MACBENCH: A multimodal chemistry and materials science benchmark

Nawaf Alampara[1] **Indrajeet Mandal**[2] **Pranav Khetarpal**[3] **Hargun Singh Grover**[4]

Mara Schilling-Wilhelmi[1] **N. M. Anoop Krishnan** [3,4*]

Kevin Maik Jablonka[1,5,6,†]
[1]FSU Jena[‡] [2] IIT Delhi [§]
[3] IIT Delhi [¶]
[4] IIT Delhi [‖]
[5]CEEC Jena[**] [6] HIPOLE Jena [††]
nawaf.alampara@uni-jena.de
krishnan@iitd.ac.in mail@kjablonka.com

Abstract

We present the multimodal Materials and Chemistry Benchmark (MACBENCH), a benchmark for evaluating multimodal capabilities of AI models in chemistry and materials science tasks. This benchmark addresses the lack of comprehensive, domain-specific evaluation tools for multimodal AI in scientific contexts. MACBENCH encompasses tasks across three key areas: fundamental scientific understanding, data extraction from visual information, and practical laboratory knowledge, totaling 628 questions. It includes diverse visual inputs such as laboratory images, band structures, crystal structures, and atomic force microscopy images paired with multiple-choice questions. We evaluate state-of-the-art multimodal AI models (GPT-4o, Claude-3.5-Sonnet, Gemini-1.5-Pro) on MACBENCH, revealing significant performance variations across tasks and skills. While models excel at basic pattern recognition and information retrieval, they struggle with complex reasoning and applying scientific principles to novel situations. Notably, we observe a disconnect between object recognition and contextual understanding in laboratory safety scenarios. MACBENCH provides crucial insights into the capabilities and limitations of multimodal AI in chemistry and materials science, serving as a valuable tool for guiding the development of more capable AI systems for scientific research.

[*]corresponding author
[†]corresponding author
[‡]Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldstr. 10, 07743 Jena, Germany
[§]School of Interdisciplinary Research, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[¶]Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[‖]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
[**]Center for Energy and Environmental Chemistry Jena, Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany
[††]Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstraße 12-14, 07743, Jena, Germany

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

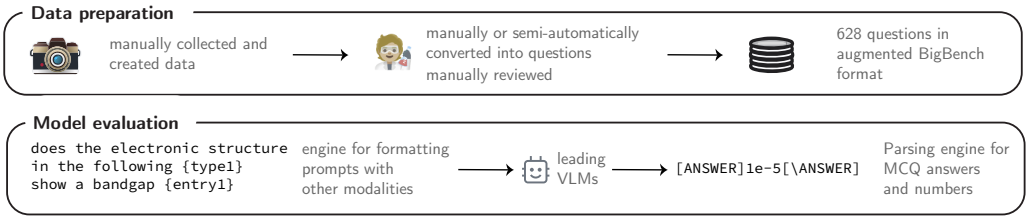

Figure 1: **Overview of MACBENCH.** We introduce MACBENCH, a benchmark for materials science and chemistry. We manually and semi-automatically gathered data and curated questions in different categories, using a format similar to the BigBench library. With this corpus, we developed an engine to automatically evaluate models, and used it to assess frontier VLMs.

1 Introduction

The ability of large language models (LLMs) to assimilate large-scale information, ground them into a given context, and apply numerical computations to make real-time decisions has made them an essential tool in accelerating science and engineering. In materials science and chemistry, these models have shown promise in various applications, including property prediction [1–3], inverse design of materials [4], automating experiments, and data extraction from scientific literature [5–8]. However, much information in scientific literature is spread across multiple modalities, including text, tables, figures, and even videos, that need to be processed together in the context to make a meaningful inference. Thus, the inherently multimodal nature of scientific research, particularly in chemistry and materials science, necessitates developing and evaluating models that seamlessly integrate textual and visual information.

While LLMs have shown promising performance for materials science and chemistry [9, 10], they are insufficient for many real-world applications that require the incorporation of visual information such as plots, figures, and tables [11, 12]. Moreover, the visual modality allows probing for tacit knowledge [13] of laboratory environments, which is crucial for automating experiments and developing AI-powered laboratory assistants [14–17]. Importantly, one of the main ways that frontier models are thought to pose chemical or biosafety risks is by bridging tacit knowledge gaps. This involves filling the gap left by the scarcity of experts with hands-on laboratory experience [18].

Addressing this key challenge has been the focus of recent research on multimodal LMs (MLMs)[19, 20]. However, standardized benchmarks for chemistry and materials science for these MLMs are lacking. This gap makes it challenging to assess and develop multimodal models in scientific contexts and to understand their visual capabilities in these fields.

To address this, we present MACBENCH, a multimodal benchmark for evaluating vision-language models in chemistry and materials science (Figure 1). It includes various visual inputs: laboratory images, band structure diagrams, crystal structures, tables, atomic force microscopy images, and hand-written molecular structures.

