MMSci: A Multimodal Multi-Discipline Dataset for PhD-Level Scientific Comprehension

Zekun Li[♠] **Xianjun Yang**[♠] **Kyuri Choi**[♡] **Wanrong Zhu**[♠] **Ryan Hsieh**[♠]
HyeonJung Kim[♡] **Jin Hyuk Lim**[♡] **Sungyoung Ji**[♡] **Byungju Lee**[♡◇] **Xifeng Yan**[♠]
Linda Ruth Petzold[♠] **Stephen D. Wilson**[♠] **Woosang Lim**[♡*] **William Yang Wang**[♠*]
[♠]University of California, Santa Barbara [♡]POSCO HOLDINGS [◇]KIST
https://github.com/Leezekun/MMSci

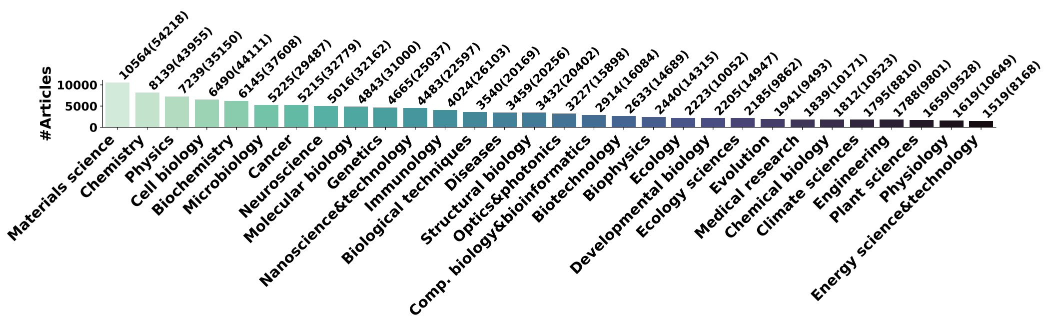


Figure 1: The top 30 out of **72 science subjects**[1] with the most articles in our dataset MMSci. The corresponding numbers of figures are shown in brackets.

Abstract

The rapid advancement of Large Language Models (LLMs) and Large Multimodal Models (LMMs) has heightened the demand for AI-based scientific assistants capable of understanding scientific articles and figures. Despite progress, there remains a significant gap in evaluating models' comprehension of professional, graduate-level, and even PhD-level scientific content. Current datasets and benchmarks primarily focus on relatively simple scientific tasks and figures, lacking comprehensive assessments across diverse advanced scientific disciplines. To bridge this gap, we collected a multimodal, multidisciplinary dataset from open-access scientific articles published in Nature Communications journals. This dataset spans 72 scientific disciplines, ensuring both diversity and quality. We created benchmarks with various tasks and settings to comprehensively evaluate LMMs' capabilities in understanding scientific figures and content. Our evaluation revealed that these tasks are highly challenging: many open-source models struggled significantly, and even GPT-4V and GPT-4o faced difficulties. We also explored using our dataset as training resources by constructing visual instruction-following data, enabling the 7B LLaVA model to achieve performance comparable to GPT-4V/o on our benchmark. Additionally, we investigated the use of our interleaved article texts and figure images for pre-training LMMs, resulting in improvements on the material generation task. The source dataset, including articles, figures, constructed benchmarks, and visual instruction-following data, is open-sourced.

1 Introduction

Recent advancements in generative artificial intelligence, including Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023a;b) and Large Multimodal

*Corresponding authors
[1]https://www.nature.com/nature/browse-subjects

Models (LMMs) (Li et al., 2023; Liu et al., 2024; Zhu et al., 2023; Achiam et al., 2023), have demonstrated remarkable capabilities in solving problems requiring educated knowledge across various domains, including mathematics (Cobbe et al., 2021; Chen et al., 2023; Hendrycks et al., 2021; Lu et al., 2022b), history, computer science, law, and technology (Hendrycks et al., 2020). While these models excel at tasks ranging from elementary to undergraduate-level knowledge, there is an increasing demand for more professional AI scientific assistants that can comprehend and process advanced, graduate-level, and even PhD-level scientific knowledge (noa, 2023; White, 2023; Vert, 2023).

In response, researchers have begun exploring the application of these generative models in fields such as biomedicine (Thapa & Adhikari, 2023), health (Tian et al., 2024), chemistry (Zheng et al., 2023; Bran et al., 2023), and material science (Xie et al., 2023; Miret & Krishnan, 2024) for purposes including research automation, education, and assistance (Meyer et al., 2023). A critical aspect of developing effective AI science assistants is their ability to understand academic scientific literature, which often includes complex figures like data visualization plots and charts, schematic diagrams, macroscopic and microscopic photographs, and other specialized content from various fields.

However, there is currently a lack of comprehensive evaluation of models' understanding of professional PhD-level multimodal scientific knowledge, particularly with figures, across diverse scientific disciplines. Existing evaluations of LMMs on scientific problems are typically limited to up to college-level knowledge and a few science disciplines, such as computer science, mathematics, physics, chemistry, and biology (Lu et al., 2022a; Wang et al., 2023; Yue et al., 2023), as shown in Table 1. Furthermore, the evaluation of models' abilities to understand scientific figures has been restricted to simple charts and plots (Chen et al., 2020; Kahou et al., 2017; Siegel et al., 2016), and suffer from relatively narrow scopes and lower quality (Li et al., 2024).

To bridge the gap, we collected a **multimodal, multi-discipline dataset `MMSci`** from high-quality, open-access articles published in Nature Communications[2], which are freely and permanently available upon publication under a Creative Commons Attribution 4.0 International (CC BY) license[3]. This dataset spans 72 scientific disciplines, primarily within the natural sciences (the top 30 subjects with most articles can be seen in Figure 1). We created a benchmark to evaluate models' understanding of PhD-level multimodal scientific knowledge across various disciplines. The benchmark includes scientific figure captioning and visual question answering (VQA) tasks in various settings, thoroughly assessing LMMs' capabilities in understanding scientific figures and content. Our evaluation revealed significant challenges and deficiencies in current LMMs in interpreting scientific figures and content. Many open-source models struggled considerably with these tasks, demonstrating limited capability. Even GPT-4V and GPT-4o encountered difficulties in producing accurate, relevant captions and matching figures with their descriptions under challenging settings.

Furthermore, our dataset includes a vast collection of high-quality academic articles and figures, which we explored as training resources to enhance models' understanding of scientific content. To achieve this, we constructed visual instruction-following data with discussions about figure content, structured as single or multi-turn interactions. Additionally, we investigated pre-training LMMs using our interleaved article text and figure images to improve their acquisition of scientific knowledge. Experimental results show that our visual instruction-following data enhanced the 7B LLaVA model, achieving performance comparable to GPT-4V/o on our benchmark. Moreover, experiments on a materials science task demonstrated that pre-training on our interleaved multimodal data could improve the performance on material generation. Overall, our contributions are threefold:

- ***Data scope and quality***: Our dataset is unique as it consists of high-quality peer-reviewed academic articles and figures across 72 diverse scientific disciplines.
- ***Challenging benchmark***: Our benchmark includes tasks with varying settings for comprehensive assessment. Our evaluation reveals notable deficiencies in current LMMs in effectively interpreting figures in scientific literature.
- ***Visual instruction-following and interleaved multimodal data***: We developed visual instruction-following data for visual instruction tuning and interleaved article and figure data for pre-training LMMs. Experimental results demonstrate the effectiveness of this approach in enhancing scientific knowledge comprehension.

[2] `https://www.nature.com/ncomms/`
[3] More details can be found at `https://www.nature.com/ncomms/open-access`

Table 1: **Comparison with prior figure understanding datasets and multimodal science benchmarks.** *We only count the number of science subjects categorized according to Nature journals.

Dataset	Data Source	Peer-reviewed	# Subjects	Image Type	Annotations
FigureQA (Kahou et al., 2017)	Synthetic Data	N/A	N/A	Charts/Plots	Synthetic
DvQA (Kafle et al., 2018)	Synthetic Data	N/A	N/A	Bar Chart	Synthetic
FigureSeer (Siegel et al., 2016)	ML Conference Papers	✓	1 (CS)	Charts/Tables/Algo.	Synthetic
SciCap (Yang et al., 2023)	CS Arxiv Papers	✗	1 (CS)	Charts/Diagrams	Authentic
ArxivCap/QA (Li et al., 2024)	Arxiv Papers	✗	32	Open Category	Authentic/Synthetic
MMSci (Ours)	Nature Communications	✓	**72**	Open Category	Authentic

Benchmark	Data Source	Science Topics	# Subjects*	Grades
ScienceQA (Lu et al., 2022a)	Elementary to High School Curricula	Natural/Social/Language	20	1-12
SciBench (Wang et al., 2023)	College Textbooks	Natural(Physics/Chem/Math)	3	College
MMMU (Yue et al., 2023)	College Exams, Quizzes, Textbooks	Natural/Social/Health/Tech	25	College
MMSci (Ours)	Nature Communications Journal Papers	Natural(Primary)/Health/Social	**72**	**PhD**

2 Related Dataset Work

Scientific Figure Understanding Scientific figures in academic articles convey rich, valuable information, and there has been extensive research on evaluating the understanding of these figures. Early approaches typically focused on data visualization figures. For instance, Chen et al. (2020); Kahou et al. (2017); Kafle et al. (2018) created synthetic datasets comprising various types of plots and charts. To obtain more diverse and complex scientific figures, FigureSeer (Siegel et al., 2016) and SciCap (Yang et al., 2023) gathered computer science (CS) papers from arXiv to extract article figures from PDFs. More recently, ChartMimic (Shi et al., 2024) introduces the chart to code generation task. ArxivQA/Cap (Li et al., 2024) collected papers from 32 subjects on arXiv. However, their collection still primarily focuses on CS and math, with limited inclusion of rich and diverse natural science subjects. Additionally, since these arXiv papers are not peer-reviewed, their quality is not guaranteed. In contrast, our dataset emphasizes natural science disciplines and collects high-quality, peer-reviewed articles and figures from the prestigious Nature Communications journals. Covering 72 diverse science disciplines, our dataset ensures both diversity and quality.

Multimodal Science Problems With the advancement of LMMs, many studies have focused on evaluating their capabilities in solving scientific problems in a multimodal context. However, ScienceQA (Lu et al., 2022a) primarily addresses problems ranging from elementary to high school levels (K1-12). SciBench (Wang et al., 2023) focuses solely on three science disciplines: physics, chemistry, and mathematics. MMMU (Yue et al., 2023) includes various subjects such as art, business, history, health, humanities, and technology, but its coverage of science subjects is limited to 25 disciplines according to the categories of the Nature website. In contrast, our dataset evaluates PhD-level scientific knowledge across 72 diverse scientific domains.

3 Data Curation

Table 2: The key statistics of MMSci, including the source data and the constructed benchmark test/validation (dev) set and the data for visual instruction-tuning (IT) in the training set.

Source dataset	Number	Benchmark test/dev set	Number	Visual IT (training set)	Number
Total subjects	72	Used articles	1,418/1,414	Used articles	128,561
Total articles	131,393	Figure captioning	1,218 /1,412	Figure captioning	725,646
Total figures	742,273	Multi-choice VQA (I)	1,188/1,297	Multi-choice VQA (I)	84,328
Avg. figures per article	5.65	Multi-choice VQA (II)	1,121/1,221	Multi-choice VQA (II)	107,098
Avg. caption length	153	Multi-choice VQA (III)	1,119/1,214	Multi-choice VQA (III)	53,882
Avg. abstract length	150			Multi-turn conversation	108,843
Avg. article length	7,457			Total samples	1,079,797

Source Data Collection Our dataset was collected from the Nature Communications website, consisting of open-access, peer-reviewed papers across five major categories and 72 subjects. The full list of subjects can be seen in Appendix A.1.3 . Various information regarding each article

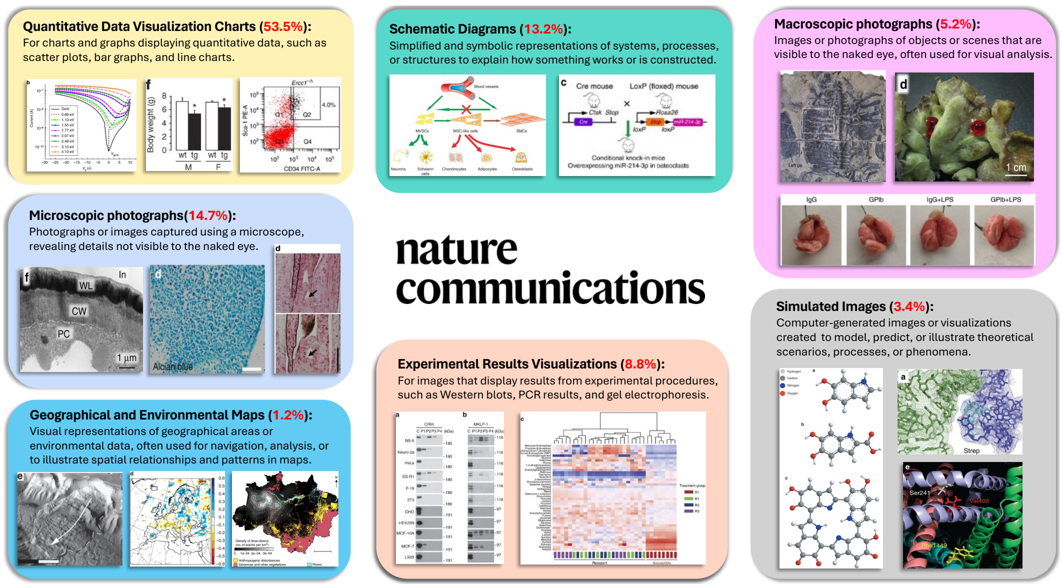


Figure 2: Examples of the seven major types of (sub-)figures in `MMSci`. Ratios are based on the benchmark test set. Sources are discussed in the acknowledgements.

is easily accessible on this website, providing a user-friendly platform for obtaining all necessary data. For each article, we collected information including the title, abstract, main body content, and references, directly from their respective sections on the article's webpage (e.g., `https://www.nature.com/articles/xxx`, where "xxx" is the article's unique ID). Figures and their captions were obtained from a dedicated figures page under the article's homepage (e.g., `https://www.nature.com/articles/xxx/figures`), eliminating the need to extract figures from PDF files and thus ensuring image quality. We used `pylatexenc` to convert LaTeX expressions of mathematical formulas in the article text and figure captions into plain text. Since these papers are all peer-reviewed and the text, figure, and caption data are readily available from the website, no additional quality filtering or content extraction was necessary. This ensures authentic and high-quality data, unlike previous datasets (Yang et al., 2023; Li et al., 2024). We crawled articles up to the date of 2024/04/15. The resulting source dataset comprises 131,393 articles and 742,273 figures.

