nature communications

Article

https://doi.org/10.1038/s41467-026-76378-7

Scaling multimodal materials representation learning with synthetic narratives

Received: 15 April 2025

Accepted: 27 July 2026

Published online: 26 August 2026

Check for updates

Yang Jeong Park[1,2,3], Mayank Kumaran[4], Chia-Wei Hsu[2], Elsa Olivetti[2,5] & Ju Li[1,2,5]

Artificial intelligence is increasingly applied to inverse materials design, including crystals and molecules. For molecular systems, multimodal models that integrate chemical structures with natural language have enabled flexible, instruction-driven reasoning. However, approaches for crystalline materials remain limited by the scarcity and imbalanced bias of available crystal datasets and the weak semantic supervision provided by peer-reviewed literature. We introduce a contrastive language-crystals model (CLaC), a multimodal contrastive learning framework that aligns crystal structures with natural language, pre-trained on property-conditioned synthetic narratives of crystal structure-text pairs generated by language models. Here we show that, compared with models trained on literature-derived text, this property-grounded synthetic supervision enables semantic alignment unachievable with existing academic corpora. Embedding-space analyses reveal structured associations between language and crystal structures. These results establish a foundation for zero-shot, instruction-driven reasoning in crystalline materials.

The search for materials that exceed existing performance limits is a fundamental objective of materials science. Over the past decade, advances in computing and artificial intelligence (AI) have reshaped this pursuit, with machine learning (ML) accelerating tasks ranging from virtual screening[1,2] and property prediction[3–5] to generative modeling[6–12] and active learning[13,14]. These developments have been driven by models that treat crystals as 3D graphs, enabling neural networks to capture structural features with minimal manual feature engineering.

While structure-centric approaches have driven substantial progress, the recent emergence of large language models (LLMs) suggests the advent of a fifth research paradigm in materials science, often referred to as intelligent science[15–17]. In this paradigm, AI systems integrate human knowledge, symbolic reasoning, and data-driven patterns. Yet, the potential of such systems remains constrained by their limited sensory modalities. Text-only LLMs, for example, exhibit only rudimentary visual abilities and lack a fundamental understanding of visual geometry. Self-supervised representation learning provides a promising pathway to overcome these limitations. In particular, contrastive learning has proven effective at bridging different modalities, as exemplified by CLIP[18], which forms the foundation of many contemporary generative and vision-language models[19–22]. Vision-language models demonstrate how multimodality can expand perceptual capability; extending this idea to materials science raises the prospect of enabling AI systems to perceive chemical and crystal spaces in ways that go beyond human intuition.

Early efforts toward multimodal representations of crystalline materials reflect the growing interest in language-driven materials modelling. Previous studies have investigated the alignment of crystal structures with electronic descriptors and textual annotations[23], the translation of atomic structures into material strings for large language model (LLM) based inference[24], and the generation of crystal structures from textual or compositional constraints[25,26]. Together, these works demonstrate the feasibility of

[1]Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Materials Science and Engineering, Ulsan National Institute of Science and Technology, Ulju-gun, Ulsan, Republic of Korea. [4]Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA. [5]MIT-IBM Watson AI Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. e-mail: elsao@mit.edu; liju@mit.edu

incorporating natural language into materials representation and generation.

Despite this progress, existing approaches exhibit several fundamental limitations. In most cases, textual information is used to encode low-level structural attributes, including elemental composition, coordination environments, and interatomic distances, rather than to capture higher-level, human-centric knowledge related to materials properties, functional behavior, or potential applications. In addition, text-based representations alone struggle to preserve crystallographic symmetry and geometric equivariances with high fidelity. As a result, generated structures often remain close to known prototypes and display limited structural diversity[27]. These observations underscore the need for a unified representation that combines precise, equivariant structural encodings with rich natural language descriptions grounded in materials knowledge.

Despite this promise, multimodal learning for crystalline materials remains comparatively underexplored. Two major challenges contribute to this gap. First, the availability of crystal data is limited. Vision-language models are typically trained on hundreds of millions of image-text pairs, and molecular research benefits from access to billions of molecular structures[28]. In contrast, existing crystal databases contain only millions of entries[29,30]. Second, the distribution of crystal data is heavily biased, reflecting sociological tendencies that concentrate research on a small set of highly studied, so-called hot materials, limiting linguistic diversity and masking the vast combinatorial landscape permitted by first-principles thermodynamics. Most existing models are trained directly on such biased and limited datasets, and therefore inherit these constraints without explicitly addressing them.

Here we introduce the Contrastive Language-Crystal model (CLaC), a multimodal framework that aligns 3D atomic structures with rich linguistic descriptions. To address data scarcity and distributional bias, we adopt a synthetic narrative generation strategy inspired by Park et al.[31], in which textual descriptions are generated by grounding language models in computed material properties. This property-conditioned approach enables the synthesis of diverse and informative text that extends beyond structural templates to include functional behavior, synthesis considerations, and potential applications. Crystal structures are encoded using a graph neural network architecture that preserves geometric invariance, and contrastive learning is used to align structural and textual representations within a shared latent space.

We evaluate CLaC with a particular focus on its capacity for zero-shot multimodal understanding and zero-shot inverse design, defined here as the ability to retrieve crystal structures from property-guided natural language queries. Analysis of the multimodal latent space shows that models trained on physically grounded synthetic narratives learn a structured alignment between linguistic descriptions and geometric representations. By contrast, models trained solely on raw literature text fail to exhibit comparable semantic organization, underscoring the importance of synthetic data quality rather than scale alone. Together, these results demonstrate the feasibility of language-guided materials search and position CLaC as a step toward instruction-based materials discovery.

Results

Contrastive language-crystals pre-training model

The core of this research approach is the idea of learning the perception of the chemical space composed of crystalline substances through supervised learning embedded in synthetic narrative data. Synthetic narrative learning offers several potential advantages over natural language learning. Compared to the challenges of collecting and curating natural language data, synthetic narrative learning offers significantly greater scalability. This is because it can leverage various computational chemistry techniques, the vast material databases available online, and the inference capabilities of language models to expand the data. Furthermore, it offers the significant advantage of linking material properties and language, enabling flexible zero-shot transfer learning, rather than simply learning vocabulary.

In this section, we introduce the overall data pipeline, model architecture and the pre-training process as illustrated in Fig. 1. We designed a foundation model for a comprehensive understanding of both natural language and crystal structures. The model consists of two neural network branches, a crystal graph encoder and a text encoder, which encodes the graphs and texts into a joint representation space, respectively. Although many successful multimodal learning strategies have emerged since CLIP, research on applying these strategies remains a future work. To pre-train our model for a multimodal alignment, we use GPT-synthesized narratives[31] for the Materials Project as well as academic paper paragraphs. These narratives are generated based on publicly available materials databases, leveraging the reasoning power of LLMs (Fig. 1a). As an example, we present a generated narrative for a $KNiIO_6$ crystal (Suppl. Fig. 1). We also perform

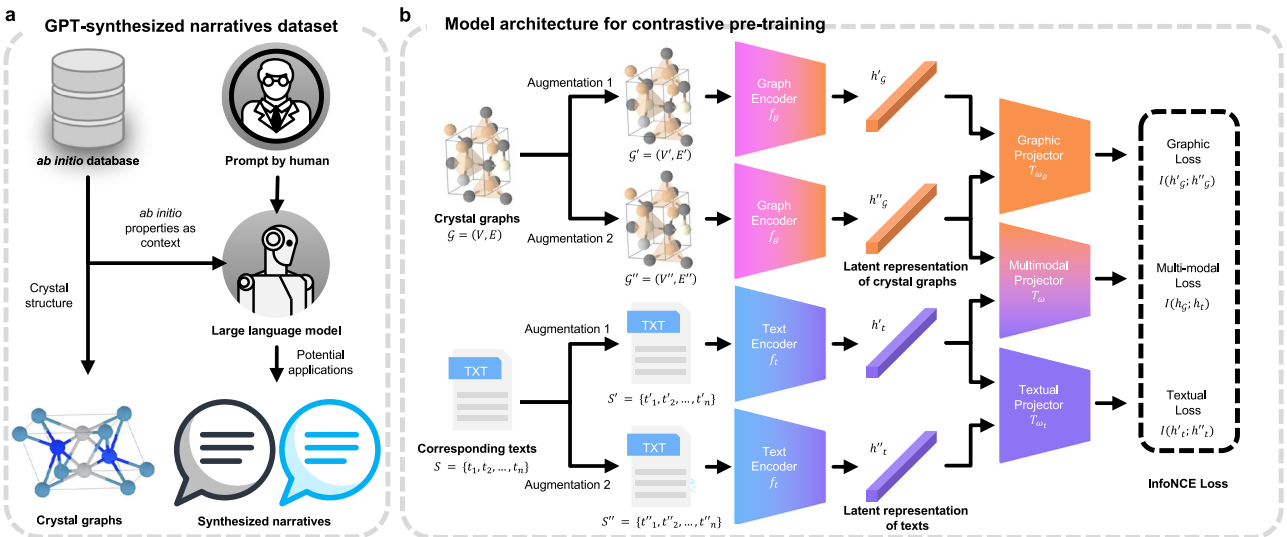


Fig. 1 | Conceptual schematic diagram of the data pipeline and model architecture of the contrastive language-crystals multimodal pretraining. a The data processing flow for generating GPT-synthesized narrative multimodal pair data. **b** The overall architecture of the model. The original graph, text, and corresponding augmented entities pass through graph and text encoders, which feed each projector, leading to a multimodal projector that aligns both data types.

https://doi.org/10.1038/s41467-026-76378-7

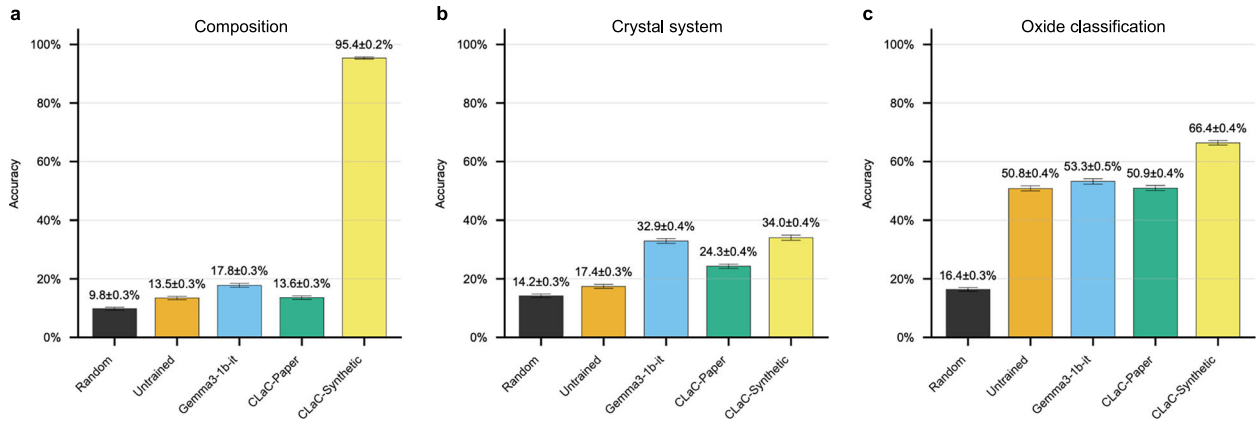


Fig. 2 | Zero-shot multimodal understanding across composition, crystal system, and oxide type. Accuracy of zero-shot crystal-text retrieval evaluated on three classification tasks: (**a**), elemental composition; (**b**), crystal system; and (**c**), oxide classification. Models are evaluated without task-specific fine-tuning using a multiple-choice retrieval protocol. Baselines include random selection, an untrained multimodal model, and a text-only large language model (Gemma-3-1b-it). CLaC models trained on literature-derived text (CLaC-Paper) and property-conditioned synthetic narratives (CLaC-Synthetic) are compared. Error bars denote standard errors estimated via bootstrapping. Source data are provided as a Source Data file.

a brute-force search to extract paragraphs containing specific chemical formulas from the academic literature (Suppl. Fig. 2). GPT narratives contain 300 times more diverse crystals than can be found in the academic literature (details in Section 4.1) written by humans. This mitigates biases present in existing academic literature, as most researchers chase after a small number of highly studied materials and neglect other possibilities on the thermodynamic convex hull, and allows the model to easily generalize its knowledge to a variety of downstream tasks. The dataset is split into 8:1:1 ratios for training, validation, and testing, respectively.

