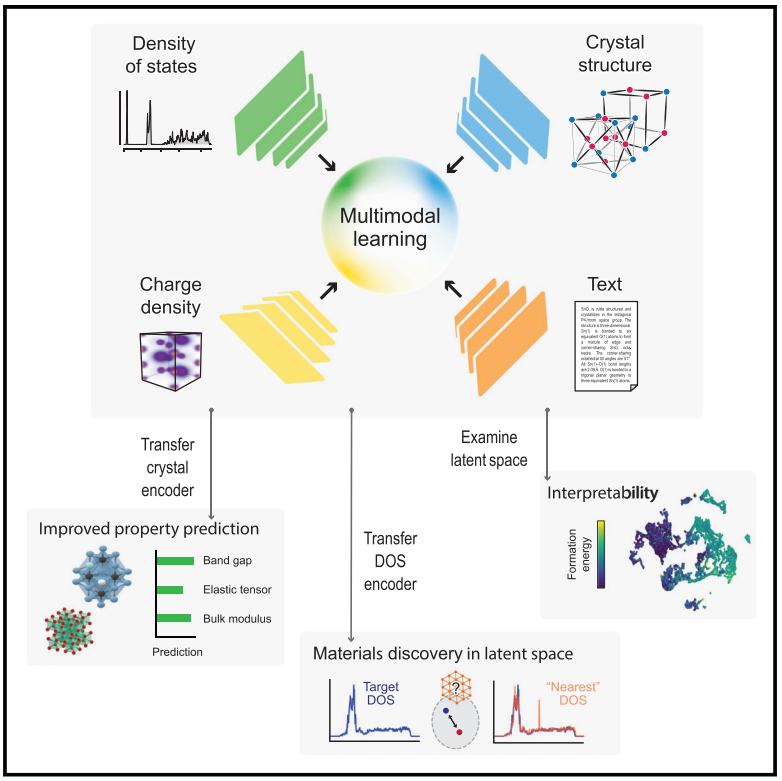
Newton

Article

Multimodal foundation models for material property prediction and discovery

Graphical abstract



Highlights

- Proposed a framework for training multimodal foundation models for materials
- Achieved state-of-the-art performance for material property prediction tasks
- Enabled novel and accurate material discovery via latent-space similarity
- Extended multimodal pre-training to handle more than two modalities

Authors

Viggo Moro, Charlotte Loh, Rumen Dangovski, ..., Peter Y. Lu, Thomas Christensen, Marin Soljačić

Correspondence

vmoro@mit.edu (V.M.), soljacic@mit.edu (M.S.)

In brief

Recently, machine-learning-based methods have been proposed to speed up the search for novel materials with specific properties. However, existing methods are typically highly task specific and struggle to utilize the rich diversity of material data. Moro et al. propose MultiMat, a framework for training general-purpose machine-learning models (foundation models) on large amounts of diverse material data. The utility and potential of the MultiMat framework is demonstrated on the tasks of material property prediction as well as material discovery.

Moro et al., 2025, Newton *1*, 100016
March 3, 2025
https://doi.org/10.1016/j.newton.2025.100016

CellPress

Newton

CellPress
OPEN ACCESS

Article

Multimodal foundation models for material property prediction and discovery

Viggo Moro,[1,6,7,*] Charlotte Loh,[2,6] Rumen Dangovski,[2,6] 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, Cambridge, MA 02139, USA
[2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Research and Exploratory Development, John Hopkins University Applied Physics Laboratory, Laurel, MD 20723, USA
[4]Data Science Institute, University of Chicago, Chicago, IL 60637, USA
[5]Department of Electrical and Photonics Engineering, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
[6]These authors contributed equally
[7]Lead contact
*Correspondence: vmoro@mit.edu (V.M.), soljacic@mit.edu (M.S.)
https://doi.org/10.1016/j.newton.2025.100016

ACCESSIBLE OVERVIEW Given the vast combinatorial search space of possible configurations, searching for new materials with specific properties has traditionally been a computationally intensive and time-consuming task. Recently, machine-learning-based approaches have been proposed to speed up this search. However, existing solutions are typically highly task specific and struggle to utilize the rich diversity of material information available when training the machine-learning models. To address these challenges, this work proposes MultiMat, a framework for training general-purpose machine-learning models (foundation models). Specifically, MultiMat is pre-trained on large amounts of diverse material data and can then be fine-tuned on various, more specific, tasks. The utility and potential of the MultiMat framework is demonstrated on the tasks of material property prediction (used to search for desirable materials during high-throughput screening) as well as direct material discovery. Additionally, it is shown that the MultiMat features correlate well with material properties, indicating that the model has effectively captured material information and could potentially provide new scientific insights in the future.

SUMMARY

Artificial intelligence is transforming computational materials science by improving property prediction and accelerating the discovery of novel materials. Recently, publicly available material data repositories have grown rapidly, encompassing 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 multimodal data available. Here, we introduce multimodal learning for materials (MultiMat), a framework enabling self-supervised multimodal training of foundation models for materials. Using the Materials Project database, we demonstrate the potential of MultiMat by: (1) achieving state-of-the-art performance for challenging material property prediction tasks; (2) enabling novel and accurate material discovery via latent-space similarity, allowing screening for stable materials with desired properties; and (3) encoding emergent features that correlate with material properties and may provide novel scientific insights.

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 speed-up, 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 with 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 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: for example, besides the crystal structure, the density of states (DOS)[53,54] and charge density[55] convey rich information about materials. Textual descriptions of the crystal, which can be machine generated,[56] offer a fourth modality that is additionally computationally cheap to acquire. It is important to point out that the aforementioned material modalities are not information independent, since they can be computed from the crystal structure. The same holds true for image-caption pairs that were used in CLIP. Therefore, the point of contrastive multimodal pre-training is not to leverage modalities with independent information but rather to learn better representations by integrating different perspectives of the same underlying data.[57–60] 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[61] and a structure-agnostic multitask learning approach for crystals.[62]

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 Figure 1A. This alignment process produces shared latent spaces and effective material representations, which can then be leveraged for a series of downstream tasks (Figures 1B–1D). 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 how the MultiMat approach enables the features to be understood through correlations with material properties by exploring the latent space from MultiMat using a dimensionality-reduction approach.