Overall, our main contributions are as follows:

1. *Multimodal dataset:* We introduce MACBENCH, a manually-curated multimodal benchmark dataset designed explicitly for evaluating vision-language models in chemistry and materials science, covering three key aspects of scientific research: (i) Fundamental materials science and chemistry understanding, (ii) Data extraction capabilities from visual scientific information, and (iii) Practical laboratory and experimentation knowledge.

2. *Benchmarking:* We evaluate state-of-the-art multimodal AI models on MACBENCH, providing insights into their strengths and limitations in scientific contexts.

The rest of this paper is organized as follows. Section 2 discusses related work in multimodal AI and scientific benchmarks. Section 3 and Section 4 describes the evaluation methodology and the MACBENCH dataset in detail, respectively. Section 5 presents our experimental results and analysis. Finally, Section 7 concludes the paper and discusses future directions.

2 Related Work

Significant progress has been made in developing benchmarks for multimodal machine learning and scientific reasoning in recent years. However, existing work falls short of addressing the specific needs of materials science and chemistry, particularly in capturing tacit knowledge and laboratory skills.

2.1 Multimodal Benchmarks in Science

Several benchmarks have been developed to evaluate multimodal models in scientific contexts. LabBench [21] includes FigQA and TableQA tasks, focusing on interpreting scientific figures and tables without additional context. ScienceQA [22] covers elementary to high school topics across 20 subjects in natural, social, and language sciences, while SciBench [23] focuses on college-level physics, chemistry, and math, including a subset of multimodal problems.

More comprehensive benchmarks have also emerged. MMMU [24] covers college-level exams and textbooks across 25 subjects in various scientific disciplines, and MMSci [25] uses PhD-level content from Nature Communications, covering 72 subjects primarily in natural sciences, health, and social sciences. OlympiadBench [26] provides challenging, diverse multimodal scientific problems at an Olympiad level, while SciFIBench [27] focuses on scientific figure interpretation in computer science.

Specialized benchmarks like MathVista [28] integrate multiple datasets for mathematical reasoning in visual contexts, and DesignQA [29] evaluates engineering documentation understanding. While these benchmarks cover various aspects of scientific reasoning and multimodal understanding, they do not specifically address the unique challenges in materials science and chemistry, particularly in terms of tacit knowledge and laboratory skills.

2.2 Materials Science and Chemistry Benchmarks

Several benchmarks have been explicitly developed for materials science and chemistry but primarily focus on text-based tasks or property prediction. ChemBench [30] provides thousands of text-based question-answer pairs covering various chemistry topics, while MaScQA [31] tests materials science knowledge through challenging questions. ChemistryQA [32] focuses on real-world chemical calculation questions, and MatSci-NLP [33] encompasses seven different NLP tasks specific to materials science.

In the realm of property prediction, benchmarks such as MoleculeNet [34], Therapeutics Data Commons (TDC) [35], Matbench [36], and MatText [3] focus on molecular and materials property prediction tasks. While these benchmarks are valuable for evaluating certain aspects of machine learning models in materials science and chemistry, they do not address the multimodal nature of scientific research in these fields nor capture the tacit knowledge and laboratory skills crucial for practical applications.

2.3 Limitations of Existing Benchmarks

While valuable, existing benchmarks have several limitations when evaluating machine learning models for materials science and chemistry. They lack a focused approach to materials science reasoning in multimodal contexts and provide insufficient coverage of tacit knowledge and laboratory skills. Moreover, there is limited integration of diverse visual inputs specific to materials science and chemistry, such as band structures, AFM images, and hand-written molecular structures. Notably absent are tasks that simulate real-world laboratory scenarios and safety considerations, which are crucial for practical applications in these fields.

MACBENCH addresses these limitations by providing a comprehensive, multimodal benchmark specifically designed for materials science and chemistry. It incorporates diverse visual inputs and assesses fundamental understanding, data extraction capabilities, and practical laboratory knowledge. By leveraging the ChemBench engine for running the benchmarks, MACBENCH ensures compatibility with existing evaluation frameworks while extending their capabilities to multimodal scientific evaluations.

3 Methods

3.1 Dataset

To create the MACBENCH corpus, we used the BigBench-inspired format from the ChemBench framework [37]. Using a templating syntax, we also expanded the framework to easily include other modalities, like images. To prevent leakage into foundation model training corpora, we ask for examples from this dataset not to be shared in plain text or images online. We also include the canary string from Srivastava et al. [37] in our corpus to help training corpora filter out our benchmark.

3.2 Model Evaluation

We used the ChemBench framework to evaluate models and serialize questions. For all models, we applied default settings and set the inference temperature to zero. To extract answers, we prompted the model to return them in `[ANSWER]` `[\ANSWER]` tags. We then used regular expressions for extraction, with an option for LLM-based parsing not utilized in our examples.