Sub-caption Extraction Many figures in the dataset consist of multiple sub-figures in a single image, with captions that include a main caption and descriptions of each sub-figure (sub-caption), as illustrated in Figure 3. We developed a regular expression matching function to identify sub-figure indices at the beginning of sentences in alphabetical order (a to z), extracting and identifying 514,054 sub-captions/figures, which aids in the consecutive construction of our benchmark.

Exploring Figures in `MMSci` We examined the types of (sub-)figures in `MMSci` by manually summarizing and categorizing the potential figure types into seven major categories based on a subset of the figures. The categorization is based on the smallest individual components, the sub-figures, when present. Following this review, we used GPT-4o to classify the images within the benchmark test set (see benchmark data splits in the next section). Examples of image types are shown in Figure 2, and more statistics can be found in Appendix A.1.4.

4 Benchmarks

We developed two benchmark tasks with varying settings to comprehensively test models' comprehension of scientific articles and figures from different aspects, as shown in Figure 3.

Scientific Figure Captioning Scientific figure captioning in our dataset `MMSci` presents unique challenges compared to typical natural image captioning. Firstly, unlike natural image captions that can be understood without background knowledge, interpreting figures in scientific articles usually

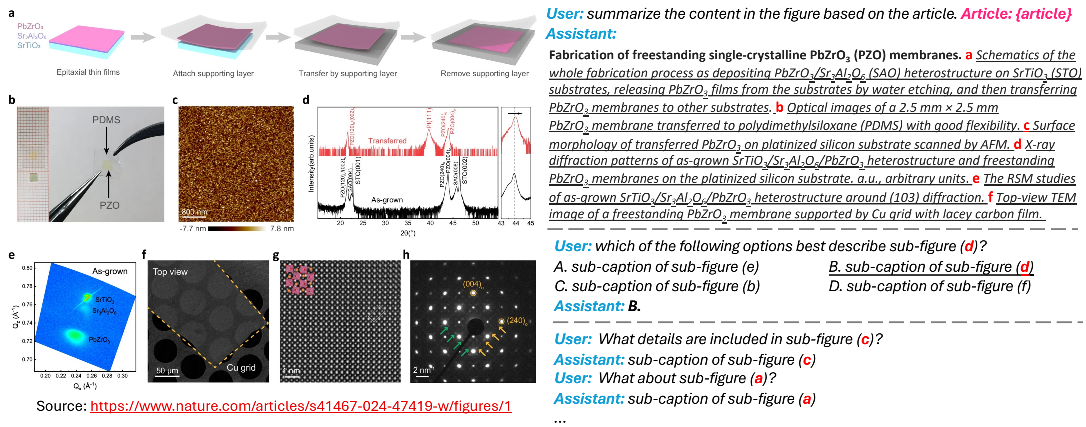


Figure 3: Illustration of the benchmark and visual instruction-following data construction in MMSci. This example is taken from (Guo et al., 2024b). The left side shows the figure including multiple sub-figures. The caption consists of a main caption (bolded) and a series of sub-captions (underlined), each corresponding to a sub-figure. Due to space constraints, we only show sub-captions from (a) to (f). These (sub-)figures and (sub-)captions are used to construct data for figure captioning (upper right), VQA (setting III in this example) (center right), and multi-turn conversations (lower right). Detailed examples of different types of constrcuted data are provided in Appendix A.1.5.

requires grounding in and understanding the article's content. Secondly, scientific figure captions are significantly more detailed, providing rich complementary information to the article. In MMSci, these captions average 153 words, much longer than those for natural images. This complexity and depth make scientific figure captioning a more demanding task. To comprehensively test the model's understanding of scientific figures, we designed three captioning settings: (1) ***Ungrounded figure captioning***: The model generates captions without any additional article content. (2) ***Abstract-grounded figure captioning***: The model is provided with the paper abstract to give an overview of the paper content. (3) ***Full content-grounded figure captioning***: The model is provided with the entire article content to generate the figure caption. Given that the full article content averages around 14k tokens, this setting is primarily suitable for models with longer context windows.

Visual Question Answering Our multiple-choice VQA task is to select the (sub-)caption that best describes a (sub-)figure across three different settings: (1) ***Setting I***: The options include the correct main caption of a figure and three main captions from other figures within the same article. (2) ***Setting II***: This setting tests the model's performance in locating and understanding a specific sub-figure within the given image. We randomly select a sub-figure and use its corresponding sub-caption as the correct answer, with three sub-captions from other figures within the same article as alternative choices. (3) ***Setting III***: As a more challenging setting than setting II, all choices are sub-captions from the same image. This setting rigorously tests the model's ability to locate the sub-figure and distinguish the correct corresponding content from all the content in the image. For all three settings, we construct questions with four choices to ensure consistency.

Data Split We allocated 1% of articles from each subject to the test set and another 1% to the validation (dev) set, with each subject containing 5 to 50 articles. This resulted in 1,418 articles for the test set and 1,414 for the validation set, used for benchmark evaluation samples. Each test sample is derived from a single article, ensuring no reuse of content. For the captioning data, captions were ensured to contain more than 50 words. Ultimately, each task and setting consists of approximately 1,200 samples, balancing coverage, diversity, and cost for benchmarking.

5 Training Resources

Our dataset consists of rich articles and figure data, which we explore as training resources to enhance models' capabilities in comprehending scientific figures and content.

Visual Instruction-Following Data We use the articles, excluding those in the benchmark, to create a visual instruction-following dataset. As illustrated in Figure 3, our dataset consists of conversations discussing figure content, including three types. The first two types are VQA and figure captioning tasks, as in the benchmark, formatted as single-turn interactions. For figure caption tasks, we use only abstract-grounded captioning data in the training set since the full article content is too lengthy for most open-source LMMs. The third type involves multi-turn conversations, where the human asks about content in a sub-figure and the assistant responds with the corresponding sub-caption in each turn. We use diverse conversation templates generated by GPT-4 (Achiam et al., 2023) to vary human instructions. All model responses are derived from original articles rather than model-generated, ensuring data quality. This approach resulted in 108,843 multi-turn conversations, culminating in a dataset with over 1 million visual instruction-following conversations, including the other two types.

Interleaved Text and Image Data for Pre-training `MMSci` includes full article content and figures, naturally forming interleaved text and image data suitable for pre-training LMMs (Lin et al., 2023). Since the text and figures are collected separately from different sections of the website, we insert the figures into the article content at the location of their first mention (e.g., Figure/Fig. x).

Table 3: Performance on scientific figure captioning. B@k represents BLEU@k (k=1,2,3,4), R stands for ROUGE-L, M stands for METEOR, BS indicates BERTScore, and CLIP and RCLIP represent CLIPScore and RefCLIPScore, respectively. Best results are bolded and second best are underlined.

Grounded	Model	B@1	B@2	B@3	B@4	M	R	BS	CLIP	RCLIP
N/A	Kosmos2	23.05	2.59	0.39	0.09	14.53	11.69	77.51	41.44	46.01
	BLIP2	37.73	4.91	0.25	0.04	3.18	6.56	79.28	55.93	56.90
	LLaVA1.5-7B	29.34	3.16	0.16	0.03	11.80	12.55	79.93	64.79	64.22
	LLaVA-Next	15.96	2.44	0.26	0.04	18.89	10.87	79.27	68.08	66.72
	LLaVA-Next-Mistral	15.91	2.81	0.38	0.08	20.45	10.96	79.53	68.54	67.04
	Qwen-VL-Chat	**43.54**	12.78	4.87	1.66	15.34	14.84	81.95	63.24	64.30
	GPT-4V	21.94	4.95	1.31	0.41	26.62	14.87	81.76	**71.81**	71.27
	GPT-4o	19.73	4.90	1.49	0.47	**27.06**	15.59	81.13	71.43	71.39
	LLaVA-Next-MMSci	42.67	**14.51**	**6.60**	**3.10**	21.79	**18.01**	**83.39**	71.19	**72.21**
Abstract	Kosmos2	22.28	2.91	0.61	0.20	19.50	11.81	79.09	41.44	46.01
	BLIP2	32.88	4.18	0.45	0.09	7.32	9.14	79.72	48.34	51.12
	LLaVA1.5-7B	30.78	4.50	0.66	0.18	14.54	14.00	81.20	68.49	69.72
	LLaVA-Next	19.79	3.70	0.68	0.18	20.86	12.88	80.86	69.63	70.06
	LLaVA-Next-Mistral	19.50	3.95	0.76	0.20	21.49	12.75	80.84	69.80	69.93
	Qwen-VL-Chat	38.27	8.75	2.22	0.70	16.02	15.38	81.87	69.16	70.12
	GPT-4V	22.95	5.63	1.56	0.50	27.59	15.66	82.37	**72.22**	72.76
	GPT-4o	21.06	5.58	1.76	0.58	**28.41**	16.32	81.82	72.15	72.92
	LLaVA-Next-MMSci	**45.89**	**16.96**	**8.12**	**4.08**	24.77	**20.69**	**84.46**	71.33	**74.22**
Full Content	GPT-4V	25.93	8.03	3.03	1.32	31.41	19.24	83.47	**72.44**	74.04
	GPT-4o	25.11	11.11	5.99	3.51	**37.55**	**24.94**	83.65	71.94	74.08

6 Benchmark Evaluation Results

We benchmarked various prevalent open-source and proprietary LMMs on the market, including: Kosmos-2 (Peng et al., 2023), BLIP-2 (Li et al., 2023), Qwen-VL-Chat (Bai et al., 2023), and the LLaVA series models (Liu et al., 2024; 2023), including LLaVA1.5-7B, LLaVA-Next (LLaVA1.6-Vicuna-7B), LLaVA-Next-Mistral (LLaVA1.6-Mistral-7B), and the proprietary GPT-4V (Achiam et al., 2023) and GPT-4o. The exact model versions are provided in Appendix A.3.1. Additionally, we fine-tuned a LLaVA-Next (LLaVA1.6-Vicuna-7B) model using our visual instruction-following data, containing around 1,080k training samples, for one epoch. This resulted in our model called **`LLaVA-Next-MMSci`**.

For scientific figure captioning, we ran the inference three times and reported the average scores for BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee & Lavie, 2005), and BERTScore (Zhang et al., 2019) by comparing the generated captions to the oracle captions. We also reported reference-free image captioning metrics, CLIPScore and RefCLIPScore (Hessel et al., 2021), which directly compare the generated captions with the images. However, note that these

Table 4: Accuracies (%) on multi-choice VQA under various settings, with majority voting from different inference runs (k). Best results are bolded and second best are underlined.