RESULTS

Modalities and architectures

To illustrate the MultiMat framework, we consider four modalities for each material, all from the Materials Project database: (1) 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 ith atom in the unit cell, and $\{\mathbf{R}_j\}_j$ is the set of unit cell lattice vectors; (2) the DOS, $\rho(E)$, as a function of energy E; (3) the charge density $n_e(\mathbf{r})$ as a function of position $\mathbf{r}$; and (4) a textual description T of the crystal obtained from Robocrystallographer.[63] 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[64] and 3D-CNN architectures.[65] The T encoder uses a frozen MatBERT[66] model, a bidirectional encoder representations from Transformers (BERT) textual model[67] that has been pre-trained on materials science literature. A key advantage of the T modality is that its data collection is relatively inexpensive, 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 information as rich as that in other "high-cost" modalities such as $\rho(E)$ and $n_e(\mathbf{r})$, which are usually obtained from *ab initio* simulations. Additional modality and architecture details are provided in the methods section.

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 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 Equation 1a; see methods).

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

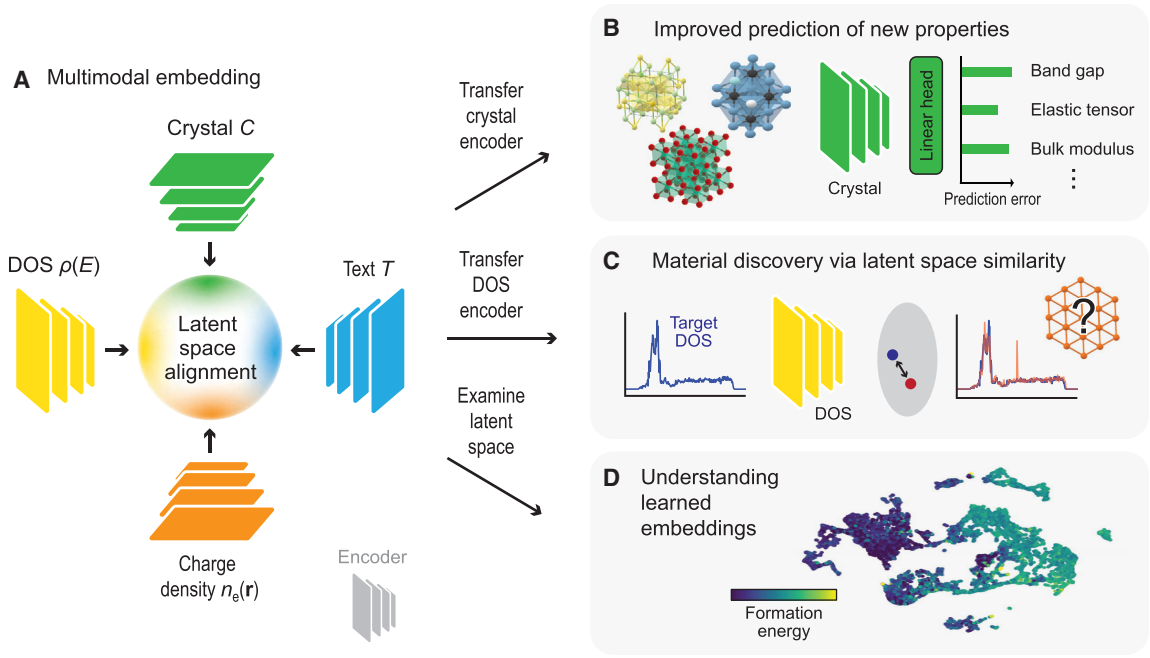


Figure 1. The multimodal learning for materials (MultiMat) approach

(A) Crystal (C), density of states (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 the understanding of its features based on correlations with material properties. 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.

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 Table S1 and supplemental methods, 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: for example, 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.

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] These tasks were chosen because they have relatively few labeled data points compared to the number of data points used during pre-training. In particular, roughly 154,000 data points are used during pre-training compared to roughly 7,000 data points for the bulk modulus, shear modulus, and elastic tensor tasks and roughly 10,000 data points for the band-gap task. Note that for the crystal property prediction tasks, only the crystal structure is used and not any of the other modalities used for multimodal pre-training.

Figure 2 compares MultiMat using multimodal pre-training with 2–4 modalities against baselines without multimodal pre-training. The two baselines are CGCNN,[25] the first method using

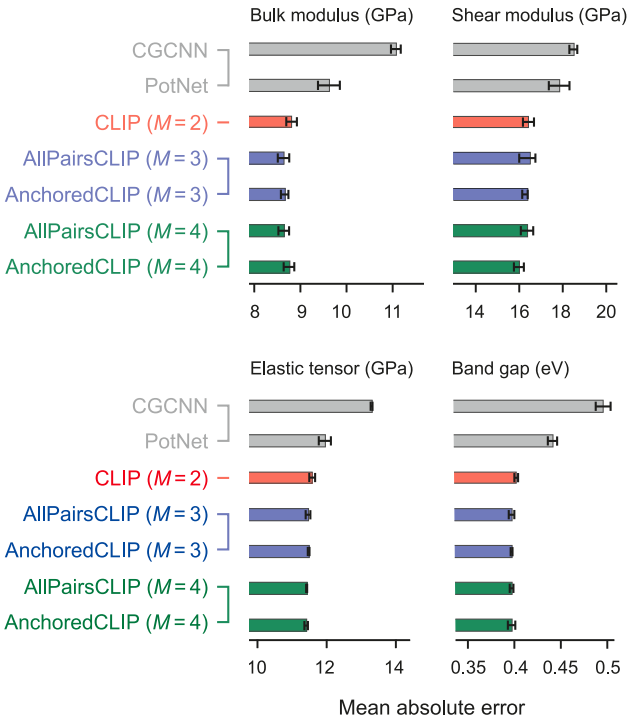


Figure 2. Crystal property prediction

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 are reported in Table 1), and error bars give the standard deviation over three random seeds, averaged over all experiments within that category.

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 Figure 2, the results shown 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 (since we are interested in predicting material properties form the crystal structure). For example, for two modalities, results are the average mean absolute error (MAE) over the combinations $\{C, \rho(E)\}$, $\{C, n_e(\mathbf{r})\}$, and $\{C, T\}$ (results of individual experiments are shown in Table 1). MultiMat pre-training significantly improves predictive performance compared to the baselines that do not make use of any pre-training. In particular, MultiMat reduces the MAE by up to $\sim 10\%$ compared to PotNet, which is the current state-of-the-art and the C architecture used in MultiMat. This performance improvement is comparable to the improvement when going from CGCNN to PotNet, methods that are separated by 5 years that represent the first method and state-of-the-art method for crystal property prediction, respectively. Moreover, note that MultiMat is a pre-training method that can be used on top of any existing or future crystal encoder to substantially improve its performance. Thus, the primary focus is on the performance difference between PotNet and MultiMat, with the CGCNN baseline serving to contextualize the overall performance advancements.