For questions requiring floating-point answers, we employed a regular-expression based parser that accommodates scientific notation. We applied a tolerance of $1\,\%$ to determine if a numeric answer is correct. All the questions were evaluated using three models, namely, GPT-4o, Claude-3.5-Sonnet, Gemini-1.5-Pro.

4 Benchmark Corpus

Figure 2 displays our benchmarking corpus. It outlines the topics and the difficulty levels of the questions, which were assigned through manual labeling. We show examples of questions in different categories in Figure 3.

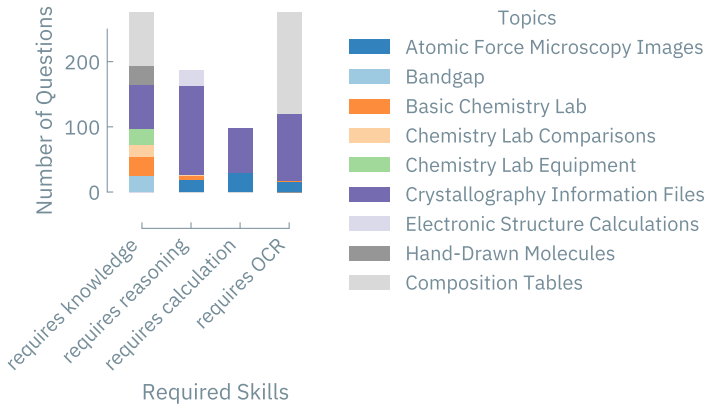

Figure 2: **Distribution of chemistry-related topics across different required skills.** The x-axis shows four categories of required skills: knowledge, reasoning, calculation, and optical character recognition (OCR). The y-axis represents the number of questions or tasks. Each colored bar corresponds to a specific chemistry topic, with Crystallography Information Files and Composition Tables being the most prevalent across skill categories.

4.1 Fundamental Materials Science Understanding

Electronic Structure We selected 15 materials each from the Materials Project database [38] representing metallic, indirect, and direct bandgap systems. Using the Materials Project IDs for these materials, we curated images of their band structures from the Materials Project AWS OpenData. We then formulated two types of questions: one to determine the presence of a bandgap and the other to classify the electronic structure as metallic, indirect, or direct bandgap semiconductors. We generated 25 and 24 questions for each type, ensuring a balanced sampling across all categories.

