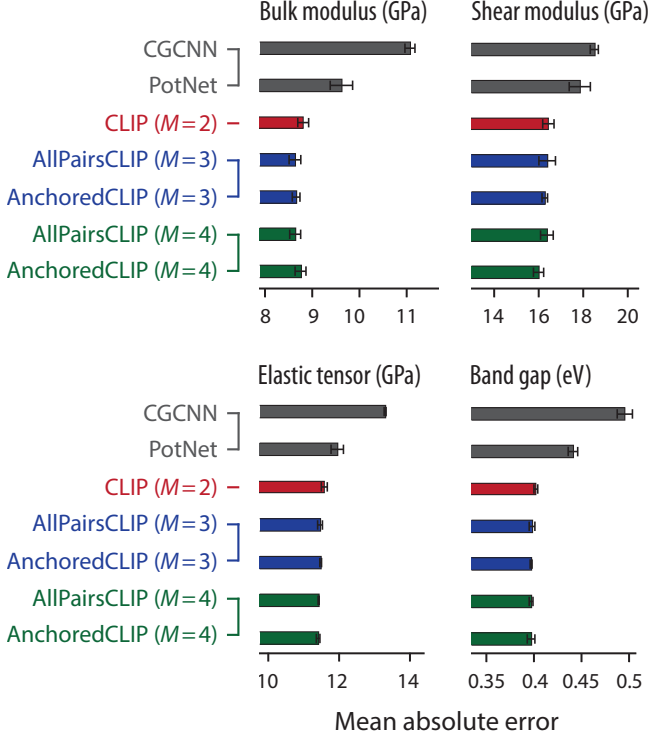
Figure 2. Crystal property prediction. Mean absolute error (MAE) for the prediction of various crystal properties across baseline methods and MultiMat. Methods are grouped by color according to the number of modalities, M, selected from the set of all modalities $\{C, \rho(E), n_e(\mathbf{r}), T\}$ (with C always selected). Results for the $M = 2$ and $M = 3$ cases show the average performance over all allowed combinations for each category (individual experiments reported in the Supplementary Information) and error bars give the standard deviation over 3 random seeds, averaged over all experiments within that category.

details):

CLIP: Applies to two modalities. We adapt CLIP to materials science by replacing the traditional image–text pairs with C paired with one other modality in $\{\rho(E), n_e(\mathbf{r}), T\}$. CLIP encourages alignment between the embeddings of a pair of modalities (via the loss term in Eq. (1a); Methods)

AllPairsCLIP: When there are more than two modalities involved, multiple pairs of modalities can be created. All-PairsCLIP includes the pairwise CLIP loss between *all* possible pairwise combinations of modalities; the loss is averaged over all such pairs.

AnchoredCLIP: Because AllPairsCLIP considers all possible pairwise combinations, the number of loss terms increases significantly with more modalities. A cheaper alternative is to only average over pairwise combinations that include C, i.e., the 'anchor', since the C encoder is arguably the most crucial for transferring to other downstream tasks (e.g., prediction tasks typically use the crystal structure as inputs).

The loss terms of both AllPairsCLIP and AnchoredCLIP are aggregates of pairwise loss terms; for n modalities, they

feature $n(n - 1)/2$ and $n - 1$ individual pairwise loss terms, respectively. In the Supplementary Information, we explore other methods that align three or more modalities without pairwise decomposition (i.e., there is only a single loss term regardless of the number of modalities).

A central advantage of pairwise alignment is the ability to exploit all available modality pairs, even when some pairs may be missing for certain database entries (since these loss terms can simply be set to zero). This is an important feature since the coverage of material databases is often incomplete: e.g., some entries may only have information for C and $\rho(E)$, others only for C and $n_e(\mathbf{r})$. A pairwise multimodal loss allows MultiMat to take advantage of a greater total amount of data than would be possible with non-pairwise methods since the information of incompletely covered entries can still be incorporated.

C. Crystal Property Prediction

After the multimodal alignment stage in MultiMat, the C encoder can be fine-tuned on various predictive tasks by attaching a randomly initialized linear head and fine-tuning end-to-end. We explore the tasks of predicting the bulk modulus, shear modulus, elastic tensor, and band gap corresponding to a crystal input. The mechanical property tasks use the Materials Project database [10] and the band gap task uses the SNUMAT semiconductor database [11]. In all tasks, the fine-tuning stage has access to only relatively few labeled data points compared to the roughly 154 000 data points used during pre-training; specifically, there are roughly 7000 data points for the bulk modulus, shear modulus, and elastic tensor tasks and roughly 10 000 data points for the band gap task.

Figure 2 compares MultiMat using multimodal pre-training with 2 – 4 modalities against baselines without multimodal pre-training. The two baselines are CGCNN [25] (a pioneering method in the use of GNNs for crystal property prediction) and PotNet [30], the current state-of-the-art method for crystal property prediction using GNNs. Note that MultiMat also uses the architecture of PotNet for the C encoder. For the two- and three-modality cases in Fig. 2, the shown results are averages of experiments over all possible two- or three-modality combinations from the set $\{C, \rho(E), n_e(\mathbf{r}), T\}$ with C always chosen. For example, for two modalities, results are the average MAE over the combinations $\{C, \rho(E)\}$, $\{C, n_e(\mathbf{r})\}$, and $\{C, T\}$ (results of individual experiments are shown in the Supplementary Information). MultiMat significantly improves predictive performance over the baselines, with up to $\sim 10\%$ reduction in MAE over the best baseline (PotNet).