We observe that including three or more modalities during pre-training marginally improves performance over just two modalities (Figure 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 (1) the fourth modality offers only a marginally different perspective on the material compared to the other three modalities or (2) the current implementation is not able to take advantage of the additional modality 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).

In Table 1, we show the individual experiments used to compute the averages for $M = 2$ and $M = 3$ in Figure 2. Non-crystal combinations are omitted from Table 1, as we are interested in predicting material properties from the crystal structure, necessitating the inclusion of the crystal structure in MultiMat. This is the standard setting for material property prediction, since the crystal structure is always readily available for materials. In contrast, other modalities need to be computed from the crystal structure. Furthermore, the crystal structure contains all information necessary to predict any material property, since the many-body Hamiltonian is uniquely determined by the crystal structure.[68]

By analyzing the individual experiments in Table 1, we can learn which combinations of modalities tend to perform better and worse, respectively. In particular, for $M = 2$, the combination $\{C, \rho(E)\}$ performs comparatively or slightly better than the other modality combinations (which themselves perform roughly the same). However, the best-performing combination is still task dependent (i.e., it depends on the material property being predicted). For $M = 3$, AnchoredCLIP with $\{C, \rho(E) \quad n_e(\mathbf{r})\}$ performs comparatively or somewhat better than the other combinations (although the modality combination performing the best varies for the different material properties). For AllPairsCLIP there is no combination of modalities that tend to work better than others across all material properties, also underlining that the best combination of modalities varies across the material properties.

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 that 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 $n_e(\mathbf{r})$.

Table 1. Detailed results for crystal property prediction

		Bulk modulus	Shear modulus	Elastic tensor	Band gap (SNUMAT)
Without MultiMat					
CGCNN		11.069 ± 0.106	18.508 ± 0.182	13.299 ± 0.027	0.495 ± 0.008
PotNet		9.618 ± 0.237	17.845 ± 0.475	11.947 ± 0.175	0.441 ± 0.005
Two-modality MultiMat					
Crystal, DOS		8.757 ± 0.050	16.143 ± 0.171	11.538 ± 0.089	0.390 ± 0.000
Crystal, charge density		8.932 ± 0.164	16.466 ± 0.221	11.619 ± 0.050	0.407 ± 0.002
Crystal, text		8.738 ± 0.133	16.694 ± 0.372	11.588 ± 0.113	0.409 ± 0.003
Three-modality MultiMat					
Crystal, DOS, charge density	AllPairsCLIP	8.652 ± 0.110	16.253 ± 0.239	11.476 ± 0.104	0.391 ± 0.002
Crystal, DOS, text	AllPairsCLIP	8.708 ± 0.095	16.571 ± 0.520	11.525 ± 0.048	0.397 ± 0.002
Crystal, charge density, text	AllPairsCLIP	8.541 ± 0.169	16.318 ± 0.358	11.382 ± 0.052	0.405 ± 0.004
Crystal, DOS, charge density	AnchoredCLIP	8.531 ± 0.097	16.045 ± 0.123	11.424 ± 0.013	0.396 ± 0.002
Crystal, DOS, text	AnchoredCLIP	8.631 ± 0.058	16.218 ± 0.108	11.543 ± 0.041	0.396 ± 0.000
Crystal, charge density, text	AnchoredCLIP	8.814 ± 0.091	16.555 ± 0.144	11.469 ± 0.021	0.399 ± 0.001
Four-modality MultiMat					
Crystal, DOS, charge density, text	AllPairsCLIP	8.639 ± 0.116	16.368 ± 0.285	11.421 ± 0.015	0.397 ± 0.002
Crystal, DOS, charge density, text	AnchoredCLIP	8.760 ± 0.115	15.994 ± 0.229	11.407 ± 0.051	0.397 ± 0.004

Results of individual experiments for crystal property prediction used to compute the averages presented in Figure 2. Prediction error is measured in MAE and standard deviation is shown over three random seeds.

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 Figure 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 training 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 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 five neighbors, suggesting that a consideration of approximately five nearest neighbors 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, Figure 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 suggests 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 followed 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 optimized by MultiMat. 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

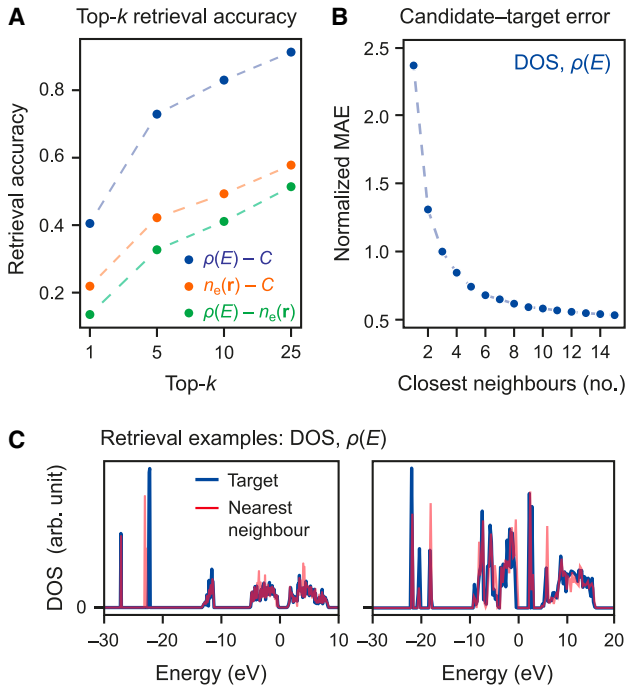


Figure 3. Material discovery via latent-space similarity

(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 $(-5$ eV to 5 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.

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 Bang et al.[69]; 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.

By design, the cosine similarity guides the contrastive loss, and its use for retrieval in the latent space has shown promising results for representation learning of images and text.[40,58] However, it is important to point out that the choice of cosine similarity and the sampling of the pre-training materials can be improved, as materials exist on high-dimensional, non-linear manifolds with large curvatures. Likewise, the alignment of different modalities in MultiMat could potentially benefit from a more suitable distance metric. We leave this for future work.