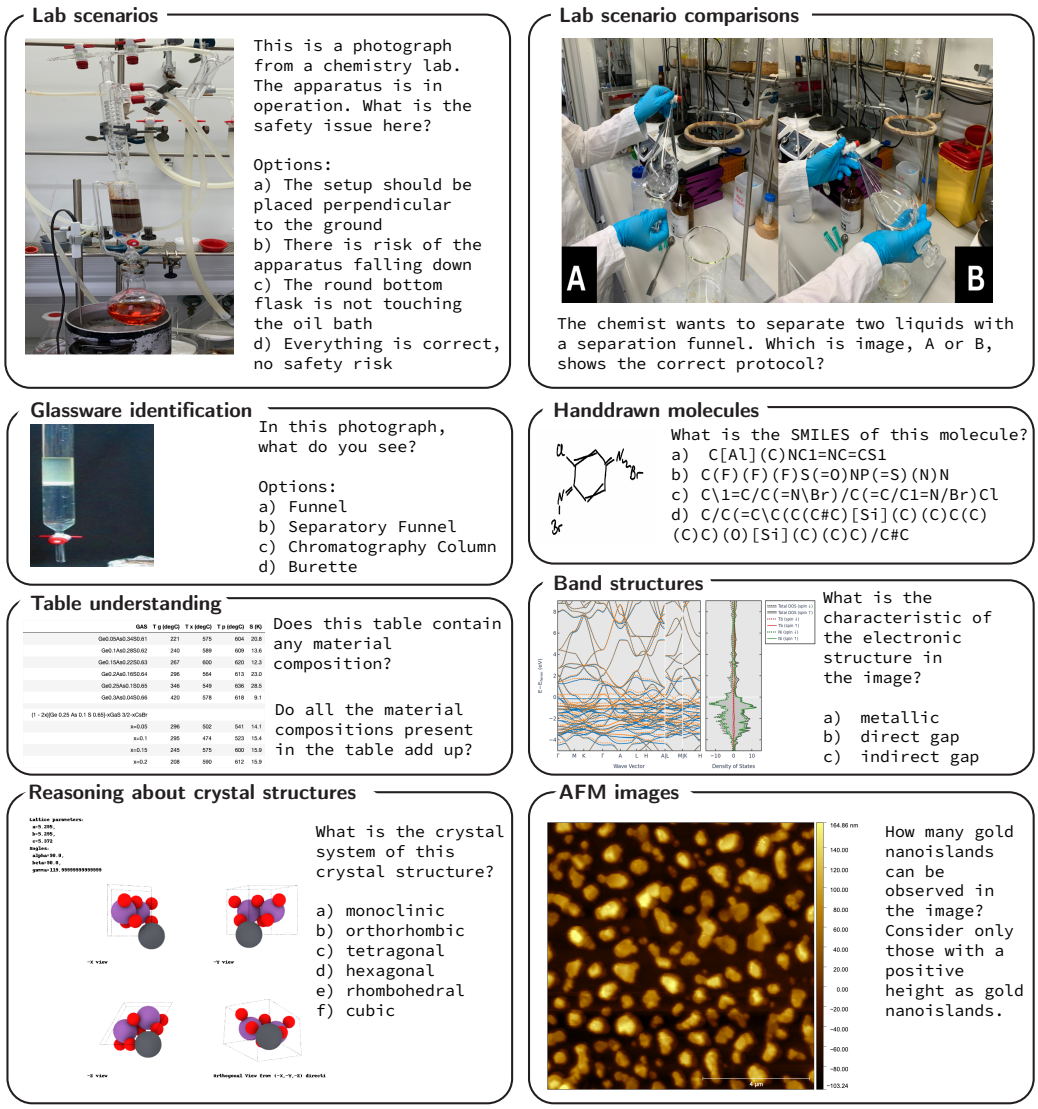

Figure 3: **Samples of questions in the different categories of MACBENCH.** Note that each category contains multiple question types but that we typically only show one sample in this figure.

Crystal Structures Crystalline materials are defined by their unit cells. Fundamental properties like density, elastic modulus, and thermal expansion depend on the crystal structure. To examine how well MLMs interpret crystal structures, we built questions around reasoning based on crystal structure renderings. We use crystallographic information files (CIF) from the American Mineralogist Crystal Structure Database (AMCSD) [39]. To ensure diversity, we select binary, ternary, quaternary, quinary, and decanary, each with diverse space groups.

We generated 170 question-answer pairs for 34 crystal structures. The questions evaluate how well MLMs understand crystal structures and their unit cells. We assess models on tasks of varying complexity, including OCR (lattice parameters are also shown along with the crystal structure), counting ability to identify the number of atoms in a crystal, knowledge of the crystal systems, and advanced reasoning.

4.2 Practical Laboratory and Experimentation

Novel Corpus of Chemistry Lab Images We staged various scenarios in a university chemistry lab to build our corpus about practical laboratory settings. In this initial version of MACBENCH, we focused on safety problems frequently observed in university lab courses and research labs.

We created two types of questions based on the images. The first set includes 38 multiple-choice questions about safety protocol violations, and the second set contains 17 questions. These questions present images and ask respondents to select the one that correctly shows a specific process.

Chemical Lab Equipment We used the LabPics image dataset by Eppel et al. [40] to create multiple-choice questions about the types of chemical glassware (Pipette, Test Tube, Beaker, Round Flask, Cylindrical Beaker, Separatory Funnel, Funnel, Burette, Chromatography Column, Condenser) shown in an image. The model was given 3 other randomly sampled glassware apart from the correct answer as options. 25 images were randomly selected for this task.

AFM Images MLMs might have a significant role in experimental planning and analysis. We utilize atomic force microscopy (AFM) images and manually developed questions based on them. These images are created in-house to prevent data leakage. Additionally, the AFM images are processed with a scale bar to provide information on length scale and image size. A set of 50 curated questions were used to assess the capability of MLMs to comprehend AFM images.

4.3 Data Extraction Capabilities

A subset of our questions focused on data extraction capabilities in settings relevant to materials science and chemistry.

Material Tables Tables are commonly used to present information about the compositions and properties of materials [41]. Understanding these tables can greatly benefit the use of MLMs in the materials domain. To test this understanding, we gathered 120 tables that included both composition and property information. These tables were formatted in various ways, such as compositions written in one cell, multiple cells, or as an equation with a variable like Si_xGe_{1-x}. We manually selected questions to assess the table understanding and materials knowledge of MLMs.

Molecule Recognition To excel in chemistry, one must be able to comprehend molecular drawings. Therefore, we designed 29 multiple-choice questions that require selecting the correct SMILES string for a hand-drawn 2D molecule. These drawings were taken from the DECIMER dataset [42], and the SMILES options were randomly selected from our subset of images.

5 Model Evaluation Results

Our evaluation of MLMs on the MACBENCH benchmark reveals striking performance variations across different categories (Figure 4). GPT-4o, Gemini-1.5-Pro, and Claude-3.5-Sonnet each excel in at least one area, but Claude-3.5-Sonnet leads in most categories.

Claude-3.5-Sonnet shines in questions requiring scientific understanding, particularly in the crystal structure, AFM image analysis, and electronic structure categories. Gemini-1.5-Pro lags in these areas, suggesting a gap in scientific reasoning.

The models perform best at simple tasks like identifying lab glassware, identifying handwritten molecules, or extracting basic data (Figure 5). Here, they approach perfection. However, they stumble on real-world lab scenarios, raising doubts about their readiness for use as assistants in real-world scenarios.