We use inter-modal and intra-modal contrastive learning in this study by following a data-efficient multimodal contrastive pre-training approach[32] to maximize the mutual information (Fig. 1b). Due to the scarcity of crystal-to-text data compared to image-to-text data, training both crystal-graph and text encoders from scratch is insufficient. To address this issue, Su et al.[33] used the weights of the GraphCL[34], pre-trained on biomedical molecules, and SciBERT[35], pre-trained on biomolecular corpora, as the initial weights of the graph and text encoder, respectively. However, unlike the biochemical molecular domain, where pre-trained GNNs trained on over 10 million unlabelled data[36–38], research on crystal graphs[39,40] have relied on only 126k Materials Project data. Therefore, we performed multimodal joint pre-training for both encoders on the prepared text-graph dataset: the graph encoder was trained from scratch, while the text encoder was initialized from a pre-trained model.

In computer vision, text-based image generation refers to the direct synthesis of visual content from natural language descriptions. The success of this paradigm has been largely driven by CLIP, which learns shared representations across modalities through large-scale contrastive learning[19–21]. Inspired by this paradigm, we contextualize our study by noting that text-based materials generation—from natural language descriptions to crystal structures—requires a model capable of predicting whether a crystal and a text fragment form a paired representation.

A multimodal model pretrained on both crystals and text provides several utilities for crystal discovery. First, by jointly processing crystal graphs and text tokens, the model enables flexible information retrieval for complex natural language queries, without being limited to predefined crystal labels. Second, intermodal alignment during pretraining supports large language models in grounding materials science knowledge in three-dimensional structural information.

Unlike vision–language learning, however, the materials science domain lacks standardized multimodal benchmarks for evaluating language–structure alignment or text-driven materials discovery. Existing benchmarks are typically designed for supervised property prediction with fixed labels, and do not assess whether a model can interpret open-ended natural language queries or generalize to unseen tasks. As a result, direct quantitative comparison across models is challenging, and evaluation protocols must be carefully designed to probe specific capabilities.

To address this gap, we focus on zero-shot evaluation settings that do not rely on task-specific fine-tuning or predefined label spaces. Specifically, we evaluate the model on zero-shot multimodal tasks that require joint understanding of both crystals and text, including cross-modal retrieval and language-guided screening. This evaluation strategy enables us to assess the extent to which the learned representations capture semantically meaningful alignment across modalities. To isolate the impact of textual supervision, we further compare models trained on synthetic narratives with those trained on text extracted directly from the literature.

Zero-shot multimodal understanding

Zero-shot understanding refers to the ability of an AI model to solve a new problem without prior specific training for that problem. For example, our model was not explicitly trained for crystal system classification or oxide type classification, but it is expected to perform those tasks due to the comprehensive pre-training.

To evaluate zero-shot understanding, we designed zero-shot classification problems on the test dataset from the Materials Project dataset as described above. Here we extracted several statements with 4 subjects: composition, structure, composition-structure, and oxide type, which together describe the essential solid-state chemical properties of the crystal. The zero-shot multimodal retrieval task can be interpreted as a choose-one-multiple-choice problem. For details about generating multiple choices, please refer to Supplementary Note 4. The model is required to retrieve the textual description for a given crystalline structure. The accuracy of the choose-one-from-multiple-choice problem is used as the evaluation metric. Several baseline strategies are proposed, including Gemma-3-1b-it[41] (Supplementary Note 5). Baseline models are also evaluated to compare zero-shot understanding performances (Suppl. Fig. 4). A bootstrapping test was used to ensure statistical robustness.

The zero-shot classification results are summarized in Fig. 2. First, in the composition classification task, the CLaC-Synthetic model achieves a high accuracy of 95.4 ± 0.2%. In contrast, the untrained model and the Gemma3-1b-it-based model exhibit substantially lower

https://doi.org/10.1038/s41467-026-76378-7

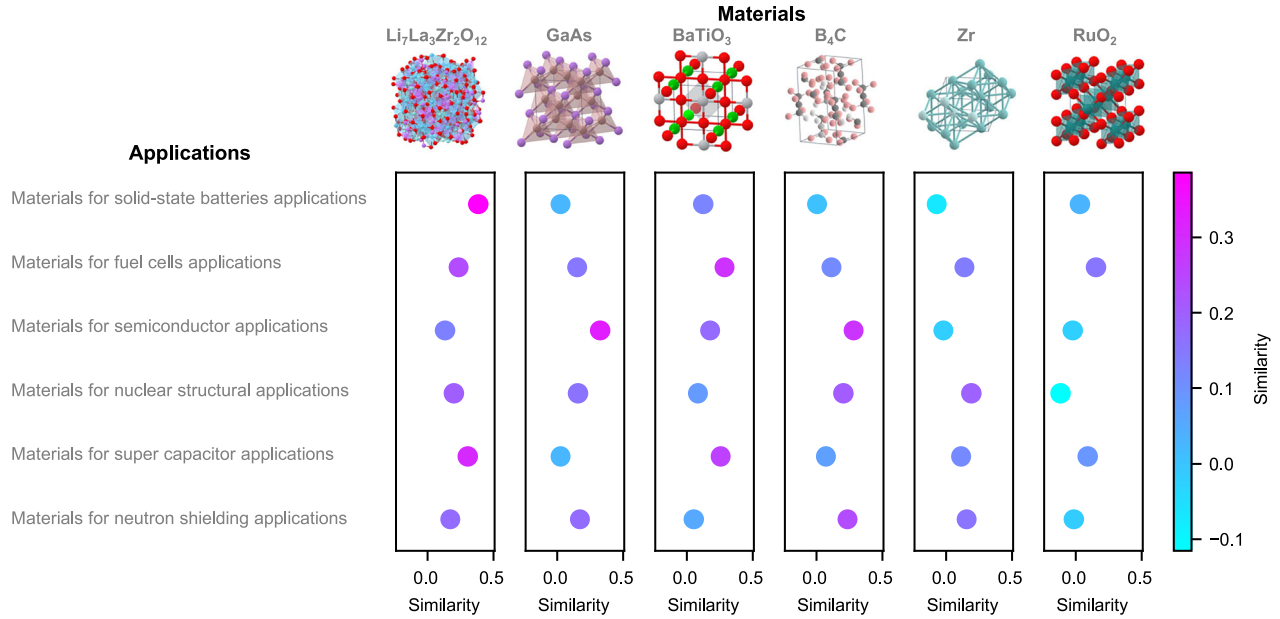


Fig. 3 | Zero-shot generalization ability. A graphical representation of the material-application similarity matrix. The matrix showcases six different materials - $Li_7La_3Zr_2O_{12}$, GaAs, $BaTiO_3$, B_4C, Zr, and RuO_2 - and their associations with six distinct application categories: solid-state batteries, fuel cells, semiconductors, nuclear structural materials, supercapacitors, and neutron shielding. Colored dots indicate the relevance of each material to the respective application, with the color coding corresponding to the degree of suitability based on underlying material properties. Source data are provided as a Source Data file.

performances of 13.5 ± 0.3% and 17.8 ± 0.3%, respectively, indicating that contrastive learning effectively captures compositional patterns. Moreover, the improvement of CLaC-Synthetic over CLaC-Paper suggests that synthetic data augmentation enhances the diversity of the compositional space.

In contrast, the crystal system classification task shows consistently lower performance across all models. Although CLaC-Synthetic achieves an accuracy of 34.0 ± 0.4%, this result remains comparable to that of Gemma3-1b-it. As discussed above, this limitation is likely related to the constraints of contrastive learning under small batch sizes (batch size = 16 in this work) combined with the highly imbalanced distribution of crystal systems. Under these conditions, negative samples are more likely to differ in chemical composition rather than in symmetry or subtle geometric features, increasing the prevalence of false negatives and reducing the model's ability to learn fine-grained structural distinctions. We further note that disabling graph augmentation did not lead to a meaningful improvement in crystal system classification accuracy, suggesting that the observed performance limitation is unlikely to be driven by augmentation-induced perturbations, but rather reflects intrinsic challenges related to negative sample diversity and data imbalance.

For oxide classification, we observe a distinct and intriguing pattern. Both the untrained model and Gemma3-1b-it substantially outperform random guessing, suggesting that pretrained language models already encode general chemical knowledge related to oxide formation. CLaC-Synthetic further improves performance, demonstrating that contrastive learning on materials data enhances discrimination of oxidation states. Interestingly, CLaC-Paper underperforms the base model, implying that paper-derived textual data alone may lack sufficient diversity for robust oxide classification.

We illustrate our model's zero-shot performance through additional case studies. Instead of performing a classification task in the multiple-choice paradigm, we extend it to explore potential applications of materials. We can screen several materials suitable for specific applications by calculating similarities. The visual matrix in Fig. 3 positions six materials along the vertical axis, each linked horizontally to a range of specialized application domains through a systematic array of colored dots. Each material is evaluated against six application categories: solid-state batteries, fuel cells, semiconductor devices, nuclear structural components, supercapacitors, and neutron shielding. The materials featured are $Li_7La_3Zr_2O_{12}$, a lithium lanthanum zirconium oxide (LLZO) often utilized as a solid electrolyte in solid-state battery technology; Gallium arsenide (GaAs), a compound semiconductor renowned for its superior electron mobility compared to silicon; Barium titanate ($BaTiO_3$), widely acknowledged for its use in multilayer ceramic capacitors and electroceramics; Boron carbide (B_4C), one of the hardest materials available, serving as an ideal candidate for armor and neutron shielding due to its high hardness and low density; Zirconium (Zr), a lustrous metal that finds its applications in nuclear reactors owing to its low neutron-capture cross-section; and Ruthenium dioxide (RuO_2), a compound known for its applications in electrochemical capacitors, also known as supercapacitors, due to its excellent conductivity and durability. This zero-shot learning capability for specific industry sectors has potential applications in natural language-based target material generation. Our model represents the first step in this research direction. Realizing instruction-based material discovery using the developed model remains a challenge for future work.