We observe that including three or more modalities during pre-training improves performance over just two modalities (see Fig. 2). On the other hand, we found no significant gains in using MultiMat with $M = 4$ modalities over $M = 3$ modalities. Speculatively, this might reflect that: (i) there is only marginal additional information from the fourth modality or (ii) the current implementation is not able to take advantage of the additional information due to model capacity limitations (to ensure fair comparison, we use a fixed architecture across all experiments and do not increase the neural network capacity to match the corresponding increased complexity).
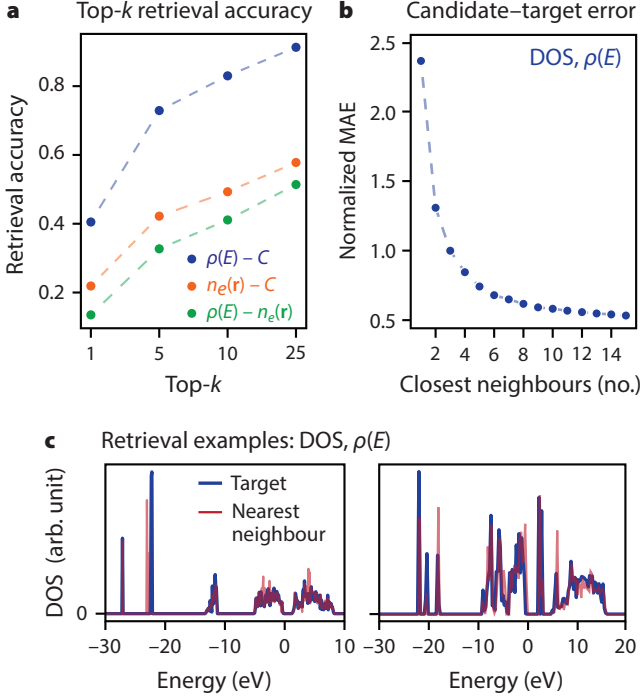
a Top-*k* retrieval accuracy
b Candidate–target error
c Retrieval examples: DOS, $\rho(E)$

Figure 3. Material discovery via latent space similarity. a, Top-*k* accuracies for cross-modality retrieval using encoders pre-trained with AnchoredCLIP, averaged over the test set. **b**, Normalized MAE between the target $\rho(E)$ from the test set and the $\rho(E)$ corresponding to the best crystal candidate from the training set, identified through our latent space similarity approach when the number of closest neighbors considered is varied. The best crystal candidate is selected from a set of crystals whose embeddings are the closest neighbors to the target $\rho(E)$ in the shared latent space, where the chosen crystal has a $\rho(E)$ with the smallest normalized MAE compared to the target $\rho(E)$. MAE values are normalized by the area of target $\rho(E)$ (both computed in the $(-5 \text{ eV}, 5 \text{ eV})$ range) and the values reported here are averaged over the whole test set. **c**, Two examples of the $\rho(E)$ corresponding to the best *C* candidate found via latent space similarity overlaid with the target $\rho(E)$ of the material discovery process.

D. Material Discovery via Latent Space Similarity

A key motivation for building fast surrogate predictive models is to enable accelerated design or identification of materials with specified properties. In this section, we demonstrate an example of how MultiMat can achieve this goal via latent-space similarity, by screening a large material database and selecting the candidate which possesses the highest similarity to the desired property. Taking the example of identifying a material with a specific "target" DOS, we proceed by: (1) embedding the target DOS using the $\rho(E)$ encoder; (2) embedding each crystal in the database of candidate materials using the *C* encoder; and (3) identifying the top-*k* crystals that maximize the (cosine) similarity between the $\rho(E)$ and *C* embeddings. Figure 3 presents results for material discovery via latent space similarity for a MultiMat model trained using AnchoredCLIP with the three modalities *C*, $\rho(E)$, and $n_e(\mathbf{r})$.

We first investigate how well the latent spaces of the different encoders are aligned since good alignment is crucial for selecting good candidate materials. To this end, we explore the cross-modality retrieval performance of the model, i.e., how often the model given a sample of a certain modality was able to retrieve the sample of another modality corresponding to the same material; the results are shown in Fig. 3a. Retrieval performance was measured on a test set containing roughly 15 000 materials (that all have *C*, $\rho(E)$, and $n_e(\mathbf{r})$ entries in the Materials Project database). "DOS–crystal" at top-*k* refers to the average accuracy (over all DOS samples in the test set) that the correct crystal structure is present within the top-*k* samples retrieved given a DOS sample in the test set. The challenge of the retrieval task depends on the size of the dataset for which retrieval is performed (in our case consisting of roughly 15 000 materials) and can be viewed as a classification task where the number of classes equals the number of samples in the dataset. Considering the size of our dataset, the strong retrieval performance demonstrates that MultiMat achieves effective alignment between the encoders of the different modalities. It is also worth noting that in AnchoredCLIP, $\rho(E)$ and $n_e(\mathbf{r})$ are never explicitly aligned (since the pairwise losses are computed only on combinations that include *C*; see Methods); "DOS–charge density" nevertheless achieves reasonably good retrieval performance.

Next, we explore how MultiMat can be used to discover materials when the desired target is not contained within the search space, by considering all DOS samples in the test set to be targets and all crystals in the train set to be potential candidates. Since a material with the exact target property does not exist in the search space, effectiveness is measured by how well the property of a selected material resembles the desired target. Figure 3b shows the error between the DOS corresponding to the best candidate material and the desired target for this task, averaged over all targets. When picking the best crystal candidate out of the *n* closest neighbors in the shared latent space, we see that the normalized mean absolute error (MAE) between the target $\rho(E)$ and the $\rho(E)$ corresponding to the best crystal structure decreases as more neighbors are considered, as expected. There are diminishing improvements in normalized MAE beyond 5 neighbours, suggesting that a consideration of approximately 5 nearest neighbours would give a reasonably good candidate for the desired property. We expect this general trend to roughly hold when scaling to larger databases with the aim of discovering new materials with suitable properties. Finally, Fig. 3c provides a visualization of two examples from our material discovery pipeline, showing a relatively good fit between the selected material and the target $\rho(E)$.