All models struggle with questions demanding calculation or scientific reasoning (Figure 8). They can recognize and recall but often fail to apply scientific principles or deduce logically.

It is interesting to observe the trade-off between precision and recall. Gemini-1.5-Pro, on average, adds 0.53 additional (incorrect) options when answering MCQ questions, compared to 0.37 for Claude-3.5-Sonnet (Table 1). Overall, Gemini-1.5-Pro and GPT-4o tend to have a recall slightly

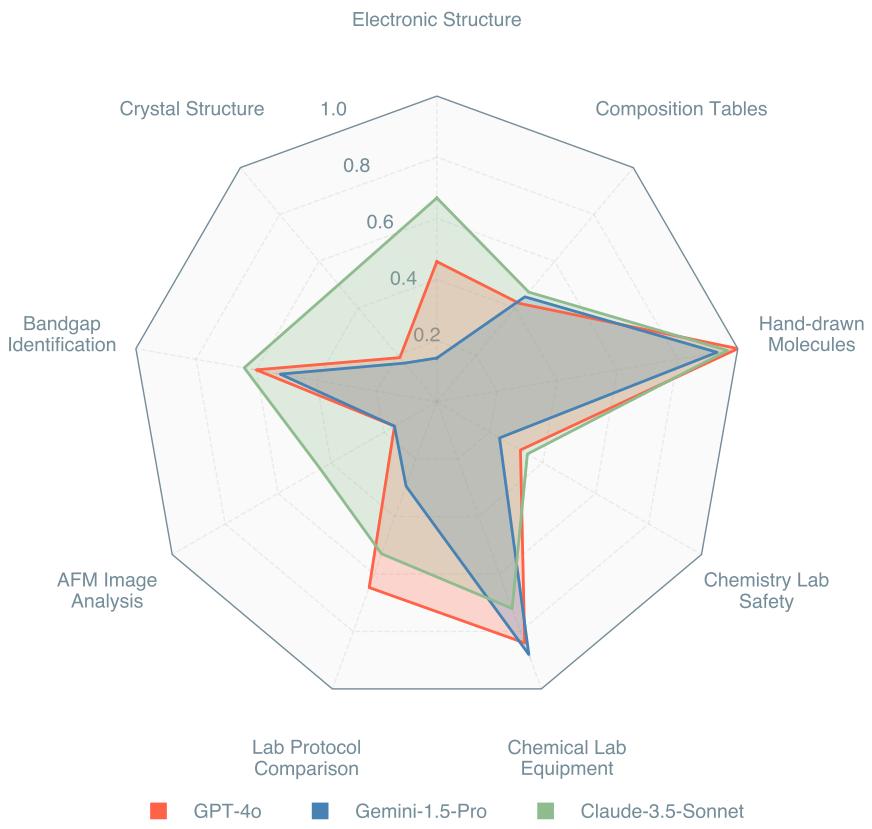

Figure 4: **Performance of MLMs on different categories of questions in MACBENCH.** The chart shows the performance of MLMs in different categories based on the fraction of questions they answered completely correctly. It is observable that the performance varies widely between categories, with Claude-3.5-Sonnet often being the top-performing model.

higher than precision, meaning they capture more relevant classes but at the cost of introducing more errors.

5.1 Common Error Modes

For 233 out of 628 questions, all three models failed to provide a correct answer. A significant portion of these occurred in questions related to reasoning about crystal structures (81) and extracting information from tables (96). However, many other questions requiring basic reasoning also seem to trip up the models. Anecdotally, on the two questions when no safety protocols were violated, all three models consistently failed to vote "everything is correct."

A striking disconnect emerges between object recognition and contextual understanding in laboratory scenarios. While models excel at identifying individual lab equipment, they struggle to grasp the broader implications of complex setups. This limitation is particularly evident in safety assessments, where models often miss subtle yet critical issues that human experts would readily identify. For instance, the models failed to recognize potential hazards in scenarios involving improperly positioned equipment, such as clamped setups not kept perpendicular or separation funnels placed at unusual heights. This suggests a gap between the models' ability to process visual information and their capacity to apply domain-specific knowledge in practical contexts.

Quantitative analysis and scale interpretation prove to be significant challenges across various tasks. In AFM image analysis, all models surprisingly frequently failed to accurately estimate widths from image legends. This difficulty extends to other quantitative tasks, such as counting specific features in images (e.g., islands or grid lines).

Figure 5: **Performance of models as a function of the required skill.** One question might be assigned to multiple required skills. Skills have been manually assigned. Models perform well based on questions requiring text recognition but struggle with questions involving calculations or reasoning.

Perhaps most surprisingly, the models exhibit unexpected difficulties in areas considered foundational in materials science and chemistry. A notable example is their frequent misclassification of electronic band structures, particularly their tendency to erroneously identify band gaps in metallic systems. This error is especially concerning given that distinguishing between metals, semiconductors, and insulators based on band structure is a basic skill for chemists and materials scientists. Such misconceptions suggest that despite their broad knowledge base, these AI models may lack a deep, principled understanding of core scientific concepts (when displayed in visual form that might not occur frequently in the training corpus).

