nature machine intelligence

Article

https://doi.org/10.1038/s42256-026-01214-y

A multimodal large language model for materials science

Received: 4 April 2025

Accepted: 6 March 2026

Published online: 24 April 2026

Check for updates

Yingheng Tang[1,8] ✉, Wenbin Xu[2,8] ✉, Jie Cao[3], Weilu Gao[4] ✉, Steven Farrell[2], Benjamin Erichson[5,6], Michael W. Mahoney[5,6,7], Andy Nonaka[1] & Zhi Jackie Yao[1] ✉

Understanding and predicting the properties of inorganic materials is crucial for accelerating advancements in materials science and driving applications in energy, electronics and beyond. Integrating material structure data with language-based information through multimodal large language models (LLMs) offers great potential to support these efforts by enhancing human–artificial intelligence interaction. However, a key challenge lies in integrating atomic structures at full resolution into LLMs. In this work, we introduce MatterChat, a versatile structure-aware multimodal LLM that unifies material structural data and textual inputs into a single cohesive model. MatterChat uses a bridging module to effectively align a pretrained universal machine learning interatomic potential with a pretrained LLM, reducing training costs and enhancing flexibility. Our results demonstrate that MatterChat greatly improves performance in material property prediction and human–artificial intelligence interaction, surpassing general-purpose LLMs such as GPT-4. We also demonstrate its usefulness in applications such as more advanced scientific reasoning and step-by-step material synthesis.

In silico material discovery traditionally relies on high-fidelity methods like density functional theory[1] and ab initio molecular dynamics[2]. However, prohibitive computational costs limit their scalability for high-throughput screening. Moreover, many advanced materials lack the mechanistic understanding due to complex compositions and phase instabilities. Consequently, breakthroughs in functional materials, such as correlated oxides[3,4] and quantum materials[5,6], have often been serendipitous rather than driven by theory. Achieving reliable, scalable and predictive design of materials requires a paradigm shift.

With the rise of AI in materials science, there has been a surge of methods aiming to overcome these limitations, ranging from surrogate models[7,8] to MLIPs[9–13] and generative models[14,15]. These models enable rapid predictions, accelerate large-scale simulations and facilitate the generation of novel materials. As a result, they have greatly advanced fields such as energy storage[16], electronics[17], catalysis[18] and biomedical applications[19]. Among these promising ML approaches, graph-based models in materials science have become increasingly popular due to their versatile graph representation of atomistic systems in which each atom is represented as a node and chemical bonds to neighbouring atoms are represented as edges. Although these graph-based methods have shown success in accurately predicting material properties, they typically lack the capacity to handle tasks that require understanding scientific context, literature-based insights and domain-specific language[20]. In particular, these models do not support human–AI

[1]Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [2]National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [3]NSF National AI Institute for Student-AI Teaming, University of Colorado at Boulder, Boulder, CO, USA. [4]Department of Electrical and Computer Engineering, The University of Utah, Salt Lake City, UT, USA. [5]Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [6]International Computer Science Institute, Berkeley, CA, USA. [7]Department of Statistics, University of California at Berkeley, Berkeley, CA, USA. [8]These authors contributed equally: Yingheng Tang, Wenbin Xu. ✉e-mail: ytang4@lbl.gov; wenbinxu@lbl.gov; weilu.gao@utah.edu; jackie_zhiyao@lbl.gov

interaction through user prompts or textual descriptions, making it difficult to incorporate expert domain knowledge and user-specified requests to close the feedback loop.

This bottleneck has inspired exploration into large language models (LLMs). LLMs like BERT[21], GPT[22], Mistral[23], Llama[24] and DeepSeek[25] have shown promise in scientific question–answer[26] and information retrieval[27]. Recent efforts have incorporated LLMs to solve materials problems[28,29] by leveraging pretrained or multimodal architectures.

Recent benchmarks, including MatSci-NLP[30], MaScQA[31], HoneyBee[32] and others[33–37], provide valuable baselines for evaluating domain-specific reasoning. However, these methods primarily rely on text-based representations—such as chemical formulas[28], SMILES strings[29,38], and Crystallographic Information Files (CIF)[39]. While informative, these textual inputs often fail to explicitly capture the complex 3D spatial relationships and local environments inherent in atomic structures. Consequently, they exhibit inferior property prediction performance compared with graph-based models[40]. Universal MLIPs[11] now allow for the extraction of rich structural information from atomistic embeddings, offering a feasible pathway for multimodal integration.

In this work, we present MatterChat, a multimodal LLM for materials science. MatterChat utilizes a modular framework that bridges pretrained language and materials models. By freezing the weights of the LLM and the material encoder, our system enables plug-and-play flexibility with components like CHGNet[41] or many-body atomic cluster expansion (MACE)[11]. This design preserves foundation model generalization and facilitates future extensions without retraining the entire architecture. MatterChat integrates structure data with textual queries, overcoming traditional LLM limitations in quantitative prediction. It maintains robust human–AI interaction and enables advanced reasoning for synthesis guidance. Embedding analysis confirms that MatterChat effectively preserves structure–property information, supporting a multimodal retrieval-augmented generation (RAG) approach to enhance inference robustness.

Results

Overview of MatterChat

Figure 1a presents the architecture of MatterChat, designed to process both material structures and user requests as inputs to generate text-based outputs for tasks such as material property prediction, structural analysis and descriptive language generation. MatterChat consists of three core components: the material processing branch, the language processing branch and the bridge model. The material processing branch extracts atomic-level embeddings from material structures represented as graphs. These embeddings are then processed by the bridge model, which uses trainable queries to produce language model-compatible embeddings. Finally, the language processing branch processes the user's text-based prompt (for example, 'What is the formation energy of the material?') into language embeddings. These embeddings are then combined with the query embeddings generated by the bridge model and fed into the LLM to produce the final output in text format. Below, we provide the details of each component.

Material processing branch. The material processing branch encodes material structures as graphs that capture the atomic local environment. We specifically utilize the encoder modules of state-of-the-art graph-based universal MLIP models, such as CHGNet[41] and MACE[11], as feature extractors to process these graphs. These encoders are pretrained on a diverse dataset of materials, encompassing a wide range of symmetries, compositions and bonding types, enabling it to effectively model complex atomic interactions and structural details. By capturing essential compositional features, such as atomic types and chemical bonds, along with spatial features like bond angles, these pretrained encoders generate high-quality atom embeddings that are both physically meaningful and well suited for downstream tasks.

Language processing branch. The language processing branch is used to process the user's text-based prompts, such as requests for property predictions, chemical formulas, space group information or other material characteristics. We use the Mistral 7B LLM[23], one of the latest open-source LLMs, chosen for its exceptional performance across a wide range of scientific and non-scientific tasks. This branch processes each prompt, transforming it into dense embeddings that capture the semantic content of the enquiry. These embeddings are then combined with the query embeddings processed by the bridge model using a structured fusion approach, allowing the model to effectively incorporate both textual and material information. This integration enables the LLM to generate precise and contextually relevant responses tailored to the user's specific material-related prompts.

Bridge model. To facilitate the integration between atom embeddings and the language processing branch, we developed a bridge model inspired by the BLIP2 architecture[42] based on a multilayer transformer framework. This bridge model includes 32 trainable query vectors that interact with atom embeddings using an alternating attention mechanism. Cross-attention in even-numbered layers extracts key features from the atom embeddings, whereas self-attention in odd-numbered layers enhances representational depth. This approach refines the atom embeddings into query embeddings that are most connected to text (Fig. 1a). Finally, these refined representations are mapped to LLM-compatible embeddings via a linear projection layer.

Figure 1b,c provides an overview of the dataset of crystalline structures used in our training set. Figure 1b visualizes the material distribution on the periodic table, highlighting that the dataset evenly spans a diverse range of elements up to plutonium. Figure 1c depicts the distribution of crystalline structures by space group across the dataset. The dataset was curated from the Materials Project[43] and contains 142,899 material structures. For each structure, we generated a corresponding text-based dataset encompassing 12 tasks: three descriptive tasks (chemical formula, space group and crystal system) and nine property prediction tasks. These property prediction tasks include metallicity, direct bandgap, stability, experimental observation, magnetic status, magnetic order, formation energy, energy above the hull and bandgap (Fig. 1a). Further details regarding the training scheme, hyperparameters and dataset curation are provided in Methods.

Figure 2 illustrates examples of a human–AI interaction with MatterChat across a diverse range of material property prediction and analysis tasks. It shows MatterChat's ability to effectively address a broad spectrum of user prompts ranging from fundamental material attributes (for example, chemical formulas, space groups and crystal system) to complex material properties (for example, thermal stability, bandgaps, formation energies and energy above the hull). Figure 2a shows three interactive examples of material property prompts from randomly selected materials from the Materials Project database. The top left panel presents a human–AI query interface with MatterChat for the material with an mp-id of mp-1001021. It provides a detailed profile including the chemical formula $Y_2Zn_4Se_2$, its crystalline structure denoted by the space group *Fd*-3*m*, and electronic properties such as a bandgap of 0.23870 eV. The interface also addresses the material's lack of thermal stability. The top middle panel shows the interaction example with the material with an mp-id of mp-1028281. It provides a comprehensive breakdown of the material's composition attributes, including its chemical formula ($Mg_{14}VSb$) and its space group (*Amm*2). The interaction further predicts that the material is both magnetic and metallic, and its formation energy is estimated at 0.07219 eV per atom. The top right panel provides an interaction example with MatterChat of the material with an mp-id of mp-10198. This panel informs the user's query about the chemical composition Mn_3PdN and its cubic crystal