The alignment between modalities MultiMat optimizes for ensures that a close match between *C* and $\rho(E)$ embeddings in the multimodal space signifies similarity in the physical space between the candidate material corresponding to the *C* embeddings and the material corresponding to the target $\rho(E)$. This proposed material discovery approach leverages the extensive scale of *C* databases, which typically exceeds the number of entries for other modalities by at least an order of magnitude, and thus allows one to identify existing materials that would have a $\rho(E)$ very similar to the target, had it been computed. This constitutes an accelerated form of material design, which only uses inference through the neural network encoders fol-

lowed by a nearest neighbor search to find materials likely to exhibit certain desired properties. This material discovery approach is enabled by the alignment between encoders that MultiMat optimizes. In contrast "forward-only" approaches to material design are based on an encoder-decoder structure (e.g., where C is first encoded to a latent space and then decoded to predict $\rho(E)$). A potential benefit of our latent-based similarity approach lies in the fact that searching for candidates in a low-dimensional latent space compared to the physical space is likely easier for high-dimensional properties such as $\rho(E)$. A related work focusing on $\rho(E)$ was introduced in Ref. 64; however, it differs from ours by only working for binary composition materials and focusing on material design through chemical composition for a fixed atomic structure, thus neglecting the structural information of materials. Note that while our results focus on $\rho(E)$, the approach is applicable to other modalities, provided that the respective encoders are trained with MultiMat.

E. Interpretability of MultiMat Features

Finally, we explore the interpretability of the MultiMat latent space. Specifically, we use Uniform Manifold Approximation and Projection (UMAP) to transform the high-dimensional learned features from the crystal encoder into a more visualizable two-dimensional space [65], as shown in Figure 4. Note that for the results in this section, we used AnchoredCLIP trained with $\{C, \rho(E), n_e(\mathbf{r})\}$. We see that the 2D features reveal that materials with similar properties tend to be close together, and thus the features learned from MultiMat can be easily interpreted in a physically meaningful way.

In Fig. 4a, we color-code each embedding by one of the seven possible crystal systems—cubic, hexagonal, monoclinic, orthorhombic, tetragonal, triclinic, and trigonal. Each crystal system are collections of space groups that are typically similar to each other, thus demonstrating clustering by the spatial structure of the material. There is some broad color-based clustering, such as cubic crystals (red) concentrating near the top and monoclinic crystals (blue) concentrating near the bottom of the 2D plot. Additionally, some of the smaller clusters of points tend to be the same color, such as the pink cluster on the left representing a trigonal lattice.

Furthermore, we explore how the MultiMat embeddings cluster based on formation energy (a continuous property) and whether a crystal is a metal (a discrete property). Specifically, in Fig. 4b–c, we color code the dimensionality-reduced embeddings based on the value of the respective property for each material. Although the pre-trained model has never seen labels indicating a material's formation energy or whether a material is a metal, materials with similar (different) properties are still close together (far apart) in the 2D space. This suggests that the model is not merely learning random abstract features or memorizing data; it is learning features that capture information about materials' physical properties. In future work, insights derived from these features may be used to guide the search and discovery of materials with particular optical or electronic properties without the need for costly beyond-DFT methods [66, 67].

III. DISCUSSION

The incorporation of additional modalities into MultiMat improves its predictive performance. In particular, there is a big jump in performance between one and two modalities (i.e., between the baseline and MultiMat with two modalities) and a smaller jump between two and three modalities at which point the performance improvements due to incorporating additional modalities saturates. This also points to promising future research opportunities in using more than two modalities for multimodal learning in domains outside of materials.

The material property prediction tasks considered in this work have less available data than what is used for the multimodal pre-training phase. This significant difference in dataset size underscores the robust representation MultiMat develops during its pre-training phase, which likely contributes to its strong performance in crystal property prediction tasks, even with relatively limited fine-tuning data. Small datasets are of particular interest in materials science, since many open questions in the field concern specific classes of materials with few known data points [68–70]. MultiMat could potentially alleviate some problems of traditional data-driven ML methods for materials that typically require large quantities of data.

The methodological innovations introduced in this work may also have applications beyond the domain of materials science. The field of multimodal learning has so far been predominantly centered on integrating just two modalities, stemming partly from the limited methodologies capable of scaling to more than two modalities [40, 71, 72]. Prior research in the area has largely focused on working with image–text pairs scraped from the web, thereby reducing the need for multimodal methods that go beyond two modalities [46, 47]. In this work, we made use of CLIP but also introduced novel extensions for multimodal pre-training that were specifically tailored to handle more than two modalities. Furthermore, we extend our contributions by detailing two additional pre-training methods in the Supplementary Information, employing simultaneous rather than pairwise alignment of modalities, and performing on par.