The performance gaps are mainly seen in tasks that involve multi-step reasoning, combining domain knowledge with visual interpretation, and applying scientific principles to new situations. The models' performance decreases significantly when tasks require more complex cognitive processes beyond simple recognition or recall.

6 Discussion

Our evaluation of MLMs on MACBENCH reveals both promising capabilities and significant limitations in their application to chemistry- and materials-science-related tasks.

Claude-3.5-Sonnet's superior performance, especially in tasks requiring deeper scientific understanding, suggests that some models may be developing a more robust grasp of scientific concepts. However, even the best-performing model falls short in many areas, indicating that current MLMs are not yet ready for unsupervised use in real-world chemistry and materials science settings.

6.1 Limitations

While our benchmark provides valuable insights, it has several limitations:

- *Limited evaluations of models and agents:* We currently only evaluated a limited number of proprietary MLMs. We focused on those because they present the leading edge, but one might expect different performances of models fine-tuned for scientific figure understanding. Similarly, we would expect better performance if the models were given access to specialized tools.

- *Dealing with uncertainty:* In our current settings, models did not have the option to decline an answer. While we have done this to measure the performance across a wide variety with low variance, it is also relevant to understand how models would perform only on the subset of questions on which they are certain [21].
- *Limited number of question types:* At the moment, MACBENCH only supports multiple-choice questions and questions with numeric output.
- *Human baseline:* We did not include a human expert baseline for comparison, which could provide additional context for interpreting model performance.
- *Prompting:* Evaluations of language models are often affected by prompting. We do not claim that our prompting strategies are the best, but we offer our entire pipeline for reproducibility.

7 Conclusions

MACBENCH is a significant step towards evaluating the capabilities of multimodal language models in chemistry and materials science. Our benchmark reveals both promising advancements and critical limitations in current state-of-the-art models.

The results show that while MLMs excel in identifying lab equipment and extracting basic data, they struggle with tasks that require deeper scientific understanding, complex reasoning, and practical application of knowledge.

Our findings highlight the need for further development in AI systems to bridge the gap between broad knowledge acquisition and the nuanced, context-aware reasoning characteristic of human expertise in scientific domains.

Moving forward, it is crucial to address the limitations identified in current models, particularly in their ability to integrate visual information with domain-specific knowledge and perform quantitative analysis. MACBENCH sets a new standard for evaluating MLMs in chemistry and materials science, paving the way for more capable and reliable AI assistants in scientific research and education.

References

[1] Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **2024**, *6*, 161–169.

[2] Ramos, M. C.; Michtavy, S. S.; Porosoff, M. D.; White, A. D. Bayesian optimization of catalysts with in-context learning. *arXiv preprint arXiv:2304.05341* **2023**,

[3] Alampara, N.; Miret, S.; Jablonka, K. M. MatText: Do Language Models Need More than Text & Scale for Materials Modeling? *arXiv preprint arXiv:2406.17295* **2024**,

[4] Gruver, N.; Sriram, A.; Madotto, A.; Wilson, A. G.; Zitnick, C. L.; Ulissi, Z. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379* **2024**,

[5] Schilling-Wilhelmi, M.; Ríos-García, M.; Shabih, S.; Gil, M. V.; Miret, S.; Koch, C. T.; Márquez, J. A.; Jablonka, K. M. From Text to Insight: Large Language Models for Materials Science Data Extraction. *arXiv preprint arXiv:2407.16867* **2024**,

[6] Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications* **2024**, *15*, 1–13.

[7] Foppiano, L.; Lambard, G.; Amagasa, T.; Ishii, M. Mining experimental data from materials science literature with large language models. *arXiv preprint arXiv:2401.11052* **2024**,

[8] Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A. S.; Ceder, G.; Persson, K. A.; Jain, A. Structured information extraction from scientific text with large language models. *Nature Communications* **2024**, *15*, 1–12.

[9] Lei, G.; Docherty, R.; Cooper, S. J. Materials science in the era of large language models: a perspective. *Digital Discovery* **2024**,

[10] Jablonka, K. M.; others 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* **2023**, *2*, 1233–1250.

[11] Miret, S.; Krishnan, N. Are LLMs Ready for Real-World Materials Discovery? *arXiv preprint arXiv:2402.05200* **2024**,

[12] Hira, K.; Zaki, M.; Sheth, D.; Krishnan, N. A.; others Reconstructing the materials tetrahedron: challenges in materials information extraction. *Digital Discovery* **2024**, *3*, 1021–1037.

[13] Polanyi, M.; Sen, A. *The tacit dimension*; University of Chicago Press: Chicago; London, 2009.