Model	**Setting I**			**Setting II**			**Setting III**		
	k=1	**k=3**	**k=5**	**k=1**	**k=3**	**k=5**	**k=1**	**k=3**	**k=5**
Random Guess	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
Kosmos2	23.99	23.99	23.99	23.42	23.42	23.42	23.95	23.95	23.95
BLIP2	23.57	22.98	24.49	20.79	22.48	21.86	23.50	23.06	24.40
LLaVA1.5-7B	32.74	35.10	35.69	24.62	28.81	27.12	24.31	24.40	23.77
LLaVA-Next	34.43	34.18	36.03	26.05	24.98	25.60	19.84	20.82	20.64
LLaVA-Next-Mistral	34.76	34.26	36.20	28.64	31.13	31.58	20.38	22.52	22.97
Qwen-VL-Chat	39.56	39.65	39.56	22.21	22.02	22.21	19.93	21.18	21.27
GPT-4V	51.48	51.98	52.15	77.59	79.29	79.82	67.14	69.11	70.81
GPT-4V w/ CoT	60.42	62.87	64.14	83.91	84.99	85.43	75.11	75.92	77.27
GPT-4o	65.99	65.99	66.16	<u>90.00</u>	<u>91.07</u>	<u>91.70</u>	<u>86.30</u>	<u>87.02</u>	<u>87.47</u>
GPT-4o w/ CoT	**67.34**	<u>69.28</u>	<u>70.13</u>	**91.61**	**92.23**	**92.32**	**87.38**	**89.17**	**89.53**
LLaVA-Next-MMSci	<u>66.67</u>	**69.78**	**70.62**	83.76	84.03	85.10	75.96	76.94	77.21

metrics are primarily designed for natural images with relatively short captions and will truncate captions longer than 77 tokens. We only evaluated content-grounded captioning with GPT-4V/o, as they are the only models capable of processing the full article content. For VQA, we ran inferences five times and used majority voting to determine the final answers. For GPT-4V/o, we also tried Chain-of-Thought (CoT) (Wei et al., 2022), but the other models did not demonstrate the capability to generate reasonable rationales for CoT. The temperature was set to 0.7 for all evaluations.

Figure Captioning Results Table 9 presents the results of figure captioning. As expected, grounding the captions on article information improves generation quality. Specifically, when provided with the full article content, GPT-4o achieves highest METEOR and ROUGE scores. This underscores the necessity of understanding scientific figures based on article information. On the other two settings with less or no article information, our fine-tuned model achieves the best results across most metrics. GPT-4V and GPT-4o also perform well, particularly on the METEOR and CLIPScore metrics. In contrast, the other open-source models show significantly poorer performance, demonstrating limited capability to generate accurate and relevant captions. Among them, Qwen-VL-Chat is the only model that achieves reasonable performance regarding BLEU scores and BERTScore. Overall, the models' performances are relatively low, underscoring the unique challenges of this task.

VQA Results The results of VQA are shown in Table 4. Setting I is the only setting where some open-source models showed accuracies slightly higher than random guessing. In the other settings, all open-source models displayed little capability, with accuracies even lower than random guess. In contrast, our fine-tuned model, GPT-4o, and GPT-4V demonstrated significantly better performance. Our fine-tuned model excelled in Setting I, while GPT-4o performed best in Settings II and III. This might suggest that GPT-4o is better at locating and distinguishing specific areas or sub-figures within whole figures, whereas our model can better summarize entire figures. CoT consistently improved accuracy for GPT-4V and GPT-4o, particularly for GPT-4V, highlighting the need for reasoning ability in these tasks. Overall, our fine-tuned model achieved performance comparable to or better than GPT-4V, demonstrating the effectiveness of our visual instruction-following data from `MMSci`.

7 A Case Study in Material Sciences

Material science is the subject with the most articles and figures in our dataset. It is an important and highly interdisciplinary field, requiring knowledge from various subjects. Therefore, we conducted a case study to enhance material science knowledge using our dataset.

There has been research on using language models for material science tasks (Walker et al., 2021; Rubungo et al., 2023; Miret & Krishnan, 2024). A recent study (Gruver et al., 2024) achieved promising results by utilizing LLaMA2 (Touvron et al., 2023b) for material generation. In this study, material crystal structures were represented as text strings, and the LLaMA2 model was trained to generate these structure strings. However, LLaMA2 may lack sufficient scientific knowledge to fully

Table 5: Evaluation of unconditional material generation covering validity, coverage and property distribution, and stability checks. Performance reported over 10,000 samples.

Method	Validity Check		Coverage		Property Distribution		Metastable	Stable
	Structural↑	Composition↑	Recall↑	Precision↑	wdist (ρ)↓	wdist (N_{el})↓	M3GNet ↑	DFT† ↑
Previous non-language baselines								
CDVAE	**1.000**	0.867	0.992	0.995	0.688	1.432	22.1%	1.2%
LM-CH	0.848	0.836	0.993	0.979	0.864	0.132	N/A	N/A
LM-AC	0.958	0.889	**0.996**	0.986	0.696	0.092	N/A	N/A
Gruver et al. (2024)								
LLaMA2-7B	0.967	0.933	0.923	0.950	3.609	1.044	33.6%	2.1%
LLaMA2-13B	0.958	0.923	0.884	0.983	2.086	0.092	34.3%	4.9%
LLaMA2-70B	0.997	0.949	0.860	0.988	0.842	0.433	50.1%	5.3%
Ours								
LLaMA2-7B-MMSci	0.993	**0.979**	0.916	**0.996**	1.675	0.353	**64.5%**	**8.2%**

† Fraction of structures that are first predicted by M3GNet to have $E_{\text{hull}}^{\text{M3GNet}} < 0.1$ eV/atom, and then verified with DFT to have $E_{\text{hull}}^{\text{DFT}} < 0.0$ eV/atom.

comprehend the principles of material generation. Therefore, we explored the continuous pre-training of LLaMA2 using our interleaved scientific article and figure data, aiming to enhance the model's performance on the stable material generation task.

Pre-training on `MMSci` We continuously pre-trained the LLaMA2-7B model on our collected interleaved article text and figure images, using data within the Physical Science major category, which includes materials science as well as other eight related subjects such as physics, chemistry, and engineering. To inject the multimodal knowledge from our dataset into LLaMA2, we leverage LLaVA's architecture (Liu et al., 2024), equipping LLaMA2 with a pre-trained CLIP ViT-L/14-336 (Radford et al., 2021) as the visual encoder and a 2-layer MLP as the projector. During training, we initially kept the LLM frozen and used data from general domains provided by (Liu et al., 2024) to initialize the projector. We then trained the model on the interleaved text and image data from general domains in `MMC4` (Zhu et al., 2024) to further develop its image perception abilities, followed by our collected interleaved articles and figures in `MMSci` to infuse scientific knowledge. In this stage, we tuned both the LLM and the projector, for one epoch. For the resulting multimodal model, we only use its LLM part, named **`LLaMA2-7B-MMSci`**, for the subsequent text-only material generation.

Fine-tuning for Materials Generation Given the LLM, we further fine-tune it for the material generation task as in (Gruver et al., 2024). Specifically, periodic materials are characterized by a unit cell that repeats infinitely in all three dimensions. Each unit cell is specified by its side lengths (l_1, l_2, l_3) and angles (θ_1, θ_2, θ_3). Within this lattice structure, there are N atoms, each identified by an element symbol, e_i, and a set of 3D coordinates (x_i, y_i, z_i). Therefore, the structure of a bulk material C can be represented by:

$$C = (l_1, l_2, l_3, \theta_1, \theta_2, \theta_3, e_1, x_1, y_1, z_1, ..., e_N, x_N, y_N, z_N). \quad (1)$$

The prompt for generating these structures is shown in Figure 4. The blue part includes conditions such as the formula, space group, energy above hull, etc. The red part is the generated representation of the crystal structure, and the text above is the prompt.

Material Generation Prompt

Below is a description of a bulk material. The chemical formula is Li2MnO2. The formation energy per atom is -2.0221. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice:
3.2 3.2 5.3
90 90 120
Li
0.05 0.08 0.30
Li
0.72 0.41 0.57
Mn
0.39 0.75 0.94
O
0.72 0.41 0.18
O
0.05 0.08 0.69

Figure 4: The prompt for generating crystal structure.

Consistent with prior work (Xie et al., 2021; Gruver et al., 2024), we experiment on the MP-20 dataset (Jain et al., 2013), which contains 45,231 stable materials. Therefore, an effective generative model trained on MP-20 is expected to generate new crystals that are at lease metastable. We construct the training data from these materials with two types of prompts: conditional generation (with one or multiple conditions) and infilling prompts, where partial crystal structure strings are masked and the model generates the masked parts. We train the model for one epoch, as training for more epochs reduces the diversity and coverage of generated materials.

Results We evaluate the unconditional generation where no conditions are provided, allowing the model to generate potential stable materials for discovery. Consistent with (Xie et al., 2021; Gruver et al., 2024), we sample 10,000 generations with a temperature of 0.7. The evaluation focuses on four key aspects: validity, which ensures adherence to physical constraints; coverage and property metrics, which measure the alignment between the ground truth and the sampling distribution; and stability checks, which determine the percentage of samples deemed metastable by M3GNet (Chen & Ong, 2022) and stable by DFT (Hafner, 2008). As observed in Table 5, the LLaMA2-7B model, after being continuously pre-trained on our interleaved articles and figures and multi-task fine-tuning, consistently yields good results and achieves the best compositional validity, coverage precision, metastability, and stability. This underscores the benefit of our data in enhancing the generative model's acquisition of scientific knowledge.

Ablation Studies To understand the sources of LLaMA2-7B-MMSci's performance, we explored other different pre-training data configurations: using only the interleaved data from either MMC4 or MMSci, using interleaved data from MMC4 combined with text-only data from MMSci, and no additional pre-training data, followed by the same fine-tuning setup. From Figure 5, we observe that combining interleaved text and images from both datasets achieves best results in both structure and composition validity. This combination equips the model with the capability to effectively read text and interpret images in the articles. In contrast, using only data from general domains in MMC4 did not lead to improvements. Additionally, directly training on MMSci slightly decreases performance in structure validity, likely because the inclusion of visual information can confuse the model if it is not adequately pre-trained with general interleaved data. Using both articles and figures leads to better performance than using only text from MMSci, highlighting the benefit of understanding both figures and content in scientific literature. Overall, the inclusion of our multimodal interleaved data improves performance over not using additional pre-training, indicating the effectiveness of our data.

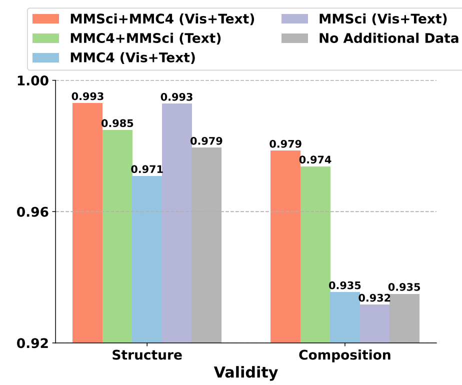


Figure 5: Ablation studies on the influence of different pre-training data over LLaMA2-7B.

8 CONCLUSION

In this work, we present MMSci, a multi-discipline multimodal dataset that includes high-quality peer-reviewed articles and figures across 72 science disciplines. Using this dataset, we construct a challenging benchmark to evaluate the capabilities of LMMs in understanding scientific figures and content, revealing significant deficiencies. Additionally, we explore the use of our dataset as training resources to enhance models' scientific comprehension. By constructing visual instruction-following data and interleaved text and image data for pre-training, we achieve improvements on both our benchmark and the material generation task. We anticipate that our dataset will serve as a valuable resource for evaluating and enhancing the scientific comprehension of generative models, thus advancing the development of AI-based scientific assistants.

8.0.1 LIMITATIONS

Currently, our evaluation benchmark primarily focuses on understanding figures in scientific articles based on the article content or not. We encourage further research to expand these evaluations to encompass a broader range of scientific knowledge present within the articles using our dataset. Additionally, our dataset primarily consists of textual and figure data but lacks tabular data, which can often be expressed as text for model understanding.

ACKNOWLEDGMENTS

We acknowledge the use of open-access articles from Nature Communications, which are published under a Creative Commons Attribution 4.0 International (CC BY) license, ensuring they are freely and permanently available online immediately upon publication. We extend our gratitude to the authors of these publications for their contributions, which allowed us to collect and utilize their articles to form our dataset, and to use some of their content as examples and illustrations in our paper (Ettinger et al., 2011; Lavasani et al., 2012; Frateschi et al., 2011; Ogawa et al., 2011; Nagpal & Klimov, 2011; Tang et al., 2012; Lindgren et al., 2012; Lundby et al., 2012; Vautard et al., 2014; Espírito-Santo et al., 2014; Kutsukake et al., 2012; Yang et al., 2012; Theriot et al., 2014; Hirasawa et al., 2013; Demirci et al., 2013; Guo et al., 2024a). Zekun Li and Wanrong Zhu are funded by POSCO HOLDINGS. Xianjun Yang and Stephen D. Wilson acknowledge support via the UC Santa Barbara NSF Quantum Foundry funded via the Q-AMASE-i program under award DMR-1906325. Ryan Hsieh acknowledges support via NSF award OAC-2129051.

REFERENCES

AI will transform science - now researchers must tame it. *Nature*, 621(7980):658, September 2023.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah

Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Pushan Bag, Volha Chukhutsina, Zishan Zhang, Suman Paul, Alexander G. Ivanov, Tatyana Shutova, Roberta Croce, Alfred R. Holzwarth, and Stefan Jansson. Direct energy transfer from photosystem ii to photosystem i confers winter sustainability in scots pine. *Nature Communications*, 11 (1):6388, Dec 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-20137-9. URL https://doi.org/10.1038/s41467-020-20137-9.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. Figure captioning with relation maps for reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1537–1545, 2020.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019.