An advantage of our screening approach to material discovery is the ability to constrain the search space to materials that are known to be stable. This is a practical strategy since crystalline structures for stable materials are abundant compared to other modalities that are collected via computational methods (e.g., charge density or DOS). Our material discovery approach provides a rapid solution to material discovery and mitigates the large computational costs otherwise required in traditional simulation and experimental procedures when searching over these crystal databases. The screening approach can also be extended to incorporate multiple modalities simultaneously—this multimodality conditioning could e.g., be leveraged to identify materials with desirable properties of multiple modalities simultaneously (e.g., the DOS and charge density). Future work could explore building generative models from MultiMat's latent space to harness its effective learned representations. Moreover, we expect the results to further improve if the search for candidate materials is extended to larger databases of stable materials. Consequently, this represents an interesting direction for future work in materials
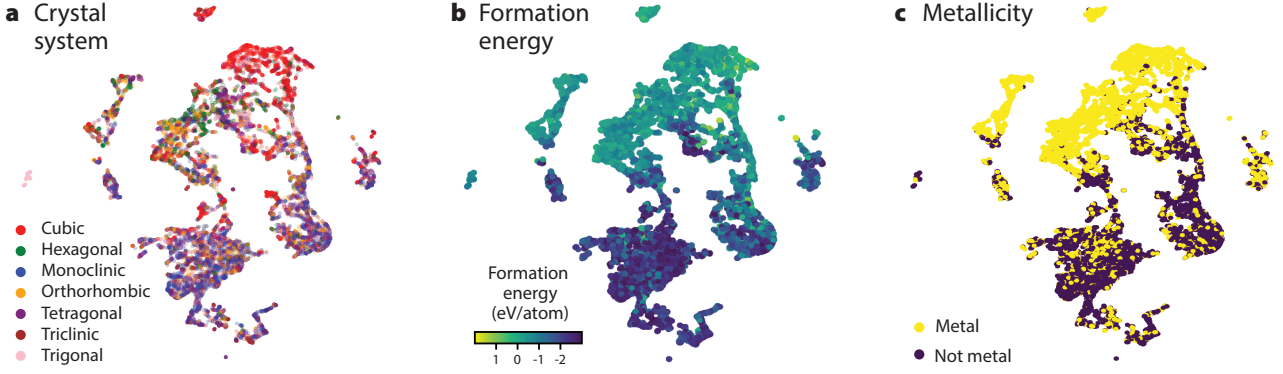
Figure 4. Interpretability of crystal embeddings. a, Crystal embeddings after dimensionality reduction by UMAP are shown, with each embedding color-coded by one of the seven crystal systems. Some clustering based on the crystal system can be observed. **b,** Visualization of these dimensionality-reduced embeddings after color-coding according to each material's formation energy. **c,** Visualization of these dimensionality-reduced embeddings after color-coding based on whether each material is a metal or not.

discovery and design. E.g., the recent GNoME database [73], consisting of 2.2 million materials predicted to be stable by ML, is particularly well-suited for this purpose. Other suitable databases include the Crystallography Open Database [74] and the Inorganic Crystal Structure Database [9], with roughly 500 000 and 280 000 entries, respectively.

IV. METHODS

A. Encoder Architectures

Here we describe the encoder architectures used for the various modalities.

Crystal structure encoder. For the C encoder, we adopted the PotNet architecture [30], the state-of-the-art for predicting properties of crystalline materials. PotNet represents the crystal structure data as a graph, where the nodes are atoms and the edges are interatomic potentials. In contrast to other methods, PotNet accounts for the complete set of interatomic potentials, enabling it to learn powerful representations of crystal structures.

Density of states (DOS) encoder. The data for each material consists of a list of energies E and the corresponding DOS $\rho(E)$. We utilized a Transformer architecture to encode $\rho(E)$ [60]. Because the energies E for which $\rho(E)$ is measured can vary between different samples in the data, we removed the positional encoding traditionally used in Transformers and instead introduced a learnable embedding layer for the energies. Specifically, we separately embedded the $\rho(E)$ values and their corresponding energies E, followed by concatenating these embeddings along the embedding dimension (thus doubling the effective embedding dimension). Subsequently, a linear layer was employed to mix the embeddings for each token. This was then followed by another layer, which down-sampled the embeddings for each token back to the original embedding dimension (i.e., the embedding dimension is halved). This adaptation allows the $\rho(E)$ encoder to adeptly handle $\rho(E)$ samples with variable energy ranges since it accounts for continuous inputs (instead of discrete) and has a notion of where a particular $\rho(E)$ lies along the energy axis.

Charge density encoder. The $n_e(\mathbf{r})$ is represented as a three-dimensional tensor corresponding to the voxelized $n_e(\mathbf{r})$ (i.e., a three-dimensional array of real numbers corresponding to the charge density per unit volume). For the $n_e(\mathbf{r})$ encoder, we utilized a 3D ResNext architecture [61] which, due to its 3D convolutions, can capture spatial patterns in all three dimensions of the three-dimensional $n_e(\mathbf{r})$ tensor.

Text encoder. The textual descriptions of the crystal structure are machine-generated by Robocrystallographer [59] and are available in the Materials Project database, similar to that used in [56]. Each crystal is described in a paragraph containing natural language and chemical symbols. For better contextual understanding (in contrast to regular text models pre-trained on the Internet), we use MatBERT [62], which has been pre-trained on a large corpus of material science literature, to generate embeddings for each textual description. MatBERT has a context window of 512 tokens; thus we truncate samples with more tokens to fit within the context window. Note that approximately 66% of the dataset has less than or equal to 512 tokens and does not require truncation. As with most classification applications of BERT [63] models, the model outputs a "CLS" token that is typically used for downstream tasks. In this work, we use the embedding of the "CLS" token output as the embedding. Its embedding dimension is 768; to align it with the embeddings from the other encoders, we use a two-layer trainable MLP to project it down to a dimension of 128. Note that during MultiMat pre-training, the MatBERT model is frozen (weights are not trained) and the only trainable parameters are those of the projection MLP.

B. Multimodal Pre-training Methods

CLIP [40] is a pre-training method that makes use of image–caption pairs from the web to build effective visual representations of input text. CLIP makes use of a contrastive loss function to pull matching image–caption pairs (pairs where

the caption corresponds to the image) closer in the embedding space whilst pushing non-matching pairs (pairs where the caption does not correspond to the image) further apart, thereby aligning the image encoder with the text encoder. Here, alignment refers to the degree to which embeddings of a matching pair of modalities are similar in the embedding space. This alignment results in effective visual representations that can be used for a variety of tasks [40, 75].

Here, we describe the methods for multimodal pre-training that we used; first, we explain how we adapted CLIP [40] to the domain of crystalline materials. After that, we describe our novel methods to handle multimodal pre-training with more than two modalities. In particular, we show how CLIP, which is limited to pre-training with two modalities, can be generalized to handle more than two modalities [40].

