

Supplementary Information

Supplementary Note 1 Related work

Supervised learning for molecules and crystals

Supervised property prediction for atomistic systems has progressed rapidly from hand-crafted descriptors to end-to-end representation learning on molecular graphs and periodic crystals. Early milestones include SchNet, which introduced continuous-filter convolutions to learn chemically meaningful embeddings for both molecules and materials directly from atomic species and coordinates[1]. For crystalline solids, CGCNN established a widely adopted paradigm that represents periodic structures as graphs and performs message passing to predict formation energies, band gaps, and other properties while retaining interpretability at the atom/bond level[2]. Subsequent work highlighted the importance of explicitly encoding higher-order geometric information beyond pairwise distances: DimeNet incorporated directional message passing using angular information with spherical basis functions, substantially improving molecular quantum-property benchmarks[3, 4], while ALIGNN introduced atomistic line-graph message passing to efficiently model bond-angle interactions and improve accuracy across diverse materials-property tasks[5]. More recently, symmetry-aware and equivariant architectures such as PaiNN and NequIP have demonstrated that enforcing rotational equivariance can markedly improve data efficiency and fidelity for predicting tensorial quantities and interatomic potentials[6, 7]. In parallel, large-scale supervised training on extensive relaxation trajectories has enabled broadly transferable models such as M3GNet, which leverages explicit many-body interactions for universal interatomic potentials and downstream materials screening[8].

Self-supervised learning for molecules and crystals

GNNs have significantly impacted material property predictions [2, 9] by providing an alternative way for complex quantum chemical computations. However, the costly labeling process requiring expensive quantum chemical calculations is the bottleneck in GNN application. To address this challenge, various self-supervised learning (SSL) methods have been proposed. Wang et al. [10] reduced the need for expensive quantum chemical labeling by applying contrastive learning to unlabeled 10 million molecules using molecular GNNs. Koker et al. [11] emphasized the importance of augmentation techniques when applying contrastive learning to encoding periodic crystals as graphs with CGCNNs as encoders. Magar et al. [12] improved the performance of CGCNNs by applying Barlow twins [13], a self-supervised learning method that does not require negative pairs.

Scientific learning

A growing body of work explores how foundation models can be adapted to scientific learning, where the goal is not only linguistic competence but also faithful modeling of scientific entities, modalities, and decision workflows. In biomedicine, BioGPT demonstrated that domain-specific generative pretraining on large-scale biomedical literature can substantially improve performance on generation-centric and information extraction tasks, motivating specialized LMs for scientific corpora[14]. Complementarily, multimodal scientific understanding has been studied through figure-centric supervision: SciCap introduced a large-scale scientific figure-caption dataset and benchmarked end-to-end caption generation, highlighting both the feasibility and challenges of grounding scientific language in visual evidence[15]. In materials science, recent systems are attempting to connect language models to downstream scientific

workflows, but significant limitations remain. CSLLM presents a coordinated language model pipeline for crystal synthesizability and synthetic pathway prediction[16]. MatLLMSearch integrates large-scale language models into an evolutionary crystal structure search loop, demonstrating that LLM can guide exploration[17]. However, existing methods lack an explicit alignment between language and material representations of structure-property relationships.

Multimodal learning for molecules

In the molecular domain, performing multimodal contrastive learning between graphs and text has been demonstrated to improve representation ability by encoding knowledge from an unstructured domain-specific corpus. Zeng et al. [18] proposed to encode both SMILES [19, 20] representation and text corpus of molecules simultaneously using a BERT [21] model as backbone. To train the proposed model, they extracted a corpus from the semantic scholar [22], a database of over 136M published scientific literature, to construct 10k molecule-text pairs. Su et al. [23] proposed to apply the contrastive learning paradigm with distinguishing graph and text branches. To overcome limited training data, authors utilized pre-trained models in the biochemical molecular domain for both GNN, GraphCL [24], and text encoder, SciBERT [25]. Liu et al. [26] pre-trained their model on a larger training set of 280k pairs, and both previous studies enabled zero-shot tasks and showed improved performance compared to unimodal models on downstream tasks.

Multimodal learning for crystals

Multimodal learning in crystalline materials has not been explored extensively. Das et al. [27] proposed CrysMMNet, which simultaneously integrates graph inputs and text inputs describing the structure of materials generated by a Robocrystallographer [28]. However, this approach focuses on creating joint representations and using them only to predict material properties and does not address issues related to zero-shot tasks and downstream tasks using natural language. More recently, Moro et al. introduced MultiMat [29], a multimodal foundation model that aligns multiple material modalities—including crystal structures, electronic density of states, charge density, and textual descriptions—into a shared latent space via contrastive learning. While MultiMat demonstrates strong performance in property prediction and latent-space-based material discovery, its use of text is primarily auxiliary and it is not explicitly designed for natural-language-driven querying or zero-shot language-based tasks. In contrast, Suzuki et al. proposed Contrastive Language-Structure Pre-training (CLaSP) [30], which directly aligns crystal structures with natural language derived from the materials science literature. By leveraging publication titles and abstracts as supervision, CLaSP enables text-based retrieval and zero-shot screening of crystal structures using human-interpretable language, highlighting a complementary direction that emphasizes language-centric downstream tasks over multimodal property alignment.

Supplementary Note 2 GPT-synthesized narratives

We introduce an example of GPT-synthesized narratives of KNiIO_6 . These textual data were synthesized based on realistic properties from publicly available materials databases. This is a promising alternative to address the data scarcity and bias issues we currently face.

The substance in question has the chemical formula KNiIO_6 . It forms with a total energy per atom of -2.96134 eV/atom and an energy above hull of 1.77847 eV/atom. Its formation energy per atom is -0.85383 eV/atom. The substance has a band gap of 1.196 eV, which is the amount of energy required to move an electron from the top of the valence band to the bottom of the conduction band. The scintillation attenuation length of the substance is 2.722038 cm, which means it can absorb and re-emit light. The space group of the substance is symbolized by P312, with a crystal system of trigonal. It has a density of 4.09 g/cm³, with a volume of 129.62222 Å³. The substance has a total magnetization of 0.872899 $\mu\text{B}/\text{f.u.}$, indicating its magnetic properties. Its enthalpy per atom is -4.552144 eV/atom, which is the heat generated or absorbed in chemical reactions. The substance is an oxide, and it is unstable (stable: False). The substance in question is a rare-earth metal oxide known as potassium nickelate. Given its unique properties, there are a number of possible application areas for this material, including:

1. Photovoltaics: With a band gap of 1.196 eV, potassium nickelate could be utilized for solar energy conversion. The material could potentially absorb light in the visible and near-infrared range, generating electron-hole pairs that could be collected to produce an electrical current. The scintillation attenuation length of 2.722038 cm should also make it efficient at converting light to electrical energy.
2. Magnetic storage: With a total magnetization of 0.872899 $\mu\text{B}/\text{f.u.}$, potassium nickelate could be utilized for magnetic storage. For example, the material could be used in the fabrication of hard disk drives, where data is stored magnetically. The magnetic properties of the material would allow it to be magnetized and demagnetized easily, enabling data to be written and read.
3. Catalysis: Given its enthalpy per atom of -4.552144 eV/atom, potassium nickelate could act as a catalyst to facilitate chemical reactions. The heat generated or absorbed in chemical reactions could be used to drive forward or reverse reactions, respectively. For example, the material could be used to facilitate the conversion of biomass into biofuels by breaking down complex organic molecules into simpler components.
4. Magneto-optical devices: With its unique optical and magnetic properties, potassium nickelate could be utilized in the fabrication of magneto-optical devices. For example, the material could be used in the production of magneto-optical sensors, which are used for magnetic field sensing. Magneto-optical devices could also be used to manipulate light using magnetic fields.