https://doi.org/10.1038/s42256-026-01214-y

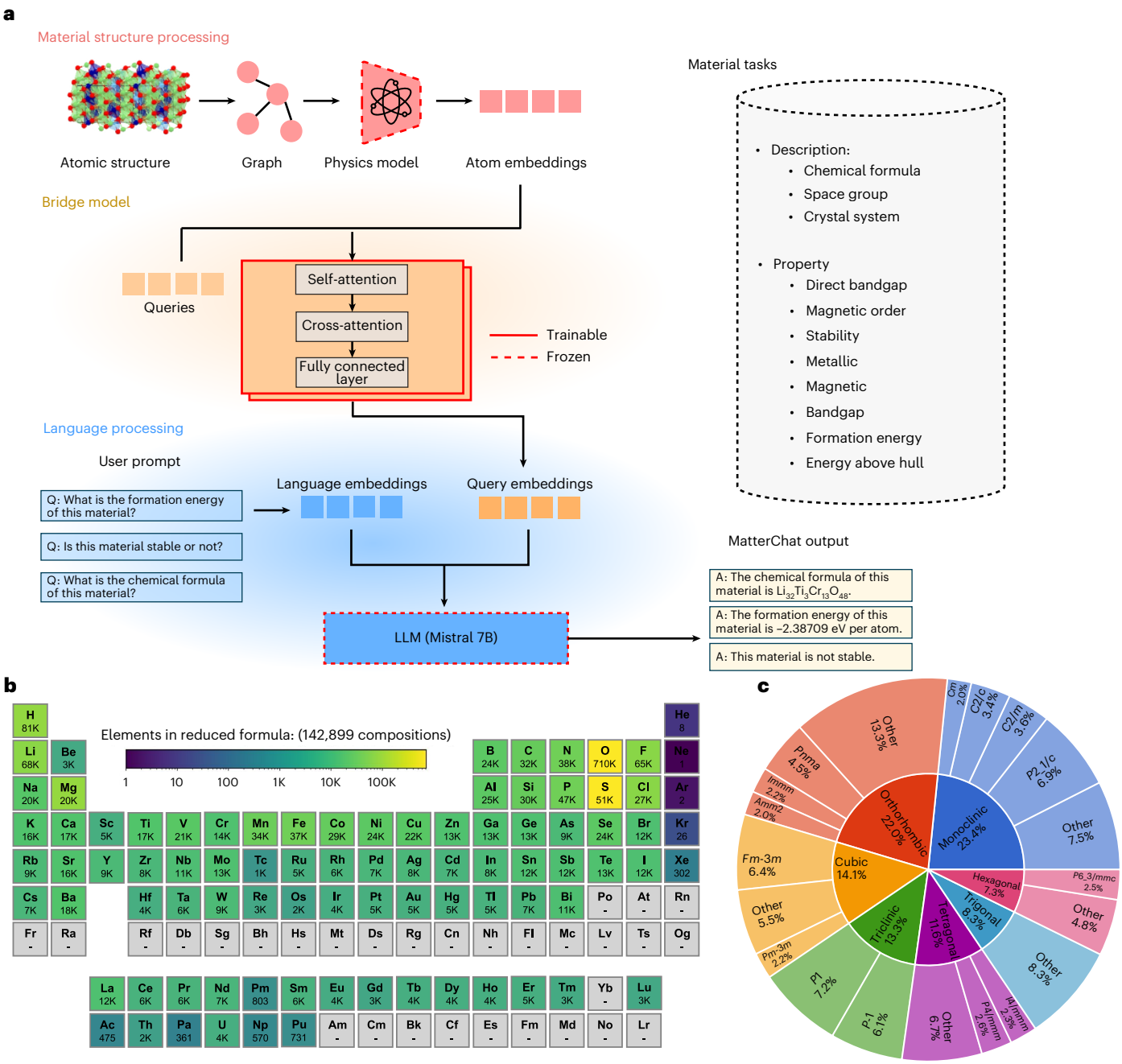

Fig. 1 | Overview of MatterChat: a modular multimodal LLM for material-based question answering. a, MatterChat architecture: the system includes a material encoder that generates atom embeddings and an LLM that processes language data. These components are connected by a trainable bridge model, which aligns material structure with natural language to support tasks such as material description and property prediction. **b**, Elemental distribution across 142,899 compositions, representing the dataset's compositional diversity. **c**, Dataset distribution shown by space groups (outer ring) and crystal systems (inner ring), illustrating structural variation within the dataset.

structure, with the space group classified as *Pm*-3*m*. Additionally, it estimated that the material possesses an indirect bandgap, which is an important characteristic for applications in electronics. MatterChat also accurately predicts the ferromagnetic magnetic behaviours that the material exhibits, and it mentions its energy above hull value at 0.01357 eV per atom. In the bottom panel, we present a comparative evaluation of MatterChat's performance on formation energy evaluation tasks for newly discovered materials from GNoME[44]. The model was compared against commercial LLMs, like Gemini[45], GPT-4o[46] and DeepSeek[25]. The results show MatterChat's superior accuracy in estimating formation energies, consistently delivering predictions closer to the ground truths. For example, MatterChat's formation energy predictions for mp-3202380 and mp-3206774 show a remarkable alignment with the ground-truth values. These results demonstrate MatterChat's ability to integrate structural and textual data seamlessly for a wide range of material property tasks.

Figure 3 demonstrates MatterChat's advanced reasoning capabilities, showing how it leverages the comprehensive knowledge base of LLMs to address complex materials science challenges. By using a multimodal query system, MatterChat effectively combines material structure data with textual reasoning. This integration facilitates a working memory scheme[47], which enables the model to provide

https://doi.org/10.1038/s42256-026-01214-y

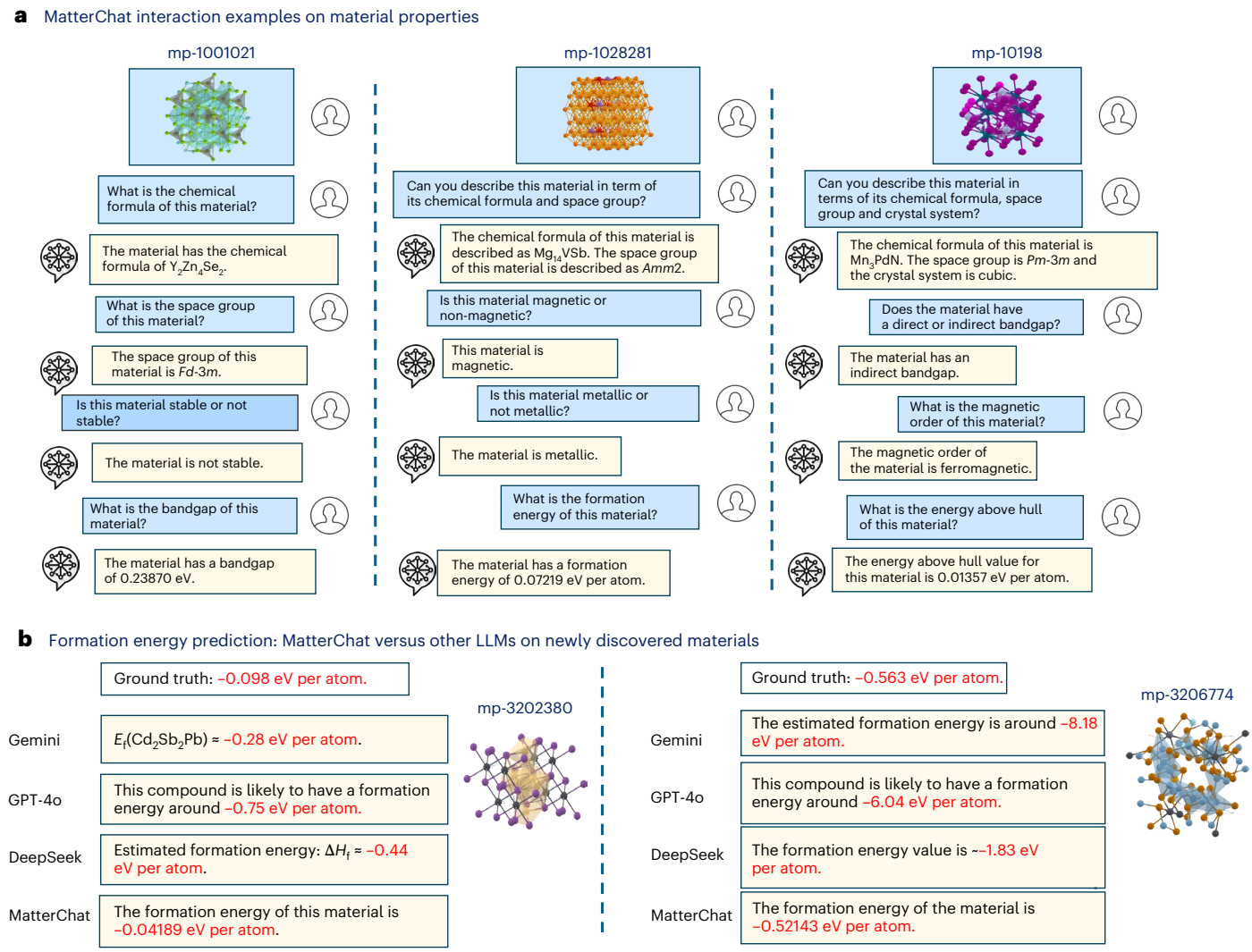

Fig. 2 | MatterChat accurately predicts material properties and outperforms state-of-the-art LLMs. a, Illustration of multimodal material property queries using MatterChat. The model accurately interprets user prompts to predict chemical formulas, crystallographic properties, stability, electronic bandgap, magnetic order and energy metrics of materials. The three panels demonstrate the framework's ability to address diverse materials science enquiries, showing its alignment of graph-based and textual embeddings for precise question answering. **b**, Comparative evaluation of formation energy predictions for newly discovered material from GNoME[44]. Predictions from MatterChat compared against the ground-truth values along with evaluations from commercial LLMs (Gemini[45], GPT-4o[46] and DeepSeek[25]). The results show the accuracy and stability of the MatterChat in quantitative material evaluation tasks, which closely aligns with the ground truth, demonstrating its ability to integrate material graph embeddings for precise property prediction.