Understanding MultiMat features through correlations with material properties

Finally, we explore how the MultiMat latent space can be understood through correlations with material properties. Specifically, we use uniform manifold approximation and projection (UMAP) to transform the high-dimensional learned features from the crystal encoder (after MultiMat but before fine-tuning on property prediction tasks) into a more visualizable 2D space,[70] as shown in Figure 4. For the results in this section, we used AnchoredCLIP trained with $\{C, \rho(E), n_e(\mathbf{r})\}$. The 2D features reveal that materials with similar properties tend to be close together, suggesting that the features learned from MultiMat can be understood based on physical properties. Again, it is important to point out that the use of the cosine similarity as a distance metric for UMAP is likely not optimal. What is more, the dimensionality-reduced embeddings might be discontinuous. Nevertheless, we still observe a clear correlation between the dimensionality-reduced embeddings and various physical properties.

In Figure 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 comprises 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 orthorhombic crystals (yellow) concentrating in the upper left part 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. However, it is difficult to clearly show this many discrete properties in one plot, so the embeddings of different crystal systems are also shown separately in Figure S1.

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 Figures 4B and 4C, 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 still tend to be close together (far apart) in the 2D space. This is particularly the case when color coding based on the formation energy. When color coding based on whether a material is a metal or not, the clustering is not as apparent, although metals still tend to be concentrated more toward the top and non-metals more toward the bottom. 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 methods beyond density functional theory (DFT).[71,72]

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

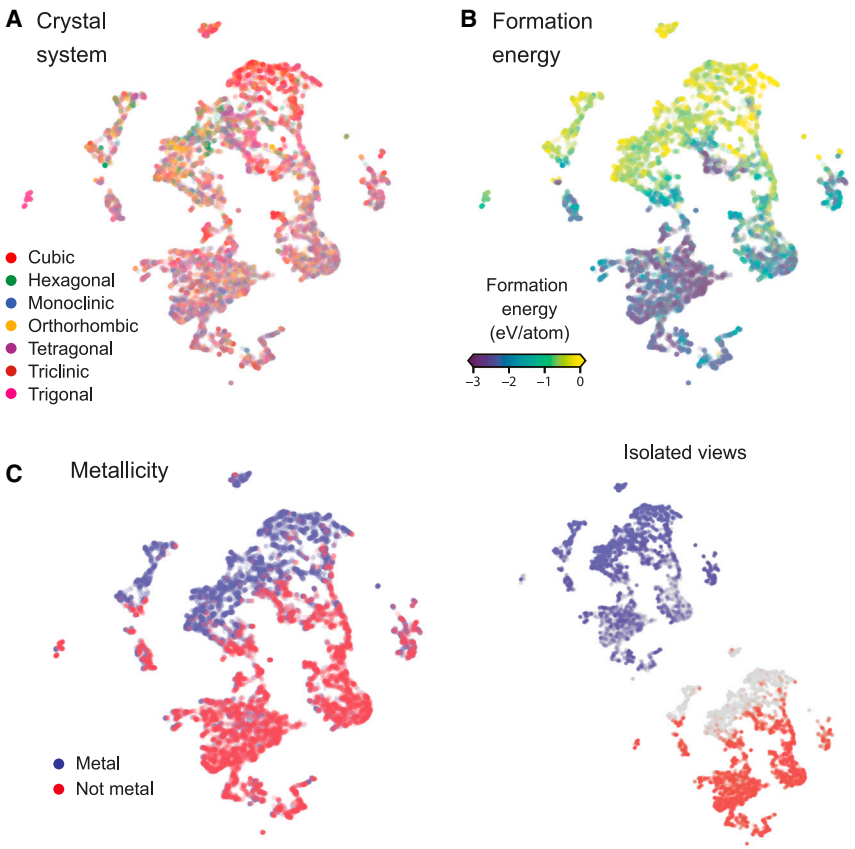


Figure 4. Understanding the crystal embeddings after MultiMat

(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. The metal and non-metal embeddings are also shown separately.

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 saturate. 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 fewer available data than is used for the multimodal pre-training phase. This significant difference in dataset size underscores the robust representation developed by MultiMat 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.[73–75] 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.[40,76,77] 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 (see Table S1 and supplemental methods) 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, for example, 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 discovery and design. For example, the recent GNoME database,[78] 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[79] and the Inorganic Crystal Structure Database,[9] with roughly 500,000 and 280,000 entries, respectively.

METHODS

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 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 $\rho(E)$.[64] 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 downsampled 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 (instead of discrete) inputs 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 (3D) 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[65] which, due to its 3D convolutions, can capture spatial patterns in all three dimensions of the 3D $n_e(\mathbf{r})$ tensor.

Text encoder

The textual descriptions of the crystal structure are machine generated by Robocrystallographer[63] and are available in the Materials Project database, similar to that used in Rubungo et al.[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,[66] 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[67] 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.

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 while 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.[40,58]

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. Thereafter, 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

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

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

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

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

$$L(C,\rho),(\text{crystal}-\text{DOS}) \qquad \text{(Equation 2a)}$$

$$L(C,n_e),(\text{crystal}-\text{charge density}) \qquad \text{(Equation 2b)}$$

$$L(C,T),(\text{crystal}-\text{text}) \qquad \text{(Equation 2c)}$$

where the loss function L is given by Equation 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

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

where each term in the total loss is the individual CLIP for two modalities given by Equation 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. 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

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

where both terms in the total loss objective are again given by the CLIP loss function in Equation 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 pairwise loss while setting the loss terms of the missing entries to zero; thus, batchwise training can be performed as per normal.

Material discovery via latent-space similarity and understanding the MultiMat embeddings

Here, we elaborate on the experimental procedures undertaken for the results pertaining to material discovery and the understanding of embeddings following multimodal pre-training. For the retrieval and material-discovery experiments illustrated in Figure 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 Figure 3A was computed on the test set (i.e., consisting of samples not in the training set, which was used for the multimodal pre-training). Regarding the experiments showcased in Figures 3B and 3C, 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 training 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 Figure 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 (1) helps to account for the varying energy ranges of different materials in the Materials Project data, obviating a need for extrapolation, and (2) covers the energy range of primary physical interest, since most electrical and optical properties are influenced mainly by electrons near the Fermi level.[68,80,81] 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

$$\text{nMAE} = \frac{\int_{-5\text{ eV}}^{5\text{ eV}} \left|\rho_{\text{target}}(E) - \rho_{\text{nearest neighbour}}(E)\right| \, dE}{\int_{-5\text{ eV}}^{5\text{ eV}} \rho_{\text{target}}(E) \, dE} \quad \text{(Equation 5)}$$

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

For the results presented in Figure 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 2D space through UMAP dimensionality reduction.[70] In Figure 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.