Overall, the properties of potassium nickelate make it a promising candidate for a variety of applications, ranging from solar energy conversion to magnetic storage and catalysis. Its unique combination of optical and magnetic properties makes it particularly appealing for use in magneto-optical devices.

Supplementary Figure 1 Generated description about KNiIO_6 . The narrative describes the material property itself and infers potential applications of it.

Supplementary Note 3 Brute-force extracted narratives

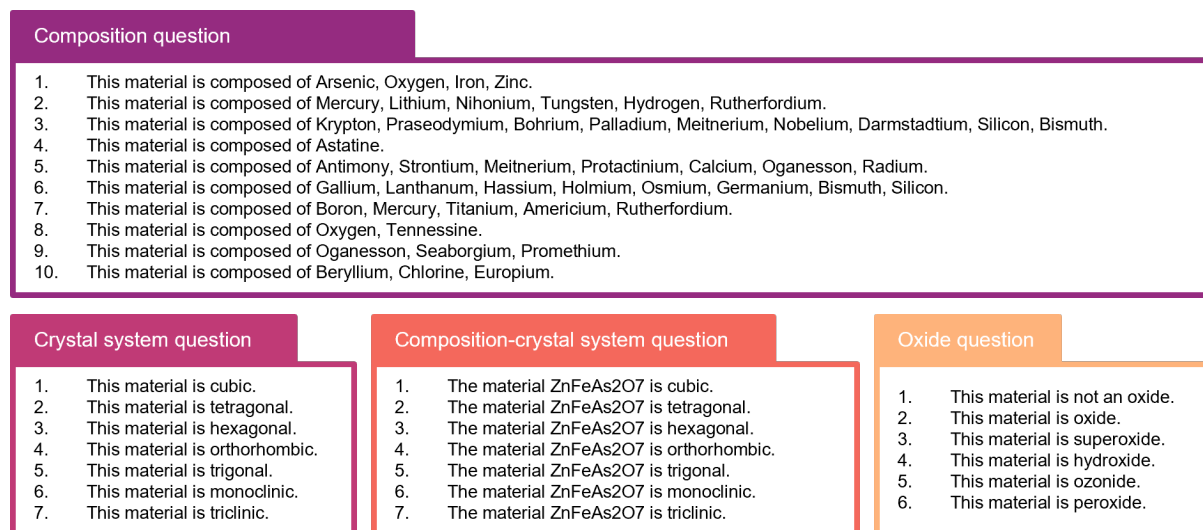
We introduce an illustrative example of a brute-force extracted textual description of CaSeO_3 . This text was reconstructed based on the content and style of descriptions commonly used in prior materials studies, rather than being directly copied from original source.

X-ray diffraction (XRD) analysis is a reliable technique for identifying surface crystalline phases and atomic arrangements. The XRD patterns of all adsorption products exhibit similar features (Fig. S4), and here we focus on the results obtained for CaO powder and the sample denoted as "CaO-75." As shown in Fig. 8, the CaO (001) plane is the dominant exposed facet in the CaO powder, further validating the use of a CaO (001) surface slab model to represent the CaO surface. Both the adsorption product (red) and the pristine CaO powder (black) show diffraction peaks corresponding to CaO (PDF card No. 48-1467), along with minor contributions from $\text{Ca}(\text{OH})_2$ (PDF card No. 44-1481). The XRD data indicate that the adsorption of SeO_2 does not significantly alter the bulk crystal structure of CaO. Nevertheless, subtle differences in the diffraction patterns can be observed in the 10° – 30° 2θ range, which are attributed to surface-adsorbed SeO_2 molecules inducing slight changes in the atomic distribution near the surface. In contrast, XRD patterns of adsorption products obtained after high-temperature treatment (500°C for 48 h) reveal the formation of crystalline CaSeO_3 . This observation suggests that CaO surfaces bearing adsorbed SeO_2 can undergo recrystallization into a new phase, namely CaSeO_3 , when assisted by reaction-enhancing conditions such as elevated temperature and prolonged reaction time. Further investigation is required to elucidate the detailed mechanisms underlying this transformation.

Supplementary Figure 2 Illustrative brute-force extracted description of CaSeO_3 . The text is synthetically reconstructed and does not reproduce any original description verbatim[31].

Supplementary Note 4 Generating multiple-choice problems for zero-shot multimodal understanding evaluation

To evaluate the multimodal recognition performance of our model, we proposed multiple-choice problems for various subjects. To evaluate the compositional understanding of the developed model, we generate 10 sentences including only one correct answer. The correct answer is generated from the given materials while other incorrect answers are randomly synthesized as described in Supplementary Fig. 3.



Supplementary Figure 3 Examples of multiple choice problems for evaluating various understandings of the CLaC model. Multiple-choice problems are designed to evaluate the basic zero-shot understanding for materials science.

Supplementary Note 5 Baseline models

To systematically evaluate the zero-shot multimodal capabilities of the proposed CLaC model, we compare against several baseline approaches that represent distinct modeling paradigms.

Random baseline

As a lower-bound reference, we include a random baseline in which answers are selected uniformly at random from the candidate set.

Multimodal contrastive baselines

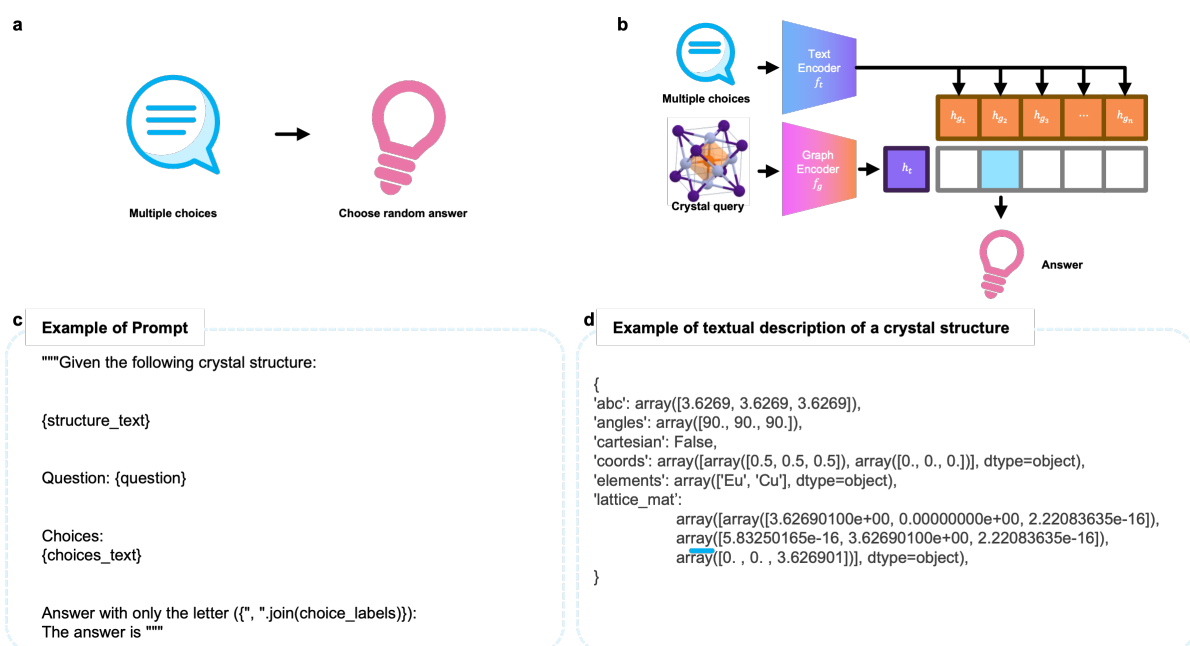
We evaluate three variants of CLIP-style multimodal models that share the same architecture but differ in training status and textual supervision:

1. **Untrained model**, where both the graph encoder and text encoder are randomly initialized and no multimodal training is performed. This baseline assesses whether the task can be solved without learned cross-modal alignment.
2. **CLaC-Paper**, trained using text extracted from academic literature via brute-force matching of chemical formulas. This setting reflects prior approaches that rely on naturally occurring, literature-derived text for multimodal supervision.
3. **CLaC-Synthetic**, trained using property-conditioned synthetic narratives generated by large language models. These narratives explicitly encode quantitative material properties and associated semantic descriptions.

For all multimodal baselines, crystal structures are encoded as graphs using a graph neural network, and textual inputs are encoded using a Transformer-based text encoder. Similarity between crystals and texts is computed in a shared embedding space learned via contrastive training.

Text-only large language model baseline

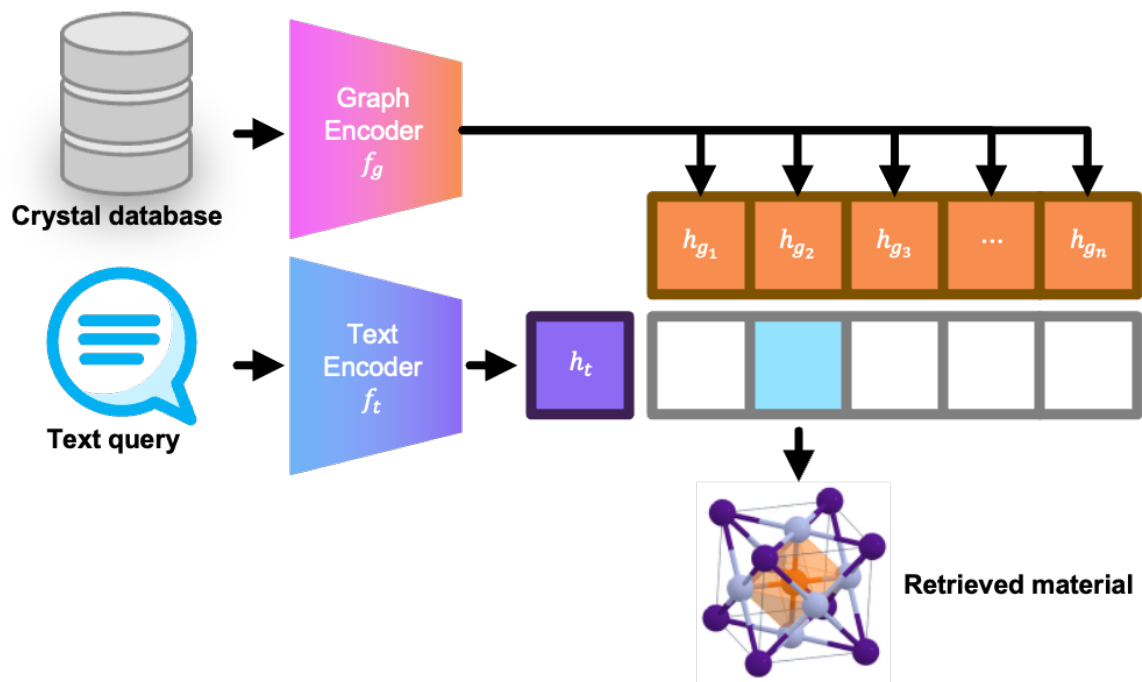
To compare against unimodal language-based reasoning, we include a text-only large language model baseline using Gemma-3-1b-it. In this setting, the model is evaluated via in-context instruction following. The crystal structure to be queried and the candidate answer set are serialized into textual form and provided within the prompt. The model is instructed to select the most appropriate answer based solely on the textual description.



Supplementary Figure 4 Baseline and comparison models for zero-shot multimodal retrieval. Schematic illustration of the evaluation baselines and comparison settings used in this study. **a**, Random baseline, in which answers are selected uniformly at random from the candidate set. **b**, Multimodal contrastive models, including an untrained model, CLaC-Paper trained on literature-derived text, and CLaC-Synthetic trained on property-conditioned synthetic narratives. In these models, crystal structures are encoded by a graph encoder and textual descriptions are encoded by a text encoder, with similarities computed in a shared embedding space. **c,d**, Prompt-based evaluation of a text-only large language model (Gemma-3-1b-it), where natural-language instructions are provided to guide selection among candidate materials. Panels **c** and **d** illustrate examples of the prompts and textual inputs used for instruction-following inference.

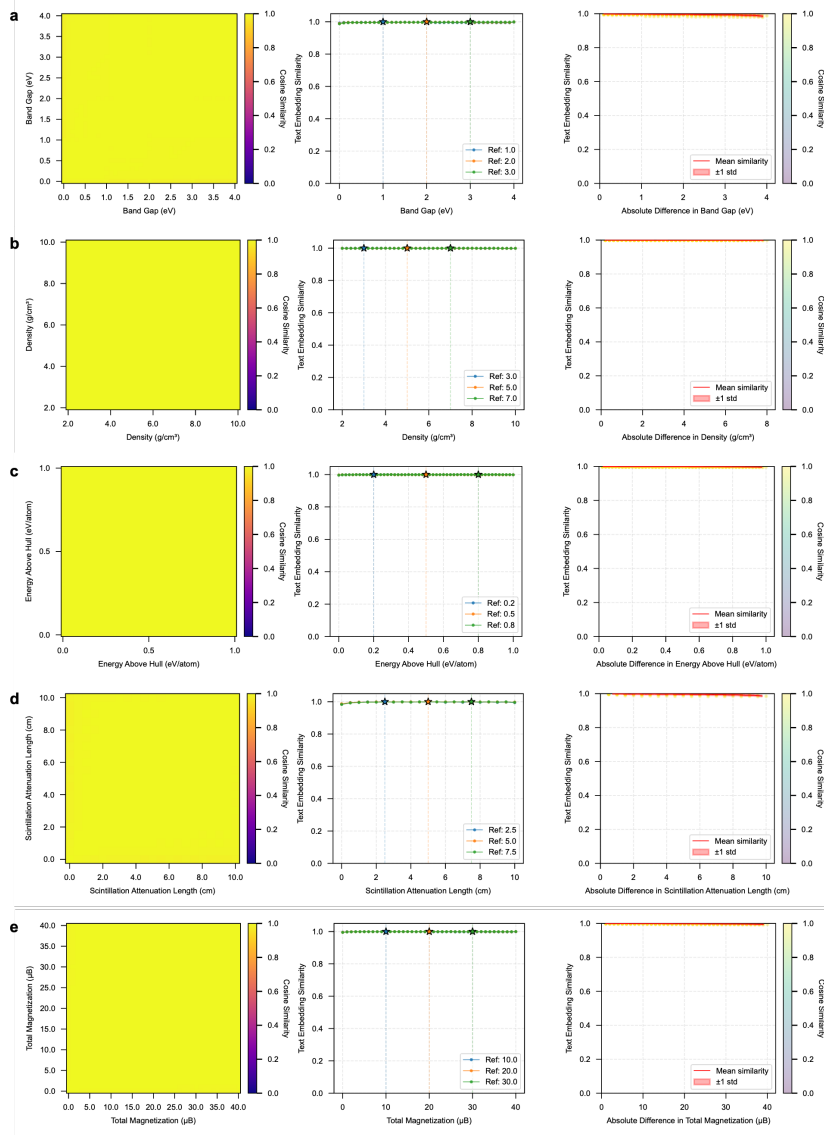
Supplementary Note 6 Zero-shot retrieval task