domain-specific reasoning, detailed synthesis procedures and explanations that are deeply grounded in the structural properties of materials. Figure 3a presents the chat log for silicon with the space group of *cmcm*. MatterChat not only retrieves the chemical formula and the correct space group but it also provides a rationale for the structural instability of this silicon phase. The model explains that the *cmcm* space group exhibits a higher energy per unit cell compared with the thermodynamically stable cubic diamond structure of silicon, making it less likely to occur under standard conditions. Figure 3b illustrates an interaction regarding a popular semiconductor material gallium nitride (GaN). Here MatterChat accurately identifies the chemical formula and space group (*P63mc*), and generates a detailed metal–organic chemical vapour deposition synthesis protocol that aligns with established experimental standards. Specifically, the model identifies trimethylgallium and ammonia as precursors within an 800–1,000 °C temperature window, directly matching landmark methods such as those reported elsewhere[48,49]. This demonstrates the model's ability to leverage inherited knowledge to provide practical, grounded and experimentally viable scientific reasoning. Figure 3c explores an interaction for a widely used ferrite material, yttrium iron garnet. MatterChat is able to take the structure and generate detailed text descriptions. Additionally, MatterChat can further generate a synthesis protocol for YIG that aligns with established experimental procedures[50]. By identifying the correct 3:5 mixing ratio of Y_2O_3 and Fe_2O_3 and specifying critical parameters like the 5 °C min^{-1} thermal rate, the model demonstrates its capability to apply domain-specific knowledge in accordance with standard practices and characterization techniques like X-ray diffraction and scanning electron microscopy[50]. MatterChat generates synthesis guidance via a modular two-stage process without task-specific supervision. First, structural attributes—including formula, space group and crystal system—are extracted via a frozen encoder and tokenized to form a persistent working memory. Second, the LLM generates responses conditioned on this context, aligning with a symbolic memory framework[47] in which the inferred material facts anchor reasoning. By utilizing the LLM's inherited knowledge with explicit structural signals, MatterChat produces physically plausible, literature-aligned synthesis

https://doi.org/10.1038/s42256-026-01214-y

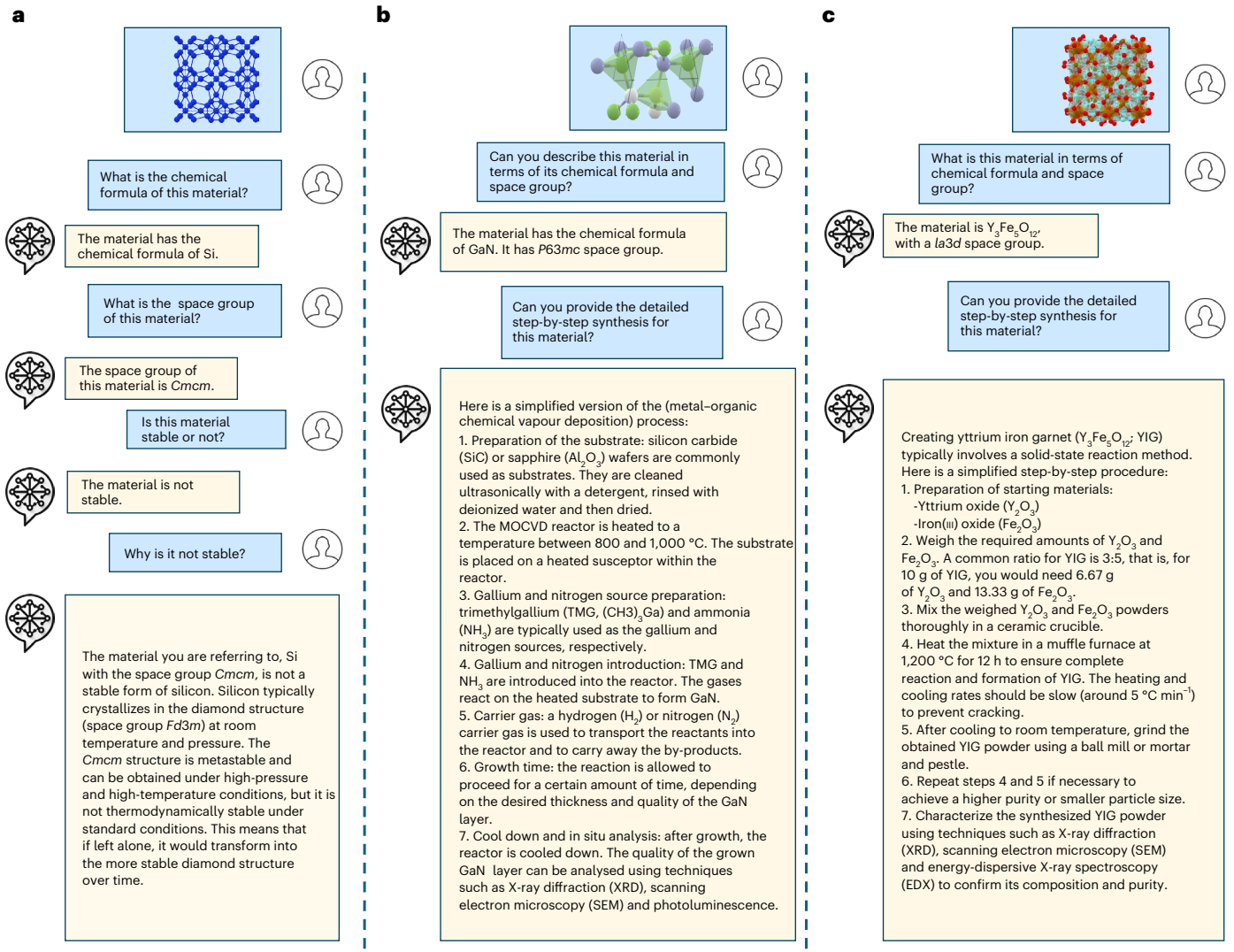

Fig. 3 | MatterChat has the ability to solve more sophisticated tasks inherited from the pretrained LLM. a, Material property query for silicon (Si), including its chemical formula, space group, stability, and the reasoning for why it is not stable under standard conditions. **b**, Highlights a material query for GaN, providing its chemical formula, space group, and a step-by-step synthesis procedure using methods like hydride vapour phase epitaxy, metal–organic chemical vapour deposition and molecular-beam epitaxy. **c**, Material query interaction, yttrium iron garnet (YIG; $Y_3Fe_5O_{12}$), detailing its chemical formula, space group and a simplified step-by-step synthesis procedure using the solid-state reaction method.

outputs. This modularity ensures a clear boundary between material perception and linguistic reasoning, enhancing both interpretability and structure-conditioned generation.

MatterChat-extracted embeddings contain structural and property information

We further explore MatterChat's ability to leverage material structural information by providing a detailed visualization/clustering analysis with the uniform manifold approximation and projection (UMAP) dimension reduction technique[51]. Figure 4a–e shows comprehensive visualizations of embeddings processed by the bridge model, with all material samples that contain silicon (Si), carbon (C) and their composites compounds (for example, SiC and Si_xC_y) from the Materials Project database[52]. UMAP was used to reduce the embeddings from an original 4,096 dimensions to two dimensions, with the x and y axes corresponding to the first and second reduced dimensions, respectively.

Figure 4a presents the visualizations containing all the selected materials; each sample is colour coded with a structure similarity score[53]. The clustering generally follows distinctions in chemical compositions. Additionally, materials with the same atomic composition are grouped into separate clusters based on crystalline structural differences (for example, carbon with diamond versus graphite crystalline structure). Figure 4b,d shows the zoomed-in visualizations of clustering results for materials consisting exclusively of Si and SiC compositions. Figure 4d shows the gradient of structure similarity scores, ranging from blue (low similarity) to red (high similarity), demonstrating how closely related structural features result in spatial proximity within the embedding space. However, an interesting exception is observed with SiC (Fig. 4b): despite its identical composition and similar structural phases, two distinct clusters of SiC emerge, suggesting that factors beyond composition and structure alone influence their separation. To further explore factors that influence clustering, we labelled the samples according to their formation energy, with results displayed for SiC (Fig. 4c) and Si (Fig. 4e). These figures clearly show a trend from low to high formation energy. This analysis reveals that clusters grouped by structural similarity also align closely in terms of formation energy. Such findings indicate the model's ability to produce embeddings that not only differentiate