[14] Ramos, M. C.; Collison, C. J.; White, A. D. A Review of Large Language Models and Autonomous Agents in Chemistry. *arXiv preprint arXiv:2407.01603* **2024**,

[15] Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **2024**, 1–11.

[16] Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578.

[17] Darvish, K.; Skreta, M.; Zhao, Y.; Yoshikawa, N.; Som, S.; Bogdanovic, M.; Cao, Y.; Hao, H.; Xu, H.; Aspuru-Guzik, A.; others ORGANA: A Robotic Assistant for Automated Chemistry Experimentation and Characterization. *arXiv preprint arXiv:2401.06949* **2024**,

[18] Barrett, A. M.; Jackson, K.; Murphy, E. R.; Madkour, N.; Newman, J. Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *arXiv preprint arXiv:2405.10986* **2024**,

[19] Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* **2023**,

[20] Zhang, D.; Yu, Y.; Li, C.; Dong, J.; Su, D.; Chu, C.; Yu, D. MM-LLMs: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* **2024**,

[21] Laurent, J. M.; Janizek, J. D.; Ruzo, M.; Hinks, M. M.; Hammerling, M. J.; Narayanan, S.; Ponnapati, M.; White, A. D.; Rodriques, S. G. LAB-Bench: Measuring Capabilities of Language Models for Biology Research. *arXiv preprint arXiv:2407.10362* **2024**,

[22] Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; Kalyan, A. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *arXiv preprint arXiv:2209.09513* **2022**,

[23] Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; Wang, W. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. *arXiv preprint arXiv:2307.10635* **2024**,

[24] Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; others MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502* **2024**,

[25] Li, Z.; Yang, X.; Choi, K.; Zhu, W.; Hsieh, R.; Kim, H.; Lim, J. H.; Ji, S.; Lee, B.; Yan, X.; others MMSci: A Multimodal Multi-Discipline Dataset for PhD-Level Scientific Comprehension. *arXiv preprint arXiv:2407.04903* **2024**,

[26] He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z. L.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; others OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008* **2024**,

[27] Roberts, J.; Han, K.; Houlsby, N.; Albanie, S. SciFIBench: Benchmarking Large Multimodal Models for Scientific Figure Interpretation. *arXiv preprint arXiv:2405.08807* **2024**,

[28] Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; Gao, J. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. *arXiv preprint arXiv:2310.02255* **2024**,

[29] Doris, A. C.; Grandi, D.; Tomich, R.; Alam, M. F.; Ataei, M.; Cheong, H.; Ahmed, F. DesignQA: A Multimodal Benchmark for Evaluating Large Language Models' Understanding of Engineering Documentation. *arXiv preprint arXiv:2404.07917* **2024**,

[30] Mirza, A.; Alampara, N.; Kunchapu, S.; Emoekabu, B.; Krishnan, A.; Wilhelmi, M.; Okereke, M.; Eberhardt, J.; Elahi, A. M.; Greiner, M.; others Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475* **2024**,

[31] Zaki, M.; Jayadeva; Mausam; Krishnan, N. M. A. MaScQA: investigating materials science knowledge of large language models. *Digital Discovery* **2024**, *3*, 313–327.

[32] Wei, Z.; Ji, W.; Geng, X.; Chen, Y.; Chen, B.; Qin, T.; Jiang, D. ChemistryQA: A Complex Question Answering Dataset from Chemistry. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. 2021.

[33] Song, Y.; Miret, S.; Liu, B. MatSci-NLP: Evaluating Scientific Language Models on Materials Science Language Tasks Using Text-to-Schema Modeling. *arXiv preprint arXiv:2305.08264* **2023**,

[34] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018**, *9*, 513–530.

[35] Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y. H.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. 2021.

[36] Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials* **2020**, *6*, 1–12.

[37] Srivastava, A.; others Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* **2023**,

[38] Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; others Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

[39] Downs, R. T.; Hall-Wallace, M. The American Mineralogist crystal structure database. *American Mineralogist* **2003**, *88*, 247–250.

[40] Eppel, S.; Xu, H.; Bismuth, M.; Aspuru-Guzik, A. Computer Vision for Recognition of Materials and Vessels in Chemistry Lab Settings and the Vector-LabPics Data Set. *ACS Central Science* **2020**, *6*, 1743–1752.

[41] Gupta, T.; Zaki, M.; Khatsuriya, D.; Hira, K.; Krishnan, N. A.; others DiSCoMaT: Distantly Supervised Composition Extraction from Tables in Materials Science Articles. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023; pp 13465–13483.

[42] Rajan, K.; Brinkhaus, H. O.; Agea, M. I.; Zielesny, A.; Steinbeck, C. DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature Communications* **2023**, *14*, 1–13.