Supplementary Fig. 5 schematically illustrates the zero-shot crystal-text retrieval procedure described in the Methods (Zero-shot evaluation).

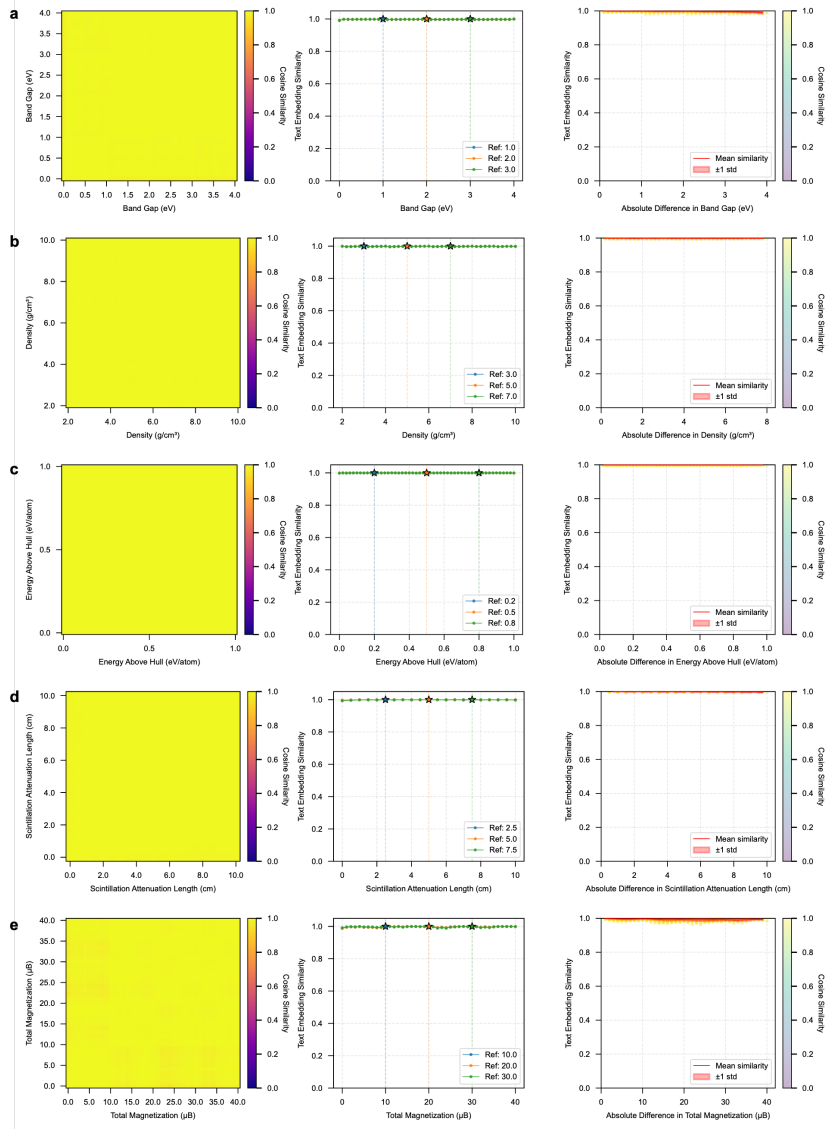


Supplementary Figure 5 Schematic illustration of zero-shot crystal-text retrieval via multimodal contrastive representations. Crystal structures and textual descriptions are independently encoded using a graph encoder and a text encoder, respectively. The graph encoder produces a set of crystal embeddings for candidate materials, while the text encoder maps a natural-language query to a text embedding. Retrieval is performed by computing similarities between the text embedding and all crystal embeddings in the shared latent space. The crystal structure with the highest similarity score is selected as the retrieval result.

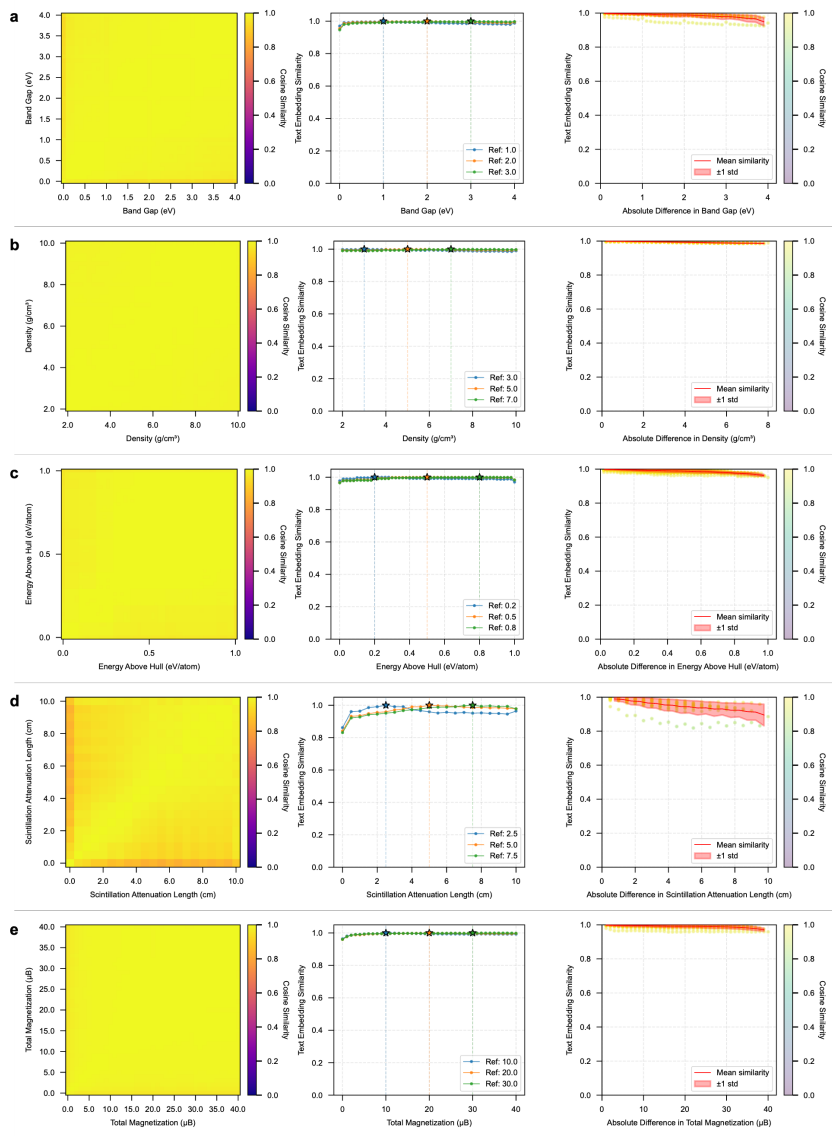
Supplementary Note 7 Text embedding similarity



Supplementary Figure 6 Semantic structure of numerical values within the text embedding space for untrained CLaC model. 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.