Zero-shot inverse design

Leveraging the capabilities of multimodal models offers the potential to design materials with target properties directly from natural language instructions. To explore this possibility, we propose a workflow in which natural language queries are used to screen a candidate pool containing multiple materials (Suppl. Fig. 5). Such candidate pools can now be generated at low computational cost using diffusion-based generative models for crystalline materials. By combining these models, it becomes possible to identify crystalline materials with desired properties without any additional post hoc training.

Our evaluation of CLaC in a text-to-materials retrieval setting reveals a clear dependence on the source and quality of textual supervision (Fig. 4). Given a natural-language query describing a target property, the model retrieves a small set of candidate materials ranked by text-materials similarity. However, when CLaC is trained on text

https://doi.org/10.1038/s41467-026-76378-7

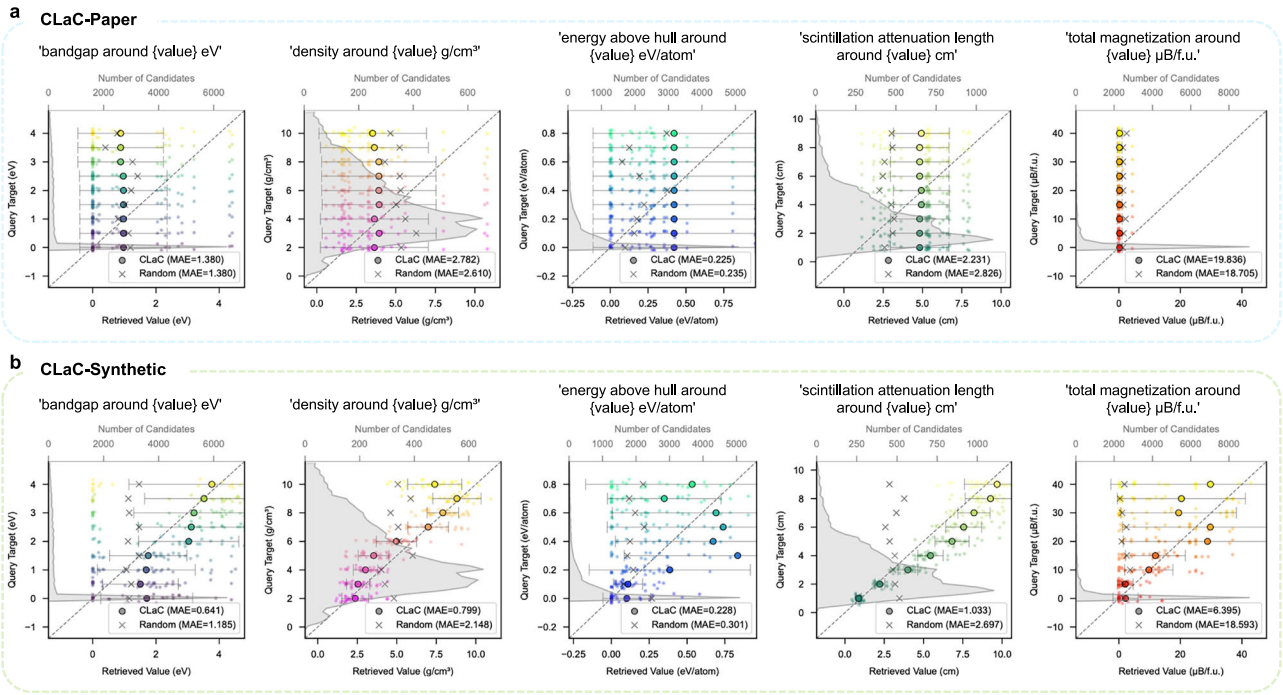


Fig. 4 | Zero-shot inverse design performance for text-to-materials property-guided retrieval. For each textual query describing a target property, the top 30 most similar candidate materials are retrieved and evaluated against the true target value. **a** CLaC trained on text extracted from literature (CLaC-Paper) is compared to (**b**), CLaC trained on synthetic narratives (CLaC-Synthetic). Each panel reports retrieved versus target values for different material properties. Error bars denote ± 1 standard deviation of the property values of the top 30 retrieved candidates (n = 30), and dashed lines denote ideal agreement. Grey marginal distributions show the overall candidate pool. Mean absolute errors (MAE) are reported for CLaC and for a random retrieval baseline. Source data are provided as a Source Data file.

fragments brute-force-extracted from scientific papers (CLaC-Paper), its retrieval performance is largely indistinguishable from random selection across all properties considered. The retrieved candidates show little correlation with the target values, indicating that noisy, weakly aligned literature text is insufficient to induce meaningful multimodal grounding in this setting.

In contrast, CLaC-Synthetic, trained on LLM-generated synthetic descriptions, consistently retrieves candidates that are more closely aligned with the target properties and substantially outperforms the random baseline. The improvement is evident in both reduced mean absolute error and stronger monotonic trends between retrieved and target values. These results suggest that synthetic text, despite being artificially generated, provides clearer and more structured supervision that better aligns language with graph-based material representations.

Taken together, these findings demonstrate serious limitations in building multimodal material models using the full text of scientific papers, highlighting the importance of text curation and generation strategies. Furthermore, multimodal learning using graph neural networks in the CLIP approach enables effective zero-shot queries of material databases only when the text modalities contain sufficient information. As with multimodal models in general, careful consideration of data structure and evaluation is essential. Continued research is needed to understand how dataset composition, text generation strategies, and deployment environments impact retrieval behavior and potential biases, and to identify optimal approaches for reliable and responsible use in materials discovery.

We further evaluate CLaC under a more linguistically realistic setting, where queries resemble qualitative natural-language descriptors rather than explicit numerical property specifications. This regime probes whether the model can associate abstract vocabulary—closer to how researchers might describe materials informally—with meaningful trends in the underlying materials space. As shown in Fig. 5, CLaC-Paper again fails to demonstrate meaningful structure: the distributions of retrieved candidates closely mirror those of the full database, indicating that text fragments extracted directly from the literature do not provide sufficient semantic grounding for qualitative language.

In contrast, CLaC-Synthetic exhibits markedly different behavior. When trained on LLM-generated descriptions, the model retrieves candidates whose property distributions shift systematically with the intended semantic meaning of the query, producing clear ordering and separation across qualitative categories. This result suggests that synthetic text enables the model to internalize latent mappings between natural-language concepts and material attributes, even when those concepts are not explicitly defined in quantitative terms. Together with the quantitative-query results, these findings reinforce that effective multimodal alignment in materials science depends less on the volume of textual data than on its semantic clarity and consistency, and that synthetic language can play a crucial role in bridging human vocabulary and structured scientific representations.

This section presents an example of applying a more complex text-based query. For instance, the development of novel solid-state electrolytes for lithium-ion secondary batteries requires the identification of materials that exhibit not only high ionic conductivity but also a sufficiently large band gap to suppress electronic conductivity. However, since ionic conductivity information is not included in our synthetic narrative dataset, this criterion was excluded, and materials were retrieved solely based on a high band gap condition. As a result shown in Table 1, the top 30 retrieved materials ranked by similarity all contain lithium, with an average band gap of 3.23 eV. Notably, Li_6PClO_5, which ranks fourth in the list, has been identified as a promising candidate in independent studies[42], while Li_4SiO_4, ranked ninth, has previously been synthesized as a thin film for use as a stable solid electrolyte[43]. These findings provide additional validation for the reliability and effectiveness of the proposed methodology.

https://doi.org/10.1038/s41467-026-76378-7

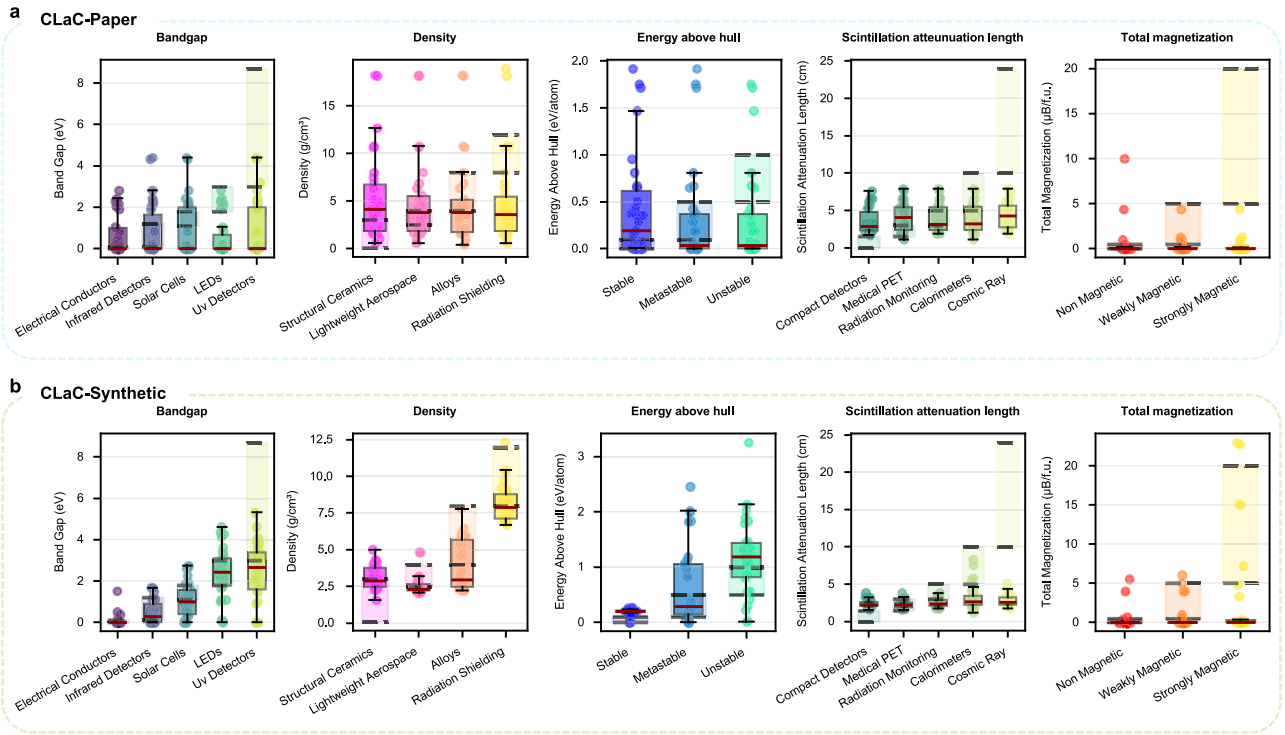


Fig. 5 | Zero-shot inverse design performance for text-to-materials language-guided retrieval. Performance of text-to-materials retrieval when the queries are expressed using qualitative, natural-language-like descriptors rather than explicit quantitative property values. **a** CLaC trained on text extracted from literature (CLaC-Paper) is compared to (**b**), CLaC trained on synthetic narratives (CLaC-Synthetic). For each query, the distributions of the top-ranked retrieved materials are shown as box plots across different semantic categories, with individual points indicating retrieved candidates. In each box plot, the center line denotes the median, the box limits denote the lower and upper quartiles (interquartile range, IQR), and the whiskers extend to 1.5 × IQR; outliers beyond the whiskers are not shown, as all individual retrieved candidates are overlaid as points. Shaded regions indicate the range of the full candidate pool. Source data are provided as a Source Data file.