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, DOS, charge densities, and textual descriptions; those four modalities were used for multimodal pre-training. In addition to those modalities, we 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 visualization results, whereby we color coded the UMAP embeddings 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 root-mean-square error (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 Heyd-Scuseria-Ernzerhof (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[82,83] have explored using ML to predict band gaps of the SNUMAT database.

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[66] (together with a two-layer MLP) for the T encoder. Each encoder produces an embedding with dimension $d = 128$. We use the AdamW optimizer[84] 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 is 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 with no weight decay, and the peak learning rate is 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 Figure 3, we used a slightly smaller batch size of 100 for MultiMat pre-training, as we observed that this resulted in slightly better performance.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Viggo Moro (vmoro@mit.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This article uses data that is all open-sourced and publicly available. Specifically, the data were downloaded from the Materials Project (https://next-gen.materialsproject.org/) and SNUMAT (https://www.snumat.com/) databases.
- MultiMat was developed using the PyTorch framework. All source code used for training and generating the results is publicly available at https://github.com/vmoro1/multimat.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank Sean Mann, Michael Huang, Donato Jimenez Beneto, Di Luo, Owen Dugan, Li Jing, Jasper Snoek, and Jamie Smith for fruitful discussions. This research was sponsored in part by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This material is also based upon work sponsored in part by the US Army DEVCOM ARL Army Research Office through the MIT Institute for Soldier Nanotechnologies under Cooperative Agreement number W911NF-23-2-0121 and in part by the Air Force Office of Scientific Research under the award number FA9550-21-1-0317. T.C. acknowledges the support of a research grant (project no. 42106) from Villum Fonden. C.L. received support from DSO National Laboratories - Singapore. A.M. received support from the National Science Foundation Graduate Research Fellowship under grant no. 1745302. P.Y.L. gratefully acknowledges the support of the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Sciences program.

AUTHOR CONTRIBUTIONS

V.M., C.L., R.D., and M.S. conceived the research. V.M., C.L., and R.D. designed and implemented the framework and performed the experiments. A.G., A.M., Z.C., S.K., P.Y.L., and T.C. contributed to the development of the framework. T.C. and M.S. supervised the research. V.M., C.L., and R.D. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

M.S. is an advisory board member for *Newton*.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, we used ChatGPT for minor polishing of the writing. After using this tool, we reviewed and edited the content as needed and take full responsibility for the content of the published article.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.newton.2025.100016.

Received: September 6, 2024
Revised: November 1, 2024
Accepted: January 21, 2025
Published: February 13, 2025

REFERENCES

1. Ghiringhelli, L.M., Vybiral, J., Levchenko, S.V., Draxl, C., and Scheffler, M. (2015). Big data of materials science: critical role of the descriptor. Phys. Rev. Lett. *114*, 105503.
2. Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput. Mater. *2*, 16028. https://doi.org/10.1038/npjcompumats.2016.28.
3. Sun, W., Bartel, C.J., Arca, E., Bauers, S.R., Matthews, B., Orvañanos, B., Chen, B.R., Toney, M.F., Schelhas, L.T., Tumas, W., et al. (2019). A map of the inorganic ternary metal nitrides. Nat. Mater. *18*, 732–739.
4. Deringer, V.L., Bernstein, N., Csányi, G., Ben Mahmoud, C., Ceriotti, M., Wilson, M., Drabold, D.A., and Elliott, S.R. (2021). Origins of structural and electronic transitions in disordered silicon. Nature *589*, 59–64. https://doi.org/10.1038/s41586-020-03072-z.
5. Zhong, M., Tran, K., Min, Y., Wang, C., Wang, Z., Dinh, C.T., De Luna, P., Yu, Z., Rasouli, A.S., Brodersen, P., et al. (2020). Accelerated discovery of CO_2 electrocatalysts using active machine learning. Nature *581*, 178–183. https://doi.org/10.1038/s41586-020-2242-8.
6. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. Nature *559*, 547–555.
7. Damewood, J., Karaguesian, J., Lunger, J.R., Tan, A.R., Xie, M., Peng, J., and Gómez-Bombarelli, R. (2023). Representations of materials for machine learning. Annu. Rev. Mater. Res. *53*, 399–426.
8. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (MIT press).
9. Hellenbrandt, M. (2004). The inorganic crystal structure database (ICSD)—present and future. Crystallogr. Rev. *10*, 17–22.
10. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. Apl. Mater. *1*, 011002. https://doi.org/10.1063/1.4812323.
11. Kim, S., Lee, M., Hong, C., Yoon, Y., An, H., Lee, D., Jeong, W., Yoo, D., Kang, Y., Youn, Y., and Han, S. (2020). A band-gap database for semiconducting inorganic materials calculated with hybrid functional. Sci. Data *7*, 387. https://doi.org/10.1038/s41597-020-00723-8.
12. Tang, F., Po, H.C., Vishwanath, A., and Wan, X. (2019). Comprehensive search for topological materials using symmetry indicators. Nature *566*, 486–489. https://doi.org/10.1038/s41586-019-0937-5.
13. Zhang, T., Jiang, Y., Song, Z., Huang, H., He, Y., Fang, Z., Weng, H., and Fang, C. (2019). Catalogue of topological electronic materials. Nature *566*, 475–479. https://doi.org/10.1038/s41586-019-0944-6.
14. Vergniory, M.G., Elcoro, L., Felser, C., Regnault, N., Bernevig, B.A., and Wang, Z. (2019). A complete catalogue of high-quality topological materials. Nature *566*, 480–485. https://doi.org/10.1038/s41586-019-0954-4.
15. Schleder, G.R., Padilha, A.C.M., Acosta, C.M., Costa, M., and Fazzio, A. (2019). From DFT to machine learning: recent approaches to materials science—a review. J. Phys. Mater. *2*, 032001.
16. Axelrod, S., Schwalbe-Koda, D., Mohapatra, S., Damewood, J., Greenman, K.P., and Gómez-Bombarelli, R. (2022). Learning matter: Materials design with machine learning and atomistic simulations. Acc. Mater. Res. *3*, 343–357.
17. Huang, B., von Rudorff, G.F., and von Lilienfeld, O.A. (2023). The central role of density functional theory in the AI age. Science *381*, 170–175.
18. Saal, J.E., Oliynyk, A.O., and Meredig, B. (2020). Machine learning in materials discovery: Confirmed predictions and their underlying approaches. Annu. Rev. Mater. Res. *50*, 49–69.
19. Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T.D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M.A., Chae, H.S., Einzinger, M., Ha, D.G., Wu, T., et al. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat. Mater. *15*, 1120–1127.
20. Lu, S., Zhou, Q., Ouyang, Y., Guo, Y., Li, Q., and Wang, J. (2018). Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. Nat. Commun. *9*, 3405.
21. Ma, A., Zhang, Y., Christensen, T., Po, H.C., Jing, L., Fu, L., and Soljačić, M. (2023). Topogivity: A machine-learned chemical rule for discovering topological materials. Nano Lett. *23*, 772–778. https://doi.org/10.1021/acs.nanolett.2c03307.
22. Fuhr, A.S., and Sumpter, B.G. (2022). Deep generative models for materials discovery and machine learning-accelerated innovation. Front. Mater. *9*, 865270.
23. Anstine, D.M., and Isayev, O. (2023). Generative models as an emerging paradigm in the chemical sciences. J. Am. Chem. Soc. *145*, 8736–8750.
24. Yao, Z., Sánchez-Lengeling, B., Bobbitt, N.S., Bucior, B.J., Kumar, S.G.H., Collins, S.P., Burns, T., Woo, T.K., Farha, O.K., Snurr, R.Q., and Aspuru-Guzik, A. (2021). Inverse design of nanoporous crystalline reticular materials with deep generative models. Nat. Mach. Intell. *3*, 76–86. https://doi.org/10.1038/s42256-020-00271-1.
25. Xie, T., and Grossman, J.C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett. *120*, 145301. https://doi.org/10.1103/physrevlett.120.145301.
26. Schütt, K.T., Sauceda, H.E., Kindermans, P.J., Tkatchenko, A., and Müller, K.R. (2018). Schnet–a deep learning architecture for molecules and materials. J. Chem. Phys. *148*.
27. Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S.P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. Chem. Mater. *31*, 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294.
28. Choudhary, K., and DeCost, B. (2021). Atomistic line graph neural network for improved materials property predictions. npj Comput. Mater. *7*, 185. https://doi.org/10.1038/s41524-021-00650-1.
29. Yan, K., Liu, Y., Lin, Y., and Ji, S. (2022). Periodic graph transformers for crystal material property prediction. Adv. Neural Inf. Process. Syst. *35*, 15066–15080.
30. Lin, Y., Yan, K., Luo, Y., Liu, Y., Qian, X., and Ji, S. (2023). Efficient approximations of complete interatomic potentials for crystal property prediction. In Proceedings of the 40th International Conference on Machine Learning, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds. (PMLR), pp. 21260–21287. https://proceedings.mlr.press/v202/lin23m.html.

