A Appendix

Contents

A Appendix 1
A.1 Dataset Description 1
A.1.1 Dataset Summary 1
A.1.2 Data and Code Access 1
A.1.3 Subjects 2
A.1.4 Image Types 2
A.1.5 Case Study 5
A.2 Datasheet 11
A.2.1 Motivation 11
A.2.2 Intended Use 11
A.2.3 Data Collection 11
A.2.4 Social Impact and Ethical Considerations 11
A.2.5 Limitations 12
A.2.6 Author Statement 12
A.2.7 Hosting, Licensing, and Maintenance Plan 12
A.3 Experimental Setup 12
A.3.1 Evaluated Model 12
A.3.2 Evaluation Setup and Results 13
A.3.3 Visual Instruction Tuning 13
A.3.4 Visual Language Pre-training 13
A.3.5 Materials Generation 14

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.[1] 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`.

---

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

- **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://anonymous.4open.science/r/MMSci-2321`

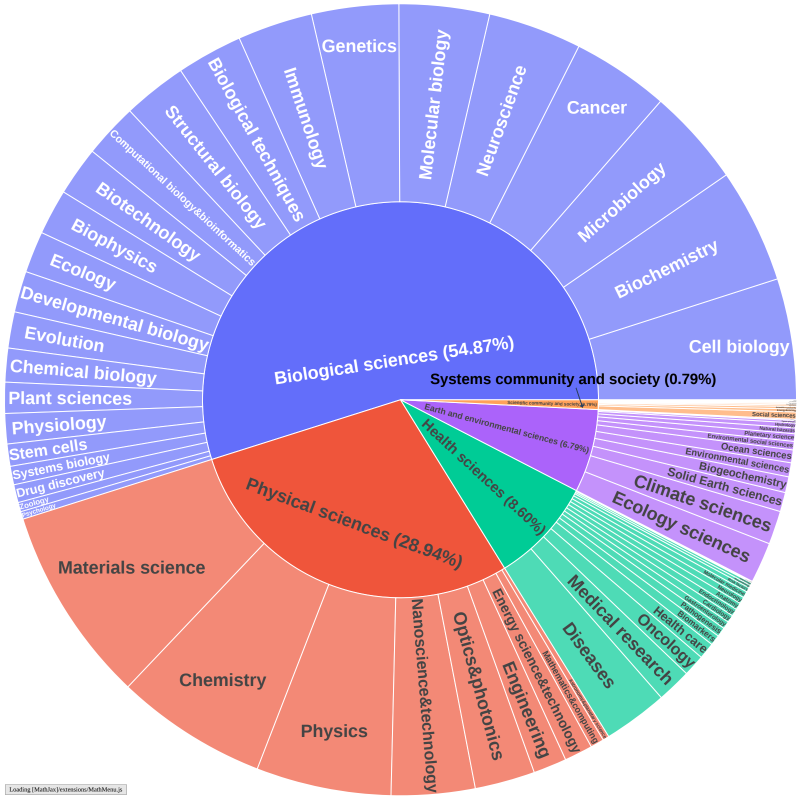


Figure 1: The five major categories and 72 subjects in our dataset.

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.[2]. The visualizations are shown in Figure 1, and detailed statistics of these subjects are provided in Table 1. 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

---

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

Table 1: 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

summarized in Table 2. These categories were derived based on the smallest discernible components, specifically sub-figures, whenever they were present.

Table 2: 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

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:

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.

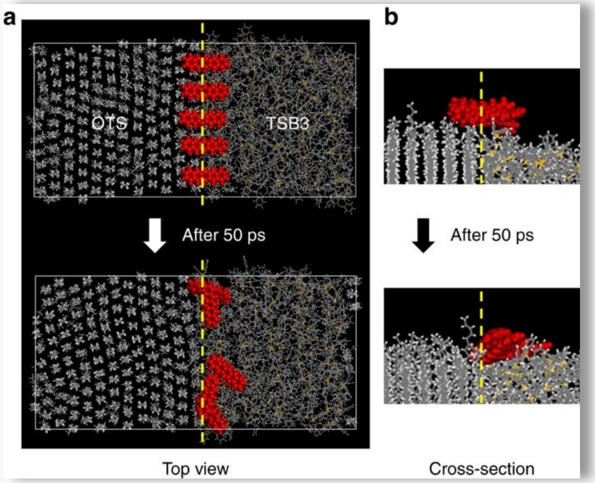
Manual Annotation for Unclassified Images Our authors performed manual annotations for 17 images in cases where GPT-4o could not classify images due to OpenAI's policy restrictions. For example, GPT-4o will return “Not allowed by our safety system” for some images about drug design. This ensured comprehensive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

Final Results The final classification results are presented in Table 2. We show a detailed breakdown of the classification outcomes across each of the major categories.

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 2, 3, 4, abstract-grounded figure captioning in Figure 5, and the multi-turn visual instruction following data in Figure 6.

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.

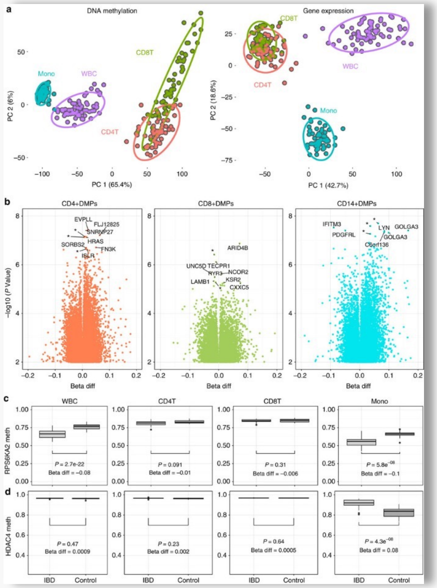
This choice is supported by the presence of a specific molecular structure labeled as TSB3 and the visual representation of molecular interactions or changes occurring around this label, which suggests a focus on the growth mechanism involving pentacene on TSB3.