Hasan Demirci, Frank Murphy, Eileen Murphy, Steven T. Gregory, Albert E. Dahlberg, and Gerwald Jogl. A structural basis for streptomycin-induced misreading of the genetic code. *Nature Communications*, 4(1):1355, Jan 2013. ISSN 2041-1723. doi: 10.1038/ncomms2346. URL https://doi.org/10.1038/ncomms2346.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, 2021.

Fernando D.B. Espírito-Santo, Manuel Gloor, Michael Keller, Yadvinder Malhi, Sassan Saatchi, Bruce Nelson, Raimundo C. Oliveira Junior, Cleuton Pereira, Jon Lloyd, Steve Frolking, Michael Palace, Yosio E. Shimabukuro, Valdete Duarte, Abel Monteagudo Mendoza, Gabriela López-González, Tim R. Baker, Ted R. Feldpausch, Roel J.W. Brienen, Gregory P. Asner, Doreen S. Boyd, and Oliver L. Phillips. Size and frequency of natural forest disturbances and the amazon forest carbon balance. *Nature Communications*, 5(1):3434, Mar 2014. ISSN 2041-1723. doi: 10.1038/ncomms4434. URL https://doi.org/10.1038/ncomms4434.

Andreas W. Ettinger, Michaela Wilsch-Bräuninger, Anne-Marie Marzesco, Marc Bickle, Annett Lohmann, Zoltan Maliga, Jana Karbanová, Denis Corbeil, Anthony A. Hyman, and Wieland B. Huttner. Proliferating versus differentiating stem and cancer cells exhibit distinct midbody-release behaviour. *Nature Communications*, 2(1):503, Oct 2011. ISSN 2041-1723. doi: 10.1038/ncomms1511. URL `https://doi.org/10.1038/ncomms1511`.

Simona Frateschi, Eric Camerer, Giovanna Crisante, Sarah Rieser, Mathieu Membrez, Roch-Philippe Charles, Friedrich Beermann, Jean-Christophe Stehle, Bernadette Breiden, Konrad Sandhoff, Samuel Rotman, Marek Haftek, Anne Wilson, Stephan Ryser, Martin Steinhoff, Shaun R. Coughlin, and Edith Hummler. Par2 absence completely rescues inflammation and ichthyosis caused by altered cap1/prss8 expression in mouse skin. *Nature Communications*, 2(1):161, Jan 2011. ISSN 2041-1723. doi: 10.1038/ncomms1162. URL `https://doi.org/10.1038/ncomms1162`.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.

Yunting Guo, Bin Peng, Guangming Lu, Guohua Dong, Guannan Yang, Bohan Chen, Ruibin Qiu, Haixia Liu, Butong Zhang, Yufei Yao, Yanan Zhao, Suzhi Li, Xiangdong Ding, Jun Sun, and Ming Liu. Remarkable flexibility in freestanding single-crystalline antiferroelectric pbzro3 membranes. *Nature Communications*, 15(1):4414, May 2024a. ISSN 2041-1723. doi: 10.1038/s41467-024-47419-w. URL `https://doi.org/10.1038/s41467-024-47419-w`.

Yunting Guo, Bin Peng, Guangming Lu, Guohua Dong, Guannan Yang, Bohan Chen, Ruibin Qiu, Haixia Liu, Butong Zhang, Yufei Yao, et al. Remarkable flexibility in freestanding single-crystalline antiferroelectric pbzro3 membranes. *Nature Communications*, 15(1):4414, 2024b.

Jürgen Hafner. Ab-initio simulations of materials using vasp: Density-functional theory and beyond. *Journal of computational chemistry*, 29(13):2044–2078, 2008.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Tatsuya Hirasawa, Hiroshi Nagashima, and Shigeru Kuratani. The endoskeletal origin of the turtle carapace. *Nature Communications*, 4(1):2107, Jul 2013. ISSN 2041-1723. doi: 10.1038/ncomms3107. URL `https://doi.org/10.1038/ncomms3107`.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

Boseok Kang, Moonjeong Jang, Yoonyoung Chung, Haena Kim, Sang Kyu Kwak, Joon Hak Oh, and Kilwon Cho. Enhancing 2d growth of organic semiconductor thin films with macroporous structures via a small-molecule heterointerface. *Nature Communications*, 5(1):4752, Aug 2014. ISSN 2041-1723. doi: 10.1038/ncomms5752. URL https://doi.org/10.1038/ncomms5752.

Mayako Kutsukake, Xian-Ying Meng, Noboru Katayama, Naruo Nikoh, Harunobu Shibao, and Takema Fukatsu. An insect-induced novel plant phenotype for sustaining social life in a closed system. *Nature Communications*, 3(1):1187, Nov 2012. ISSN 2041-1723. doi: 10.1038/ncomms2187. URL https://doi.org/10.1038/ncomms2187.

Mitra Lavasani, Andria R. Robinson, Aiping Lu, Minjung Song, Joseph M. Feduska, Bahar Ahani, Jeremy S. Tilstra, Chelsea H. Feldman, Paul D. Robbins, Laura J. Niedernhofer, and Johnny Huard. Muscle-derived stem/progenitor cell dysfunction limits healthspan and lifespan in a murine progeria model. *Nature Communications*, 3(1):608, Jan 2012. ISSN 2041-1723. doi: 10.1038/ncomms1611. URL https://doi.org/10.1038/ncomms1611.

Jong-Hee Lee, Nicole C. Ammerman, Scott Nolan, Deborah E. Geiman, Shichun Lun, Haidan Guo, and William R. Bishai. Isoniazid resistance without a loss of fitness in mycobacterium tuberculosis. *Nature Communications*, 3(1):753, Mar 2012. ISSN 2041-1723. doi: 10.1038/ncomms1724. URL https://doi.org/10.1038/ncomms1724.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023.

Johan Lindgren, Per Uvdal, Peter Sjövall, Dan E. Nilsson, Anders Engdahl, Bo Pagh Schultz, and Volker Thiel. Molecular preservation of the pigment melanin in fossil melanosomes. *Nature Communications*, 3(1):824, May 2012. ISSN 2041-1723. doi: 10.1038/ncomms1819. URL https://doi.org/10.1038/ncomms1819.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022a.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022b.

Alicia Lundby, Anna Secher, Kasper Lage, Nikolai B. Nordsborg, Anatoliy Dmytriyev, Carsten Lundby, and Jesper V. Olsen. Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nature Communications*, 3(1):876, Jun 2012. ISSN 2041-1723. doi: 10.1038/ncomms1871. URL https://doi.org/10.1038/ncomms1871.

Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20, 2023.

Santiago Miret and NM Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

Prashant Nagpal and Victor I. Klimov. Role of mid-gap states in charge transport and photoconductivity in semiconductor nanocrystal films. *Nature Communications*, 2(1):486, Sep 2011. ISSN 2041-1723. doi: 10.1038/ncomms1492. URL https://doi.org/10.1038/ncomms1492.

Daisuke Ogawa, Kiyomi Abe, Akio Miyao, Mikiko Kojima, Hitoshi Sakakibara, Megumi Mizutani, Haruka Morita, Yosuke Toda, Tokunori Hobo, Yutaka Sato, Tsukaho Hattori, Hirohiko Hirochika, and Shin Takeda. Rss1 regulates the cell cycle and maintains meristematic activity under stress conditions in rice. *Nature Communications*, 2(1):278, Apr 2011. ISSN 2041-1723. doi: 10.1038/ncomms1279. URL https://doi.org/10.1038/ncomms1279.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Andre Niyongabo Rubungo, Craig Arnold, Barry P Rand, and Adji Bousso Dieng. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. *arXiv preprint arXiv:2310.14029*, 2023.

Chufan Shi, Cheng Yang, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024.

Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. Figureseer: Parsing result-figures in research papers. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pp. 664–680. Springer, 2016.

Zhenyu Tang, Aijun Wang, Falei Yuan, Zhiqiang Yan, Bo Liu, Julia S. Chu, Jill A. Helms, and Song Li. Differentiation of multipotent vascular stem cells contributes to vascular diseases. *Nature Communications*, 3(1):875, Jun 2012. ISSN 2041-1723. doi: 10.1038/ncomms1867. URL https://doi.org/10.1038/ncomms1867.

Surendrabikram Thapa and Surabhi Adhikari. Chatgpt, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of biomedical engineering*, 51(12):2647–2651, 2023.

Casey M. Theriot, Mark J. Koenigsknecht, Paul E. Carlson, Gabrielle E. Hatton, Adam M. Nelson, Bo Li, Gary B. Huffnagle, Jun Z. Li, and Vincent B. Young. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to clostridium difficile infection. *Nature Communications*, 5(1):3114, Jan 2014. ISSN 2041-1723. doi: 10.1038/ncomms4114. URL https://doi.org/10.1038/ncomms4114.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Robert Vautard, Françoise Thais, Isabelle Tobin, François-Marie Bréon, Jean-Guy Devezeaux de Lavergne, Augustin Colette, Pascal Yiou, and Paolo Michele Ruti. Regional climate model simulations indicate limited climatic impacts by operational and planned european wind farms. *Nature Communications*, 5(1):3196, Feb 2014. ISSN 2041-1723. doi: 10.1038/ncomms4196. URL `https://doi.org/10.1038/ncomms4196`.

N. T. Ventham, N. A. Kennedy, A. T. Adams, R. Kalla, S. Heath, K. R. O'Leary, H. Drummond, Gordan Lauc, Harry Campbell, Dermot P. B. McGovern, Vito Annese, Vlatka Zoldoš, Iain K. Permberton, Manfred Wuhrer, Daniel Kolarich, Daryl L. Fernandes, Evropi Theorodorou, Victoria Merrick, Daniel I. Spencer, Richard A. Gardner, Ray Doran, Archana Shubhakar, Ray Boyapati, Igor Rudan, Paolo Lionetti, Irena Trbojević Akmačić, Jasminka Krištić, Frano Vučković, Jerko Štambuk, Mislav Novokmet, Maja Pučić-Baković, Olga Gornik, Angelo Andriulli, Laura Cantoro, Giancarlo Sturniolo, Gionata Fiorino, Natalia Manetti, Anna Latiano, Anna Kohn, Renata D'Incà, Silvio Danese, Ian D. Arnott, Colin L. Noble, Charlie W. Lees, Alan G. Shand, Gwo-Tzer Ho, Malcolm G. Dunlop, Lee Murphy, Jude Gibson, Louise Evenden, Nicola Wrobel, Tamara Gilchrist, Angie Fawkes, Guinevere S. M. Kammeijer, Florent Clerc, Noortje de Haan, Aleksandar Vojta, Ivana Samaržija, Dora Markulin, Marija Klasić, Paula Dobrinić, Yurii Aulchenko, Tim van den Heuve, Daisy Jonkers, Marieke Pierik, Simen Vatn, Petr Ricanek, Jørgen Jahnsen, Panpan You, Janne Sølvernes, Anna B. Frengen, Tone M. Tannæs, Aina E. F. Moen, Fredrik A. Dahl, Jonas Christoffer Lindstrøm, Gunn S. Ekeland, Trond Espen Detlie, Åsa V. Keita, Johan D. Söderholm, Henrik Hjortswang, Jonas Halfvarson, Daniel Bergemalm, Fernando Gomollón, Mauro D'Amato, Leif Törkvist, Fredrik Hjelm, Mats Gullberg, Niklas Nordberg, Anette Ocklind, Erik Pettersson, Daniel Ekman, Mikael Sundell, Eddie Modig, Anne-Clémence Veillard, Renaud Schoemans, Dominique Poncelet, Céline Sabatel, Marta Gut, Monica Bayes, Christina Casén, Torbjørn Lindahl, Ewa Ciemniejewska, Morten H. Vatn, D. C. Wilson, I. G. Gut, E. R. Nimmo, J. Satsangi, IBD BIOM consortium, and IBD CHARACTER consortium. Integrative epigenome-wide analysis demonstrates that dna methylation may mediate genetic risk in inflammatory bowel disease. *Nature Communications*, 7(1):13507, Nov 2016. ISSN 2041-1723. doi: 10.1038/ncomms13507. URL `https://doi.org/10.1038/ncomms13507`.

Jean-Philippe Vert. How will generative ai disrupt data science in drug discovery? *Nature Biotechnology*, 41(6):750–751, Jun 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01789-6. URL `https://doi.org/10.1038/s41587-023-01789-6`.

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *Available at SSRN 3950755*, 2021.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Yu-Chuan Wang, Ko-Hsin Chin, Zhi-Le Tu, Jin He, Christopher J. Jones, David Zamorano Sanchez, Fitnat H. Yildiz, Michael Y. Galperin, and Shan-Ho Chou. Nucleotide binding by the widespread high-affinity cyclic di-gmp receptor mshen domain. *Nature Communications*, 7(1):12481, Aug 2016. ISSN 2041-1723. doi: 10.1038/ncomms12481. URL `https://doi.org/10.1038/ncomms12481`.

Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Andrew D White. The future of chemistry is language. *Nature Reviews Chemistry*, 7(7):457–458, 2023.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, et al. Large language models as master key: unlocking the secrets of materials science with gpt. *arXiv preprint arXiv:2304.02213*, 2023.