31. Oviedo, F., Ferres, J.L., Buonassisi, T., and Butler, K.T. (2022). Interpretable and explainable machine learning for materials science and chemistry. Acc. Mater. Res. *3*, 597–607.

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

33. Wang, A.Y.T., Mahmoud, M.S., Czasny, M., and Gurlo, A. (2022). CrabNet for explainable deep learning in materials science: bridging the gap between academia and industry. Integr. Mater. Manuf. Innov. *11*, 41–56.

34. Hargreaves, C.J., Dyer, M.S., Gaultois, M.W., Kurlin, V.A., and Rosseinsky, M.J. (2020). The earth mover's distance as a metric for the space of inorganic compositions. Chem. Mater. *32*, 10610–10620.

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

36. Muckley, E.S., Saal, J.E., Meredig, B., Roper, C.S., and Martin, J.H. (2023). Interpretable models for extrapolation in scientific machine learning. Digital Discovery *2*, 1425–1435.

37. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2108.07258.

38. OpenAI; Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., and Altman, S. (2023). GPT-4 technical report. Preprint at arXiv. https://arxiv.org/abs/2303.08774.

39. Team Gemini; Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., and Silver, D. (2023). Gemini: a family of highly capable multimodal models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.11805.

40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (PMLR), pp. 8748–8763.

41. Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (IEEE), pp. 11975–11986. https://doi.org/10.1109/ICCV51070.2023.01100.

42. Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (PMLR), pp. 12888–12900.

43. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., and Gao, J. (2022). Regionclip: Region-based language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 16793–16803. https://doi.org/10.1109/CVPR52688.2022.01629.

44. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. (2024). Clip-adapter: Better vision-language models with feature adapters. Int. J. Comput. Vis. *132*, 581–595.

45. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv. https://doi.org/10.48550/arXiv.2204.06125.

46. Kim, W., Son, B., and Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In International conference on machine learning (PMLR), pp. 5583–5594.

47. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., and Cao, Y. (2022). SimVLM: Simple visual language model pretraining with weak supervision. In International Conference on Learning Representations. URL: https://openreview.net/forum?id=GUrhfTuf_3

48. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022). Coca: Contrastive Captioners Are Image-Text Foundation Models (Transactions on Machine Learning Research). https://openreview.net/forum?id=Ee277P3AYC.

49. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. (2021). Florence: A new foundation model for computer vision. Preprint at arXiv. https://doi.org/10.48550/arXiv.2111.11432.

50. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., and Misra, I. (2023). Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457.

51. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., and Savarese, S. (2023). ULIP: Learning a Unified Representation of Language, Images, and Point Clouds for 3D Understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 1179–1189. https://doi.org/10.1109/CVPR52729.2023.00120.

52. Guzhov, A., Raue, F., Hees, J., and Dengel, A. (2022). Audioclip: Extending clip to image, text and audio. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), pp. 976–980. https://doi.org/10.1109/ICASSP43922.2022.9747631.

53. Toriyama, M.Y., Ganose, A.M., Dylla, M., Anand, S., Park, J., Brod, M.K., Munro, J.M., Persson, K.A., Jain, A., and Snyder, G.J. (2022). How to analyse a density of states. Materials Today Electronics *1*, 100002. https://doi.org/10.1016/j.mtelec.2022.100002.

54. Lee, N., Noh, H., Kim, S., Hyun, D., Na, G.S., and Park, C. (2024). Density of states prediction of crystalline materials via prompt-guided multi-modal transformer. In Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 2694. https://doi.org/10.5555/3666122.3668816.

55. Dos Santos, L.H. (2020). Applications of charge-density analysis to the rational design of molecular materials: A mini review on how to engineer optical or magnetic crystals. J. Mol. Struct. *1203*, 127431. https://doi.org/10.1016/j.molstruc.2019.127431.