In the original CLIP method, we have two modalities $\mathcal{A}$ and $\mathcal{B}$, and their corresponding samples $\mathbf{A}_i$ and $\mathbf{B}_i$ for a batch of N samples (where i is the index over the batch). After the samples are encoded using the modality-specific encoders $f_{\mathcal{A}}$ and $f_{\mathcal{B}}$, the embeddings are given by $\mathbf{a}_i = f_{\mathcal{A}}(\mathbf{A}_i)$ and $\mathbf{b}_i = f_{\mathcal{B}}(\mathbf{B}_i)$. The CLIP objective connecting $\mathcal{A}$ and $\mathcal{B}$ is then given by

$$\ell(\mathcal{A}, \mathcal{B}) = -\sum_{i=1}^{N} \log \frac{e^{\text{sim}(\mathbf{a}_i, \mathbf{b}_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{a}_i, \mathbf{b}_j)/\tau}}, \qquad (1a)$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity metric and τ is the temperature parameter. In practice, the symmetric loss

$$L(\mathcal{A}, \mathcal{B}) = \frac{1}{2}[\ell(\mathcal{A}, \mathcal{B}) + \ell(\mathcal{B}, \mathcal{A})], \qquad (1b)$$

is used. CLIP was originally introduced in the context of image–caption pairs, with $\mathcal{A}$ representing an image modality and $\mathcal{B}$ a text modality.

CLIP Adapted to Materials Science. The most straightforward approach to multimodal pre-training in materials science is the direct adaptation of two-modality CLIP to materials-specific modalities. In particular, C can be seen as analogous to an image and the $\rho(E)$, $n_e(\mathbf{r})$ or T can be seen as analogous to the caption of an image in the original formulation of CLIP. This allows us to explore three distinct options for multimodal pre-training using CLIP in materials science by making use of C and $\rho(E)$, by making use of the C and $n_e(\mathbf{r})$ or by making use of C and T. Specifically, the loss functions are

$$L(C, \rho), \qquad \text{(crystal–DOS)} \qquad (2a)$$
$$L(C, n_e), \qquad \text{(crystal–charge density)} \qquad (2b)$$
$$L(C, T), \qquad \text{(crystal–text)} \qquad (2c)$$

where the loss function L is given by Eq. (1b).

AllPairsCLIP. Apart from a direct adaptation of CLIP to the materials science context, we also introduce two methods that extend the CLIP objective to accommodate and align an arbitrary number of modalities. The first of these, AllPairsCLIP, generalizes the CLIP objective to more than two modalities by aggregating the CLIP losses between all combinations of two modalities. Specifically, to incorporate all four modalities, C, $\rho(E)$, $n_e(\mathbf{r})$, and T, the AllPairsCLIP objective is computed as

$$L_{\text{AllPairsCLIP}} = \frac{1}{6}[L(C, \rho) + L(C, n_e) + L(C, T) + \\ L(\rho, n_e) + L(\rho, T) + L(n_e, T)] \qquad (3)$$

where each term in the total loss is the individual CLIP for two modalities given by Eq. (1b). A computational challenge arises from the combinatorial nature of pairwise alignments: for n modalities, the number of pairwise alignments or terms in the loss function scales as $(n^2 - n)/2$. This scaling is increasingly burdensome as n grows.

AnchoredCLIP. To address the computational drawback posed by the AllPairsCLIP method, we propose an alternative approach, also based on CLIP, which we call AnchoredCLIP. This method introduces the concept of an *anchor modality*, a core modality, rich in information, with which every other modality shares an information-overlap with. Contrary to aligning every possible pair of modalities as in AllPairsCLIP, AnchoredCLIP only aligns pairs consisting of the anchor modality and each of the other modalities. This approach significantly reduces the number of modality-pairs being aligned, i.e., terms in the loss function. Specifically, for n modalities, the number of pairs aligned is reduced to $n - 1$. In the context of materials science, when considering C, $\rho(E)$, $n_e(\mathbf{r})$, and T, we choose as anchor modality C since it constitutes a natural representation for crystalline materials that are commonly used for downstream tasks. The AnchoredCLIP objective for these modalities is then

$$L_{\text{AnchoredCLIP}} = \frac{1}{3}[L(C, \rho) + L(C, n_e) + L(C, T)], \qquad (4)$$

where both terms in the total loss objective are again given by the CLIP loss function in Eq. (1b).

Batch masking. When using three or more modalities via AllPairsCLIP and AnchoredCLIP, some samples may not have data entries for all the modalities—e.g., some samples in the batch may have data entries for all modalities C, $\rho(E)$, $n_e(\mathbf{r})$, T, while some samples may have missing entries of $\rho(E)$ or $n_e(\mathbf{r})$ (in the Materials Project database, C and T exist for all the entries). Out of 154 718 materials in the Materials Project, there are 121 915 with entries for $n_e(\mathbf{r})$, 89 071 entries with $\rho(E)$ and 78 461 entries with both $n_e(\mathbf{r})$ and $\rho(E)$. Note that T exists for all C. To take care of this during MultiMat pre-training, for each sampled batch of size B, we create a separate binary mask of dimension B for each pair of modalities to indicate the existence of their data entries for each sample in the batch. This binary mask is then used to screen and select all existing samples within the batch to compute each pair-wise loss while setting the loss terms of the missing entries to zero, thus batch-wise training can be performed as per normal.

C. Material Discovery via Latent Space Similarity and Interpretability of Embeddings

Here, we elaborate on the experimental procedures undertaken for the results pertaining to material discovery and the interpretability analysis of embeddings following multimodal pre-training. For the retrieval and material discovery experiments illustrated in Fig. 3, we utilized encoders that were pre-trained using AnchoredCLIP on three modalities of C, $\rho(E)$ and $n_e(\mathbf{r})$. We split the pre-training dataset into train–test subsets in an 80:20 ratio (resulting in approximately 62 000 and 16 000 train and test materials respectively). MultiMat

pre-training was performed on the training set and the retrieval accuracy shown in Fig. 3a was computed on the test set (i.e., consisting of samples not in the train set which was used for the multimodal pre-training). Regarding the experiments showcased in Fig. 3b-c, the target $\rho(E)$ came from the test set, again ensuring these were not part of the pre-training dataset. We then treated all materials in the train set as potential candidate materials, aiming to identify the materials being the closest neighbors for each target $\rho(E)$.