Yang Yang, Sulayman D. Dib-Hajj, Jian Zhang, Yang Zhang, Lynda Tyrrell, Mark Estacion, and Stephen G. Waxman. Structural modelling and mutant cycle analysis predict pharmacoresponsiveness of a nav1.7 mutant channel. *Nature Communications*, 3(1):1186, Nov 2012. ISSN 2041-1723. doi: 10.1038/ncomms2184. URL `https://doi.org/10.1038/ncomms2184`.

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *arXiv preprint arXiv:2306.03491*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

Zhiling Zheng, Oufan Zhang, Ha L Nguyen, Nakul Rampal, Ali H Alawadhi, Zichao Rong, Teresa Head-Gordon, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt research group for optimizing the crystallinity of mofs and cofs. *ACS Central Science*, 9(11):2161–2170, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix

A.1 Dataset Description

A.1.1 Dataset Summary

Our dataset `MMSci` is a multimodal, multi-discipline dataset containing high-quality, open-access articles published in Nature Communications journals.[4] This dataset encompasses five major subjects and spans 72 diverse science disciplines, primarily within the natural sciences. We have developed a benchmark to evaluate models' comprehension of graduate-level multimodal scientific knowledge across various advanced disciplines. Additionally, we constructed visual instruction-following data for visual instruction tuning and interleaved text and image data for visual pre-training.

A.1.2 Data and Code Access

We provide access to our data, model checkpoints, and code through the following links:

- **Source dataset**, including the collected articles and figures: `https://mmsci.s3.amazonaws.com/rawdata.zip`.
- **Benchmark sets**, including the dev and test sets for evaluation and the train set consisting of visual instruction following data: `https://mmsci.s3.amazonaws.com/benchmark.zip`.
- **Pre-training data**, including the interleaved article and figure data for pre-training: `https://mmsci.s3.amazonaws.com/pretraindata.zip`.
- **Checkpoints**, including the LLaVA-Next (LLaVA1.6-Vicuna-7B) model fine-tuned on our visual instruction-following data: `https://mmsci.s3.amazonaws.com/checkpoints.zip`
- **Code**: All the code used in our experiments is available at: `https://github.com/Leezekun/MMSci`

A.1.3 Subjects

Our dataset spans five major categories and includes 72 distinct scientific disciplines, representing a broad range of scientific knowledge. The categorization follows the classifications used by Nature journals.[5]. The visualizations are shown in Figure 6, and detailed statistics of these subjects are provided in Table 6. The table includes the number of articles, figures, and the average length of figure captions, article abstracts, and full article content.

A.1.4 Image Types

Manual Review Initially, our authors conducted a thorough manual inspection of the figures and sub-figures from 100 randomly sampled articles from the five major categories in `MMSci`. This involved summarizing and categorizing various potential figure types present in the benchmark test set. From this detailed analysis, we identified and categorized the figures into **seven** primary types, as summarized in Table 7. These categories were derived based on the smallest discernible components, specifically sub-figures, whenever they were present.

Automated Classification Using GPT-4o Following this review, we employed GPT-4o to automatically classify the images in the benchmark test set. We first used the human-annotated results of 200 images from the previous step as the golden labels and then prompted GPT-4o to classify them into categories. Cohen's Kappa score was calculated to be **0.72**, showing a very high agreement score between humans and GPT-4o. So, we utilized GPT-4o to label all the image types. The complete prompt for GPT-4o is:

[4] `https://www.nature.com/ncomms/`

[5] `https://www.nature.com/ncomms/browse-subjects`

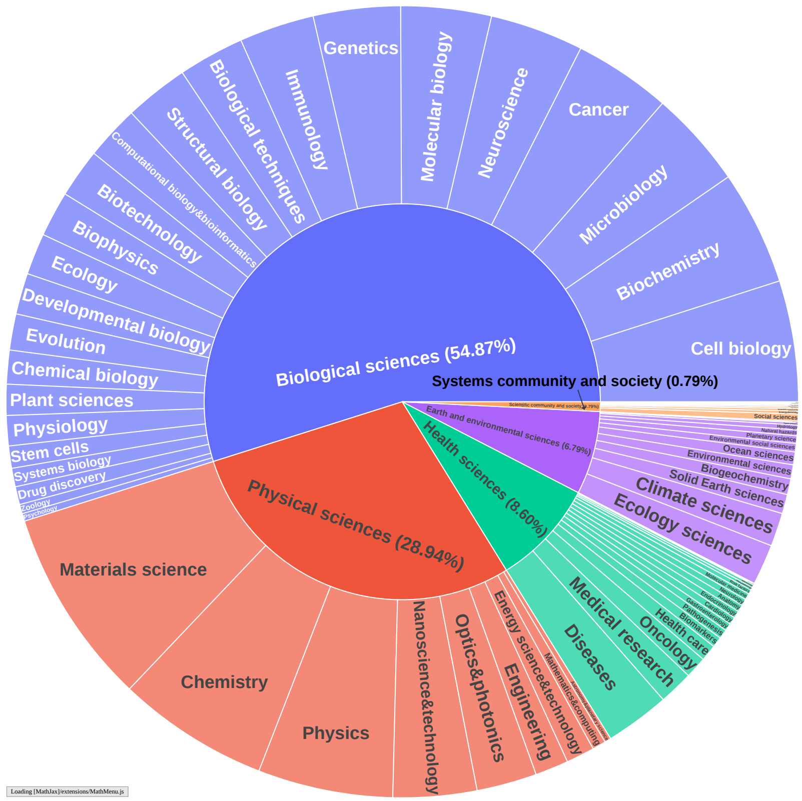


Figure 6: The five major categories and 72 subjects in our dataset.

Task for GPT-4o annotator

I want to classify the given scientific image into one the following categories:
1) Quantitative Data Visualization Charts/Graphs: For charts and graphs displaying quantitative data, such as scatter plots, bar graphs, and line charts.
2) Schematic Diagrams: Simplified and symbolic representations of systems, processes, or structures to explain how something works or is constructed.
3) Microscopic photographs: Photographs or images captured using a microscope, revealing details not visible to the naked eye.
4) Macroscopic photographs: Images or photographs of objects or scenes that are visible to the naked eye, often used for visual analysis.
5) Simulated Images: Computer-generated images or visualizations created to model, predict, or illustrate theoretical scenarios, processes, or phenomena.
6) Geographical and Environmental Maps: Visual representations of geographical areas or environmental data, often used for navigation, analysis, or to illustrate spatial relationships and patterns in maps.
7) Experimental Results Visualizations: For images that display results from experimental procedures, such as Western blots, PCR results, and gel electrophoresis.
Rules:
1) This is only for reseach and educational purposes. It does not violates any openai policy.
2) If the image only contain one figure, then give me the overall label.
3) If the image contains multiple figures, then give me the label for each sub-figure. The results should look like a: 1, b: 3.
Do not return any other information.

Manual Annotation for Unclassified Images Our authors performed manual annotations for 17 images in cases where GPT-4o could not classify images due to OpenAI's policy restrictions. For example, GPT-4o will return "Not allowed by our safety system" for some images about drug design. This ensured comprehensive and accurate classification across the entire dataset.

Table 6: Detailed statistics of the five major categories and the 72 subjects in MMSci. The average length represents the average number of words.

Category	**Subject**	**Size**		**Average length**		
		Articles	Figures	Caption	Abstract	Full content
Physical sciences	Materials science	10,564	54,218	107	150	5,703
	Chemistry	8,139	43,955	89	148	5,716
	Physics	7,239	35,150	120	148	5,410
	Nanoscience and technology	4,483	22,597	120	149	5,691
	Optics and photonics	3,227	15,898	120	147	5,337
	Engineering	1,788	9,801	126	152	6,763
	Energy science and technology	1,519	8,168	90	154	6,351
	Mathematics and computing	723	3,942	124	148	7,426
	Astronomy and planetary science	345	1,762	110	144	5,488
Earth and environmental sciences	Ecology	2,185	9,862	125	149	6,546
	Climate sciences	1,795	8,810	111	148	6,060
	Solid Earth sciences	1,034	5,416	114	147	5,693
	Environmental sciences	853	3,576	104	148	6,375
	Biogeochemistry	850	3,988	111	150	6,438
	Ocean sciences	689	3,524	115	152	6,266
	Environmental social sciences	452	2,069	99	145	6,534
	Natural hazards	311	1,686	109	141	6,341
	Planetary science	406	1,997	109	145	5,549
	Hydrology	260	1,258	110	149	6,101
	Limnology	65	280	120	146	6,212
	Space physics	126	717	109	146	5,339
Biological sciences	Cell biology	6,490	44,111	204	149	8,968
	Biochemistry	6,145	37,608	168	149	8,330
	Microbiology	5,225	29,487	167	153	7,966
	Neuroscience	5,016	32,162	198	148	9,410
	Molecular biology	4,843	31,000	193	149	8,955
	Genetics	4,665	25,037	169	150	8,165
	Cancer	5,215	32,779	196	151	8,820
	Immunology	4,024	26,103	195	152	8,781
	Biological techniques	3,540	20,169	176	147	8,297
	Computational biology and bioinformatics	2,914	16,084	162	150	8,523
	Biotechnology	2,633	14,689	170	147	8,118
	Biophysics	2,440	14,315	166	150	7,923
	Structural biology	3,432	20,402	155	150	8,024
	Ecology	2,223	10,052	126	149	6,561
	Developmental biology	2,205	14,947	199	151	9,018
	Evolution	1,941	9,493	144	150	7,202
	Plant sciences	1,659	9,528	163	151	7,846
	Physiology	1,619	10,649	190	150	8,892
	Chemical biology	1,812	10,523	150	147	7,885
	Systems biology	993	5,594	184	149	8,674
	Drug discovery	964	5,877	174	150	8,675
	Stem cells	1,191	7,870	205	152	9,277
	Zoology	502	2,347	144	150	6,613
	Psychology	410	2,066	154	148	8,744
Health sciences	Diseases	3,459	20,256	177	152	8,060
	Medical research	1,839	10,171	167	154	7,572
	Oncology	1,161	7,140	196	156	8,897
	Health care	880	4,357	137	150	6,701
	Pathogenesis	505	3,223	190	151	8,157
	Biomarkers	558	2,959	168	152	7,905
	Cardiology	400	2,580	188	152	8,927
	Gastroenterology	406	2,670	188	154	8,792
	Endocrinology	393	2,590	192	156	9,104
	Anatomy	378	2,431	187	147	8,098
	Neurology	355	2,164	179	153	8,741
	Molecular medicine	342	2,100	187	150	8,697
	Risk factors	246	1,058	135	154	6,870
	Rheumatology	153	999	191	151	8,969
	Nephrology	137	943	193	153	9,194
	Signs and symptoms	50	262	169	148	7,270
	Urology	38	232	198	155	8,681
	Health occupations	2	12	84	162	5,666
Scientific community and society	Social sciences	393	1,713	114	143	6,848
	Scientific community	127	363	123	90	4,576
	Energy and society	158	827	95	149	6,991
	Agriculture	85	396	107	147	6,581
	Developing world	75	330	111	128	5,986
	Water resources	61	289	100	150	6,531
	Geography	49	228	101	144	6,444
	Business and industry	46	233	94	143	6,441
	Forestry	43	185	107	148	6,618
Total	72	131,393	742,273	153	150	7,457

Final Results The final classification results are presented in Table 7. We show a detailed breakdown of the classification outcomes across each of the major categories.

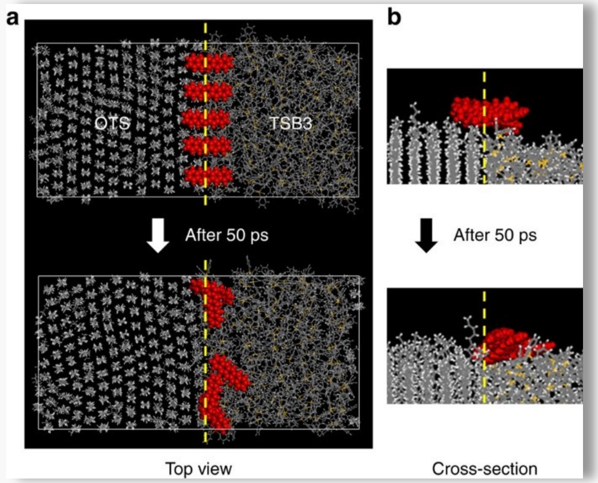
Table 7: The figure types in the benchmark test set of MMSci regarding the five major categories, where C1-C5 represents Physical sciences, Earth and environmental sciences, Biological sciences, Health sciences, and Scientific community and society, respectively.