Text embedding analysis

To evaluate whether the learned text embeddings capture continuous material property information, we performed a similarity analysis on natural language queries describing five representative material properties. For each property, we encoded templated text queries spanning the full value range and computed pairwise cosine similarities between the resulting embeddings. The results are visualized using similarity heatmaps, reference-based similarity decay curves, and similarity-difference plots (Fig. 6).

The analysis indicates that the CLaC-Synthetic model learns continuous representations of material property space. The similarity matrix heatmaps (Fig. 6, left column) exhibit a diagonal-dominant structure, with the highest similarity (cosine > 0.95) concentrated along the diagonal where property values are identical. The similarity gradually decreases for more distant value pairs. This pattern confirms that the model encodes property magnitude information rather than treating distinct numerical queries as independent discrete categories.

The reference value decay curves (middle column) demonstrate similarity degradation as property values diverge from a given reference point. For instance, in the band gap analysis, a reference value of 2.0 eV shows cosine similarity declining from 1.0 (self-match) to approximately 0.65 at 4.0 eV ($\Delta = 2.0$ eV). The decay is approximately symmetric around each reference point, indicating that the model captures bidirectional property relationships without directional bias.

Crucially, we observed that the steepness of the similarity decay in the text latent space correlates with retrieval precision. Properties exhibiting sharp similarity gradients, such as density and scintillation attenuation length (Fig. 6b, d), demonstrate high discriminative power, where embedding similarity drops rapidly as values diverge. This sensitivity effectively penalizes small deviations, translating to superior retrieval performance characterized by a tight alignment between query targets and retrieved values (Fig. 4). In contrast, properties with more gradual decays, such as energy above hull (Fig. 6c), retain high similarity even across larger value differences. This broader tolerance in the embedding space corresponds to the relatively higher variance observed in retrieval, suggesting that a sharper metric in the latent space is advantageous for precise continuous value regression.

We further compared these results with those from CLaC-Paper and Galactica-125M (Suppl. Fig. 6) to assess the necessity of our proposed alignment method (Suppl. Fig. 7). In contrast to the smooth, monotonic gradients observed in our model, these baselines failed to capture the continuous variations of physical quantities. Their embedding spaces exhibited disordered similarity patterns with negligible correlation to numerical differences, indicating that standard language modeling objectives alone are insufficient to induce a physically meaningful metric space. This comparison underscores that the proposed cross-modal training is essential for structuring the text latent space to respect physical magnitudes.

Finally, we investigated the impact of incorporating a standard language modeling (LM) loss on the embedding structure. When trained with an auxiliary next-token prediction objective, the similarity gradients became notably more gradual across all properties (Suppl. Fig. 8). This smoothing effect implies that the LM objective, which prioritizes contextual probability and semantic coherence, conflicts with the formation of a precise metric space required for continuous regression. While next-token prediction helps the model understand the semantic role of numbers within a sentence, it tends to blur the strict Euclidean or cosine distance between distinct magnitudes. This trade-off suggests that the objectives of semantic fluency and

Table 1 | Materials containing Li ions with high bandgap for solid-state electrolyte candidate (Top 30)

Rank	Formula	Similarity	Band Gap (eV)	E_{hull} (eV/atom)	Rank	Formula	Similarity	Band Gap (eV)	E_{hull} (eV/atom)
1	$Li_{10}Sn(PO_6)_2$	0.7131	3.642	0.125	16	$Li_5Cu(PO_4)_2$	0.6217	1.847	0.029
2	$Li_2SnP_2O_7$	0.6782	4.062	0.021	17	$Li_3SiBiCO_7$	0.6216	3.437	0.074
3	$Li_{10}Si_2PbO_{10}$	0.6752	2.961	0.015	18	$Li_5Cu(PO_4)_2$	0.6214	1.707	0.030
4	Li_6PClO_5	0.6493	5.276	0.006	19	$Li_3P_2WO_9$	0.6191	3.348	0.042
5	$Li_5NiP_3O_{11}$	0.6466	3.527	0.074	20	Li_3ThF_7	0.6179	5.686	0.209
6	$Li_9Sn_3P_8O_{29}$	0.6450	0.000	0.073	21	Li_3TaF_6	0.6167	0.000	0.500
7	Li_4ZrO_4	0.6423	4.323	0.037	22	$Li_3Ti(PO_4)_2$	0.6153	0.242	0.074
8	$Li_5MnP_3O_{11}$	0.6381	4.163	0.284	23	$Li_3Sn_2(PO_4)_3$	0.6145	2.928	0.079
9	Li_4SiO_4	0.6343	4.656	0.058	24	Li_3AsO_4	0.6145	4.151	0.004
10	$Li_4Ga_3Si_3BrO_{12}$	0.6333	3.951	0.000	25	$Li_5V(PO_4)_3$	0.6137	1.584	0.093
11	$LiSnPO_4$	0.6331	3.290	0.082	26	Li_8BiO_3	0.6116	0.000	0.449
12	Li_2SO_5	0.6293	0.000	0.363	27	$Li_4Zn(PO_4)_2$	0.6114	4.686	0.001
13	$Li_4Ga_3Si_3ClO_{12}$	0.6292	4.412	0.000	28	$Li_4Fe(PO_4)_2$	0.6079	4.207	0.053
14	$Li_3Bi(BO_3)_2$	0.6243	2.937	0.086	29	Li_2SO_4	0.6065	6.220	0.055
15	$Li_3Sn_2P_5O_{16}$	0.6235	4.095	0.038	30	$Li_2S_2O_7$	0.6063	5.536	0.000

continuous physical encoding may be mutually interfering, highlighting a critical design consideration for developing future language models specialized in materials science.

Visualization of latent space

To qualitatively validate the representation learning capabilities of our model, we visualized the high-dimensional embeddings of the test dataset using Uniform Manifold Approximation and Projection (UMAP)[44]. As illustrated in Fig. 7, the projection reveals two distinct global clusters: the left cluster represents the text embeddings processed by the language model, galactica-125M, while the right cluster corresponds to the crystal structure embeddings encoded by our GNN backbone, PaiNN. While this separation indicates a modality gap often observed in contrastive learning[45], the two modalities share a common neighborhood structure over the same materials. On average, 57% of the fifty nearest neighbours of a crystal in the graph embedding space are also among its fifty nearest neighbours in the text embedding space, compared with 0.4% expected by chance. This alignment implies that, although the contrastive objective enforces alignment only for paired examples, the learned representations generalize to the neighborhood structure of the test set. The structure is also chemically meaningful, as oxide type, which is never used as a training label, can be recovered from either modality with accuracies of 0.95 and 0.96 using a nearest-neighbor classifier, compared with a majority-class rate of 0.51 (Supplementary Note 8 and Supplementary Fig. 9).

We analyzed the model's capability to capture continuous physical magnitudes by probing the latent space with numerical text prompts. Figure 7a and b present the distributions for bandgap and density, respectively. The leftmost panels map the ground truth values onto the crystal embeddings as a reference. Subsequent panels illustrate the cosine similarity between the embeddings and specific numerical queries, such as "2.0 eV" or "4.0 g/cm^3." The regions exhibiting high similarity shift spatially as the queried value increases, following the distribution of the ground truth data. This spatial correspondence indicates that the model encodes continuous physical properties through cross-modal alignment without explicit regression objectives.

Furthermore, we examined the representation of categorical material classes and functional applications (Fig. 7c). Queries using semantic descriptors (e.g., "Ceramic," "Conductor") or functional keywords (e.g., "Li-ion battery," "Thermoelectric") resulted in localized activation patterns within the crystal embedding cluster. These patterns suggest that materials sharing similar chemical compositions or application domains are grouped in the latent space. For instance, the prompt "Li-ion battery" activates specific sub-regions associated with relevant structural features. Combined with previous evaluations on zero-shot tasks, these results highlight the superior embedding quality of CLaC caused by joint contrastive learning. Future research could focus on enhancing data generation methodologies and refining training processes to achieve a more precise alignment between the linguistic hypothesis space of scientists and the materials space, thereby enabling novel approaches for exploring this intersection.

Ablation study

To better understand the contribution of each component in our proposed model, we conduct an ablation study across four key design dimensions: graph encoder architecture, contrastive loss formulation, language model backbone, and intramodal loss incorporation. We evaluate all component combinations on zero-shot retrieval and inverse design tasks spanning five material properties (band gap, density, energy above hull, scintillation attenuation length, and total magnetization) with retrieval depth $k = 10$, yielding 200 experimental configurations (Table 2 and Suppl. Figs. 10, 11).

Cross-modal retrieval and inverse design tasks impose distinct representational requirements that may lead to trade-offs. Retrieval emphasizes global pattern matching by capturing holistic material similarity across compositional, structural, and symmetry-related features, enabling the identification of semantically related materials irrespective of specific property values. In contrast, property prediction relies on local pattern matching, which requires precise encoding of fine-grained structural motifs and bonding environments that are strongly correlated with target properties. Architectural choices that improve global semantic alignment can inadvertently obscure local structural details that are essential for accurate property prediction, and the converse can also occur. This tension between global and local feature preservation is consistently observed across multiple design dimensions in the ablation study.