https://doi.org/10.1038/s42256-026-01214-y

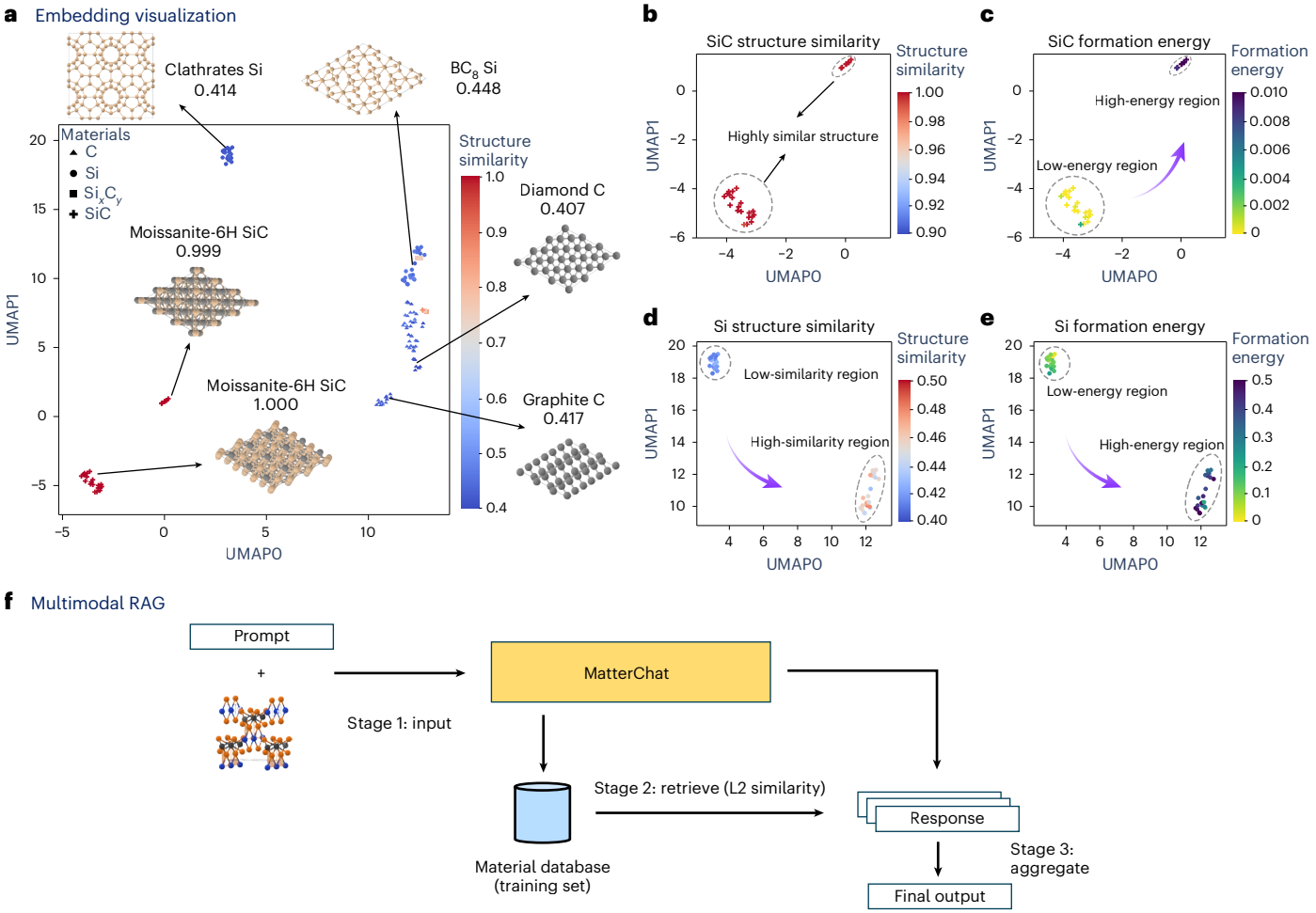

Fig. 4 | UMAP visualization of structural embeddings extracted from the bridge model. a, Visualization of samples containing Si and C elements from the Materials Project database, showing how materials cluster based on their structural embeddings extracted from the bridge model. The value indicates the structural similarity calculated using the SOAP descriptor in combination with the REMatch kernel (Methods). **b**,**c**, Visualizations of the SiC subgroup colour coded by structural similarity (**b**) and formation energy (**c**). The two clusters exhibit high structural similarity, with formation energy further assisting in distinguishing between them. **d**,**e**, Visualizations of Si subgroup colour coded by structural similarity (**d**) and formation energy (**e**). The two clusters demonstrate a smooth transition in both structural similarity and formation energy, indicating that both factors captured by the structural embeddings contribute to the observed clustering. **f**, Proposed multimodal RAG for robust prediction.

structural characteristics but also correlate with key material properties. To evaluate the generalization ability of MatterChat across a broader chemical space, we extended the structural embedding analysis beyond the initial silicon–carbon system to diverse material families (Supplementary Figs. 1–4). These include various iron-based compounds (oxides, sulfides, nitrides and carbides), as well as transition metal oxides containing iron, copper, cobalt and molybdenum. Similar trends are observed. The UMAP visualizations of the learned embeddings demonstrate that the model effectively captures the distinctive characteristics of different inorganic compounds. Distinct compound types form well-separated clusters in terms of both average structural similarity and formation energy similarity, whereas smooth transitions are observed within individual clusters. These findings suggest that both structural and property-related information are encoded in the learned representations, which is consistent with the property-supervised training of the model. Overall, the results indicate that the representations learned by the bridge model are robust and exhibit strong discriminative power across diverse material classes. Given that the embeddings derived from the bridge model preserve both material structure and property-relevant information, we implemented a multimodal RAG mechanism during inference (Fig. 4f). Instead of relying solely on a single output from MatterChat for each query–sample pair, we now retrieve additional information of two more samples from the material pool (training set). This retrieval is based on the L2 similarity between the embeddings of the sample material and those in the pool. After that, we aggregate all three results to get the final output by applying a majority-voting strategy for classification tasks and averaging for quantitative tasks. Such a method could further enhance the overall robustness of MatterChat across different tasks. The details of the visualization method are provided in Methods.

Comprehensive quantitative analysis for all material tasks

To evaluate MatterChat, we benchmarked its performance across nine tasks on the evaluation set (14,290 samples) against open-source LLMs (Vicuna[54], Mistral[23]) and physical ML models (SchNet[55], CHGNet[41]) and MACE[11]. For LLM baselines, material structures were serialized as CIF-derived text within identical prompt structures (Methods).

In classification (Fig. 5a–f), including metallicity, stability and magnetism, MatterChat consistently outperformed all baselines. In particular, it achieved higher accuracy than specialized physical models like CHGNet, demonstrating that integrating graph-based data with natural language reasoning provides a more holistic representation of material chemistry.

https://doi.org/10.1038/s42256-026-01214-y

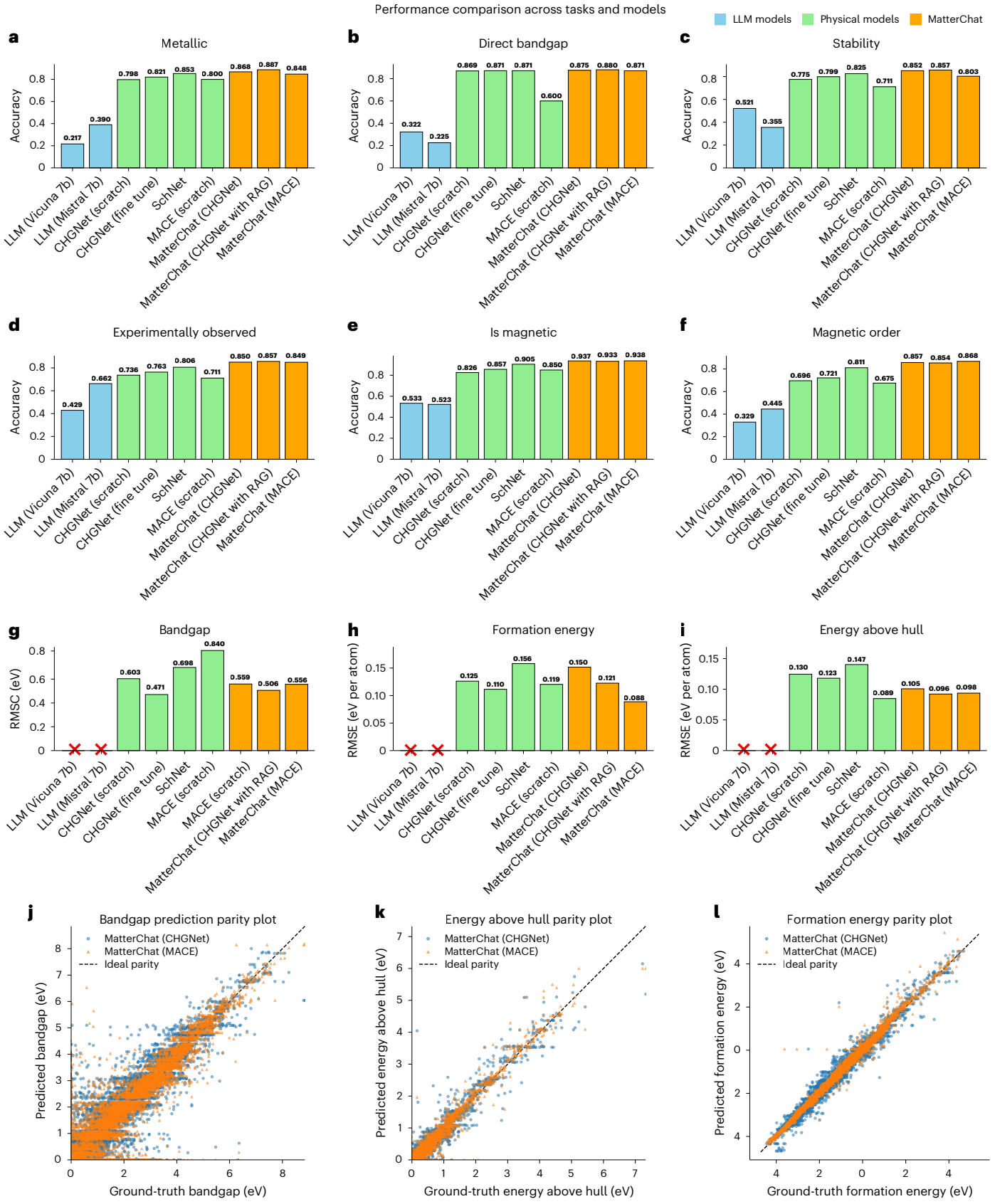

Fig. 5 | Performance comparison of MatterChat, open-source LLMs and physical pretrained models across nine material property tasks. **a**–**f**, Classification task accuracies for predicting whether a material is metallic (**a**), has a direct bandgap (**b**), is thermodynamically stable (**c**), is experimentally observed (**d**), is magnetic (**e**) and is of magnetic ordering type (**f**), in which MatterChat consistently outperforms other models. **g**–**i**, RMSE results for numerical property predictions, demonstrating MatterChat's superior precision in bandgap (**g**), formation energy (**h**) and energy above the hull (**i**) tasks. **j**–**l**, Parity plots for bandgap (**j**), energy above the hull (**k**) and formation energy (**l**), illustrating the alignment between predicted values from MatterChat (with both CHGNet and MACE encoders) and ground-truth values.