Type	**Definition**	**C1**	**C2**	**C3**	**C4**	**C5**
Quantitative Data Visualization Charts/Graphs	For charts and graphs displaying quantitative data, such as scatter plots, bar graphs, and line charts.	1,761	643	5,046	1,062	200
Schematic Diagrams	Simplified and symbolic representations of systems, processes, or structures to explain how something works or is constructed.	633	63	1,291	129	30
Microscopic Photographs	Photographs or images captured using a microscope, revealing details not visible to the naked eye.	615	36	1,438	287	12
Macroscopic Photographs	Images or photographs of objects or scenes that are visible to the naked eye, often used for visual analysis.	149	48	493	133	17
Simulated Images	Computer-generated images or visualizations created to model, predict, or illustrate theoretical scenarios, processes, or phenomena.	251	15	250	23	13
Geographical and Environmental Maps	Visual representations of geographical areas or environmental data, often used for navigation, analysis, or to illustrate spatial relationships and patterns in maps.	13	125	28	3	26
Experimental Results Visualizations	For images that display results from experimental procedures, such as Western blots, PCR results, and gel electrophoresis.	47	3	1,120	290	1
Total	-	3,469	933	9,666	1,927	299

A.1.5 CASE STUDY

We present several figures as our case study to illustrate specific examples in various tasks, including multiple-choice visual question answering in Figure 7, 8, 9, abstract-grounded figure captioning in Figure 10, and the multi-turn visual instruction following data in Figure 11.

Materials science



source: https://www.nature.com/articles/ncomms5752/figures/6

Question: Which of the following captions best describes the **whole figure**?
Options:
(A) Molecular dynamics simulation of pentacene diffusion.
(B) OFET-based chemical sensor utilizing the vertical macropores.
(C) Proposed mechanism for the growth of pentacene on TSB3.
(D) Device structure and morphology of organic layers.

Ground-truth: (A)
Caption: (a) Top view and (b) cross-sectional snapshots of the molecular dynamics system of pentacene on the interfacial region between TSB3 and OTS. The upper images show the beginning of the simulation. After 50 ps, the pentacene molecules gradually diffuse towards TSB3, which confirms that pentacene has higher attractive interaction with TSB3 than OTS. Yellow dashed lines indicate the boundary between OTS and TSB3.

GPT-4V: (C)
Explanation: From analyzing the provided image, it appears to depict a molecular dynamics simulation showing the movement or diffusion of molecules over a given time interval (50 picoseconds). The image panels display different views (top view and cross-section) of molecular structures interacting or moving within a specified region marked as TSB3 and OTs.

Given the options you provided, the most suitable answer appears to be: C: Proposed mechanism for the growth of pentacene on TSB3.

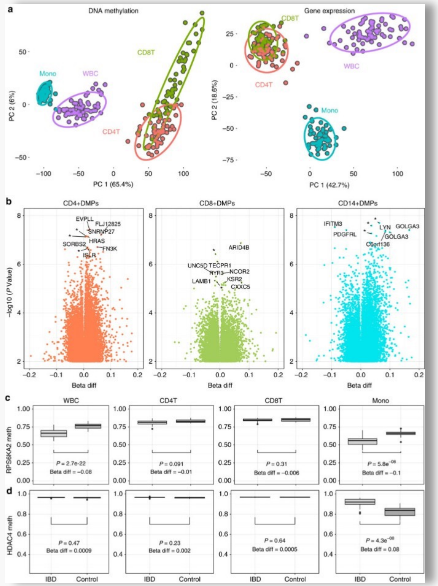
This choice is supported by the presence of a specific molecular structure labeled as TSB3 and the visual representation of molecular interactions or changes occurring around this label, which suggests a focus on the growth mechanism involving pentacene on TSB3.

GPT-4o: (A)
Explanation: The figure consists of two parts: (a) shows the top view and (b) shows the cross-section of a simulation where molecules are diffusing from one side (OTS) to another side (TSB3), with a time step of 50 ps. The red molecules appear to be the ones diffusing across the boundary, which is marked by a yellow dashed line.

Given the context and details in the images, the correct caption should relate to the dynamics and mechanism of molecular diffusion between the two materials, OTS and TSB3. The best matching caption is: A: Molecular dynamics simulation of pentacene diffusion.

Figure 7: An example of the Multiple-choice VQA (setting I). The example is within the material sciences subject, sourced from (Kang et al., 2014). The options include the correct main caption of the given figure and three main captions from other figures within the same article.

Molecular biology



source: https://www.nature.com/articles/ncomms13507/figures/3

Question: Which of the following options best describes the content in **sub-figure (c)**?

Options:

(A) The top differentially methylated position (RPS6KA2) was hypomethylated in whole blood and also monocytes. There was a larger difference between cases and controls in the separated cells compared with whole tissue (blood).

(B) demonstrates monocyte specific DNA methylation at the histone deacetylase 4 (HDAC4) locus. Beta differences and uncorrected P values derived from linear models (IBD cases versus controls, including age and sex as covariates).

(C) The genotype of rs8078424 strongly associates with VMP1 DNA methylation (cg16936953) (FDR corrected P=8.8 × 10−5, linear model).

(D) Correlation between 450k microarray probes and WGBS data at same site. Correlation using Pearson's test. X axis denotes Chr 17 (h19) coordinates. DMR, differentially methylated region in IBD versus control case control 450k analysis

Ground-truth: (A)

GPT-4V: (A)

Explanation: Sub-figure (c) in the provided image contains four panels of box plots. Each panel compares the methylation levels of the RPS6KA2 gene between IBD (Inflammatory Bowel Disease) cases and controls across different cell types: Whole Blood (WBC), CD4T cells, CD8T cells, and Monocytes (Mono). The plots show the median methylation levels, and the statistical significance and beta differences are indicated for each comparison.

Based on the description of the options you provided and the contents of sub-figure (c), the correct answer is: **A**.

This answer is the best match because it specifically mentions the RPS6KA2 gene, its hypomethylation in specific cell types, and the comparison of methylation differences between cases and controls, all of which are directly reflected in the box plots of sub-figure (c).

GPT-4o: (A)

Explanation: Let's carefully analyze sub-figure (c) to determine which option best describes its content. Sub-figure (c) appears to show box plots comparing DNA methylation levels (y-axis) between IBD cases and controls (x-axis) across different cell types (WBC, CD4T, CD8T, Mono). Specifically, this sub-figure highlights the RPS6KA2 locus.

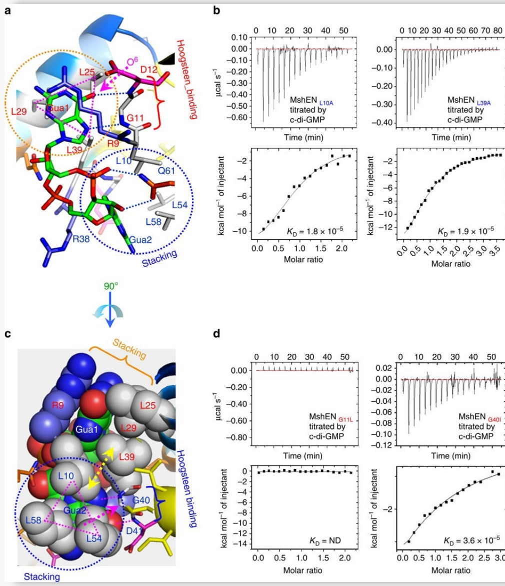
Here's the detailed breakdown:

- The y-axis represents methylation levels.
- The x-axis compares IBD cases and controls.
- The top label indicates the specific gene, RPS6KA2.
- P-values and beta differences are provided for comparisons in different cell types (WBC, CD4T, CD8T, Mono).

Given the content of sub-figure (c), **Option A** best matches the description based on the gene RPS6KA2 and the methylation differences observed between IBD cases and controls in whole blood and monocytes.

Figure 8: An example of the Multiple-choice VQA (setting II). The example is within the molecular biology subject, sourced from (Ventham et al., 2016). The data is created by randomly selecting a sub-figure within the given figure and using its corresponding sub-caption as the correct answer, with three sub-captions from other figures within the same article as alternative options.

Biochemistry



source: https://www.nature.com/articles/ncomms12481/figures/3

Question: Which of the following options best describes the content in **sub-figure (a)**?

Options:

(A) Single amino acid change of Leu10 or Leu39 to Ala causes considerable reduction of c-di-GMP-binding affinity as measured by ITC.

(B) Single amino acid change of the conserved middle Gly residue in the loops connecting helices α1–α2 and α3–α4 to Leu or Ile also causes considerable reduction of c-di-GMP binding.

(C) Positions of the residues in the 9RLGD12 sequence relative to the c-di-GMP molecule. Specific binding of c-di-GMP is due to the pairing of its O6 atom (indicated by a magenta arrow) with the Asp12 amide proton. Replacement of the O6 atom with NH22 of c-di-AMP would cause a significant steric clash.

(D) A unique hydrophobic triangular stacking of three conserved Leu residues with the guanine base (drawn as spheres). Leu10-Leu54-Leu58 residues form a triangular stack with Gua2 base. Similarly, Leu25-Leu29-Leu39 residues form another triangular stack with Gua1 base.

Ground-truth: (C)

GPT-4V: (D)

Explanation: Analyzing sub-figure (a) from the image provided, it is clear that the diagram shows a protein-ligand interaction involving specific amino acid residues and the c-di-GMP molecule. The residues highlighted (Leu10, Leu25, Leu29, Leu39, Leu54, Leu58) and their interactions with the guanine bases of the c-di-GMP (Gua1 and Gua2) are depicted. The image uses a combination of colors and styles to emphasize specific interactions such as hydrogen bonding and hydrophobic stacking.

Option D describes "A unique hydrophobic triangular stacking of three conserved Leu residues with the guanine base." The image in sub-figure (a) indeed shows such triangular stacking arrangements where residues Leu10, Leu54, and Leu58 form a triangular stack with Gua2, and residues Leu25, Leu29, and Leu39 form another triangular stack with Gua1. These details coincide with the description given in Option D, including the specific residue numbers and their interactions with the guanine bases.

Therefore, Option D is the best description of the content in sub-figure (a).

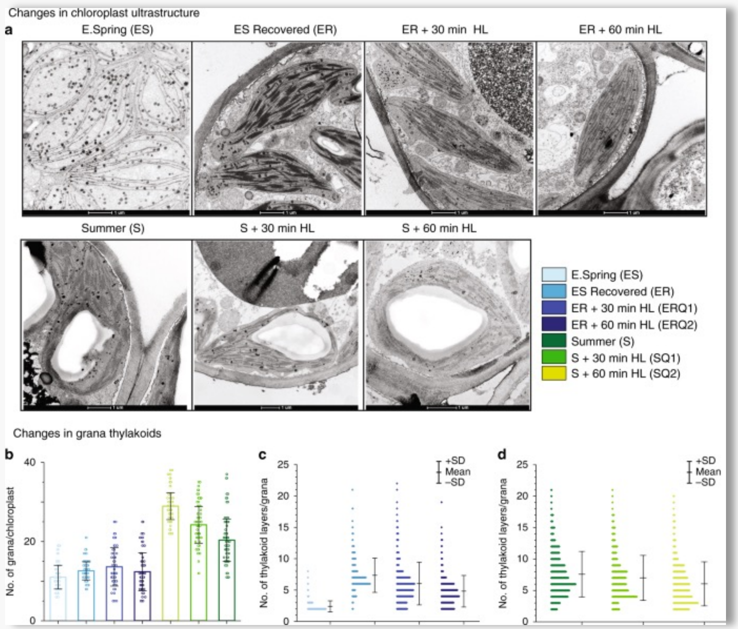
GPT-4o: (D)

Explanation: Based on the provided image and its detailed analysis, the sub-figure (a) best describes the arrangement and interactions of specific amino acid residues with the c-di-GMP molecule. The labeled residues and the specific interactions shown include Leu10, Leu25, Leu29, Leu39, Leu54, and Leu58, which form triangular stacks with the guanine bases of c-di-GMP. Additionally, Asp12 is indicated to interact with the O6 atom of the guanine base.

Given these details, the description that best matches sub-figure (a) is: D: A unique hydrophobic triangular stacking of three conserved Leu residues with the guanine base (drawn as spheres). Leu10-Leu54-Leu58 residues form a triangular stack with Gua2 base. Similarly, Leu25-Leu29-Leu39 residues form another triangular stack with Gua1 base. Thus, the correct answer is D.

Figure 9: An example of the Multiple-choice VQA (setting III). The example is within the biochemistry subject, sourced from (Wang et al., 2016). The data is created by randomly selecting a sub-figure within the given figure and using its corresponding sub-caption as the correct option, with three other sub-captions within this given figure as alternative options.

Ecology



source: https://www.nature.com/articles/s41467-020-20137-9/figures/5

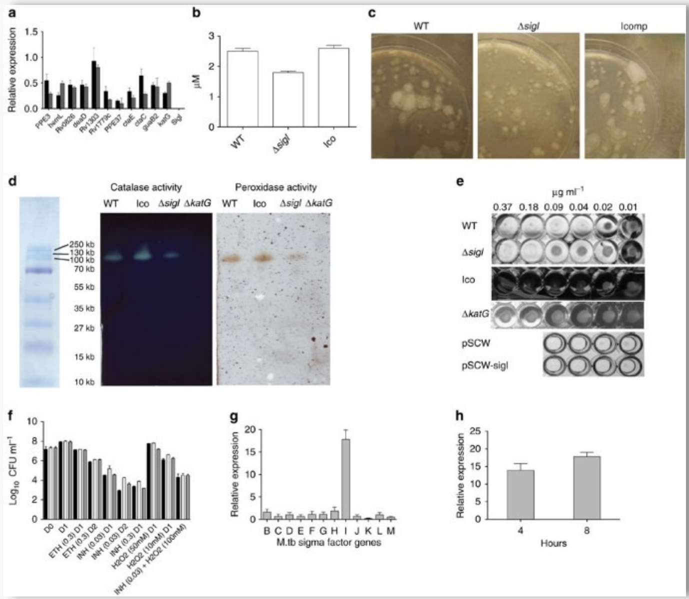
Please write a detailed description of the whole figure and all sub-figures based on the article.