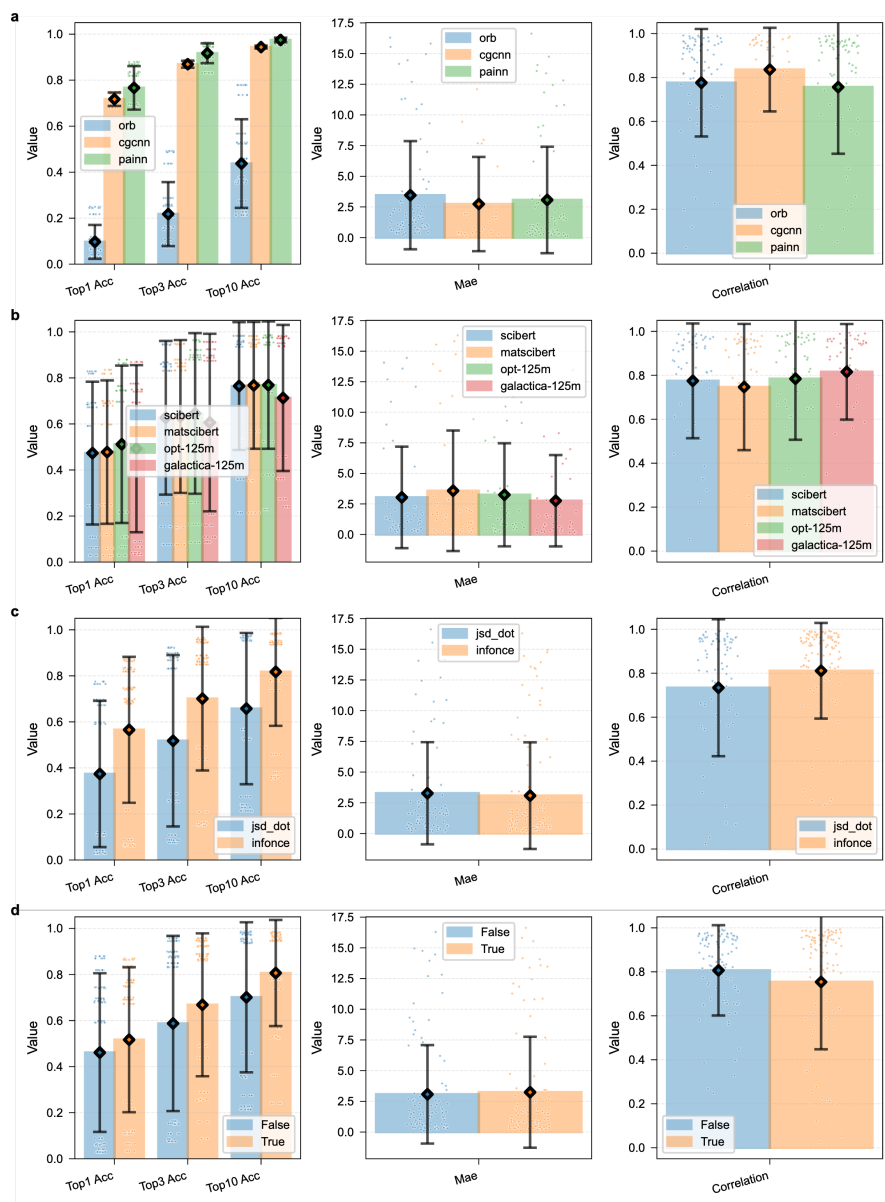


Supplementary Figure 7 Semantic structure of numerical values within the text embedding space for CLaC-Paper. Cosine similarity analysis of text embeddings generated using the prompt “material with [property] of [value] [unit]” for five continuous properties: **a**, Band Gap, **b**, Density, **c**, Energy 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.