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 2: 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.

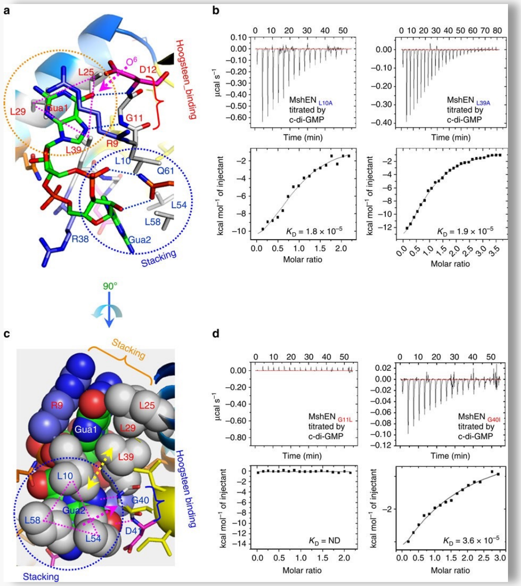
Here's the detailed breakdown:

- 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 3: 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.

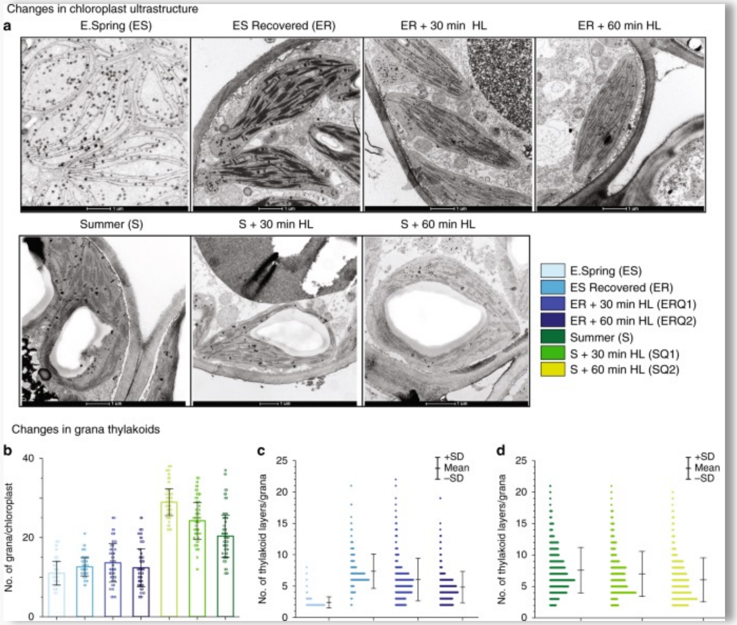
Therefore, Option D is the best description of the content in sub-figure (a).

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 4: 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



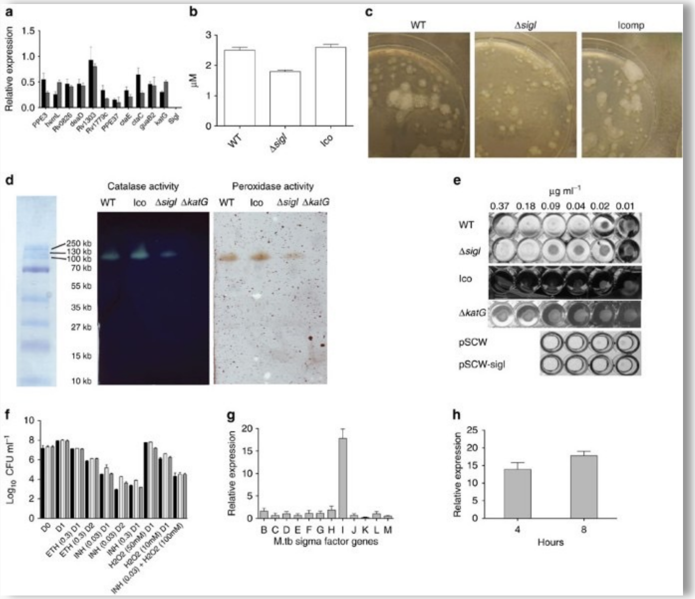
source: https://www.nature.com/articles/s41467-020-20137-9/figures/5

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−2s−1, ER samples treated with 800 µmol m−2s−1high light for 30 min (ERQ1), for 60 min (ERQ2). Summer (S), Summer samples treated with 1200 µmol m−2s−1high 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 5: 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 6: 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[3]. 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

---

[3] `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 3: Evaluated LMMs in our experiments with their versions or Huggingface model paths.

(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.

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[4]. The specific versions of proprietary models and paths for open-source models are detailed in Table 3. All inferences for the open-source models were executed on a computing cluster equipped with eight NVIDIA A100 GPUs, each with 40GB of memory.

---

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

Table 4: 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

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 standard deviations are reported in Table 4. 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 5: 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 5. 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

Table 6: 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	✓

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 6.

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.

Table 7: 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%

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 7, LLaMA2-7B-MMSci, after 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.

References

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

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`.

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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

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.

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.

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.

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.