https://doi.org/10.1038/s42256-026-01214-y

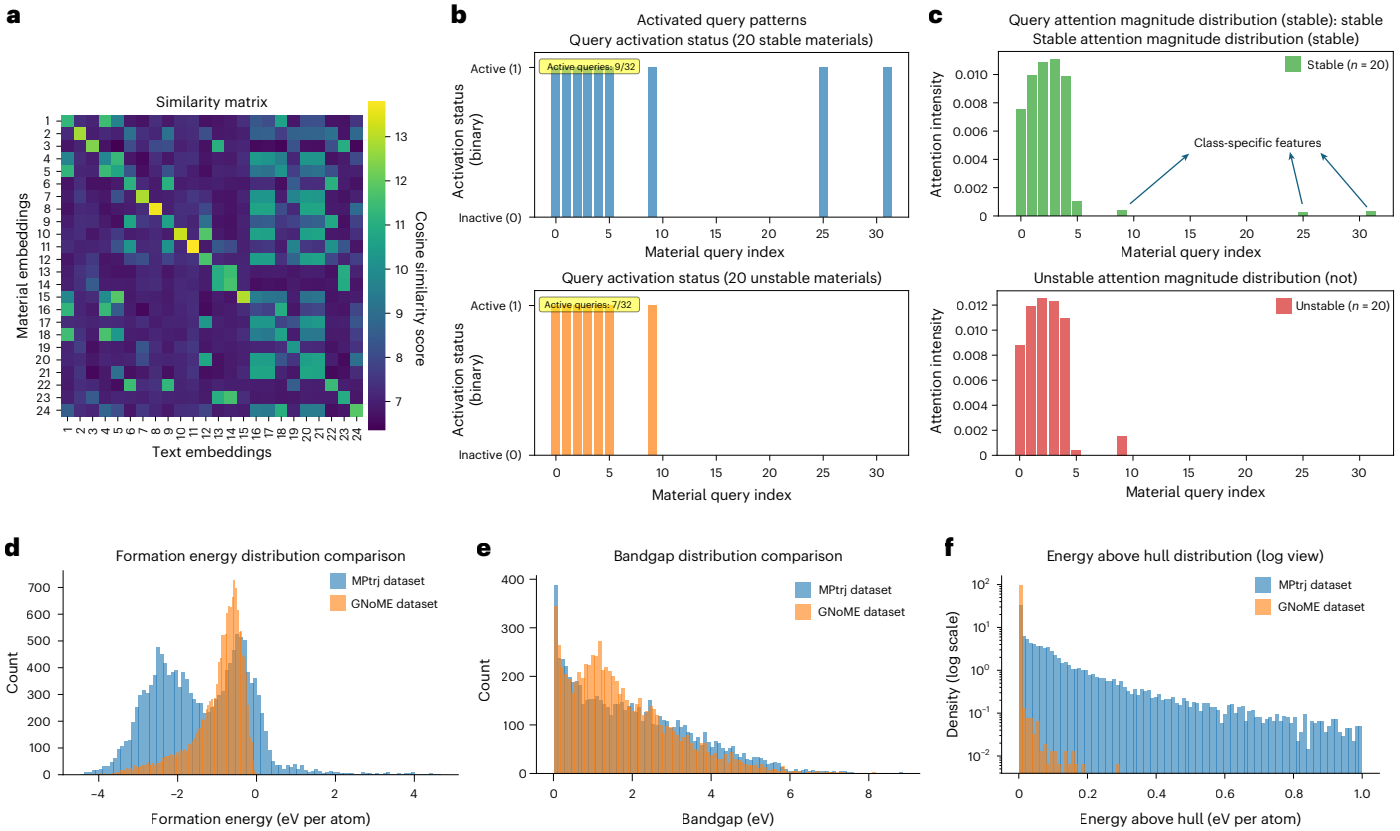

Fig. 6 | Visualization of structure–text alignment in MatterChat's bridge model. a, Cosine similarity matrix between 24 material query embeddings and 24 text token embeddings, showing structured alignment patterns across different modalities. A complete list of the materials corresponding to indices 1–24, along with their text token embeddings, is provided in Supplementary Table 4. **b**, Material query activated during stability classification (across random 20 stable and 20 unstable material examples). A query is defined as activated if it ranks among the top-5 ($k = 5$) most-attended embeddings for key linguistic tokens. The union of these activations across each class reveals that although foundational structural features are concentrated in indices 0–5 and 9, indices 25 and 31 are selectively utilized for stable materials. **c**, Detailed attention distribution values of certain tokens 'stable' and 'not' tokens across material query indices ($n = 20$ per material class). Both tokens prioritize indices 0–4 as core structural descriptors. An asymmetric pattern emerges: stable materials exhibit distinct attention to indices 25 and 31, whereas 'not' shows elevated intensity at index 9. **d**–**f**, Distribution comparisons between the MPtrj test dataset and the GNoME[44] out-of-distribution dataset for three key properties: formation energy (**d**), bandgap (**e**) and energy above hull (**f**) (log scaled). These histograms show clear distributional differences between the MPtrj test set and GNoME datasets across all three properties.

For numerical property prediction (Fig. 5g–i), including formation energy, energy above hull and bandgap, MatterChat yielded the lowest root mean squared error (RMSE), whereas pure LLMs were excluded from comparison due to inherent limitations in quantitative precision[56]. The framework's robustness was further validated through fivefold cross-validation (Supplementary Figs. 7 and 8). Although the raw performance values of cross-validation decreased slightly across folds due to reduced training data, results remained consistent with the original train/test data split. These findings demonstrate that MatterChat effectively bridges qualitative scientific reasoning with quantitative atomistic characterization across diverse material domains.

Comparative study and visual attention analysis

To evaluate MatterChat's architectural effectiveness, we compared it against established baseline strategies across all material property tasks (Extended Data Table 1). Our multimodal bootstrapping approach[42] notably outperforms both the Simple Adapter[57,58] and pure LLM baselines, achieving superior accuracy and maintaining the efficiency of frozen pretrained components. Extensive ablation studies on bridge configurations, encoder selection and pretraining strategies further confirm that optimal cross-attention frequency and bridge pretraining are critical for model convergence and predictive precision (Methods). Ablation studies across different LLM backbones (e.g., Llama 3 and DeepSeek R1) and GNN encoders further demonstrate the architectural flexibility of MatterChat (Supplementary Table 3). Integrating a multimodal RAG module further enhances performance, reducing regression RMSE by ~12% and improving the classification accuracy by ~0.6%. This improvement is achieved with negligible computational overhead (latency, ~0.7%), demonstrating a favourable speed–accuracy trade-off for large-scale screening. Unless otherwise stated, baseline figures (for example, Figs. 2 and 3) reflect performance without RAG.

To assess cross-dataset generalization, we evaluated MatterChat on an external resource from the GNoME project[44]. Despite considerable distributional shifts in target properties relative to our training data (Fig. 6d–f), MatterChat—particularly the MACE-based variant—demonstrates robust transferability, achieving superior accuracy across all tasks without additional fine tuning (Extended Data Table 2). These results indicate that equivariant structural representations generalize more effectively across diverse data sources. Furthermore, these gains underscore the advantage of MatterChat's modular framework, which enables strong performance on external benchmarks without full-model retraining.

To further investigate the interpretability of structure–text alignment, we analysed both similarity matrix between materials and text embeddings and the attention behaviour of the bridge model. We randomly selected 35 materials and computed the cosine similarity between the 24 structure embeddings (queries) and 24 token embeddings from the paired textual descriptions (chemical formula,

space group and crystal system). This reveals consistent diagonal alignment in the embedding space (Fig. 6a), suggesting that specific structural slots are consistently linked with semantically meaningful linguistic features. The structural embeddings (indices 1–24) represent the graph-based representations of the materials listed in Supplementary Table 4, whereas the corresponding text embeddings represent their linguistic descriptors comprising chemical formula, space group and crystal system.

Beyond the diagonal alignment shown in Fig. 6a, off-diagonal patterns reveal a structured embedding space. Indices 16–23 show that complex multicomponent systems (for example, $Li_5La_4TiNb_7O_{28}$) cluster through shared coarse-grained characteristics rather than strictly element-specific distinctions, though index 19 remains distinct, preserving compositional specificity. Similarly, strong mutual similarities for indices 13 and 14 (cubic, *Fm*-3*m*) and 20 and 21 (monoclinic, 2/*m*) reflect the influence of shared structural symmetry on the joint representation. Although supporting physically meaningful clustering, these patterns identify a resolution limit for subtle intra-class variations, indicating the enhanced structural resolution as a priority for future refinement.

To investigate the model's internal inference mechanism, we examined the attention distributions across material query indices for 20 random sampled stable and 20 unstable samples (Fig. 6b,c). Although foundational structural features are consistently captured in indices 0–4 and 9, distinct class-specific markers emerge that guide the model's thermodynamic predictions. Specifically, stable materials uniquely activate indices 25 and 31, suggesting these embeddings key structural features associated with stability. Conversely, index 9 appears to function as a marker for instability; although it is used for both classes, its intensity is notably higher for unstable materials, suggesting it identifies energetically unfavourable atomic arrangements. These distinct patterns of query selection and attention intensity demonstrate that MatterChat does not merely recall data but effectively maps linguistic concepts onto physically relevant structural descriptors during inference.