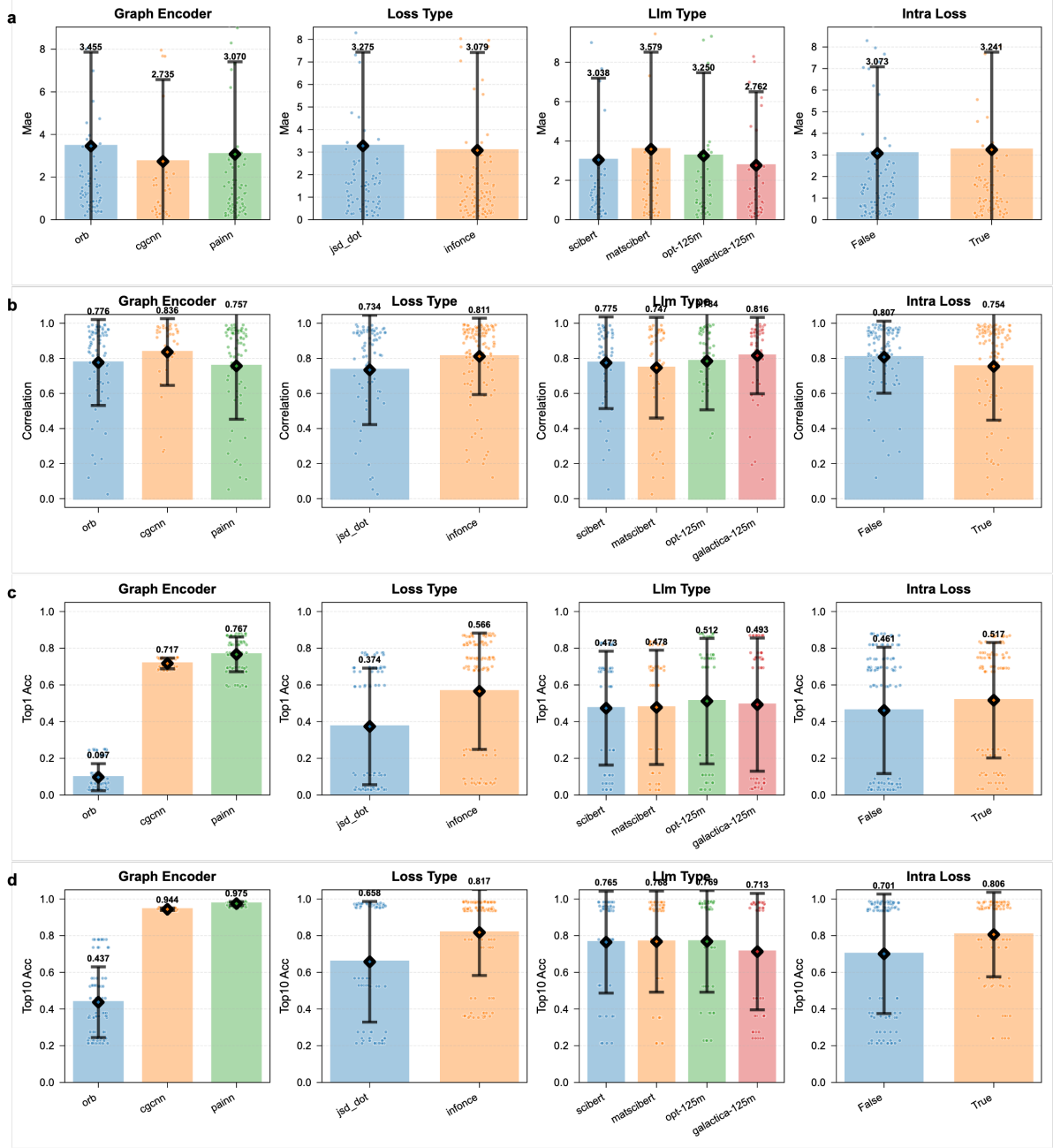


Supplementary Figure 8 Semantic structure of numerical values within the text embedding space for CLaC-Synthetic trained with LM loss. Cosine similarity analysis of text embeddings generated using the prompt template “material with [property] of [value] [unit]” for five continuous properties: **a**, 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.