We compare three graph neural network architectures for crystal structure encoding. PaINN (polarizable atom interaction neural network)[46] and CGCNN (crystal graph convolutional neural network)[3] were trained from scratch, and a machine learning interatomic potential mode, ORB (orbital-materials)[47], pretrained on large-scale potential energy surface datasets, was selected for comparison. PaINN achieves superior retrieval accuracy (76.7% top-1 vs. 71.7%) while CGCNN excels at property prediction (2.735 MAE and 0.836 correlation vs. 3.070 MAE and 0.757 correlation). This task-dependent trade-

https://doi.org/10.1038/s41467-026-76378-7

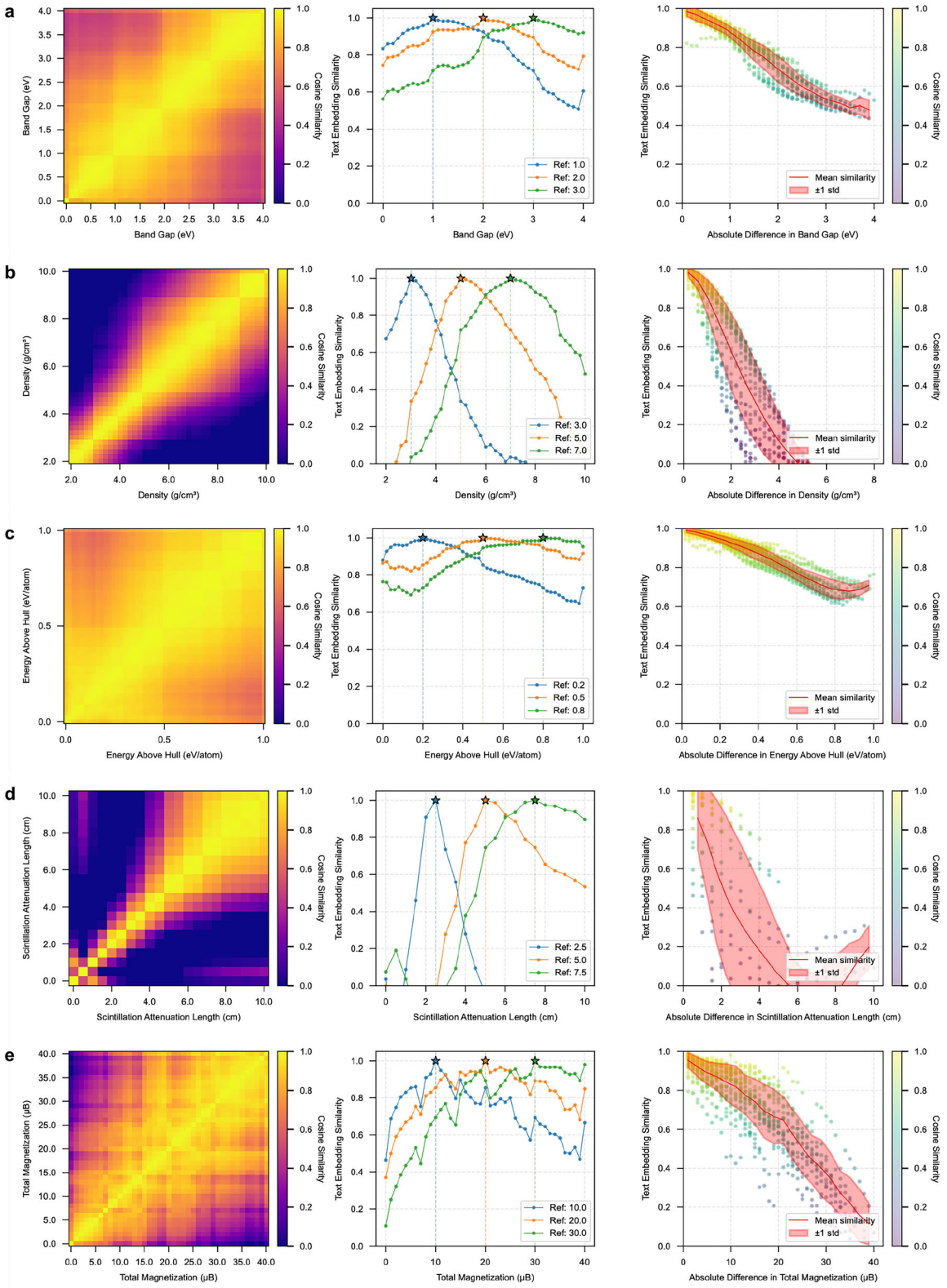


Fig. 6 | Semantic structure of numerical values within the text embedding space. Cosine similarity analysis of text embeddings generated using the prompt template "material with [property] of [value] [unit]" for five continuous properties: (**a**) Band Gap, (**b**) Density, (**c**) Energy Above Hull, (**d**) Scintillation Attenuation Length, and (**e**) Total Magnetization. The left panels display pairwise similarity heatmaps across a range of property values, where yellow indicates high similarity (1.0) and blue indicates low similarity. The middle panels illustrate similarity profiles relative to specific reference values (marked with stars). The right panels show the correlation between cosine similarity and the absolute difference between values. The red solid line represents the mean similarity, and the shaded region indicates ± 1 standard deviation. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-76378-7

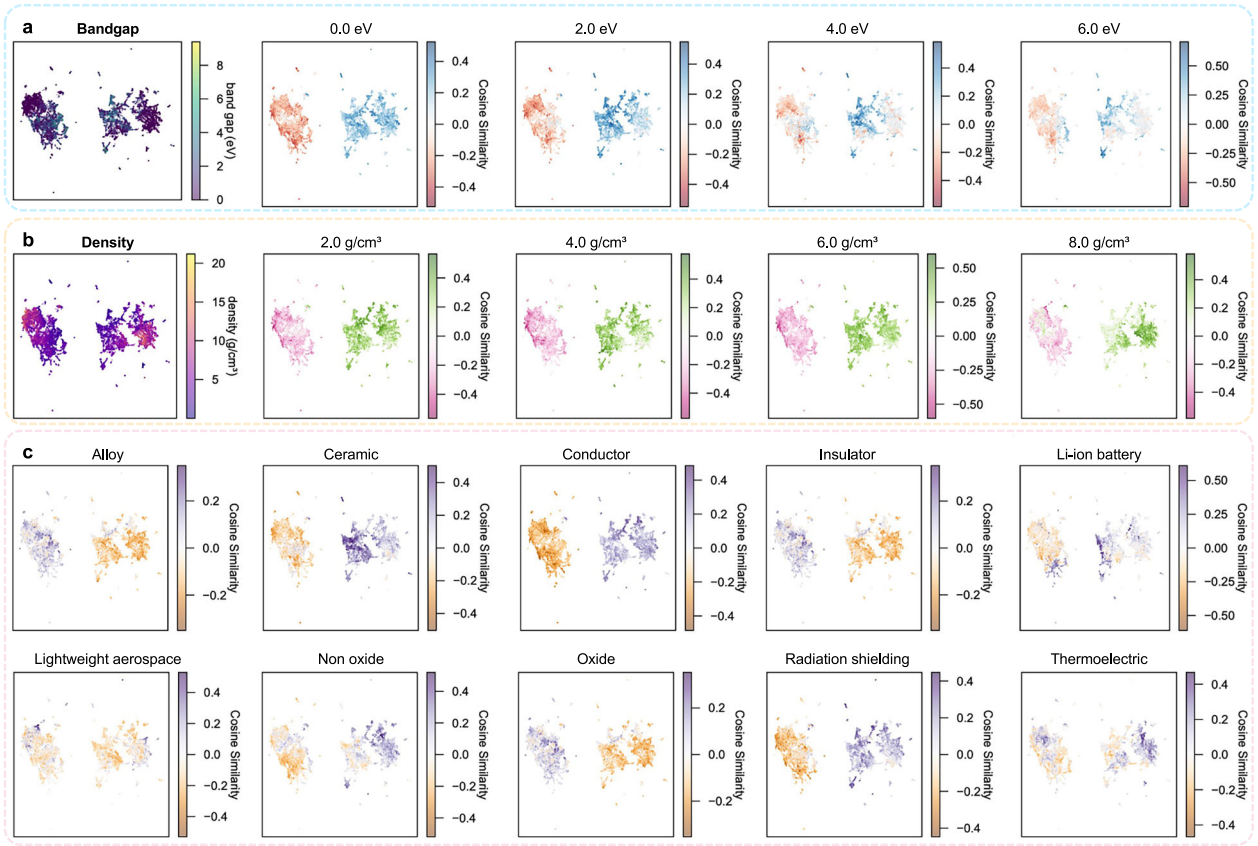


Fig. 7 | Latent space visualization of cross-modal alignment between crystal structures and text. The Uniform Manifold Approximation and Projection (UMAP) of the test dataset embeddings generated by CLaC. The visualization reveals two distinct main clusters, corresponding to the text modality (left) and the crystal graph modality (right). Visualization of continuous physical properties: (**a**) Band gap and (**b**) Density. The leftmost panels display the ground truth values mapped onto the latent space. Subsequent panels illustrate the cosine similarity between the learned embeddings and specific numerical text prompts (indicated above each plot). The spatial progression of high-similarity regions (red) aligns with the increasing values in the text prompts, demonstrating the model's ability to regress continuous physical magnitudes. **c** Visualization of categorical material classes and functional applications. Each subplot displays the cosine similarity distribution for a specific text query (e.g., "Conductor", "Li-ion battery").

Table 2 | Comprehensive ablation study across four architectural components

Component	Variant	Top-1 ↑	Top-3 ↑	Top-10 ↑	MAE ↓	Corr ↑
Graph Encoder	PaINN	**0.767**	**0.917**	**0.975**	3.070	0.757
	CGCNN	0.717	0.869	0.944	**2.735**	**0.836**
	ORB	0.097	0.218	0.437	3.455	0.776
LLM Backbone	OPT-125M	**0.512**	**0.646**	**0.769**	3.250	0.784
	Galactica-125M	0.493	0.606	0.713	**2.762**	**0.816**
	MatSciBERT	0.478	0.632	0.768	3.579	0.747
	SciBERT	0.473	0.627	0.765	3.038	0.775
Loss Type	InfoNCE	**0.566**	**0.701**	**0.817**	**3.079**	**0.811**
	JSD+Dot	0.374	0.518	0.658	3.275	0.734
Intramodal Loss	True	**0.517**	**0.668**	**0.806**	3.241	0.754
	False	0.461	0.587	0.701	**3.073**	**0.807**

All values are averaged across 200 experimental configurations (5 properties × 40 model variants, $k = 10$ retrieval depth). Within each component group, bold entries indicate the best value for each metric (highest for Top-1, Top-3, Top-10, and Corr; lowest for MAE).

off indicates that neither architecture is universally superior. Despite being pretrained on large-scale datasets, ORB dramatically underperforms both randomly initialized encoders. We hypothesize this failure stems from a fundamental mismatch between pretraining and downstream objectives: ORB's representations are optimized for potential energy surface (PES) prediction, which requires atomic force fields and local energy landscapes. During pretraining, the network likely filters out or deprioritizes structural information irrelevant to PES prediction that are critical for material-text alignment and property prediction. This suggests that task-specific pretraining may be less effective than training from scratch for cross-modal materials informatics.

We compare four pretrained language models across two architectural families: encoder-only models (SciBERT, MatSciBERT) and decoder-only models (OPT-125M, Galactica-125M). MatSciBERT and Galactica-125M are further fine-tuned on materials science and scientific literature from SciBERT and OPT-125M, respectively. They were selected under the hypothesis that domain-specific pretraining would encode material- and property-related semantics beneficial for cross-modal alignment. Decoder-only architectures consistently achieve higher retrieval accuracy than encoder-only architectures, with OPT-125M leading at 51.2% top-1 accuracy. However, within each architectural family, domain-specific pretraining shows minimal impact: the general-domain OPT-125M and scientific-domain Galactica-125M differ by only 1.9 percentage points in top-1 accuracy, while SciBERT and MatSciBERT differ by 0.5 percentage points. Despite being pretrained on materials science literature, MatSciBERT does not outperform the

general-domain OPT-125M, and Galactica-125M shows no clear advantage over OPT-125M. This indicates that contrastive fine-tuning on material-text pairs largely overcomes the gap between general and domain-specific pretrained representations.