For the quantitative evaluation of the material discovery strategy shown in Fig. 3b, we compute the MAE between the target and nearest-neighbor $\rho(E)$ in the energy range from -5 eV to $+5$ eV, using linear interpolation to map the target and nearest-neighbor $\rho(E)$ onto the same equispaced energy grid. We restrict our focus to this limited range because it (i) helps to account for the varying energy ranges of different materials in the Materials Project data, obviating a need for extrapolation, and (ii) covers the energy range of primary physical interest, since most electrical and optical properties are influenced mainly by electrons near the Fermi level [76–78]. Additionally, the MAE between the target and nearest neighbor $\rho(E)$ was normalized by the area of the target $\rho(E)$ in the -5 eV to $+5$ eV range. This normalization ensures a more equitable comparison across different targets–nearest-neighbor pairs. Mathematically, we define the normalized MAE in the energy range from -5 eV to $+5$ eV by

$$\text{nMAE} = \frac{\int_{-5\,\text{eV}}^{5\,\text{eV}} |\rho_{\text{target}}(E) - \rho_{\text{nearest neighbour}}(E)|\, \mathrm{d}E}{\int_{-5\,\text{eV}}^{5\,\text{eV}} \rho_{\text{target}}(E)\, \mathrm{d}E}. \quad (5)$$

Note that this metric, despite its relative character, may still exhibit large values (e.g., exceeding unity), even for slight misalignment of the resonance energies because the DOS frequently is a sharply peaked quantity.

For the interpretability results presented in Fig. 4, we made use of materials from the same test set that was used for the retrieval and material discovery results discussed above. The embeddings of the approximately 16 000 crystal structures in the test set were transformed into a two-dimensional space through UMAP dimensionality reduction [65]. In Fig. 4b, a few of these materials were identified as outliers in terms of their formation energy and thus removed. This was done to make the color gradient easier to interpret.

D. Data

We constructed a multimodal dataset for materials science using data from the Materials Project [10], a well-established open-source initiative. This dataset included crystal structures, density of states, charge densities, and textual descriptions; those four modalities were used for multimodal pre-training. In addition to those modalities, we also made use of the bulk modulus, shear modulus, and elastic tensor data for material property prediction performance evaluation of MultiMat (after fine-tuning on those tasks) as well as for establishing non-pre-trained baselines. We also used Materials Project data for the interpretability results where we color-coded by crystal system, formation energy, and whether the material is a metal.

Despite its comprehensiveness, the Materials Project has known data quality limitations for certain material properties, e.g., for the band gaps. Specifically, the RMSE between the Materials Project band gaps (computed using DFT) and their experimentally observed counterparts is 1.05 eV, potentially affecting the efficacy and reliability of models trained on band gaps from the Materials Project [10]. To address this, we utilized the HSE gaps in the SNUMAT semiconductor database [11], which offers more accurate band gap values (RMSE of 0.36 eV relative to experimentally determined band gaps) due to using a more accurate DFT functional. We used the version of this database where materials with a computed gap of 0 eV were filtered out; note that even this filtered version of this database does contain some large gap insulators. This SNUMAT semiconductor database contains around 10 000 materials without any multimodal information. We used it to fine-tune and evaluate models pre-trained with multimodal data from the Materials Project and also to establish baselines for models without any multimodal pre-training. Note that some previous works [79, 80] have explored using ML to predict band gaps of the SNUMAT database.

E. Implementation Details and Settings for Training and Evaluation

MultiMat pre-training. We use the PotNet architecture for the C encoder, a Transformer-based architecture for the $\rho(E)$ encoder, a 3D ResNeXt architecture for the $n_e(\mathbf{r})$ encoder, and MatBERT [62] (together with a two-layer MLP) for the T encoder. Each encoder produces an embedding with dimension $d = 128$. We use the AdamW optimizer [81] for training, with a cosine-decay learning rate schedule and a linear warm-up schedule of 10 epochs. The peak learning rate is fixed at 10^{-4} and weight-decay is fixed at 5×10^{-4}. We use a batch size of 360 across all pre-training experiments and perform pre-training for a total of 500 epochs.

Fine-tuning for prediction tasks. After pre-training, the C encoder is transferred, and a linear head is randomly initialized. The model was then fine-tuned for various material property prediction tasks. We use the AdamW optimizer with a cosine-decay learning rate schedule and linear warm-up with 10 epochs. We use a batch size of 120, no weight-decay, and the peak learning rate was swept over $\{10^{-3}, 10^{-4}, 10^{-5}\}$. From the downstream data entries available for the specific prediction task, we create a train, validation, and test split in the ratio of $60 : 20 : 20$. The pre-trained C encoder was fine-tuned on the training set and early stopping was performed based on the lowest validation error on the validation set. The best checkpoint (i.e., with the lowest validation loss) was then used to evaluate on the test set. Error bars were created by taking the standard deviation from three different experiments with different seeds.

Material discovery via latent space similarity. For the results in Fig. 3, we used a slightly smaller batch size of 100 for MultiMat pre-training as we observed that this resulted in slightly better performance.

[1] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Big data of materials science: critical role of the descriptor, Physical Review Letters **114**, 105503 (2015).

[2] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Computational Materials **2** (2016).

[3] W. Sun, C. J. Bartel, E. Arca, S. R. Bauers, B. Matthews, B. Orvañanos, B.-R. Chen, M. F. Toney, L. T. Schelhas, W. Tumas, *et al.*, A map of the inorganic ternary metal nitrides, Nature Materials **18**, 732 (2019).

[4] V. L. Deringer, N. Bernstein, G. Csányi, C. B. Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, and S. R. Elliott, Origins of structural and electronic transitions in disordered silicon, Nature **589**, 59 (2021).

[5] M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, *et al.*, Accelerated discovery of CO_2 electrocatalysts using active machine learning, Nature **581**, 178 (2020).