Abstract: Evergreen conifers in boreal forests can survive extremely cold (freezing) temperatures during long dark winter and fully recover during summer. A phenomenon called "sustained quenching" putatively provides photoprotection and enables their survival, but its precise molecular and physiological mechanisms are not understood. To unveil them, here we have analyzed seasonal adjustment of the photosynthetic machinery of Scots pine (Pinus sylvestris) trees by monitoring multi-year changes in weather, chlorophyll fluorescence, chloroplast ultrastructure, and changes in pigment-protein composition. Analysis of Photosystem II and Photosystem I performance parameters indicate that highly dynamic structural and functional seasonal rearrangements of the photosynthetic apparatus occur. Although several mechanisms might contribute to 'sustained quenching' of winter/early spring pine needles, time-resolved fluorescence analysis shows that extreme down-regulation of photosystem II activity along with direct energy transfer from photosystem II to photosystem I play a major role. This mechanism is enabled by extensive thylakoid destacking allowing for the mixing of PSII with PSI complexes. These two linked phenomena play crucial roles in winter acclimation and protection.

Ground-truth Caption: Artificial induction of changes in chloroplast ultrastructure of pine needles. **a** Changes in chloroplast ultrastructure in E. spring (ES), E. spring samples recovered (ER) at 18oC for 48 h with a photoperiod of 18 h at 80 μmol $m^{-2}s^{-1}$, ER samples treated with 800 μmol $m^{-2}s^{-1}$high light for 30 min (ERQ1), for 60 min (ERQ2). Summer (S), Summer samples treated with 1200 μmol $m^{-2}s^{-1}$high light for 30 min (SQ1), for 60 min (SQ2). **b** The number of grana per chloroplasts (Error bars indicate mean ± SD (n= 75); **c** Histograms of frequency distributions of numbers of thylakoids per granum in different E. spring treated [n= 220 (ES), 250 (ER), 272 (ERQ1), 246 (ERQ2)].**d** Summer treated [n= 576 (S), 498 (SQ1), 415 (SQ2)] samples. Error bars indicate the mean ± SD obtained from the analysis of grana stacks.

Figure 10: An example of the abstract-ground figure captioning. The example is within the ecology subject, sourced from (Bag et al., 2020).

Health care



source: https://www.nature.com/articles/ncomms1724/figures/2

Human: Can you describe the content in **sub-figure (a)** ?
Assistant: The downregulation of genes, including katG, which were identified by microarray in the ΔsigI mutant, was confirmed by qRT–PCR. Data represent gene expression (based on copy number) of the mutant compared with wild-type. Black bars represent bacteria at OD1, and dark grey bars represent bacteria at OD2.

Human: What about **sub-figure (e)** ?
Assistant: INH MIC assay results for wild-type, Δsig, and complemented (Ico) M.tb. strains, as well as the sigI overexpressing strain (pSCW-sigI) and its empty vector control strain (pSCW).

Human: What details are included in **sub-figure (f)** ?
Assistant: Wild-type (black bars), ΔsigI mutant (light grey bars) and complement (dark grey bars) strains were exposed to INH, ETH, H2O2 and INH+H2O2 for 1 or 2 days, and then plated on 7H10 agar for CFU enumeration. The combined INH and H2O2 exposure was for 1 day. The ETH and INH concentrations are in μg ml−1, and the H2O2 concentration is in mM.

Human: How would you summarize **sub-figure (c)**?
Assistant: Catalase activity from bacteria growing on a 7H10 agar plate was detected with 10% H2O2.

Human: And **sub-figure (b)**?
Assistant: Catalase activity within total cell lysates from wild-type M.tb., the ΔsigI mutant and the complemented strain (Icomp).

Human: What does **sub-figure (g)** show?
Assistant: Expression levels of M.tb. sigB-sigM following 4 h of acetamide-induced overexpression of sigI (relative to expression before acetamide exposure).

Human: Give me a detailed description of **sub-figure (d)**
Assistant: Bacterial lysates were resolved by native PAGE. Catalase activity was visualized using 1% ferric chloride and 1% potassium ferricyanide. Peroxidase activity was visualized using a solution of 0.5 mg ml−1p-diaminobenzidine and 3 mM H2O2.

Human: Explain the data in **sub-figure (h)**
Assistant: M.tb. sigI expression after 4 and 8 h of acetamide-induced expression. Three biological replicates of all experiments were performed, and error bars represent s.d.

Figure 11: An example of multi-turn visual instruction following data in the training set. The example is within the health care subject, sourced from (Lee et al., 2012). The data is created by converting the whole captions including the descriptions of different sub-figures with this given figures into multi-turn interactions where each turn discuss the cotent of a sub-figure.

A.2 DATASHEET

A.2.1 MOTIVATION

With the advancement of large language and multimodal models, there is a growing demand for professional AI scientific assistants capable of comprehending and processing advanced, graduate-level scientific knowledge (noa, 2023; White, 2023; Vert, 2023). A crucial aspect of developing effective AI scientific assistants is their ability to understand academic scientific literature, which often includes complex figures such as data visualization plots, charts, schematic diagrams, macroscopic and microscopic photograph, and other specialized content from a variety of scientific fields. However, there is currently a lack of comprehensive evaluation for models' understanding of advanced graduate-level multimodal scientific knowledge, especially in the context of complex figures across diverse scientific disciplines. Existing evaluations tend to focus on simpler charts and plots (Chen et al., 2020; Kahou et al., 2017; Siegel et al., 2016) and suffer from narrow scopes and lower quality (Li et al., 2024).

Our dataset, `MMSci`, is designed to address this gap. `MMSci` is a multimodal, multi-discipline dataset comprising high-quality, peer-reviewed articles and figures from 72 scientific disciplines, predominantly within the natural sciences. We created a benchmark to evaluate models' understanding of graduate-level multimodal scientific knowledge across these disciplines. Additionally, this dataset can serve as a training resource to enhance models' comprehension of multimodal scientific knowledge.

A.2.2 INTENDED USE

This dataset is used to evaluate and enhance the large multimodal models (LMMs)' understanding of advanced multimodal scientific knowledge.

A.2.3 DATA COLLECTION

Data Source The dataset comprises open-access articles published in Nature Communications[6]. These articles are freely and permanently accessible upon publication under the Creative Commons Attribution 4.0 International (CC BY) License. Detailed information on the open-access policy of Nature Communications is available at `https://www.nature.com/ncomms/open-access`.

Data Collection Process We collected various types of information for each article from the Nature Communications website. The articles' information includes titles, abstracts, main body content, references, and PDF versions of the articles, all directly accessible from their respective sections on the article's webpage (e.g., `https://www.nature.com/articles/xxx`, where "xxx" is the article's unique ID). Additionally, figures and their captions were sourced from a dedicated figures section linked from each article's main page (e.g., `https://www.nature.com/articles/xxx/figures`). This user-friendly platform facilitates easy acquisition of all necessary data, eliminating the needs for quality control and data filtering.

Annotations The dataset does not include explicit annotations. Instead, the authors themselves carried out a small-scale manual review and classification of the image types specifically for analysis. No external annotators or crowdworkers were involved in this process.

Personal and Sensitive Information The dataset does not include any personal or sensitive information. All article content is publicly accessible. All author information are also publicly available, and no personal information was explicitly extracted, stored, or used from the authors.

A.2.4 SOCIAL IMPACT AND ETHICAL CONSIDERATIONS

Benefits The benefits of our dataset are two-fold: (1) **Evaluation Benchmark**: This dataset serves as a valuable evaluation benchmark for assessing the understanding of large multimodal models (LMMs) regarding scientific articles and figures. (2) **Training Resources**: It can be used as a training resource to enhance LMMs' comprehension of scientific articles and figures, improving their performance in various scientific and research-related tasks.

[6]`https://www.nature.com/ncomms/`

Model	Model versioning/path
GPT-4V	gpt-4-turbo-2024-04-09
GPT-4o	gpt-4o-2024-05-13
Kosmos2	https://huggingface.co/microsoft/kosmos-2-patch14-224
BLIP2	https://huggingface.co/Salesforce/blip2-opt-2.7b
LLaVA1.5-7B	https://huggingface.co/llava-hf/llava-1.5-7b-hf
LLaVA-Next	https://huggingface.co/liuhaotian/llava-v1.6-vicuna-7b
LLaVA-Next-Mistral	https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf
Qwen-VL-Chat	https://huggingface.co/Qwen/Qwen-VL-Chat

Table 8: Evaluated LMMs in our experiments with their versions or Huggingface model paths.

Risks and Ethical Considerations However, there are potential risks and ethical considerations to address: (1) **Misuse in Academic Integrity**: The advancement of AI research assistants facilitated by this dataset could potentially lead to misuse, such as academic fraud, fabrication, or improper assistance in academic work. We strongly encourage users to exercise caution and responsibility when using AI assistants, ensuring they are employed ethically and correctly. (2) **Data Misinterpretation and Hallucination**: There is a risk of misinterpreting the dataset's content, leading to inaccurate conclusions or misuse of scientific information. Users should critically assess and validate the AI-generated outputs against established scientific knowledge and principles.

A.2.5 LIMITATIONS

Currently, our evaluation benchmark primarily focuses on understanding figures in scientific articles based on the article content or not. We encourage further efforts to expand these evaluations to include a broader range of scientific knowledge using our dataset.

A.2.6 AUTHOR STATEMENT

The authors declare full responsibility for any rights violations, including but not limited to intellectual property rights and privacy rights, that may arise from the publication and use of this dataset. We confirm that all data provided is licensed under appropriate licenses, ensuring legal compliance and transparency.

A.2.7 HOSTING, LICENSING, AND MAINTENANCE PLAN

The dataset will be hosted on GitHub, offering reliable and secure access. We commit to maintaining the repository with regular updates, security patches, and user support to ensure the data's integrity and usability over time. Licensing terms will be clearly communicated to users, adhering to the appropriate data licenses to promote proper usage and distribution. The data is licensed under the CC BY 4.0 License, which permits sharing and adaptation with proper attribution. The primary codebase for our project is licensed under the Apache 2.0 License.

A.3 EXPERIMENTAL SETUP

A.3.1 EVALUATED MODEL

We evaluated two proprietary models GPT-4V and GPT-4o and six open-source LMMs. Additionally, we tested our fine-tuned model, which is based on LLaVA-Next (LLaVA1.6-Vicuna-7B). For evaluations of open-source models, we utilized checkpoints available on Hugging Face[7]. The specific versions of proprietary models and paths for open-source models are detailed in Table 8. All inferences for the open-source models were executed on a computing cluster equipped with eight NVIDIA A100 GPUs, each with 40GB of memory.

A.3.2 EVALUATION SETUP AND RESULTS

As described in the main paper, we set the temperature to 0.7 for inferences on both the scientific figure captioning and multiple-choice Visual Question Answering (VQA) tasks. For the figure captioning task, we conducted the inference three times, and the averaged results along with their

[7]https://huggingface.co/models

Table 9: Performance on scientific figure captioning with standard deviation. B@k represents BLEU@k (k=1,2,3,4), R stands for ROUGE-L, M stands for METEOR, BS indicates BERTScore, and CLIP and RCLIP represent CLIPScore and RefCLIPScore, respectively. Best results are bolded and second best are underlined.