Discussion

In this study, we present MatterChat, a multimodal framework that achieves superior performance in material properties prediction and scientific reasoning tasks by leveraging a more effective representation of materials. A key innovation of MatterChat is its ability to leverage existing advancements in both materials science and language modelling by integrating a pretrained material foundation encoder with a pretrained LLM. Rather than training an entire model from scratch, MatterChat achieves strong performance by training only a lightweight bridge model, efficiently aligning material structure representations with textual understanding and maintaining high accuracy across diverse materials science tasks. Moreover, MatterChat is designed for multitask learning, enabling it to handle both classification and numerical property prediction. This capability allows the framework to tackle a diverse range of materials science tasks within a unified model. Another advantage of our approach is the use of graph-based structural embeddings instead of relying solely on a .cif text input. Although CIF files encode atomic structures, their text-based format relies entirely on attention mechanisms, which can struggle to explicitly capture geometric symmetries and increase computational overhead due to lengthy tokenization. By directly processing atomic graphs, MatterChat effectively preserves material symmetry and spatial relationships, leading to more accurate structure–property learning and maintaining computational efficiency. Furthermore, we evaluated its performance on the derived properties that require structural internalization, such as atom counts and density (Supplementary Table 5). Although the model accurately retrieves discrete identifiers like the number of atoms per unit cell (28 atoms), it exhibits a 'resolution gap' in predicting continuous numerical properties such as volume and density. This shows a common limitation of LLMs in high-precision zero-shot numerical regression, despite their success in structural reasoning tasks.

Limitation and future work

(1) Alignment and interpretation: MatterChat's behavioural success on property tasks may reflect learned correlations rather than deep semantic internalization of graph-based structural semantics. This limits interpretability and compositional reasoning involving structural concepts. Addressing this requires explicit representation-level alignment objectives—such as contrastive losses, modality matching or shared embedding projections—to ensure the LLM fully grounds language in atomic representations[59–62].

(2) Data and reasoning: current training relies on single-turn question–answer pairs, lacking the multistep reasoning and cross-modal inference chaining essential for expert enquiry[63–65]. Future developments should transition towards multiturn, multimodal dialogue trajectories. Techniques like phased instruction tuning[66–68] and least-to-most prompting[69] offer promising pathways for stepwise scientific problem-solving grounded in material structures.

(3) Hallucination and reliability: frozen LLM backbones are susceptible to hallucinations in which language priors dominate structural information[70–72]. Although RAG provides initial contextual grounding[73], future modular enhancements are necessary. These include multimodal fusion techniques (for example, mixture of features)[74], domain-adaptive fine tuning on expert corpora[75,76] and hallucination-aware training objectives[25,77–79]. Finally, post hoc correction frameworks—including fact-checking and self-revision loops—can further enhance the reliability of open-ended scientific responses[80,81].

Finally, although the current work prioritizes structure-informed reasoning, MatterChat's modular architecture is designed for future extensibility to text-only materials benchmarks[33,35–37]. Its interchangeable components provide a flexible framework for potential systematic evaluation on tasks like synthesis question–answer classification from abstracts in future studies, offering a pathway to further bridge the gap between linguistic and structure-aware understanding.

Methods

Dataset curation

In this work, we curated a comprehensive dataset from the Materials Project Trajectory (MPtrj) dataset[52], focusing specifically on relaxed samples. By selecting these stable configurations rather than complete trajectory data, we ensure that the dataset captures the equilibrium states of materials, which are more relevant for downstream tasks such as material property prediction. The final dataset consists of 142,899 high-quality samples, offering a rich and diverse representation of inorganic materials.

To facilitate effective model training and evaluation, we randomly shuffled the dataset and partitioned it into training and testing subsets using a 9:1 split ratio. This ensures that a substantial portion of data is available for learning, maintaining a dedicated portion for rigorous performance validation, allowing us to assess the generalization capabilities of the model.

In addition to the relaxed structural data, we retrieved detailed material property information using the Materials Project API[43]. Each material is retrieved by a unique mp-id and is enriched with a variety of key descriptors that span both structural and electronic properties. These include

- Structure: the full atomic structure of the material, detailing atomic positions and bonding.
- Chemical formula: the overall chemical composition.
- Space group: the crystallographic space group of the material, reflecting its symmetry properties.

- Crystal system: the broader classification of the material's crystal structure.
- Metallicity: an indicator of whether the material is metallic or insulating.
- Magnetic properties: whether the material is magnetic and its magnetic ordering (for example, ferromagnetic or antiferromagnetic).
- Experimental observables: properties that can be compared directly with experimental data.
- Direct bandgap: the direct bandgap energy, a key property for semiconductors.
- Stability: whether the material is thermodynamically stable.
- Energy above hull: a measure of how stable the material is compared with other phases.
- Bandgap: the electronic bandgap, an important factor in determining a material's electronic properties.
- Formation energy: energy required to form the material from its constituent elements.

These attributes offer a comprehensive view of each material, encompassing both its structural arrangement and electronic behaviour. By integrating this wealth of data, our model is capable of capturing complex material property relationships, supporting tasks such as bandgap prediction, stability analysis and metallicity determination. This dataset not only provides a robust foundation for training ML models but it also contributes to broader efforts in materials discovery and property optimization.

Training detail

MatterChat uses a bootstrapping strategy commonly used in multimodal learning for vision-language tasks, adapted here for materials science applications. The training process consists of two main stages: pretraining to align material structures with descriptive text, and fine tuning for both descriptive and property prediction tasks with the LLM module integrated (Supplementary Fig. 2). The pretraining phase aims to establish a foundational alignment between material structures and descriptive text. In this stage, the model connects a frozen graph encoder with pairs of graph data and the corresponding textual descriptions, without attaching the LLM module. Here the bridge model acts as a text generator, learning to extract descriptive graph representations that effectively capture structural information relevant to the text data. MatterChat utilizes pretrained checkpoints for graph encoders. For the invariant encoder, we use the CHGNet model pretrained on the MPTraj dataset of Materials Project structures and energies[41,52]. For the equivariant encoder, we use the publicly released MACE-MP-0 (large) model[11]. These encoders provide chemically meaningful atom-level representations and are integrated into MatterChat without additional pretraining of the graph encoder.

This stage consists of three core optimizing targets, each with distinct interaction mechanisms between graph embeddings and text, and maintain a consistent input format:

1. Graph–text correlation learning (contrastive loss). This task aligns graph and text representations by maximizing the similarity between matched graph–text pairs and minimizing it for mismatched pairs. A contrastive loss is used:

$$\mathcal{L}_{\text{correlation}} = -\sum_{i=1}^{N} \log\left[\frac{\exp(\text{sim}(q_i,t_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(q_i,t_j)/\tau)}\right], \quad (1)$$

where q_i and t_i represent the graph and text embeddings, respectively, and τ is the temperature parameter controlling the distribution's sharpness.

2. Graph-driven text prediction (conditional language modelling loss). The bridge model generates descriptive text based on graph data, conditioned through attention mechanisms. The loss function is defined as

$$\mathcal{L}_{\text{prediction}} = -\sum_{t=1}^{T} \log[P(y_t|y_{<t}, Q)], \quad (2)$$

where Q represents the graph query features, and y_t is the token at position t in the output sequence.

3. Graph–text association (binary cross-entropy loss). This task predicts whether each graph–text pair is correctly matched. A binary cross-entropy loss with hard negative sampling is applied:

$$\mathcal{L}_{\text{association}} = -\sum_{i=1}^{N}(y_i\log[s_i] + (1-y_i)\log[1-s_i]), \quad (3)$$

where s_i is the model's prediction score and y_i indicates whether the pair is matched (1) or not (0).

The total pretraining loss is the sum of the individual task losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{correlation}} + \mathcal{L}_{\text{prediction}} + \mathcal{L}_{\text{association}}. \quad (4)$$

After pretraining, the model undergoes instructive fine tuning to optimize its performance on both descriptive and property prediction tasks. In this stage, the pretrained bridge model is integrated with the LLM to enhance multimodal learning. A fully connected layer is introduced between the bridge model's output and the LLM's input. The fine-tuning phase includes 12 multimodal subtasks, including three material description tasks and nine property prediction tasks. Description tasks refine the model's ability to link structural features with detailed textual explanations, whereas property prediction tasks focus on improving quantitative accuracy in material property estimation. Fine tuning is guided by a supervised cross-entropy loss defined as

$$\mathcal{L}_{\text{fine tune}} = -\sum_{i=1}^{N}\sum_{j=1}^{T} y_{i,j}\log[P(y_{i,j}|x_i)], \quad (5)$$

where $y_{i,j}$ represents the ground-truth token for the jth position of the ith sample, and $P(y_{i,j}|x_i)$ is the model's predicted probability of the correct token given the multimodal input x_i.

In the pretraining stage, the model is trained using the AdamW optimizer with a learning rate of 2×10^{-4}, with a cosine decay scheduler and linear warm-up starting from 1×10^{-6}. A weight decay of 0.05 is applied to regularize the model, with a batch size of 32 and gradient accumulation over five steps to manage computational efficiency. Mixed-precision training is enabled to improve the performance and reduce memory usage. The model is trained for ~25 epochs, with checkpoints saved every 2,000 iterations. During the fine-tuning stage, the AdamW optimizer is again used with a learning rate of 2×10^{-4}, featuring a warm-up phase to 1×10^{-4} followed by decay to 1×10^{-5}. The batch size is set to 8, with gradient accumulation over 16 batches to effectively increase the batch size. Fine-tuning runs for around 20 epochs, with checkpoints saved every 300 steps and at the end of each epoch. Additionally, distributed training is implemented using four A100 GPUs per node across eight nodes, leveraging the distributed data parallel strategy to enhance training efficiency and scalability. It takes around 48 h to complete the training. We have also summarized the training hyperparameters used across all baseline GCN models to ensure consistent evaluation. The SchNet model was trained using consistent hyperparameters across classification and regression tasks to ensure fair comparison. We used the Adam optimizer with a learning rate of 1×10^{-5} and weight decay of 1×10^{-4}, along with a StepLR scheduler (step size of 20, $\gamma = 0.5$). Models were trained for 50 epochs with a batch size of 16. Cross-entropy loss was used for classification and mean squared error loss was used for regression. All the CHGNet models were

trained using the SchNet-style optimizer and scheduler with a learning rate of 1×10^{-5} for classification and 1×10^{-3} for regression. All models were trained for 50 epochs with a batch size of 16. Cross-entropy loss was used for classification and mean squared error loss was used for regression. These settings were applied uniformly to both pretrained and non-pretrained CHGNet variants. The MACE model was trained using consistent hyperparameters across classification and regression tasks. We used the Adam optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}, together with a StepLR scheduler (step size of 20, $\gamma = 0.5$). All models were trained for 100–200 epochs with a batch size of 256. Cross-entropy loss was used for classification tasks and the mean squared error loss was used for regression tasks.