56. Rubungo, A.N., Arnold, C., Rand, B.P., and Dieng, A.B. (2023). LLM-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.14029.

57. Wang, T., and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Proceedings of the 37th International Conference on Machine Learning, H.D. III and A. Singh, eds. (PMLR), pp. 9929–9939. https://proceedings.mlr.press/v119/wang20k.html.

58. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International conference on machine learning (PMLR), pp. 1597–1607.

59. van den Oord, A., Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1807.03748.

60. Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., and Vogt, J.E. (2023). Identifiability results for multimodal contrastive learning. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=U_2kuqoTcB.

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

62. Prein, T., Pan, E., Doerr, T., Olivetti, E., and Rupp, J.L. (2023). MTENCODER: A multi-task pretrained transformer encoder for materials representation learning. In AI for Accelerated Materials Design-NeurIPS 2023 Workshop.

63. Ganose, A.M., and Jain, A. (2019). Robocrystallographer: automated crystal structure text descriptions and analysis. MRS Communications *9*, 874–881. https://doi.org/10.1557/mrc.2019.94.

64. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Preprint at arXiv. https://doi.org/10.48550/arXiv.1706.03762.

65. Xie, S., Girshick, R., Dollar, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE). https://doi.org/10.1109/CVPR.2017.634.

66. Walker, N., Trewartha, A., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K., Ceder, G., and Jain, A. (2021). The impact of domain-specific pre-training on named entity recognition tasks in materials science. SRN Electron. J. Available at SSRN 3950755. https://doi.org/10.2139/ssrn.3950755.

67. Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.

68. Mahan, G.D. (2000). Many-Particle Physics (Springer Science & Business Media).

69. Bang, K., Kim, J., Hong, D., Kim, D., and Han, S.S. (2024). Inverse design for materials discovery from the multidimensional electronic density of states. J. Mater. Chem. A *12*, 6004–6013.

70. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. J. Open Source Softw. *3*, 861. https://doi.org/10.21105/joss.00861.

71. Knøsgaard, N.R., and Thygesen, K.S. (2022). Representing individual electronic states for machine learning GW band structures of 2D materials. Nat. Commun. *13*, 468.

72. Deslippe, J., Samsonidze, G., Strubbe, D.A., Jain, M., Cohen, M.L., and Louie, S.G. (2012). BerkeleyGW: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures. Comput. Phys. Commun. *183*, 1269–1289.

73. Zhang, Y., and Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. npj Comput. Mater. *4*, 25.

74. Xu, P., Ji, X., Li, M., and Lu, W. (2023). Small data machine learning in materials science. npj Comput. Mater. *9*, 42.

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

76. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S.C.H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Adv. Neural Inf. Process. Syst. *34*, 9694–9705.

77. Pramanick, S., Jing, L., Nag, S., Zhu, J., Shah, H.J., LeCun, Y., and Chellappa, R. (2023). VoLTA: Vision-language transformer with weakly-supervised local-feature alignment. Transactions on Machine Learning Research. https://openreview.net/forum?id=Kt2VJrCKo4.

78. Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G., and Cubuk, E.D. (2023). Scaling deep learning for materials discovery. Nature *624*, 80–85.

79. Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirs, M., Serebryanaya, N.R., Moeck, P., Downs, R.T., and Le Bail, A. (2012). Crystallography open database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Res. *40*, D420–D427. https://doi.org/10.1093/nar/gkr900.

80. Grosso, G., and Parravicini, G.P. (2013). Solid State Physics (Academic press).

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

82. Wang, T., Tan, X., Wei, Y., and Jin, H. (2021). Accurate bandgap predictions of solids assisted by machine learning. Mater. Today Commun. *29*, 102932.

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

84. Loshchilov, I., and Hutter, F. (2019). Decoupled weight decay regularization. In International Conference on Learning Representations. https://openreview.net/forum?id=Bkg6RiCqY7.

NEWTON, Volume *1*

Supplemental information

Multimodal foundation models for material property prediction and discovery

Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Samuel Kim, Peter Y. Lu, Thomas Christensen, and Marin Soljačić

Supplemental Items

Note S1. Isolated Views of Dimensionality-Reduced Embeddings for Different Crystal Systems

In Figure 4a, the dimensionality-reduced crystal embeddings where color-coded based on their crystal systems. However, it can be hard to clearly color-code based on that many discrete properties in one panel. Therefore, Figure S1 shows the embeddings corresponding to different crystal systems separately. This makes it easier to understand the full extent of the embeddings corresponding to a specific crystal system, as they are no longer obstructed by embeddings corresponding to other crystal systems. We still observe some clustering based on the crystal system. For example, cubic crystal (red) are concentrated more towards the top and monoclinic crystals (blue) are concentrated more towards the bottom.

Figure S1. Isolated views of dimensionality-deduced embeddings of different crystal systems in Figure 4a. The embeddings corresponding to different crystal systems are shown separately.

Note S2. Results for Non-pairwise Multimodal Pre-training Methods

The multimodal alignment methods in the main text, i.e., AllPairsCLIP and AnchoredCLIP, focus on aggregating pairwise alignments. Here, we explore the performance of two novel multimodal pre-training methods that align multiple modalities (three or more) without pairwise decomposition (i.e., there is only a single loss term). In particular, for $M = 3$ modalities, we explore two methods we term TensorCLIP and 3D BarlowTwins; the former is an extension of the infoNCE[1] to a 3×3 similarity tensor instead of a 2×2 similarity matrix and the latter is a 3-dimensional extension of the self-supervised learning method BarlowTwins.[2] Both methods are described in detail in the supplementary methods, see Novel Non-Pairwise Multimodal Pre-training Methods. While these methods can be straightforwardly extended beyond three modalities, in this section, we only explore three modalities alignment due to computational constraints.

For our experiments, we use Crystal, DOS, and charge density as the three modalities during multimodal pre-training. A disadvantage of the two multimodal pre-training methods described above is that the similarity tensors and correlation tensors are computed over three modalities simultaneously and, without further engineering efforts, does not take care of the scenario of missing modalities (unlike methods based on pairwise decomposition where loss terms for the missing modalities can simply be dropped). While in principle this can be resolved, e.g., by applying a mask over the missing entries in the various tensors, in this section we simplify the comparison of the different multimodal methods by considering only the intersection of the Materials Project data where all three modality entries (i.e., crystal, DOS, charge density) are present. This amounts to 78 641 crystals available for multi-modal alignment.