[6] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, Nature **559**, 547 (2018).

[7] J. Damewood, J. Karaguesian, J. R. Lunger, A. R. Tan, M. Xie, J. Peng, and R. Gómez-Bombarelli, Representations of materials for machine learning, Annual Review of Materials Research **53** (2023).

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).

[9] M. Hellenbrandt, The inorganic crystal structure database (ICSD)—present and future, Crystallography Reviews **10**, 17 (2004).

[10] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials **1**, 011002 (2013).

[11] S. Kim, M. Lee, C. Hong, Y. Yoon, H. An, D. Lee, W. Jeong, D. Yoo, Y. Kang, Y. Youn, and S. Han, A band-gap database for semiconducting inorganic materials calculated with hybrid functional, Scientific Data **7** (2020).

[12] F. Tang, H. C. Po, A. Vishwanath, and X. Wan, Comprehensive search for topological materials using symmetry indicators, Nature **566**, 486 (2019).

[13] T. Zhang, Y. Jiang, Z. Song, H. Huang, Y. He, Z. Fang, H. Weng, and C. Fang, Catalogue of topological electronic materials, Nature **566**, 475 (2019).

[14] M. Vergniory, L. Elcoro, C. Felser, N. Regnault, B. A. Bernevig, and Z. Wang, A complete catalogue of high-quality topological materials, Nature **566**, 480 (2019).

[15] G. R. Schleder, A. C. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, From DFT to machine learning: recent approaches to materials science—a review, Journal of Physics: Materials **2**, 032001 (2019).

[16] S. Axelrod, D. Schwalbe-Koda, S. Mohapatra, J. Damewood, K. P. Greenman, and R. Gómez-Bombarelli, Learning matter: Materials design with machine learning and atomistic simulations, Accounts of Materials Research **3**, 343 (2022).

[17] B. Huang, G. F. von Rudorff, and O. A. von Lilienfeld, The central role of density functional theory in the AI age, Science **381**, 170 (2023).

[18] J. E. Saal, A. O. Oliynyk, and B. Meredig, Machine learning in materials discovery: Confirmed predictions and their underlying approaches, Annual Review of Materials Research **50**, 49 (2020).

[19] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, *et al.*, Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, Nature Materials **15**, 1120 (2016).

[20] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, and J. Wang, Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning, Nature Communications **9**, 3405 (2018).

[21] A. Ma, Y. Zhang, T. Christensen, H. C. Po, L. Jing, L. Fu, and M. Soljačić, Topogivity: A machine-learned chemical rule for discovering topological materials, Nano Letters **23**, 772 (2023).

[22] A. S. Fuhr and B. G. Sumpter, Deep generative models for materials discovery and machine learning-accelerated innovation, Frontiers in Materials **9**, 865270 (2022).

[23] D. M. Anstine and O. Isayev, Generative models as an emerging paradigm in the chemical sciences, Journal of the American Chemical Society **145**, 8736 (2023).

[24] Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr, *et al.*, Inverse design of nanoporous crystalline reticular materials with deep generative models, Nature Machine Intelligence **3**, 76 (2021).

[25] T. Xie and J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material

properties, Physical Review Letters **120**, 145301 (2018).

[26] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, Schnet–a deep learning architecture for molecules and materials, The Journal of Chemical Physics **148** (2018).

[27] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chemistry of Materials **31**, 3564 (2019).

[28] K. Choudhary and B. DeCost, Atomistic line graph neural network for improved materials property predictions, npj Computational Materials **7**, 185 (2021).

[29] K. Yan, Y. Liu, Y. Lin, and S. Ji, Periodic graph transformers for crystal material property prediction, Advances in Neural Information Processing Systems **35**, 15066 (2022).

[30] Y. Lin, K. Yan, Y. Luo, Y. Liu, X. Qian, and S. Ji, Efficient approximations of complete interatomic potentials for crystal property prediction, in *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 202, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR, 2023) pp. 21260–21287.

[31] F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, Interpretable and explainable machine learning for materials science and chemistry, Accounts of Materials Research **3**, 597 (2022).

[32] A. E. Allen and A. Tkatchenko, Machine learning of material properties: Predictive and interpretable multilinear models, Science advances **8**, eabm7185 (2022).

[33] A. Y.-T. Wang, M. S. Mahmoud, M. Czasny, and A. Gurlo, CrabNet for explainable deep learning in materials science: bridging the gap between academia and industry, Integrating Materials and Manufacturing Innovation **11**, 41 (2022).

[34] C. J. Hargreaves, M. S. Dyer, M. W. Gaultois, V. A. Kurlin, and M. J. Rosseinsky, The earth mover's distance as a metric for the space of inorganic compositions, Chemistry of Materials **32**, 10610 (2020).

[35] X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, and T. Y.-J. Han, Explainable machine learning in materials science, npj Computational Materials **8**, 204 (2022).

[36] E. S. Muckley, J. E. Saal, B. Meredig, C. S. Roper, and J. H. Martin, Interpretable models for extrapolation in scientific machine learning, Digital Discovery **2**, 1425 (2023).

[37] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, On the opportunities and risks of foundation models, arXiv:2108.07258 (2021).

[38] OpenAI: J. Achiam *et al.*, GPT-4 technical report, arXiv:2303.08774 (2023).

[39] Team Gemini: A. Rohan *et al.*, Gemini: a family of highly capable multimodal models, arXiv:2312.11805 (2023).

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, Learning transferable visual models from natural language supervision (2021), arXiv:2103.00020 [cs.CV].

[41] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, Sigmoid loss for language image pre-training (2023), arXiv:2303.15343 [cs.CV].

[42] J. Li, D. Li, C. Xiong, and S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation (2022), arXiv:2201.12086 [cs.CV].

[43] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, and J. Gao, RegionCLIP: Region-based language-image pretraining (2021), arXiv:2112.09106 [cs.CV].