Embedding visualization

The visualization leverages UMAP to reveal chemical insights encoded in the material embeddings that are extracted from the bridge model in a lower-dimensional space. To prepare the data, each high-dimensional embedding, originally structured as (32, 4,096), is first flattened into a single vector, capturing the essential features of the material. UMAP is then applied to this set of vectors with number of components equals 2, reducing the data to two dimensions to enable visual interpretation, with random state is set to 1 to ensure consistency in the layout across runs.

Structural similarity scores are computed using the smooth overlap of atomic positions (SOAP) descriptor[82], combined with the regularized entropy match kernel (REMatch)[83,84] to capture the structural characteristics within material embeddings. SOAP is a local atomic environment descriptor that encodes atomic geometries by expanding a Gaussian-smeared atomic density locally, using orthonormal functions derived from spherical harmonics and radial basis functions. From local descriptors to structure matching, we use the REMatch kernel on top of the SOAP descriptor. The REMatch kernel considers the best matching of local environments and uses an averaging strategy to enhance structural comparison. For SOAP construction, we consider periodic boundary conditions. The cut-off radius for the local region (r_{cut}), the number of radial basis functions (n_{max}) and the maximum degree of spherical harmonics (l_{max}) are set to 6 Å, 8 Å and 6 Å, respectively. For the REMatch kernel, the entropic penalty (α) is set to 1, and the convergence threshold is set to 1×10^{-6}. A linear pairwise metric is used for the local similarity calculation.

Baseline and RAG configurations for comparative study

We assessed MatterChat against two primary baselines: (1) a multimodal LLM using a Simple Adapter with low-rank adaptation fine tuning[57,58], updating lightweight adapter layers and the Mistral 7B backbone; and (2) a pure LLM baseline fine tuned on serialized CIF content. Our bootstrapping strategy[42] trains only the bridge module, avoiding extensive fine tuning of the frozen graph encoder and LLM. Ablation studies (Supplementary Tables 1–3 and Fig. 6) covered variations in query token length, cross-attention frequency and pretraining strategies. Results indicate that cross-attention every two layers and query lengths as low as eight tokens maintain a strong balance between efficiency and multimodal alignment. The RAG module utilizes a Faiss (Facebook AI Similarity Search)-based[85] batched cosine similarity search over ~142,000 structural embeddings. We measured an average retrieval latency of ~12 ms per query on CPU (total of <3 min for 14,290 queries). Compared with the baseline inference time of ~1.65 s per sample, this introduces only ~0.7% latency.

Cross-dataset evaluation

We curated an external test set of ~15,000 materials from the GNoME database[44] to evaluate the transferability of our model. This subset includes available density-functional-theory-computed values for bandgap, formation energy and energy above hull, providing a benchmark comparable in scale to our original test split. Distributional differences between this external set and the MPtrj training distribution[52] were characterized via property histograms (Fig. 6c–e).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The source data generated in this study and the datasets and saved models for running codes are available via Zenodo at https://doi.org/10.5281/zenodo.18735961 (ref. 86). Source data are provided with this paper.

Code availability

The codes that support the results within this paper and other findings of this study are available via Zenodo at https://doi.org/10.5281/zenodo.18735881 (ref. 87).

References

1. Kohn, W., Becke, A. D. & Parr, R. G. Density functional theory of electronic structure. *J. Phys. Chem.* **100**, 12974–12980 (1996).
2. Marx, D. & Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods* (Cambridge Univ. Press, 2009).
3. Hwang, H. Y. et al. Emergent phenomena at oxide interfaces. *Nat. Mater.* **11**, 103–113 (2012).
4. Li, D. et al. Superconductivity in an infinite-layer nickelate. *Nature* **572**, 624–627 (2019).
5. Keimer, B. & Moore, J. E. The physics of quantum materials. *Nat. Phys.* **13**, 1045–1055 (2017).
6. Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
7. 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).
8. Xu, W., Reuter, K. & Andersen, M. Predicting binding motifs of complex adsorbates using machine learning with a physics-inspired graph representation. *Nat. Comput. Sci.* **2**, 443–450 (2022).
9. Batzner, S. et al. *E*(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
10. Gasteiger, J., Becker, F. & Günnemann, S. GemNet: universal directional graph neural networks for molecules. In *35th Conference on Neural Information Processing Systems* 6790–6802 (NeurIPS, 2021).
11. Batatia, I. et al. A foundation model for atomistic materials chemistry. *J. Chem. Phys.* **163**, 184110 (2025).
12. Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R. & Miller, T. F. OrbNet: deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **153**, 124111 (2020).
13. Yang, H. et al. MatterSim: a deep learning atomistic model across elements, temperatures and pressures. Preprint at https://arxiv.org/abs/2405.04967 (2024).
14. Zeni, C. et al. A generative model for inorganic materials design. *Nature* **639**, 624–632 (2025).
15. Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. & Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. Preprint at https://arxiv.org/abs/2110.06197 (2021).
16. Ling, C. A review of the recent progress in battery informatics. *npj Comput. Mater.* **8**, 33 (2022).
17. Liu, D.-Y. et al. Machine learning for semiconductors. *Chip* **1**, 100033 (2022).
18. Yang, W., Fidelis, T. T. & Sun, W.-H. Machine learning in catalysis, from proposal to practicing. *ACS Omega* **5**, 83–88 (2020).

19. Sen, S. K. et al. Opportunities for basic, clinical, and bioethics research at the intersection of machine learning and genomics. *Cell Genom.* **4**, 100466 (2024).
20. Birhane, A., Kasirzadeh, A., Leslie, D. & Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).
21. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (ACL, 2019).
22. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
23. Jiang, A. Q. et al. Mistral 7B. Preprint at https://arxiv.org/abs/2310.06825 (2023).
24. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
25. Guo, D. et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. Preprint at https://arxiv.org/abs/2501.12948 (2025).
26. Lu, P. et al. Learn to explain: multimodal reasoning via thought chains for science question answering. In *36th Conference on Neural Information Processing Systems* (eds Oh, A. H. et al.) 2507–2521 (NeurIPS, 2022).
27. Dunn, A. et al. Structured information extraction from complex scientific text with fine-tuned large language models. Preprint at https://arxiv.org/abs/2212.05238 (2022).
28. Kim, S., Jung, Y. & Schrier, J. Large language models for inorganic synthesis predictions. *J. Am. Chem. Soc.* **146**, 19654–19659 (2024).
29. Cavanagh, J. M. et al. SMILEyLLaMA: modifying large language models for directed chemical space exploration. Preprint at https://arxiv.org/abs/2409.02231 (2024).
30. Song, Y., Miret, S. & Liu, B. MatSci-NLP: evaluating scientific language models on materials science language tasks using text-to-schema modeling. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 3621–3639 (ACL, 2023).
31. Zaki, M. et al. MASCQA: investigating materials science knowledge of large language models. *Digit. Discov.* **3**, 313–327 (2024).
32. Song, Y., Miret, S., Zhang, H. & Liu, B. HoneyBee: progressive instruction finetuning of large language models for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2023* 5724–5739 (ACL, 2023).
33. Mirza, A. et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nat. Chem.* **17**, 1027–1034 (2025).
34. Alampara, N., Miret, S. & Jablonka, K. M. MatText: do language models need more than text & scale for materials modeling? Preprint at https://arxiv.org/pdf/2406.17295v1 (2024).
35. Mishra, V. et al. Foundational large language models for materials research. Preprint at https://arxiv.org/abs/2412.09560 (2024).
36. Miret, S. & Krishnan, N. A. Are LLMs ready for real-world materials discovery? Preprint at https://arxiv.org/abs/2402.05200 (2024).
37. Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **16**, 2514–2572 (2025).
38. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **18**, 31–36 (1988).
39. Antunes, L. M., Butler, K. T. & Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nat. Commun.* **15**, 10570 (2024).
40. Ock, J., Guntuboina, C. & Barati Farimani, A. Catalyst energy prediction with CatBERTa: unveiling feature exploration strategies through large language models. *ACS Catal.* **13**, 16032–16044 (2023).
41. Deng, B. et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).
42. Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. 40th International Conference on Machine Learning* 19730–19742 (ICML, 2023).
43. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
44. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
45. Gemini Team Google. Gemini: a family of highly capable multimodal models. Preprint at https://arxiv.org/abs/2312.11805 (2023).
46. Hurst, A. et al. GPT-4o system card. Preprint at https://arxiv.org/abs/2410.21276 (2024).
47. Wang, S., Wei, Z., Choi, Y. & Ren, X. Symbolic working memory enhances language models for complex rule application. In *Proc. 2024 Conference on Empirical Methods in Natural Language Processing* 17583–17604 (ACL, 2024).
48. Nakamura, S. N. GaN growth using GaN buff er layer. *Jpn. J. Appl. Phys.* **30**, 1705 (1991).
49. Dupuis, R. Epitaxial growth of III–V nitride semiconductors by metalorganic chemical vapor deposition. *J. Cryst. Growth* **178**, 56–73 (1997).
50. Sharma, V., Saha, J., Patnaik, S. & Kuanr, B. K. Synthesis and characterization of yttrium iron garnet (YIG) nanoparticles—microwave material. *AIP Adv.* **7**, 056405 (2016).
51. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
52. Deng, B. Materials project trajectory (MPtrj) dataset. *figshare* https://doi.org/10.6084/m9.figshare.23713842.v2 (2023).
53. Himanen, L. et al. DScribe: library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
54. The Vicuna Team. Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *LMSYS* http://lmsys.org/blog/2023-03-30-vicuna/ (2023).
55. Schütt, K. et al. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Proc. 31st International Conference on Neural Information Processing Systems* 992–1002 (NeurIPS, 2017).
56. Feng, G. et al. How numerical precision affects arithmetical reasoning capabilities of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025* 46–85 (ACL, 2025).
57. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. Preprint at https://arxiv.org/abs/2106.09685 (2022).
58. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *Proc. 37th International Conference on Neural Information Processing System* 34892–34916 (NeurIPS, 2023).
59. Shu, D. et al. Large vision-language model alignment and misalignment: a survey through the lens of explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2025* 1713–1735 (ACL, 2025).
60. Gemma Team. Gemma 3 technical report. Preprint at https://arxiv.org/abs/2503.19786 (2025).
61. Zhang, Y. et al. Meta-Transformer: a unified framework for multimodal learning. Preprint at https://arxiv.org/abs/2307.10802 (2023).