Results are shown in Table S1, where TensorCLIP and 3D BarlowTwins show comparable, but slightly worse performance compared to the pairwise methods presented in the main text (in particular for the band gap prediction task). The slight degradation of performance could be due to poor optimization; due to the increase in computational challenges, the hyperparameters involved in these methods were not thoroughly studied. Research in multimodal learning thus far has predominantly focused on dealing with two-modality alignment with few works looking beyond two modalities. These results point to promising future research directions to explore these methods more deeply.

Table S1. Prediction performance of novel non-pairwise methods. We use only the 78 641 samples that have entries for all three modalities (crystal, DOS, and charge density) for multimodal pre-training. Prediction error is measured in MAE and standard deviation is shown over 3 random seeds.

	Bulk modulus	Shear modulus	Elastic tensor	Band gap (SNUMAT)
Pairwise methods				
AllPairsCLIP	8.834 ±0.081	16.179 ±0.298	11.550 ±0.074	0.401 ±0.002
AnchoredCLIP	8.945 ±0.096	16.178 ±0.147	11.612 ±0.078	0.398 ±0.003
Novel non-pairwise methods				
TensorCLIP	8.829 ±0.071	16.432 ±0.192	11.636 ±0.085	0.424 ±0.001
3D BarlowTwins	8.914 ±0.041	16.537 ±0.217	11.588 ±0.044	0.435 ±0.002

Supplemental Methods

Novel Non-Pairwise Multimodal Pre-training Methods

Here, we detail the non-pairwise multimodal pre-training methods from Note S2. Results for Non-pairwise Multimodal Pre-training Methods.

TensorCLIP TensorCLIP extends the original CLIP objective to three or more modalities. In the case of three modalities, we compute a three-dimensional similarity matrix whose entries

are the three-way dot product of the normalized embeddings. For a batch B, the infoNCE[1] loss contrasts over B^2 terms instead of B terms as in CLIP. TensorCLIP's objective is;

$$L_{\text{TensorCLIP}} = (L_{\mathcal{M}_1,(\cdot)} + L_{\mathcal{M}_2,(\cdot)} + L_{\mathcal{M}_3,(\cdot)})/3, \tag{S1}$$

with

$$L_{\mathcal{M}_1,(\cdot)} = -\sum_{i=1}^{N} \log \frac{e^{\text{sim}(\mathbf{a}_i,\mathbf{b}_i,\mathbf{c}_i)/\tau}}{\sum_{j,k} e^{\text{sim}(\mathbf{a}_i,\mathbf{b}_j,\mathbf{c}_k)/\tau}}, \tag{S2}$$

where $\text{sim}(\mathbf{a},\mathbf{b},\mathbf{c}) = \frac{\sum_{l=1}^{d} a_l b_l c_l}{\sqrt{\sum_{l=1}^{d} a_l^2}\sqrt{\sum_{l=1}^{d} b_l^2}\sqrt{\sum_{l=1}^{d} c_l^2}}$ is the generalized three-way dot product between the three embedding vectors, each with dimension d. This operation can be efficiently computed using the `einsum` package in PyTorch.[3]

3D BarlowTwins BarlowTwins[2] is originally developed for self-supervised learning in computer vision. It aims to create embeddings invariant to distortions applied to different batches of images, while simultaneously reducing redundancy between various features of these embeddings. This was achieved by encouraging the cross-correlation matrix of the embeddings to be close to the identity matrix.

To adapt the BarlowTwins approach for multimodal pre-training, we shift the focus from embeddings of two batches of distorted images to embeddings derived from various modalities. This transition requires an extension of both the loss function and the cross-correlation matrix to handle embeddings from more than two modalities. Specifically, we adapt the method to handle three modalities, though it is worth noting that it can be easily extended to n modalities in the same way we extend it from two to three modalities. The loss function for 3D BarlowTwins for three modalities is given by

$$\begin{aligned}\mathcal{L} = &\sum_{ijk \text{ s.t. } i=j=k} (1 - C_{ijk})^2 + \sum_{\substack{ijk \text{ s.t.} \\ i=j\neq k \\ \vee\, i=k\neq j \\ \vee\, j=k\neq i}} (\tfrac{1}{2} - C_{ijk})^2 \\ &+ \lambda \sum_{ijk \text{ s.t. } i\neq j\neq k} C_{ijk}^2.\end{aligned} \tag{S3}$$

In Eq. (S3), C denotes the generalized cross-correlation matrix for three modalities which is given by

$$C_{ijk} = \frac{\sum_b z_{bi}^{M_1} z_{bj}^{M_2} z_{bk}^{M_3}}{\sqrt{\sum_b (z_{bi}^{M_1})^2}\sqrt{\sum_b (z_{bj}^{M_2})^2}\sqrt{\sum_b (z_{bk}^{M_3})^2}}, \tag{S4}$$

where M_l denotes the l-th modality from which the embeddings are derived (e.g., from the crystal encoder, the DOS encoder, or the charge density encoder) and $z_{bi}^{M_l}$ denotes the i-th feature of the embedding vector from the l-th modality from the b-th sample in the batch. Additionally, the embeddings are assumed to be mean-centered along the batch dimension and λ is a hyperparameter that balances the relative influence of the terms.

In Eq. (S3), the first term is designed to foster similarity or correlation of corresponding features across different modalities. By *corresponding features*, we refer to features from different modalities with the same index in the modality embedding vectors. The second term aims to encourage a moderate level of similarity or correlation for cases where two out of the three features correspond to each other across modalities. Lastly, the third term promotes dissimilarity or decorrelation among different features across modalities. The last term can also be interpreted as minimizing the redundancy between the features across modalities.

Computational challenges While direct alignment via TensorCLIP and 3D BarlowTwins is intuitively appealing, it presents greater computational challenges. Specifically, with a batch size

of B, an embedding dimension of D, and n modalities, the n-dimensional tensor in TensorCLIP contains B^n entries, while for 3D BarlowTwins, it contains D^n entries. Therefore, the loss computation scales roughly exponential with the increase in number of modalities, while the pairwise methods presented in the main text would scale roughly as a polynomial (by increasing the number of loss terms needed).

SUPPLEMENTAL REFERENCES

1. van den Oord, A., Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding. . arXiv:1807.03748.
2. Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning. PMLR pp. 12310–12320.
3. (2023). `torch.einsum`, PyTorch documentation. https://pytorch.org/docs/stable/generated/torch.einsum.html. . Accessed: 2023-11-30.