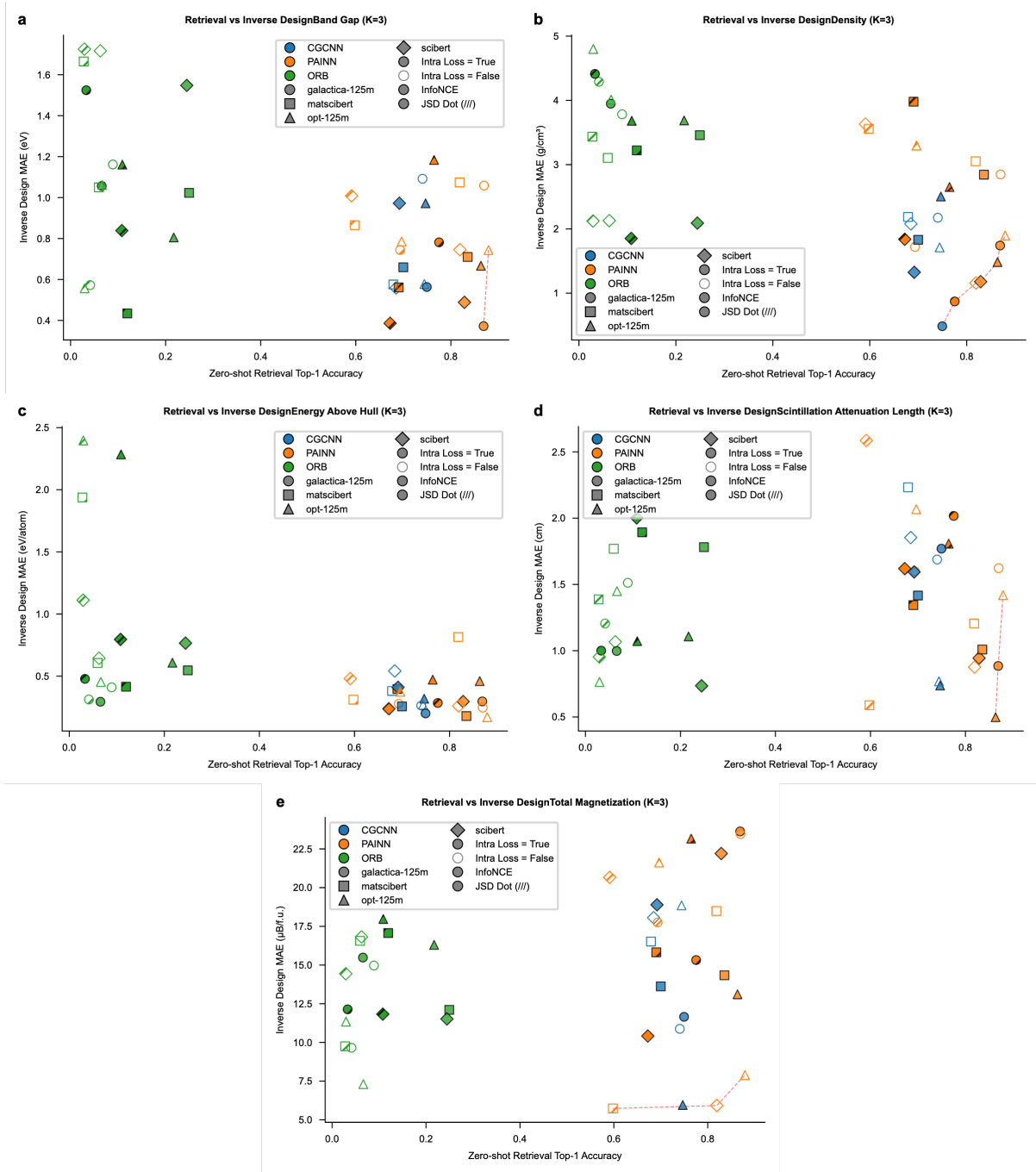
Supplementary Note 8 Ablation study



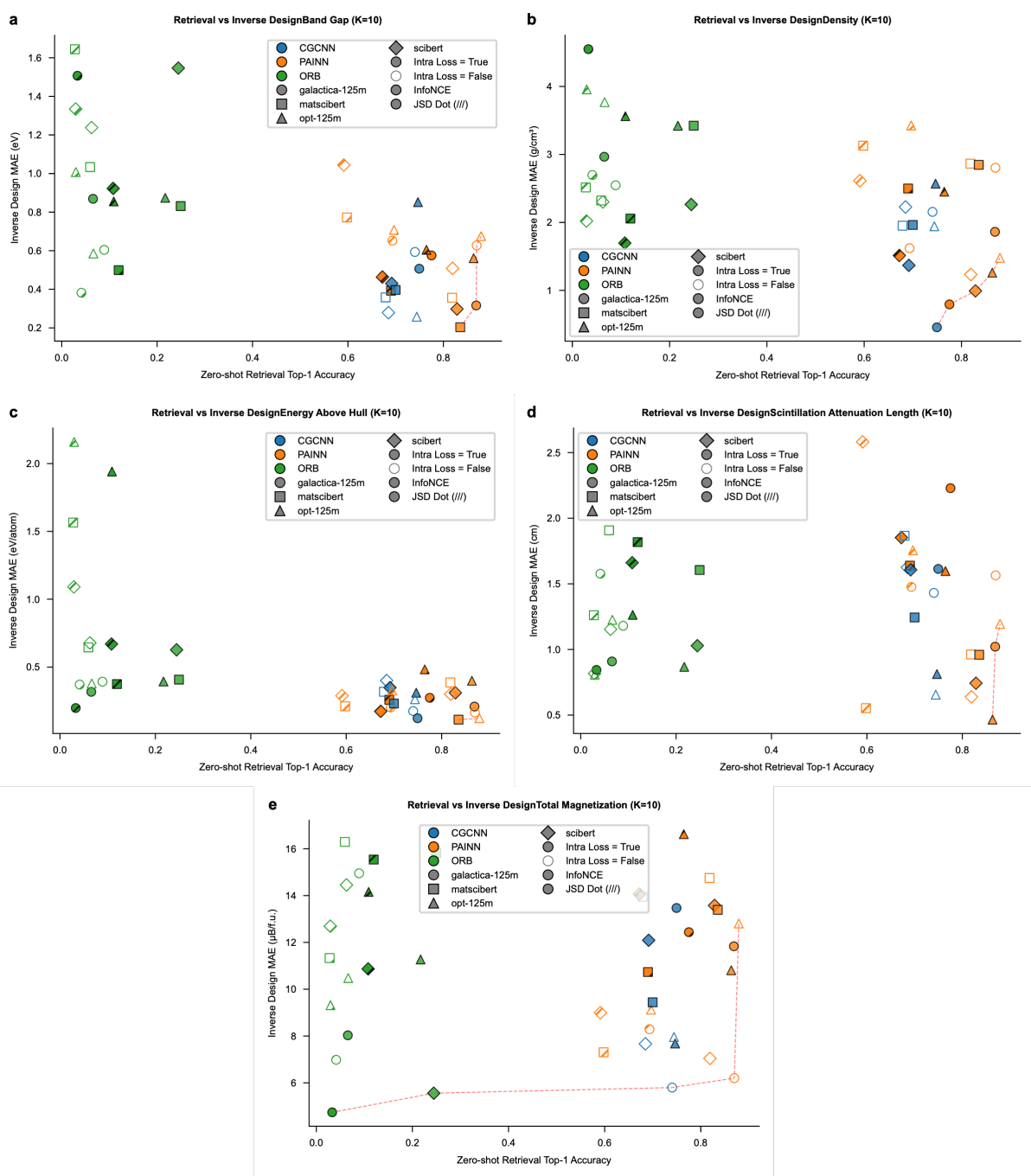
Supplementary Figure 9 Ablation studies on architectural components, text encoders, and training objectives. Rows correspond to different ablation settings, and columns report evaluation metrics: top-1, top-3, and top-10 accuracy for zero-shot retrieval (left), mean absolute error (MAE) for property regression (center), and Pearson correlation coefficient (right). **a**, Comparison of graph encoder architectures, including OrbNet-based (orb), CGCNN (cgcnn), and PaiNN (painn). **b**, Comparison of text encoders, including SciBERT, MatSciBERT, OPT-125M, and Galactica-125M. **c**, Comparison of contrastive objectives, including Jensen–Shannon divergence with dot-product similarity (jsd_{dot}) and InfoNCE loss. **d**, Effect of enabling or disabling sentence-level text augmentation during training. Bars indicate mean performance across tasks, with black diamonds denoting the average and error bars representing standard deviation. Individual dots correspond to per-task results.



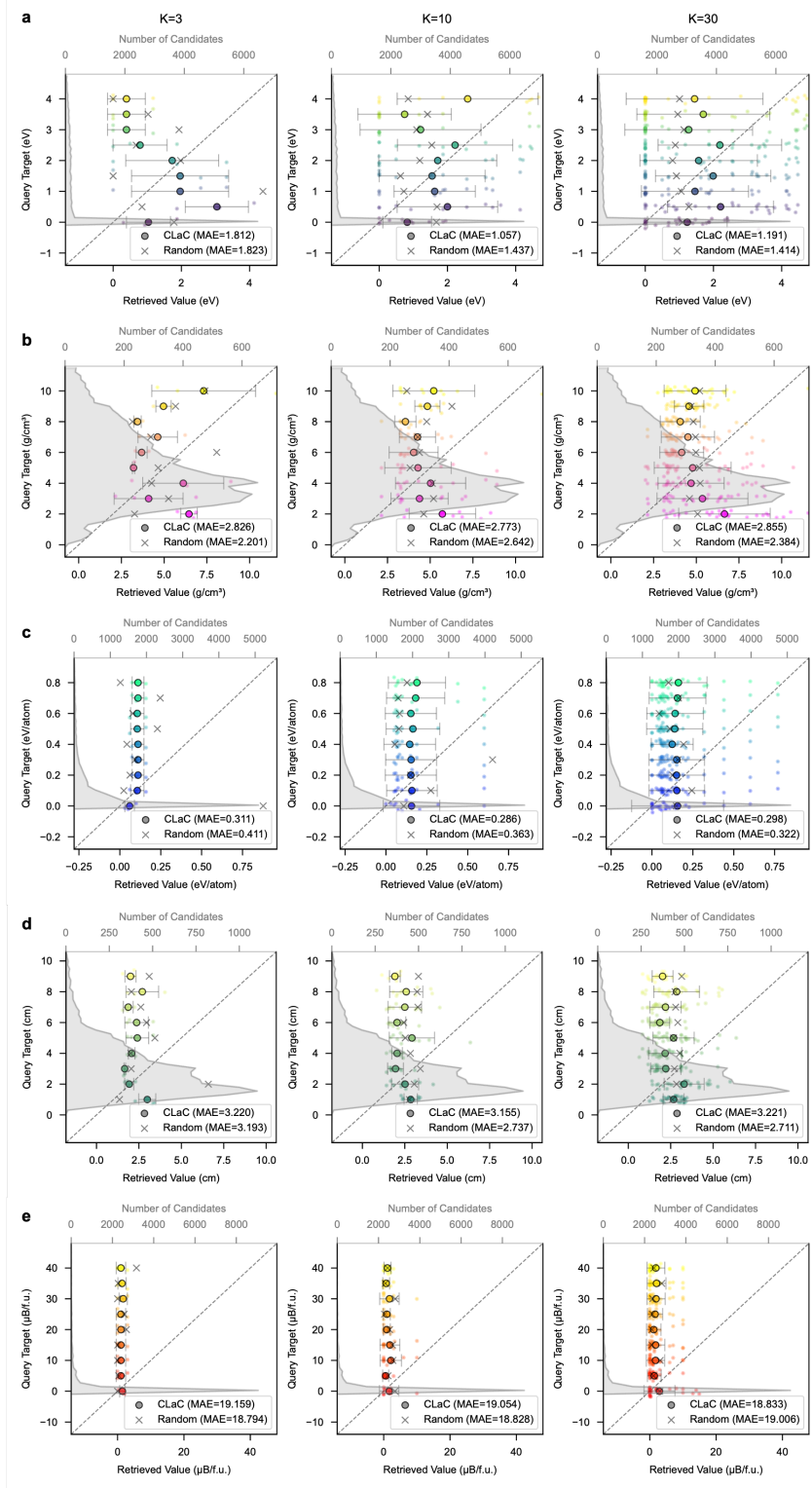
Supplementary Figure 10 Ablation results across model components and evaluation metrics. Ablation analysis evaluating the effect of different design choices on multimodal performance. Columns correspond to ablated components, including the graph encoder architecture, contrastive loss type, text encoder (LLM) choice, and the use of an intra-modality loss. Rows correspond to evaluation metrics: **a**, mean absolute error (MAE) for property regression; **b**, Pearson correlation coefficient; **c**, top-1 accuracy; and **d**, top-10 accuracy for zero-shot retrieval tasks. Bars denote mean performance across evaluation tasks, black diamonds indicate the average, and error bars represent standard deviation. Individual dots correspond to results on individual tasks or properties.