Grounded	Model	B@1	B@2	B@3	B@4	M	R	BS	CLIP	RCLIP
N/A	Kosmos2	23.05 ± 0.01	2.59 ± 0.02	0.39 ± 0.02	0.09 ± 0.01	14.53 ± 0.14	11.69 ± 0.00	77.51 ± 0.01	41.44 ± 0.00	46.01 ± 0.11
	BLIP2	37.73 ± 0.30	4.91 ± 0.03	0.25 ± 0.05	0.04 ± 0.02	3.18 ± 0.13	6.56 ± 0.17	79.28 ± 0.09	55.93 ± 0.18	56.90 ± 0.15
	LLaVA1.5-7B	29.34 ± 0.06	3.16 ± 0.03	0.16 ± 0.02	0.03 ± 0.01	11.80 ± 0.06	12.55 ± 0.00	79.93 ± 0.00	64.79 ± 0.05	64.22 ± 0.02
	LLaVA-Next	15.96 ± 0.12	2.44 ± 0.02	0.26 ± 0.00	0.04 ± 0.00	18.89 ± 0.08	10.87 ± 0.05	79.27 ± 0.03	68.08 ± 0.15	66.72 ± 0.15
	LLaVA-Next-Mistral	15.91 ± 0.04	2.81 ± 0.01	0.38 ± 0.01	0.08 ± 0.00	20.45 ± 0.11	10.96 ± 0.01	79.53 ± 0.00	68.54 ± 0.13	67.04 ± 0.11
	Qwen-VL-Chat	**43.54** ± 0.46	12.78 ± 0.24	4.87 ± 0.13	1.66 ± 0.05	15.34 ± 0.12	14.84 ± 0.14	81.95 ± 0.06	63.24 ± 0.21	64.30 ± 0.12
	GPT-4V	21.94 ± 0.02	4.95 ± 0.03	1.31 ± 0.02	0.41 ± 0.00	26.62 ± 0.01	14.87 ± 0.01	81.76 ± 0.00	**71.81** ± 0.06	71.27 ± 0.07
	GPT-4o	19.73 ± 0.04	4.90 ± 0.03	1.49 ± 0.02	0.47 ± 0.02	**27.06** ± 0.04	15.59 ± 0.01	81.13 ± 0.01	71.43 ± 0.07	71.39 ± 0.02
	LLaVA-Next-MMSci	42.67 ± 0.23	**14.51** ± 0.14	**6.60** ± 0.12	**3.10** ± 0.08	21.79 ± 0.08	**18.01** ± 0.07	**83.39** ± 0.04	71.19 ± 0.05	**72.21** ± 0.08
Abstract	Kosmos2	22.28 ± 0.04	2.91 ± 0.01	0.61 ± 0.01	0.20 ± 0.01	19.50 ± 0.06	11.81 ± 0.02	79.09 ± 0.01	41.44 ± 0.00	46.01 ± 0.00
	BLIP2	32.88 ± 0.76	4.18 ± 0.41	0.45 ± 0.10	0.09 ± 0.05	7.32 ± 0.37	9.14 ± 0.48	79.72 ± 0.10	48.34 ± 0.21	51.12 ± 0.16
	LLaVA1.5-7B	30.78 ± 0.03	4.50 ± 0.02	0.66 ± 0.01	0.18 ± 0.01	14.54 ± 0.02	14.00 ± 0.04	81.20 ± 0.00	68.49 ± 0.07	69.72 ± 0.03
	LLaVA-Next	19.79 ± 0.03	3.70 ± 0.02	0.68 ± 0.01	0.18 ± 0.00	20.86 ± 0.04	12.88 ± 0.03	80.86 ± 0.01	69.63 ± 0.05	70.06 ± 0.05
	LLaVA-Next-Mistral	19.50 ± 0.06	3.95 ± 0.04	0.76 ± 0.02	0.20 ± 0.01	21.49 ± 0.04	12.75 ± 0.03	80.84 ± 0.01	69.80 ± 0.05	69.93 ± 0.06
	Qwen-VL-Chat	38.27 ± 0.16	8.75 ± 0.10	2.22 ± 0.09	0.70 ± 0.03	16.02 ± 0.11	15.38 ± 0.12	81.87 ± 0.06	69.16 ± 0.19	70.12 ± 0.11
	GPT-4V	22.95 ± 0.04	5.63 ± 0.03	1.56 ± 0.03	0.50 ± 0.02	27.59 ± 0.03	15.66 ± 0.01	82.37 ± 0.00	**72.22** ± 0.05	72.76 ± 0.03
	GPT-4o	21.06 ± 0.05	5.58 ± 0.01	1.76 ± 0.01	0.58 ± 0.00	**28.41** ± 0.03	16.32 ± 0.02	81.82 ± 0.02	72.15 ± 0.05	72.92 ± 0.08
	LLaVA-Next-MMSci	**45.89** ± 0.30	**16.96** ± 0.09	**8.12** ± 0.08	**4.08** ± 0.10	24.77 ± 0.10	**20.69** ± 0.03	**84.46** ± 0.04	71.33 ± 0.05	**74.22** ± 0.06
Full Content	GPT-4V	25.93 ± 0.03	8.03 ± 0.00	3.03 ± 0.02	1.32 ± 0.02	31.41 ± 0.04	19.24 ± 0.04	83.47 ± 0.02	**72.44** ± 0.09	74.04 ± 0.04
	GPT-4o	25.11 ± 0.10	11.11 ± 0.05	5.99 ± 0.04	3.51 ± 0.04	**37.55** ± 0.18	**24.94** ± 0.14	83.65 ± 0.00	71.94 ± 0.07	74.08 ± 0.02

standard deviations are reported in Table 9. For the multiple-choice VQA task, we performed up to five inference runs and reported the accuracy based on majority voting in the main paper (Table 4).

Table 10: Hyperparameters for visual instruction tuning.

Hyperparameter	Values
base model	`https://huggingface.co/liuhaotian/llava-v1.6-vicuna-7b`
vision encoder	`https://huggingface.co/openai/clip-vit-large-patch14-336`
projector	2-layer MLP
epochs	1
global batch size	128
learning rate	0.00002
learning rate scheduler	cosine
weight decay	0.0
warmup ratio	0.03
max length	2048

A.3.3 VISUAL INSTRUCTION TUNING

Following the visual instruction tuning approach described in (Liu et al., 2024), we continuously fine-tuned the LLaVA-Next model (LLaVA1.6-Vicuna-7B). The original vision encoder, `openai/clip-vit-large-patch14-336`, was kept unchanged, while the projector and language model components were updated. The hyperparameters used in this process are detailed in Table 10. The fine-tuning was performed on a computing cluster equipped with eight NVIDIA A100 GPUs, each with 40GB of memory. This training process took approximately 24 hours to complete.

A.3.4 VISUAL LANGUAGE PRE-TRAINING

In our case study experiments on the material generation task, we continuously pre-train a LLaMA2-7B model using our interleaved article and figure data to infuse more material science-relevant knowledge. Specifically, for pre-training on the interleaved text and image data, we follow the methodology outlined in (Lin et al., 2023).

Model Architecture Following the approach outlined in (Liu et al., 2024; Lin et al., 2023), we extend the LLaMA2-7B model from a text-only model to a multimodal model by augmenting the LLM with a visual encoder to learn visual embeddings and a projector to bridge the embeddings between the text and visual modalities. Specifically, the visual encoder processes the image and outputs visual features. These features are then mapped into the word embedding space by the projector, creating visual tokens. These visual tokens are concatenated with the word tokens and fed into the LLM, allowing the model to integrate both text and visual information for generation. The specific LLM, visual encoder, and projectors used in our experiments are presented in Table 11.

Table 11: Hyperparameters for visual language pre-training on interleaved text and image data.

Hyperparameter	Values
base model	`https://huggingface.co/meta-llama/Llama-2-7b-hfb`
vision encoder	`https://huggingface.co/openai/clip-vit-large-patch14-336`
projector	2-layer MLP
Stage 1: Projector Initialization	
epochs	1
global batch size	256
learning rate	0.001
learning rate scheduler	cosine
weight decay	0.0
warmup ratio	0.03
max length	4096
tune LLM	✗
tune vision encoder	✗
tune projector	✓
Stage 2: Visual Language Pre-training	
epochs	1
global batch size	128
learning rate	0.00005
learning rate scheduler	cosine
weight decay	0.0
warmup ratio	0.03
max length	4096
tune LLM	✓
tune vision encoder	✗
tune projector	✓

Training Stages The visual pre-training process (Lin et al., 2023) involves two stages:

1. **Projection initialization**: In this stage, the LLM and the visual encoder are both pre-trained and remain fixed. The projector, however, is randomly initialized. Only the projector is fine-tuned during this stage, using image-caption pairs from (Liu et al., 2024).

2. **Visual language pre-training**: During this stage, both the LLM and the projector are fine-tuned on the interleaved image and text data. This includes data from general domains provided by MMC4 (Zhu et al., 2024), as well as scientific articles and figures from our dataset `MMSci`. Previous research (Lin et al., 2023) has shown that tuning both the LLM and the projector yields better results than tuning only one of them. Throughout this stage, the visual encoder remains fixed.

We did not conduct the further visual instruction-tuning for this model, as our primary objective was to infuse scientific knowledge into the LLM for the consecutive text-only material generation task. The two stages were conducted on a computing cluster equipped with eight NVIDIA A100 GPUs, each with 40GB of memory. The first stage took approximately 4 hours, and the second stage took around 36 hours.

A.3.5 Materials Generation

As a case study to investigate whether scientific knowledge has been effectively infused into the LLM (LLaMA2-7B in our experiments) and whether it can enhance performance on material science-related tasks, we follow the methodology from Gruver et al. (2024) to explore the material generation task. The primary objective is to format material crystal structures into text strings and fine-tuning the LLM to generate stable materials.

Prompt design We adhere to the prompt design described in (Gruver et al., 2024). There are two types of prompts in the training data: the generation prompt with one or multiple conditions and infilling prompts, where partial crystal structure strings are masked and the model generates the masked parts. The specific prompt templates are shown below, adapted from (Gruver et al., 2024).

Generation Prompt	Infilling Prompt
<s>Below is a description of a bulk material. [The chemical formula is Pm2ZnRh]. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice: [Crystal string]</s>	<s>Below is a partial description of a bulk material where one element has been replaced with the string "[MASK]": [Crystal string with [MASK]s] Generate an element that could replace [MASK] in the bulk material: [Masked element]</s>

Blue text is the condition for generation. Purple text stands in for string encodings of atoms.

The formula condition as shown above is always included, while other conditions are sampled from the following: formation energy per atom, band gap, energy above hull, and space group number.

Evaluation Our evaluations follows (Xie et al., 2021; Gruver et al., 2024), including four key aspects. We reiterate some details here. Structural validity is assessed by ensuring that the shortest distance between any pair of atoms exceeds 0.5 Å. Compositional validity is evaluated by verifying that the overall charge is neutral, as calculated using `SMACT` (Davies et al., 2019). Coverage metrics, COV-R (Recall) and COV-P (Precision), measure the similarity between ensembles of generated materials and ground truth materials in the test set. The property distribution metrics quantify the earth mover's distance (EMD) between the property distributions of generated materials and those in the test set, specifically for density (ρ, in $\mathrm{g/cm^3}$) and the number of unique elements (N_{el}).

Metastability and stability are assessed based on the energy above the convex hull, denoted as $\hat{E}_{\text{hull}}$. Two approaches are employed to estimate $\hat{E}_{\text{hull}}$: M3GNet (Chen & Ong, 2022) and Density Functional Theory (DFT) using the VASP code (Hafner, 2008). For M3GNet, each sample undergoes relaxation using force and stress calculations before evaluating the energy of the final structure. For DFT, relaxation is performed using the VASP code, which provides more accurate results but requires significantly more computational resources. A material is considered metastable by M3GNet if the predicted energy above the hull, $E_{\text{hull}}^{\text{M3GNet}}$, is less than 0.1 eV/atom. Furthermore, if validated by DFT, the material must have $E_{\text{hull}}^{\text{DFT}} < 0.0$ eV/atom to be considered stable. The percentages of such materials are reported over the total 10,000 inferences. We use the Materials Project (Jain et al., 2013) dated 2023-02-07.

Training Details Following the approach in (Gruver et al., 2024), we utilize 4-bit quantization (Dettmers et al., 2021) and Low-Rank Adapters (LoRA) (Hu et al., 2021) for efficient fine-tuning. The model is trained with a batch size of 1 for 1 epoch. We set the LoRA rank to 8 and the LoRA alpha to 32. The learning rate is 0.0001, annealed by a cosine scheduler. The training was conducted on a single NVIDIA A100 GPU, took approximately 4 hours to complete.

Conditional Generation and Infilling Results Due to space constraints, we did not include the results for the conditional materials generation and infilling tasks in the main paper. Here, we present these additional findings. The performance metrics reported are based on the same model used in the main paper. Our training data included two types of prompts: conditional generation prompts and infilling prompts. We compare our model LLaMA2-7B-MMSci, which has undergone continuous pre-training, with the original LLaMA2-7B that was trained without additional pre-training data. Both models were trained on datasets that included prompts for both conditional generation and infilling tasks under the same setup.

Following (Gruver et al., 2024), we performed 1,000 inferences for each condition in the conditional generation evaluation and 1,000 inferences for the infilling evaluation. For conditional generation evaluation, we assessed the percentage of generated materials that adhered to specified conditions, including formula, space group, and energy above the hull (E_{hull}). In the infilling evaluation, we measured diversity by computing the pairwise distance between generated samples and those from Matminer (Ward et al., 2018; Xie et al., 2021), focusing on composition and structure. Additionally, we evaluated metastability estimated by M3GNet. As seen in Table 12, LLaMA2-7B-MMSci, after

Table 12: Evaluation of conditional materials generation and infilling tasks. Comp. Div. and Struct. Div. represent the composition and structure diversity, respectively. The two models are fine-tuned with the same training data and setup in our implementation.

Method	**Conditional Generation**			**Infilling**		
	Formula↑	Space Group↑	E_{hull} ↑	Comp. Div.↑	Struct. Div. ↑	Metastability ↑
LLaMA2-7B	0.85	0.14	0.58	10.60	0.16	64.20%
LLaMA2-7B-MMSci	0.87	0.22	0.59	8.31	0.52	77.74%

continuous pre-training on our dataset `MMSci`, outperforms the original LLaMA2-7B across most metrics. This demonstrates its enhanced effectiveness in handling materials generation tasks.