62. Wang, Z. et al. Connecting multi-modal contrastive representations. *Adv. Neural Inf. Process. Syst.* **36**, 22099–22114 (2023).
63. Zhang, Z. et al. Multimodal chain-of-thought reasoning in language models. Preprint at https://arxiv.org/abs/2302.00923 (2023).
64. Li, Y. et al. Beyond single-turn: a survey on multi-turn interactions with large language models. Preprint at https://arxiv.org/abs/2504.04717 (2025).
65. Huang, M. et al. DialogGen: multi-modal interactive dialogue system with multi-turn text-image generation. In *Findings of the Association for Computational Linguistics: NAACL 2025* 411–426 (ACL, 2025).
66. Guan, C. et al. Multi-stage LLM fine-tuning with a continual learning setting. In *Findings of the Association for Computational Linguistics: NAACL 2025* 5499–5513 (ACL, 2025).
67. Xu, C. et al. WizardLM: empowering large pre-trained language models to follow complex instructions. In *12th International Conference on Learning Representations* (ICLR, 2024).
68. Pang, W., Zhou, C., Zhou, X.-H. & Wang, X. Phased instruction fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024* 5735–5748 (ACL, 2024).
69. Zhou, D. et al. Least-to-most prompting enables complex reasoning in large language models. Preprint at https://arxiv.org/abs/2205.10625 (2022).
70. Fu, Y., Xie, R., Sun, X., Kang, Z. & Li, X. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025* 16563–16577 (ACL, 2025).
71. Bai, Z. et al. Hallucination of multimodal large language models: a survey. Preprint at https://arxiv.org/abs/2404.18930 (2024).
72. Tonmoy, S. et al. A comprehensive survey of hallucination mitigation techniques in large language models. Preprint at https://arxiv.org/abs/2401.01313 (2024).
73. Li, J., Yuan, Y. & Zhang, Z. Enhancing LLM factual accuracy with RAG to counter hallucinations: a case study on domain-specific queries in private knowledge-bases. Preprint at https://arxiv.org/abs/2403.10446 (2024).
74. Tong, S. et al. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9568–9578 (CVPR, 2024).
75. Grayson, M., Patterson, C., Goldstein, B., Ivanov, S. & Davidson, M. Mitigating hallucinations in large language models using a channel-aware domain-adaptive generative adversarial network (CADAGAN). Preprint at https://doi.org/10.21203/rs.3.rs-5164079/v1 (2024).
76. Valentino, M., Kim, G., Dalal, D., Zhao, Z. & Freitas, A. Mitigating content effects on reasoning in language models through fine-grained activation steering. In *Proc. AAAI Conference on Artificial Intelligence* 33314–33322 (AAAI, 2026).
77. Sun, Z. et al. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024* 13088–13110 (ACL, 2024).
78. Zhao, Z. et al. Beyond multimodal hallucinations: enhancing LVLMs through hallucination-aware direct preference optimization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)* 1–6 (IEEE, 2025).
79. Ranaldi, L., Valentino, M. & Freitas, A. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 17222–17240 (ACL, 2025).
80. Zhou, Y. et al. Analyzing and mitigating object hallucination in large vision-language models. Preprint at https://arxiv.org/abs/2310.00754 (2023).
81. Lee, S., Park, S. H., Jo, Y. & Seo, M. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *Proc. the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* 391–404 (ACL, 2024).
82. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
83. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
84. Musil, F. et al. Machine learning for the structure–energy–property landscapes of molecular crystals. *Chem. Sci.* **9**, 1289–1300 (2018).
85. Douze, M. et al. The Faiss library. Preprint at https://arxiv.org/abs/2401.08281 (2024).
86. Tang, Y. Dataset for MatterChat. *Zenodo* https://doi.org/10.5281/zenodo.18735961 (2026).
87. Tang, Y. Code for MatterChat. *Zenodo* https://doi.org/10.5281/zenodo.18735881 (2026).

Acknowledgements

Y.T. and Z.J.Y. were supported by the US Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research, 'Transformational AI Model Consortium' under LAB-25-3560. This work was supported in part by previous breakthroughs obtained through the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under US Department of Energy contract number DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy, Office of Science User Facility, supported by the Office of Science, US Department of Energy, under contract number DE-AC02-05CH11231 and under NERSC GenAI award number DDR-ERCAP0030541. W.G. acknowledges support from the National Science Foundation through grant number 2235276.

Author contributions

Y.T. and W.G. conceived of the idea, with W.X. contributing enhancements and Z.J.Y. providing additional support. Z.J.Y. supervised the project. Y.T. constructed the overall ML framework and performed the ML training/inference. W.X. performed the physical model training and visualization. Y.T. and W.X. wrote the paper with the help of Z.J.Y., J.C., A.N., S.F., B.E. and M.W.M.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-026-01214-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-026-01214-y.

Correspondence and requests for materials should be addressed to Yingheng Tang, Wenbin Xu, Weilu Gao or Zhi Jackie Yao.

Peer review information *Nature Machine Intelligence* thanks Indra Priyadarsini and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-026-01214-y

Extended Data Table 1 | Comparison of material property prediction performance across different multi-modal frameworks and RAG-enhanced inference

Task	Simple Adapter w LoRA	LoRA LLM only	MatterChat	MatterChat w/ RAG
Metallic (Accuracy)	0.6373	0.6864	0.8683	**0.8873**
Direct Bandgap (Accuracy)	0.8629	0.7839	0.8753	**0.8797**
Stability (Accuracy)	0.7418	0.7944	0.8515	**0.8573**
Exp Ob (Accuracy)	0.7171	0.6549	0.8504	**0.8570**
Is Magnetic (Accuracy)	0.8339	0.6833	**0.9368**	0.9333
Magnetic Order (Accuracy)	0.7759	0.4238	**0.8570**	0.8535
Formation Energy (RMSE)	0.4105	1.8059	0.1500	**0.1212**
Energy Above Hull (RMSE)	0.4415	0.4051	0.1053	**0.0964**
Bandgap (RMSE)	1.2516	1.4725	0.5590	**0.5058**

Performance metrics for nine property prediction tasks, comparing MatterChat against various baselines. Accuracies are reported for classification, and RMSE values for regression. RMSE units are eV/Atom for formation energy and energy above hull, and eV for bandgap. All results are based on test set material (n = 14290) samples.

https://doi.org/10.1038/s42256-026-01214-y

Extended Data Table 2 | Comparison of MatterChat performance using CHGNet vs MACE (large) encoder backbones with Out Of Distribution dataset (OOD)

Task	MatterChat (CHGNet)	MatterChat (MACE - large)
Bandgap (RMSE, eV)	0.6650	**0.6599**
Energy Above Hull (RMSE, eV/atom)	0.0375	**0.0339**
Formation Energy (RMSE, eV/atom)	0.1987	**0.1376**

Comparison of model transferability using CHGNet and MACE encoders on an external dataset. RMSE units are eV/Atom for formation energy and energy above hull, and eV for bandgap. Evaluation is performed on material samples (n = 15483) from the GNoME database to assess out-of-distribution performance.

nature portfolio

Corresponding author(s): Yingheng Tang, Wenbin Xu, Weilu Gao Zhi Jackie Yao

Last updated by author(s): 23/2/26

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Custom Python scripts were used to process and curate materials datasets from public databases, including the Materials Project and GNoME. Data preprocessing was performed using established packages such as pymatgen and ASE for structure parsing, standardization, and feature extraction. The preprocessed and cleaned data is included in the submission under the data_sample directory.
Data analysis	Model training and evaluation were performed using PyTorch, PyTorch Lightning, Hugging Face Transformers, and custom implementations of the MatterChat bridge model. Code used for model training, ablation studies, and evaluation is publically available at the follow links: https://zenodo.org/records/18735961 https://zenodo.org/records/18735881

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The processed datasets used in this study publically available at the follow links:
https://zenodo.org/records/18735961
https://zenodo.org/records/18735881
These datasets include curated materials properties derived from public sources such as the Materials Project and GNoME. Raw data from these sources can be accessed directly via their respective portals (https://next-gen.materialsproject.org). All processed data necessary to reproduce the results is included in the data_sample directory.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size of n=142899 material structures was determined based on the available high-quality DFT-computed data from the Materials Project database. This scale is consistent with established benchmarks in the field of graph neural networks and multimodal learning for materials science, providing sufficient statistical power for model training and evaluatio
Data exclusions	No data were excluded from the analyses
Replication	To ensure reproducibility, we performed 5-fold cross-validation as described in the Supplementary Information. All model training runs were repeated with multiple random seeds, and the performance gains were found to be consistent and robust across different data splits.

Randomization	The dataset was randomly partitioned into training, validation, and testing sets (e.g., 80/10/10 split) during cross-validation test.
Blinding	The test/evaluate dataset was kept strictly separate and was only accessed for final performance evaluation after the model architecture and hyperparameters were finalized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*