Supplementary Figure 11 Pareto front between zero-shot retrieval accuracy and inverse design performance for 3 candidates. Scatter plots showing the Pareto relationship between zero-shot retrieval accuracy (x-axis, top-1 accuracy) and inverse design performance (y-axis, mean absolute error) across different material properties, evaluated with K=3 candidates. Panels correspond to **a**, band gap; **b**, density; **c**, energy above hull; **d**, scintillation attenuation length; and **e**, total magnetization. Each point represents a distinct model configuration, with colors indicating the graph encoder architecture, marker shapes denoting the text encoder (LLM) type, and marker styles reflecting training objectives and loss variants. Configurations located toward the lower-right region of each panel define the Pareto front, corresponding to models that jointly achieve high retrieval accuracy and low inverse design error.



Supplementary Figure 12 Pareto front between zero-shot retrieval accuracy and inverse design performance for 10 candidates. Scatter plots showing the Pareto relationship between zero-shot retrieval accuracy (x-axis, top-1 accuracy) and inverse design performance (y-axis, mean absolute error) across different material properties, evaluated with K=10 candidates. Panels correspond to **a**, band gap; **b**, density; **c**, energy above hull; **d**, scintillation attenuation length; and **e**, total magnetization. Each point represents a distinct model configuration, with colors indicating the graph encoder architecture, marker shapes denoting the text encoder (LLM) type, and marker styles reflecting training objectives and loss variants. Configurations located toward the lower-right region of each panel define the Pareto front, corresponding to models that jointly achieve high retrieval accuracy and low inverse design error.



Supplementary Figure 13 Effect of removing the intermodal loss on zero-shot inverse design retrieval. Zero-shot inverse design retrieval results for CLaC models trained without the intermodal contrastive loss, evaluated across different candidate set sizes $K=3, 10$, and 30 . Each column corresponds to a different value of K , and rows correspond to different target material properties: **a**, band gap; **b**, density; **c**, energy above hull; **d**, scintillation attenuation length; and **e**, total magnetization. For each query, the retrieved material property value is plotted against the target query value. The diagonal dashed line indicates ideal retrieval, where the retrieved value matches the target. Circles denote predictions from the CLaC model without intermodal loss, while crosses indicate a random baseline. Shaded regions represent the distribution of candidates. Mean absolute error (MAE) is reported for each setting. Across all properties and values of K , removing the intermodal loss leads to degraded inverse design performance compared to the full model (see main text).

References

- [1] Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. Schnet—a deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **148**, 241722 (2018). Publisher: AIP Publishing LLC.
- [2] Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **120**, 145301 (2018). Publisher: APS.
- [3] Gasteiger, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123* (2020).
- [4] Gasteiger, J., Giri, S., Margraf, J. T. & Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115* (2020).
- [5] Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials* **7**, 185 (2021). Publisher: Nature Publishing Group UK London.
- [6] 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). URL <https://proceedings.mlr.press/v139/schutt21a.html>.
- [7] Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications* **13**, 2453 (2022). URL <https://www.nature.com/articles/s41467-022-29939-5>. Publisher: Nature Publishing Group.
- [8] Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science* **2**, 718–728 (2022). URL <https://www.nature.com/articles/s43588-022-00349-3>. Publisher: Nature Publishing Group.
- [9] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Precup, D. & Teh, Y. W. (eds) *Neural message passing for quantum chemistry*. (eds Precup, D. & Teh, Y. W.) *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, 1263–1272 (PMLR, 2017). URL <https://proceedings.mlr.press/v70/gilmer17a.html>.
- [10] Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **4**, 279–287 (2022). Publisher: Nature Publishing Group UK London.
- [11] Koker, T., Quigley, K., Spaeth, W., Frey, N. C. & Li, L. Graph Contrastive Learning for Materials. *arXiv preprint arXiv:2211.13408* (2022).
- [12] Magar, R., Wang, Y. & Barati Farimani, A. Crystal twins: self-supervised learning for crystalline material property prediction. *npj Computational Materials* **8**, 231 (2022). Publisher: Nature

Publishing Group UK London.

- [13] Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Meila, M. & Zhang, T. (eds) *Barlow twins: Self-supervised learning via redundancy reduction*. (eds Meila, M. & Zhang, T.) *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, 12310–12320 (PMLR, 2021). URL <https://proceedings.mlr.press/v139/zbontar21a.html>.
- [14] Luo, R. *et al.* BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics* **23**, bbac409 (2022). URL <http://arxiv.org/abs/2210.10341>. ArXiv:2210.10341 [cs].
- [15] Hsu, T.-Y., Giles, C. L. & Huang, T.-H. K. SciCap: Generating Captions for Scientific Figures (2021). URL <http://arxiv.org/abs/2110.11624>. ArXiv:2110.11624 [cs].
- [16] 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? (2024). URL <http://arxiv.org/abs/2407.07016>. ArXiv:2407.07016 [cond-mat].
- [17] Gan, J. *et al.* Large Language Models Are Innate Crystal Structure Generators (2025). URL <http://arxiv.org/abs/2502.20933>. ArXiv:2502.20933 [cond-mat].
- [18] Zeng, Z., Yao, Y., Liu, Z. & Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications* **13**, 862 (2022). Publisher: Nature Publishing Group UK London.
- [19] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28**, 31–36 (1988). Publisher: ACS Publications.
- [20] Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences* **29**, 97–101 (1989). Publisher: ACS Publications.
- [21] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [22] Lo, K., Wang, L. L., Neumann, M., Kinney, R. & Weld, D. S. S2ORC: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782* (2019).
- [23] Su, B. *et al.* A Molecular Multimodal Foundation Model Associating Molecule Graphs with Natural Language (2022). URL <http://arxiv.org/abs/2209.05481>. ArXiv:2209.05481 [cs].
- [24] You, Y. *et al.* Graph contrastive learning with augmentations. *Advances in neural information processing systems* **33**, 5812–5823 (2020).
- [25] 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). URL <https://aclanthology.org/D19-1371>.
- [26] Liu, S. *et al.* Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence* **5**, 1447–1457 (2023). URL <https://www.nature.com/articles/s42256-023-00759-6>. Publisher: Nature Publishing Group.
- [27] Das, K., Goyal, P., Lee, S.-C., Bhattacharjee, S. & Ganguly, N. Evans, R. J. & Shpitser, I. (eds) *CrysMMNet: Multimodal representation for crystal property prediction*. (eds Evans, R. J. & Shpitser, I.) *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, Vol. 216 of *Proceedings of Machine Learning Research*, 507–517 (PMLR, 2023). URL <https://proceedings.mlr.press/v216/das23a.html>.
- [28] Ganose, A. M. & Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications* **9**, 874–881 (2019). Publisher: Cambridge University Press.
- [29] Moro, V. *et al.* Multimodal Foundation Models for Material Property Prediction and Discovery (2025). URL <http://arxiv.org/abs/2312.00111>. ArXiv:2312.00111 [cs].
- [30] Suzuki, Y. *et al.* Contrastive Language-Structure Pre-training Driven by Materials Science Literature (2025). URL <http://arxiv.org/abs/2501.12919>. ArXiv:2501.12919 [cs] version: 1.
- [31] Fan, Y., Zhuo, Y. & Li, L. SeO₂ adsorption on CaO surface: DFT and experimental study on the adsorption of multiple SeO₂ molecules. *Applied Surface Science* **420**, 465–471 (2017). URL <https://www.sciencedirect.com/science/article/pii/S0169433217312771>.