Two contrastive loss functions were evaluated. InfoNCE is the standard objective widely adopted in CLIP-style models, and Jensen-Shannon divergence with dot product similarity (JSD) was proposed to improve training efficiency with small batch sizes. InfoNCE consistently outperforms JSD across all metrics, achieving 56.6% top-1 accuracy compared to 37.4% for JSD—a substantial 19.2 percentage point improvement. The performance gap widens for lower retrieval depths, suggesting InfoNCE produces tighter, more discriminative clusters in the joint embedding space. These results indicate that despite JSD's theoretical advantages for small-batch training, InfoNCE's proven effectiveness in large-scale vision-language models transfers successfully to cross-modal materials informatics.

Overall, the ablation study reveals clear trade-offs between graph and text encoder architectures, with no single choice uniformly dominating all tasks. In contrast, the contrastive objective and the incorporation of intramodal loss show consistent and substantial effects on performance. Based on a Pareto analysis of all evaluated configurations (Suppl. Figs. 12, 13), PaINN-Galactica-InfoNCE with intramodal loss emerges as a Pareto-optimal configuration with a favorable balance between cross-modal retrieval accuracy and inverse design performance. We therefore adopt this configuration in all subsequent experiments presented in the main text.

Discussion

In conclusion, we developed a crystal graph-aware multimodal language model that is trained on crystal graphs and materials science corpora, including GPT-synthesized narratives. Through a series of experiments on zero-shot tasks, we confirmed that leveraging synthetic data is a promising approach to addressing data scarcity and bias within the materials science domain. The multimodal model, trained with 3D crystal structure information from synthetic data, enabled the execution of zero-shot tasks, which were challenging to achieve with previous language models. Our model can serve as a baseline for future work, for example, providing similarity metrics for evaluating text-based crystal generation tasks.

Despite the strengths exhibited by our current models, there are limitations; Our model is currently only applicable to crystalline materials. However, contemporary materials science research handles polycrystals or materials with defects that affect their properties, rather than single crystals. Furthermore, the model cannot handle metal-organic frameworks at this stage. Therefore, our next step is to cultivate models capable of identifying correlations that encompass a broader spectrum of materials and structures, thereby achieving a higher level of comprehension. Such advancements will require the development of representations that can simultaneously handle not only single crystals but also high-dimensional structures such as polycrystals and defects. Additionally, advancements in data synthesis methods are needed to minimize hallucinations and generate more informative data. Another challenge is the lack of a standardized nomenclature in the field of inorganic materials science, unlike in the molecular domain. For example, IUPAC names naturally included in academic literature provide opportunities to learn various concepts from the surrounding context. On the other hand, terms like SiO_2 in the literature cannot be distinguished by nomenclature alone as to whether they refer to a crystal or non-crystal, bulk material, or nanoparticles. Future research should carefully address these limitations.

Methods

Data collection

For training a language-materials multimodal contrastive learning model, we follow two steps: data preparation and model training. First, we use language-material pair data, which is language model-based synthetic data that Park et al.[31] have worked on previously. We also retrieve domain-specific natural language text from large-scale unlabeled materials in the SCI paper dataset. Afterward, the data of each modality passes through a graph encoder and a text encoder and is trained to be mapped to the joint representation space. In this way, our model can link materials graphs with materials engineering text descriptions.

We collect 126k materials data from Materials Project[30] to train our multimodal foundation model. To obtain crystal structure graphs of the collected compounds, we use the function to convert to crystal graphs from the Jarvis-tools library[48]. To obtain the text for each material, we use 126k material-synthesized text pairs from our previous work[31]. We also search the corpus with weak supervision using the full text of 4M academic papers exclusively available to our group using chemical formulas as queries. In the collected data, there are a total of 400k material-natural language sentence pairs corresponding to approximately 400 inorganic compounds.

To improve the robustness of multimodal contrastive learning, we employ modality-specific data augmentation strategies for both crystal graphs and textual descriptions.

For contrastive pre-training on crystalline graphs, we employ physically motivated augmentation strategies designed to preserve material identity while encouraging robustness to realistic structural variations, guided by insights from CrystalCLR[39]. While CrystalCLR identifies structure perturbation and column replacement as an effective augmentation pair, we intentionally exclude column replacement in this work. Although column replacement preserves formal valence, it can substantially alter key material properties such as ionic radius, coordination environment, lattice parameters, and electronic structure, effectively changing the semantic identity of the material. Instead, we combine structure perturbation (randomly displaces each atom by up to 50% of the minimum interatomic distance) with strain-based augmentation (scales lattice vectors by factors sampled from [1.0, 1.05]), which continuously deforms lattice parameters while preserving composition and bonding topology. This choice encourages invariance to elastic distortions and local structural noise that naturally arise in crystalline datasets, without introducing chemically distinct structures.

To enhance robustness against sentence ordering and partial information loss, we apply sentence-level augmentation to textual inputs. Given a text description composed of a sequence of sentences $S = \{s_1, s_2, \ldots, s_m\}$, we construct an augmented text sequence $\hat{S}$ through a stochastic process consisting of sentence dropping and shuffling.

Specifically, each sentence s_i is independently retained with probability $p_{\text{keep}} = 0.5$. To avoid degenerate cases where all sentences are removed, we enforce a minimum constraint such that at least one sentence is preserved by randomly selecting a sentence from the original sequence when necessary. The retained sentences are then randomly permuted to form the final augmented text:

$$\hat{S} = \text{Shuffle}\left(\{s_i | \text{Bernoulli}(p_{\text{keep}}) = 1\}\right).$$

This augmentation strategy encourages the model to learn representations that are invariant to sentence order and resilient to missing contextual information, thereby promoting semantic consistency across diverse textual realizations.

Within the contrastive learning framework, multiple augmented views of the same text are generated by independently applying the above augmentation strategies. The model is trained to align representations derived from these views, thereby enforcing invariance to sentence-level perturbations and chemical formula substitutions. This design complements structural augmentations applied to crystal graphs and contributes to learning robust and semantically meaningful multimodal representations.

Graph encoder

In the landscape of predictive analytics for material properties, Graph Neural Networks (GNNs) have emerged as one of the most prominent methodologies, demonstrating substantial success in modeling crystalline materials[3,4]. By effectively capturing local atomic environments and modeling interatomic interactions, GNNs have significantly advanced the performance of material property prediction models. A representative example is the Crystal Graph Convolutional Neural Network (CGCNN)[3], which applies graph convolution operations to crystal graphs and substantially outperforms conventional machine learning approaches.

Recent developments have further enhanced graph-based representations by incorporating higher-order interactions, such as line-graph constructions to capture three-body and four-body angular information[4], as well as equivariant vector operations for improved representational capacity[46]. In our study, to examine the impact of different graph hidden representations within the CLaC model architecture, we employ CGCNN and PaiNN as graph encoders, enabling a systematic comparison of their effectiveness.

Furthermore, to assess the effect of large-scale pre-training for graph encoders, we additionally adopt ORB[47,49], an open-source machine learning interatomic potential (MLIP) model, as a backbone graph encoder. ORB is pre-trained on large and diverse atomic configuration datasets, providing rich structural and chemical representations. By comparing ORB-based representations with those learned from CGCNN and PaiNN, we investigate whether large-scale pre-trained graph encoders offer transferable advantages in multimodal contrastive learning for crystalline materials.

The graph encoder is formulated as follows:

$$h_{\mathcal{G}} = f_g\left(\mathcal{G}, \{x_v | v \in V\}, \{e_{uv} | (u,v) \in E\}\right), \tag{1}$$

where $\mathcal{G} = (V, E)$ denotes a crystal graph with nodes V and edges E, x_v represents the feature vector of node v, e_{uv} represents the feature vector of edge (u, v), f_g denotes the graph encoder, and $h_{\mathcal{G}}$ is the pooled graph-level representation. For CGCNN-based models, CGCNN embeddings[3] are used as initial node features, and radial basis functions of interatomic distances are adopted as edge features.

Text encoder

The BERT architecture has been widely used to extract meaningful features from text and has served as a standard text encoder in vision–language multimodal models such as DeCLIP and CLIP-lite[32,50]. Beltagy et al.[35] introduced SciBERT, which was pre-trained on 1.14M scientific papers from the biomedical and computer science domains. Building upon this, Gupta et al.[51] applied domain-adaptive pre-training to SciBERT to develop MatSciBERT, a materials-aware language model trained on a materials science corpus with an expanded domain-specific vocabulary. MatSciBERT has demonstrated improved performance on downstream tasks such as named entity recognition and text classification.

In this study, SciBERT and MatSciBERT are adopted as encoder-based text encoders for comparative evaluation. To mitigate catastrophic forgetting in BERT-based models during multimodal pre-training, we additionally incorporate a masked language modeling (MLM) loss.

In addition to encoder-only architectures, we also explore decoder-based causal language models as text encoders. Specifically, we employ OPT-125M and Galactica-125M, which are lightweight autoregressive language models pre-trained on large-scale general-domain and scientific corpora, respectively. When using these models, text representations are extracted from the hidden states of the causal language model, and training is supplemented with a causal language modeling (CLM) loss instead of MLM. This allows us to systematically compare encoder-based and decoder-based text encoders within a unified multimodal contrastive learning framework.

Text encoders can be formulated as follows:

$$h_t = f_t\left(\{x_i | i \in \{1, \ldots, n\}\}\right), \tag{2}$$

where $\{x_i | i \in \{1, \ldots, n\}\}$ denotes the sequence of input token embeddings, h_t represents the resulting hidden representations produced by the text encoder, and f_t corresponds to either a BERT-based encoder or a causal language model, depending on the architecture.

Optimization

In this section, we describe our pre-training framework (Fig. 1) for crystal graphic representation learning. Given a dataset of graph–text pairs, our objective is to train a graph encoder and a text encoder such that the learned representations from each modality maximize their shared mutual information. Let f_g and f_t denote the graph encoder and text encoder with parameters θ_g and θ_t, respectively. Given a sampled graph–text pair (x_g, x_t), the corresponding representations are obtained as $h_{\mathcal{G}} = f_g(x_g)$ and $h_t = f_t(x_t)$.

The mutual information between the crystal graph x_g and the text description x_t is defined as the Kullback–Leibler (KL) divergence between their joint distribution $p(x_g, x_t)$ and the product of their marginal distributions $p(x_g)p(x_t)$:

$$I(x_g; x_t) = D_{\mathrm{KL}}\left(p(x_g, x_t) \parallel p(x_g)p(x_t)\right). \tag{3}$$

However, directly estimating mutual information for high-dimensional continuous variables is challenging. To address this issue, we adopt tractable lower-bound objectives that can be optimized efficiently in practice.