[44] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, CLIP-adapter: Better vision-language models with feature adapters (2021), arXiv:2110.04544 [cs.CV].

[45] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, Hierarchical text-conditional image generation with CLIP latents (2022), arXiv:2204.06125 [cs.CV].

[46] W. Kim, B. Son, and I. Kim, ViLT: Vision-and-language transformer without convolution or region supervision (2021), arXiv:2102.03334 [stat.ML].

[47] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, SimVLM: Simple visual language model pretraining with weak supervision (2022), arXiv:2108.10904 [cs.CV].

[48] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, CoCa: Contrastive captioners are image-text foundation models (2022), arXiv:2205.01917 [cs.CV].

[49] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang, Florence: A new foundation model for computer vision (2021), arXiv:2111.11432 [cs.CV].

[50] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, Imagebind: One embedding space to bind them all (2023), arXiv:2305.05665 [cs.CV].

[51] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese, ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding (2023), arXiv:2212.05171 [cs.CV].

[52] A. Guzhov, F. Raue, J. Hees, and A. Dengel, Audio-CLIP: Extending CLIP to image, text and audio (2021), arXiv:2106.13043 [cs.SD].

[53] M. Y. Toriyama, A. M. Ganose, M. Dylla, S. Anand, J. Park, M. K. Brod, J. M. Munro, K. A. Persson, A. Jain, and G. J. Snyder, How to analyse a density of states, Materials Today Electronics **1**, 100002 (2022).

[54] N. Lee, H. Noh, S. Kim, D. Hyun, G. S. Na, and C. Park, Density of states prediction of crystalline materials via prompt-guided multi-modal transformer (2023), arXiv:2311.12856 [cond-mat.mtrl-sci].

[55] L. H. Dos Santos, Applications of charge-density analysis to the rational design of molecular materials: A mini review on how to engineer optical or magnetic crystals, Journal of Molecular Structure **1203**, 127431 (2020).

[56] A. N. Rubungo, C. Arnold, B. P. Rand, and A. B. Dieng, LLM-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions, arXiv:2310.14029 (2023).

[57] S. Takeda, I. Priyadarsini, A. Kishimoto, H. Shinohara, L. Hamada, H. Masataka, J. Fuchiwaki, and D. Nakano, Multi-modal foundation model for material design, in *AI for Accelerated Materials Design-NeurIPS 2023 Workshop* (2023).

[58] T. Prein, E. Pan, T. Doerr, E. Olivetti, and J. L. Rupp, MTEN-CODER: A multi-task pretrained transformer encoder for materials representation learning, in *AI for Accelerated Materials Design-NeurIPS 2023 Workshop* (2023).

[59] A. M. Ganose and A. Jain, Robocrystallographer: automated crystal structure text descriptions and analysis, MRS Communications **9**, 10.1557/mrc.2019.94 (2019).

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need (2023), arXiv:1706.03762 [cs.CL].

[61] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, Aggregated residual transformations for deep neural networks (2017), arXiv:1611.05431 [cs.CV].

[62] N. Walker, A. Trewartha, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. Persson, G. Ceder, and A. Jain, The impact of domain-specific pre-training on named entity recognition tasks in materials science, Available at SSRN 3950755 (2021).

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language

understanding (2019), arXiv:1810.04805 [cs.CL].

[64] K. Bang, J. Kim, D. Hong, D. Kim, and S. S. Han, Inverse design for materials discovery from the multidimensional electronic density of states, Journal of Materials Chemistry A (2024).

[65] L. McInnes, J. Healy, and J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction (2020), arXiv:1802.03426 [stat.ML].

[66] N. R. Knøsgaard and K. S. Thygesen, Representing individual electronic states for machine learning GW band structures of 2D materials, Nature Communications **13**, 468 (2022).

[67] J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen, and S. G. Louie, BerkeleyGW: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures, Computer Physics Communications **183**, 1269 (2012).

[68] Y. Zhang and C. Ling, A strategy to apply machine learning to small datasets in materials science, npj Computational Materials **4**, 25 (2018).

[69] P. Xu, X. Ji, M. Li, and W. Lu, Small data machine learning in materials science, npj Computational Materials **9**, 42 (2023).

[70] B. Weng, Z. Song, R. Zhu, Q. Yan, Q. Sun, C. G. Grice, Y. Yan, and W.-J. Yin, Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts, Nature communications **11**, 3513 (2020).

[71] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, Align before fuse: Vision and language representation learning with momentum distillation (2021), arXiv:2107.07651 [cs.CV].

[72] S. Pramanick, L. Jing, S. Nag, J. Zhu, H. Shah, Y. LeCun, and R. Chellappa, VoLTA: Vision-language transformer with weakly-supervised local-feature alignment (2023), arXiv:2210.04135 [cs.CV].

[73] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Scaling deep learning for materials discovery, Nature , 1 (2023).

[74] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail, Crystallography open database (COD): an open-access collection of crystal structures and platform for world-wide collaboration, Nucleic Acids Research **40**, D420 (2012).

[75] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations (2020), arXiv:2002.05709 [cs.LG].

[76] G. D. Mahan, *Many-particle physics* (Springer Science & Business Media, 2000).

[77] G. Grosso and G. P. Parravicini, *Solid state physics* (Academic press, 2013).

[78] S. Kong, F. Ricci, D. Guevarra, J. B. Neaton, C. P. Gomes, and J. M. Gregoire, Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings, Nature Communications **13**, 949 (2022).

[79] T. Wang, X. Tan, Y. Wei, and H. Jin, Accurate bandgap predictions of solids assisted by machine learning, Materials Today Communications **29**, 102932 (2021).

[80] H. Choubisa, P. Todorović, J. M. Pina, D. H. Parmar, Z. Li, O. Voznyy, I. Tamblyn, and E. H. Sargent, Interpretable discovery of semiconductors with machine learning, npj Computational Materials **9**, 117 (2023).

[81] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, arXiv:1711.05101 (2018).