A Appendix

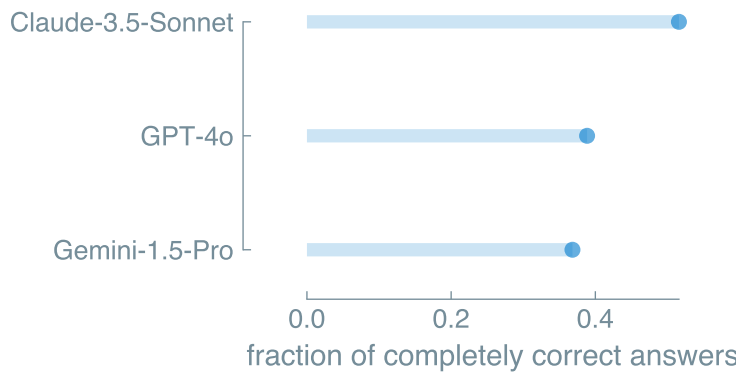

Figure 6: **Overall model performance on MACBENCH.** We compute the overall performance by measuring the fraction of questions answered completely correctly. Claude-3.5-Sonnet is the best-performing model, overall, followed by GPT-4o, with the worst model in our study being Gemini-1.5-Pro.

Table 1: **Model performance comparison on multiple choice questions.** Claude-3.5-Sonnet performs best overall, with the lowest extra classes, hamming distance, and highest rate of completely correct predictions. Gemini-1.5-Pro tends to hallucinate the most extra responses in MCQ questions. It is important to note that the standard deviations (std) are relatively high compared to the means, indicating significant variability in performance across different instances. The standard error of the mean (sem) values are generally low, suggesting that the sample means are reasonably precise estimates of the true population means.

	number of extra classes			number of missed classes			Hamming loss			frac. completely correct		
Model	mean	std	sem	mean	std	sem	mean	std	sem	mean	std	sem
Claude-3.5-Sonnet	0.37	0.50	0.03	0.39	0.54	0.03	0.70	0.94	0.05	0.52	0.50	0.02
GPT-4o	0.47	0.57	0.03	0.42	0.52	0.03	0.86	0.99	0.05	0.39	0.49	0.02
Gemini-1.5-Pro	0.53	0.68	0.04	0.40	0.50	0.03	0.88	1.04	0.06	0.37	0.48	0.02

Figure 7: **Heatmap of model performance for different question categories.** The colors and the numbers in the cells indicate the fraction of questions that have been answered completely correctly.

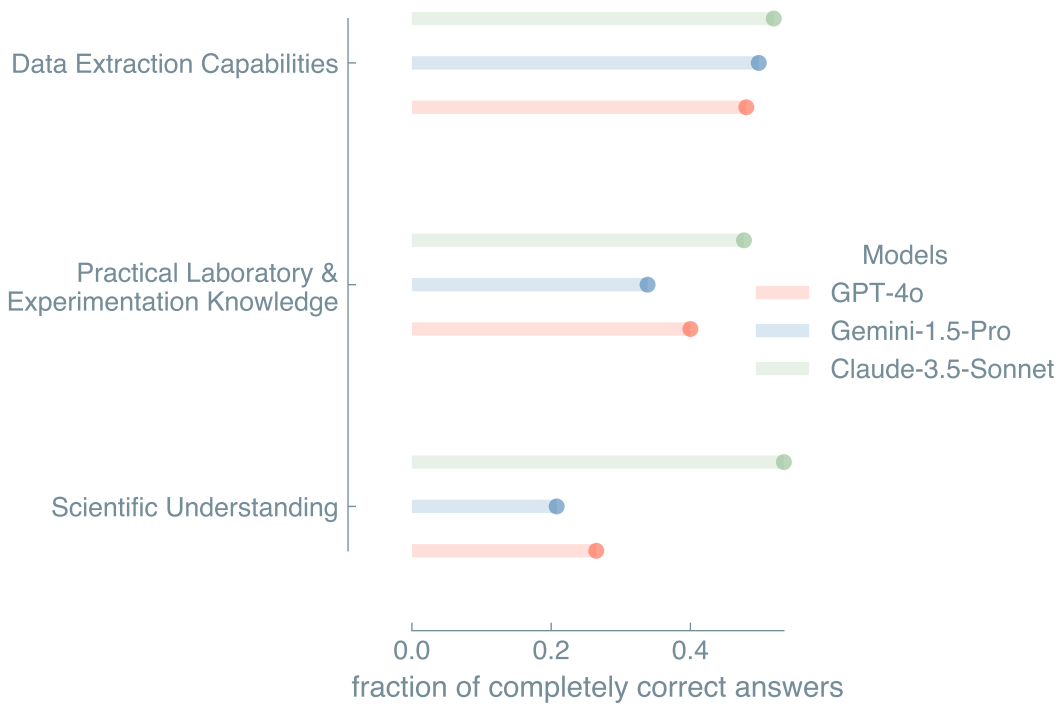

Figure 8: **MLM performance for question superclasses in MACBENCH.** We observe that models perform well for data extraction tasks but struggle with tasks requiring scientific understanding or reasoning scenarios in a chemistry lab.