We estimate and maximize mutual information using a Jensen–Shannon divergence (JSD)-based lower bound, which has been widely adopted in contrastive representation learning and generative modeling. This formulation is fully differentiable and does not explicitly depend on the number of negative samples[32].

The JSD-based mutual information estimator is given by:

$$I(h_{\mathcal{G}}; h_t) \geq \hat{I}^{\mathrm{JSD}}_{\omega}(h_{\mathcal{G}}; h_t) := \mathbb{E}_{P(h_{\mathcal{G}}, h_t)}\left[-\log\left(1 + e^{-T_\omega}\right)\right] - \mathbb{E}_{P(h_{\mathcal{G}})P(h_t)}\left[\log\left(1 + e^{T_\omega}\right)\right], \tag{4}$$

where $T_\omega : \mathcal{X}_g \times \mathcal{X}_t \to \mathbb{R}$ is a discriminator network parameterized by ω. The discriminator is jointly optimized with the encoders to distinguish between samples drawn from the joint distribution (positive pairs) and those drawn from the product of marginal distributions (negative pairs).

Accordingly, the optimization objective is defined as:

$$(\hat{\omega}, \hat{\theta}_g, \hat{\theta}_t) = \mathrm{argmax}_{\omega, \theta_g, \theta_t} \hat{I}^{\mathrm{JSD}}_{\omega}\left(f_g(x_g), f_t(x_t)\right). \tag{5}$$

Based on this formulation, we define the inter-modal and intra-modal contrastive losses as follows:

$$\mathcal{L}_{\mathrm{inter-modal}} = -\left(\mathbb{E}_{P(h_{\mathcal{G}}, h_t)}\left[-\log(1 + e^{-T_\omega})\right] - \mathbb{E}_{P(h_{\mathcal{G}})P(h_t)}\left[\log(1 + e^{T_\omega})\right]\right), \tag{6}$$

$$\mathcal{L}_{\mathrm{intra-modal,\,graph}} = -\left(\mathbb{E}_{P(h_{\mathcal{G}}, h_{\mathcal{G}'})}\left[-\log(1 + e^{-T_{\omega_g}})\right] - \mathbb{E}_{P(h_{\mathcal{G}})P(h_{\mathcal{G}'})}\left[\log(1 + e^{T_{\omega_g}})\right]\right), \tag{7}$$

$$\mathcal{L}_{\mathrm{intra-modal,\,text}} = -\left(\mathbb{E}_{P(h_t, h_{t'})}\left[-\log(1 + e^{-T_{\omega_t}})\right] - \mathbb{E}_{P(h_t)P(h_{t'})}\left[\log(1 + e^{T_{\omega_t}})\right]\right). \tag{8}$$

Table 3 | Training hyperparameters for CLaC pre-training

Hyperparameter	Value
Model Architecture	
Graph encoder	PaiNN / CGCNN
Text encoder	OPT-125M / Galactica-125M / SciBERT / MatSciBERT
Embedding dimension	512
Graph encoder hidden dim	512
Graph cutoff radius	5.0 Å
PaiNN message-passing layers	5
CGCNN convolutional layers	3
Optimization	
Optimizer	AdamW
Learning rate	2×10^{-5}
Adam β_1, β_2	0.9, 0.999
Adam ϵ	10^{-8}
Weight decay	0.0
Gradient clip norm	0.05
Training Strategy	
Batch size (per GPU)	16
Gradient accumulation steps	16
Effective batch size	256
Number of GPUs	2
Training precision	bfloat16 mixed
Distributed strategy	DDP
Maximum epochs	5
Random seed	42
Contrastive Learning	
Loss function	InfoNCE / JSD
Temperature (τ)	0.07
Visual intramodal loss	Enabled / Disabled
Textual intramodal loss	Enabled / Disabled
Data Augmentation	
Graph Augmentation:	
Structure perturbation	50% of min distance
Strain magnitude	0–5%
Text Augmentation:	
Sentence dropout rate	50%
Sentence shuffling	Enabled

In addition to the JSD-based objective, mutual information can also be optimized using the InfoNCE loss, which is a widely used contrastive objective in multimodal and self-supervised representation learning. InfoNCE formulates mutual information maximization as a multi-class classification problem, where the model is trained to correctly identify a positive pair among a set of negative samples.

Given a positive graph–text pair $(h_{\mathcal{G}}, h_t)$ and a set of $N-1$ negative text representations $\{h_{t^-}\}$, the InfoNCE loss is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\exp(\text{sim}(h_{\mathcal{G}}, h_t)/\tau)}{\exp(\text{sim}(h_{\mathcal{G}}, h_t)/\tau) + \sum_{t^-} \exp(\text{sim}(h_{\mathcal{G}}, h_{t^-})/\tau)}\right], \tag{9}$$

where $\text{sim}(\cdot,\cdot)$ denotes a similarity function (e.g., cosine similarity), and τ is a temperature hyperparameter controlling the sharpness of the distribution. By contrasting positive pairs against a set of negatives, InfoNCE provides a lower bound on mutual information that becomes tighter as the number of negative samples increases.

Overall training objective. In addition to contrastive alignment across modalities, we incorporate language modeling objectives to further enhance the quality of textual representations. Depending on the underlying text encoder architecture, we employ either a masked language modeling (MLM) loss or a causal language modeling (CLM) loss.

For encoder-based architectures, we adopt the MLM objective, which encourages the model to recover masked tokens based on their surrounding context:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i | \mathbf{x}_{\setminus i}), \tag{10}$$

where $\mathcal{M}$ denotes the set of masked positions, x_i is the ground-truth token at position i, and $\mathbf{x}_{\setminus i}$ represents the input sequence with the token at position i masked out.

For decoder-based or causal architectures, we instead employ a causal language modeling (CLM) objective, which maximizes the likelihood of each token conditioned on all preceding tokens:

$$\mathcal{L}_{\text{CLM}} = -\sum_{i=1}^{T} \log P(x_i | x_{<i}), \tag{11}$$

where $x_{<i}$ denotes the sequence of tokens preceding position i.

The choice of whether to apply a language modeling loss, as well as the selection between MLM and CLM, is treated as a hyperparameter and configured according to the model architecture and training setting. This design enables flexible integration of language modeling supervision while preserving the generality of the multimodal contrastive framework.

The final training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{inter-modal}} + \mathcal{L}_{\text{intra-modal, graph}} + \mathcal{L}_{\text{intra-modal, text}} + \lambda_{\text{LM}} \mathcal{L}_{\text{LM}}, \tag{12}$$

where $\mathcal{L}_{\text{LM}}$ denotes either $\mathcal{L}_{\text{MLM}}$ or $\mathcal{L}_{\text{CLM}}$, and λ_{LM} is a hyperparameter controlling the contribution of the language modeling loss. When $\lambda_{\text{LM}} = 0$, the language modeling objective is disabled.

Training configuration

Table 3 summarizes the hyperparameters used for model training. We optimize the model using AdamW[52] with a low learning rate suitable for pre-trained language models. To stabilize training with small batch sizes, we employ gradient accumulation over 16 batches, yielding an effective batch size of 256 samples per update. Gradient clipping with a maximum norm of 0.05 prevents exploding gradients. We use bfloat16 mixed precision training for computational efficiency and distribute training across 2 GPUs using PyTorch's Distributed Data Parallel (DDP) strategy.

The contrastive learning objective uses InfoNCE loss[53] with a temperature parameter of 0.07. To enhance representation learning, we incorporate intramodal contrastive losses alongside the standard cross-modal objective: graph-graph contrastive learning aligns augmented views of the same crystal structure, while text-text contrastive learning aligns different textual descriptions of the same material. This multi-task learning approach encourages both modality-specific invariances and cross-modal alignment.

All models were implemented using PyTorch[54] and PyTorch Lightning. Graph neural networks were built with PyTorch Geometric[55]. Text encoding leveraged the HuggingFace Transformers library[56]. Training was performed on NVIDIA GPUs with CUDA 11.8. For reproducibility, we set random seeds for Python, NumPy, PyTorch, and CUDA operations. Model checkpoints were saved based on validation loss, retaining the top 2 checkpoints and the final checkpoint.

Training CLaC for 5 epochs on the full Materials Project dataset (containing approximately 140,000 materials with paired crystal

structures and text descriptions) required approximately 20–30 GPU-hours on 4 NVIDIA A6000 GPUs (48GB each). The use of bfloat16 mixed precision training and gradient accumulation enabled efficient training without sacrificing model performance.

Zero-shot evaluation

We evaluate the zero-shot capabilities of CLaC using a multimodal similarity-based framework that requires no task-specific fine-tuning and relies solely on the representations learned during multimodal pre-training. Given a natural-language query, the text encoder produces a text representation, which is compared against precomputed crystal representations of all candidate materials in the shared latent space; the crystal structure with the highest similarity score is selected as the retrieval result (Suppl. Fig. 5). Zero-shot classification and understanding tasks are evaluated using a multiple-choice protocol, in which the correct answer is selected from a finite candidate set. For zero-shot inverse design, a natural-language query describing a target property is used to retrieve the top-k most similar candidate materials, whose property values are then compared against the target.

Text embedding analysis

To evaluate the quality of learned text embeddings and their ability to encode continuous material property information, we conducted a comprehensive similarity analysis across six material properties: band gap (0.0–4.0 eV), formation energy per atom (−3.0–1.0 eV/atom), density (2.0–10.0 g/cm^3), scintillation attenuation length (0.5–20.0 cm), total magnetization (0.0–50.0 μB), and energy above hull (0.0–1.0 eV/atom).

For each property, we generated N query texts using standardized templates (e.g., "material with band gap of X eV") spanning the property range with uniform spacing. Each text query was tokenized (max length: 128 tokens) and processed through the trained text encoder to obtain contextual features from the final hidden layer. The resulting embeddings were L2-normalized to unit vectors, ensuring that similarity measurements represent angular distance in the embedding space.

We computed the pairwise cosine similarity matrix $S \in \mathbb{R}^{N\times N}$ for all property values, where $S_{ij} = e_i \cdot e_j/(||e_i||||e_j||)$represents the similarity between embeddings for values i and j. Three complementary visualizations were generated to assess embedding quality:

(A) Similarity Matrix Heatmap: The full N × N similarity matrix was visualized using a plasma colormap (range: 0-1), with both axes representing property values. This reveals the global structure of the embedding space and identifies whether similar property values cluster together.

(B) Reference Value Decay Curves: We selected three reference values at the 25th, 50th, and 75th percentiles of each property range. For each reference point v_{ref}, we plotted the similarity decay curve showing how cosine similarity decreases as $|v - v_{ref}|$ increases. Well-trained models should exhibit smooth, monotonic decay, with self-similarity ($v = v_{ref}$) at 1.0 and gradually decreasing similarity for more distant values.

(C) Similarity vs. Value Difference: All pairwise comparisons (N(N-1)/2 pairs) were aggregated into a scatter plot with absolute property difference $\Delta v = |v_i - v_j|$ on the x-axis and cosine similarity on the y-axis. To quantify the trend, we binned the data into 20 equal-width intervals and computed the mean similarity ± 1 standard deviation within each bin. This visualization directly tests the hypothesis that embedding similarity should inversely correlate with property value difference.

Data availability

The property-conditioned synthetic narratives and the associated crystal–text pair data generated and used in this study are available on Hugging Face Datasets at https://doi.org/10.57967/hf/2770[57]. The crystal structures and materials properties were obtained from the Materials Project database (https://materialsproject.org). The literature-derived text corpus of approximately four million full-text academic papers used to train the CLaC-Paper baseline cannot be shared publicly because it is subject to copyright and licensing restrictions imposed by the respective publishers, which prohibit redistribution of the full-text content; access to these papers requires a separate licensing agreement with the corresponding publishers. Source data are provided with this paper.

Code availability

The Python code used for this work is available on GitHub at https://github.com/parkyjmit/CLaC and is archived on Zenodo[58] (https://doi.org/10.5281/zenodo.21203843).

References

1. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* **3**, 53 (2017).
2. Sorkun, M. C., Astruc, S., Koelman, J. V. A. & Er, S. An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery. *npj Comput. Mater.* **6**, 106 (2020).
3. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
4. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **7**, 185 (2021).
5. Kaundinya, P. R., Choudhary, K. & Kalidindi, S. R. Prediction of the electron density of states for crystalline compounds with atomistic line graph neural networks (ALIGNN). *JOM* **74**, 1395–1405 (2022).
6. Hoffmann, J. et al. Data-driven approach to encoding and decoding 3-d crystal structures. *arXiv preprint arXiv:1909.00949* (2019).
7. Noh, J. et al. Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
8. Court, C. J., Yildirim, B., Jain, A. & Cole, J. M. 3-D inorganic crystal structure generation and property prediction via representation learning. *J. Chem. Inf. Modeling* **60**, 4518–4535 (2020).
9. Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. & Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. *International Conference on Learning Representations* (2022).
10. Lyngby, P. & Thygesen, K. S. Data-driven discovery of 2D materials by deep generative models. *npj Comput. Mater.* **8**, 232 (2022).
11. Gong, S. et al. Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. *Sci. Adv.* **9**, eadi3245 (2023).
12. Pakornchote, T. et al. Diffusion probabilistic models enhance variational autoencoder for crystal structure generative modeling. *Sci. Rep.* **14**, 1275 (2024).
13. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, 21 (2019).
14. Chen, C.-T. & Gu, G. X. Generative deep neural networks for inverse materials design using backpropagation and active learning. *Adv. Sci.* **7**, 1902607 (2020).
15. Zubarev, D. Y. & Pitera, J. W. in *Cognitive Materials Discovery and Onset of the 5th Discovery Paradigm* (eds Pyzer-Knapp, E. O. & Laino, T.) *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*, Vol. 1326 of *ACS Symposium Series*103-120 (American Chemical Society, 2019). https://doi.org/10.1021/bk-2019-1326.ch006.
16. Leng, C. et al. Fifth paradigm in science: a case study of an intelligence-driven material design. *Engineering* **24**, 126–137 (2023).

17. Wu, M., Zhang, S. & Ren, J. AI-empowered digital design of zeolites: Progress, challenges, and perspectives. *APL Mater.* **13**, 020601 (2025).
18. Radford, A. et al. Meila, M. & Zhang, T. (eds) Learning transferable visual models from natural language supervision. (eds Meila, M. & Zhang, T.) *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, 8748-8763 (PMLR, 2021). https://proceedings.mlr.press/v139/radford21a.html.
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B.Chellappa, R., Matas, J., Quan, L. & Shah, M. (eds) High-resolution image synthesis with latent diffusion models. (eds Chellappa, R., Matas, J., Quan, L. & Shah, M.) *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684-10695 (2022).
20. Ramesh, A. et al. Meila, M. & Zhang, T. (eds) Zero-shot text-to-image generation. (eds Meila, M. & Zhang, T.) *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, 8821–8831 (PMLR, 2021). https://proceedings.mlr.press/v139/ramesh21a.html.
21. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36**, 34892–34916 (2023).
22. Li, C. et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **36**, 28541–28564 (2023).
23. Moro, V. et al. Multimodal Foundation Models for Material Property Prediction and Discovery (2025). http://arxiv.org/abs/2312.00111. ArXiv:2312.00111 [cs].
24. Song, Z., Lu, S., Ju, M., Zhou, Q. & Wang, J. Is large language model all you need to predict the synthesizability and precursors of crystal structures? http://arxiv.org/abs/2407.07016. (2024).
25. Antunes, L. M., Butler, K. T. & Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nat. Commun.* **15**, 10570 (2024).
26. Park, H., Onwuli, A. & Walsh, A. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nat. Commun.* **16**, 4379 (2025).
27. Gan, J. et al. Large Language Models Are Innate Crystal Structure Generators (2025). http://arxiv.org/abs/2502.20933. ArXiv:2502.20933 [cond-mat].
28. Irwin, J. J. & Shoichet, B. K. ZINC- a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Modeling* **45**, 177–182 (2005).
29. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr. Sect. B: Struct. Sci.* **58**, 364–369 (2002).
30. Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation, *APL Mater.* **1**, 011002 (2013).
31. Park, Y. J., Jerng, S. E., Yoon, S. & Li, J. 1.5 million materials narratives generated by chatbots. *Sci. Data* **11**, 1060 (2024).
32. Shrivastava, A., Selvaraju, R. R., Naik, N. & Ordonez, V. CLIP-Lite: Information Efficient Visual Representation Learning with Language Supervision (2023). http://arxiv.org/abs/2112.07133. ArXiv:2112.07133 [cs].
33. Su, B. et al. A Molecular Multimodal Foundation Model Associating Molecule Graphs with Natural Language (2022). http://arxiv.org/abs/2209.05481. ArXiv:2209.05481 [cs].
34. You, Y. et al. Graph contrastive learning with augmentations. *Adv. Neural Inf. Process. Syst.* **33**, 5812–5823 (2020).
35. Beltagy, I., Lo, K. & Cohan, A.Inui, K., Jiang, J., Ng, V. & Wan, X. (eds) SciBERT: A Pretrained Language Model for Scientific Text. (eds Inui, K., Jiang, J., Ng, V. & Wan, X.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620 (Association for Computational Linguistics, Hong Kong, China, 2019). https://aclanthology.org/D19-1371.
36. Fang, X. et al. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **4**, 127–134 (2022).
37. Zeng, X. et al. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nat. Mach. Intell.* **4**, 1004–1016 (2022).
38. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
39. Koker, T., Quigley, K., Spaeth, W., Frey, N. C. & Li, L. Graph Contrastive Learning for Materials. *arXiv preprint arXiv:2211.13408* (2022).
40. Magar, R., Wang, Y. & Barati Farimani, A. Crystal twins: self-supervised learning for crystalline material property prediction. *npj Comput. Mater.* **8**, 231 (2022).
41. Team, G. et al. Gemma 3 Technical Report (2025). http://arxiv.org/abs/2503.19786. ArXiv:2503.19786 [cs].
42. Lian, J., Fu, X., Gong, X., Xiao, R. & Li, H. High-Throughput NEB for Li-Ion Conductor Discovery via Fine-Tuned CHGNet Potential (2025). http://arxiv.org/abs/2507.02334.
43. Nakagawa, A., Kuwata, N., Matsuda, Y. & Kawamura, J. Thin Film Lithium Battery using Stable Solid Electrolyte Li4SiO4 Fabricated by PLD. *ECS Trans.* **25**, 155 (2010).
44. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2020). http://arxiv.org/abs/1802.03426. ArXiv:1802.03426 [stat].
45. Liang, W., Zhang, Y., Kwon, Y., Yeung, S. & Zou, J. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning (2022). http://arxiv.org/abs/2203.02053. ArXiv:2203.02053 [cs].
46. Schütt, K., Unke, O. & Gastegger, M.Meila, M. & Zhang, T. (eds) Equivariant message passing for the prediction of tensorial properties and molecular spectra. (eds Meila, M. & Zhang, T.) *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, 9377-9388 (PMLR, 2021). https://proceedings.mlr.press/v139/schutt21a.html
47. Rhodes, B. et al. Orb-v3: atomistic simulation at scale (2025). http://arxiv.org/abs/2504.06231. ArXiv:2504.06231 [cond-mat].
48. Choudhary, K. et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).
49. Neumann, M. et al. Orb: A Fast, Scalable Neural Network Potential (2024). http://arxiv.org/abs/2410.22570. ArXiv:2410.22570 [cond-mat].
50. Li, Y. et al. Supervision exists everywhere: a data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208* (2021).
51. Gupta, T., Zaki, M., Krishnan, N. A. & Mausam. MatSciBERT: a materials domain language model for text mining and information extraction. *npj Comput. Mater.* **8**, 102 (2022).
52. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization (2019). http://arxiv.org/abs/1711.05101. ArXiv:1711.05101 [cs].
53. Oord, A. v. d., Li, Y. & Vinyals, O. Representation Learning with Contrastive Predictive Coding (2019). http://arxiv.org/abs/1807.03748. ArXiv:1807.03748 [cs].
54. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Info. Proc. Syst.* **32**, (2019).
55. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric (2019). http://arxiv.org/abs/1903.02428. ArXiv:1903.02428 [cs].
56. Wolf, T. et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing (2020). http://arxiv.org/abs/1910.03771. ArXiv:1910.03771 [cs].

57. yjeong/GPT-Narratives-for-Materials Datasets at Hugging Face. https://huggingface.co/datasets/yjeong/GPT-Narratives-for-Materials.
58. Park, Y. J. parkyjmit/CLaC: Contrastive Language–Crystal Pretraining (2026). https://doi.org/10.5281/zenodo.21203843.

Acknowledgements

We gratefully acknowledge the Korea Institute of Science and Technology Information (KISTI) National Supercomputing Center for providing computational resources on the 05-Neuron supercomputer.

Author contributions

Y.J.P. conceived the study, designed and implemented the model, conducted the experiments, and wrote the manuscript. M.K. and C.-W.H. contributed to the experiments and analysis. E.O. provided part of the data and supervised the project. J.L. supervised the project and provided funding. All authors discussed the results and reviewed and revised the manuscript.

Funding

This work was supported by the Defense Threat Reduction Agency (DTRA) under Award No. HDTRA1-20-2-0002 through the Interaction of Ionizing Radiation with Matter (IIRM) University Research Alliance (URA), and by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112490369. Y.J.P. was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government, Ministry of Science and ICT (MSIT) (No. RS-2026-25497753 and RS-2026-25621049).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-76378-7.

Correspondence and requests for materials should be addressed to Elsa Olivetti or Ju Li.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

© The Author(s) 2026