Data-driven Design as a High-Impact, Ecologically Valid Benchmark for Document Understanding

Sireesh Gururaja[1] **Junwon Seo**[2] **Hung-Yi Lin**[2] **Jeremiah Milbauer**[1]
Anthony Rollett[2] **Emma Strubell**[1,2]
[1]Language Technologies Institute, School of Computer Science
[2]Department of Materials Science and Engineering
Carnegie Mellon University
sgururaj@cs.cmu.edu

Abstract

Data-driven design (DDD) is viewed in materials science as a promising avenue to accelerate materials discovery by narrowing the search space for candidate materials with desirable properties, and relies on correctly-extracted information from prior literature. Existing methods for DDD-related information extraction, however, rely on either laborious, hand-engineered pipelines, or the annotation of significant amounts of hard-to-collect data. We therefore propose DDD as a benchmark for zero- and few-shot document understanding focused on text, tables, and charts. Accurate generalization to new, unseen material domains is a way to accelerate scientific discovery by enabling the use of DDD in previously unexplored domains.

1 Introduction

Data-driven design (DDD), a process by which materials scientists use information extracted from the literature to inform future experiments, has emerged in the past decade as an important accelerator of materials discovery (Olivetti et al., 2020). As NLP methods have evolved, so too has their application to information extraction piece of data-driven design problems, from pipeline-based approaches relying heavily on rules-based, handwritten heuristics (Kim et al., 2017; Court & Cole, 2018; Jensen et al., 2019, *inter alia*) to end-to-end approaches involving fine-tuning large language models (LLMs) to act as information extractors and assistants (Zheng et al., 2023), or generate structured output describing properties directly (Dagdelen et al., 2024).

We choose to focus on information extraction, rather than end-to-end hypothesis generation, for multiple reasons. Focusing primarily on information extraction accelerates a difficult component of the DDD process, while leaving intact the rich ecosystem of specialized methods in materials science, such as physics-informed modeling Lee et al. (2024). From a machine learning perspective, information extraction is immediately verifiable in that metrics can be deterministically computed, as opposed to hypothesis generation, where evaluation remains difficult (Si et al., 2025). Further, the information extraction for DDD involves subtasks – parsing complicated information layouts and reasoning about the normalization and comparison of quantities – that are of practical use, in addition to being crucial to the success of model-based reasoning more broadly.

However, even current, LLM-based data-driven design work relies on laboriously collected annotated data. The method proposed in Dagdelen et al. (2024), for instance, suggests annotating " 100–500 text passages" in order to fine-tune an LLM to produce structured data. This type of data can be difficult to produce: it often requires domain expertise to collect, verify, and postprocess into a format that is appropriate for training such models, to say nothing of the challenges of the finetuning itself. This problem is exacerbated when considering that data-driven design efforts often seek to

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

extract information into subdomain-specific, non-overlapping schemas, limiting the possibility of data sharing or transfer learning between separate DDD efforts. Further complicating the process of DDD is the inherently multimodal nature of the extraction: information is stored across text, tables, and charts in papers that are variously available as either XML or PDF documents.

As such, the development of models capable of extracting information from these various modalities in a zero- or few-shot setting presents an exciting opportunity to accelerate data-driven design and thereby materials discovery. This work also takes advantage of increasingly multimodal LLMs, many of which now explicitly prioritize document understanding and visual question answering (VQA) as a primary objective in pretraining (Wang et al., 2024; Liu et al., 2024), and directs that development towards a task with huge potential impact; the tasks involved in DDD-related extraction, such as layout understanding, reference resolution and disambiguation, and numeric normalization also remain at the frontier of contemporary model capability (Miret & Krishnan, 2024).

2 Dataset Development

Setting and Evaluation We conceptualize this benchmark dataset as a few- and zero-shot evaluation of the task of extracting and normalizing quantities and strings from papers in their commonly distributed forms – XML and PDF. This extraction must be standardized into the form that existing DDD studies publish, typically a spreadsheet. Researchers targeting this benchmark would present a system that remains unchanged across domains, receives a description of the schema and test instances, and produces a tabular output in the presented schema.

Preliminary Dataset We piloted this task using Jensen et al. (2019). We observe a number of characteristics that place the task of IE for DDD at the frontier of contemporary model capability. These challenges include visual document understanding (particularly of tables, which we demonstrate with a worked example in appendix A, and charts, with examples in appendix B), in-document symbol and coreference resolution, and numeric reasoning to normalize quantities and units across papers. Baseline results displays significant variation based on prompt and data presentation (appendix C).

Broad Domain Coverage To accurately assess a model's degree of generalization to new domains, we wish to expand this dataset to several additional domains, including glasses Gupta et al. (2023), magnetic materials Itani et al. (2024), and metal-organic frameworks Zheng et al. (2023), *inter alia*.

Data correction and grounding. Recent DDD work increasingly trends towards automatically extracted datasets at scale with some manual validation. This extraction therefore focuses on modalities that models can handle – most often text, sometimes tables, and almost never charts. In developing this dataset further, we would like to extend the scope of the manual annotation and correction, ensuring accurate measurement, and covering all available modalities. Additionally, we propose annotating the original documents with the locations of where information was extracted and its format, allowing for fine-grained error analysis of where and why models fail.

Synthetic Data Generation. Given the existence of open knowledge bases of materials that reflect the high dimensionality of materials datasets like the Materials Project Horton et al. (2025), we additionally propose to generate new PDF and XML synthetic data that capture the variation in layouts and presentation of information such that models can be trained specifically for this task.

Open release of data. We plan to release initial version of this data as a list of DOIs, along with scripts that allow for the programmatic reconstruction of the dataset given the appropriate licenses. Longer-term, we plan to work directly with publishers for direct, scoped access to the papers in the chosen datasets, to allow open access to this benchmark.

3 Conclusion

We present information extraction (IE) for data-driven design as a high-impact benchmark task. Accurate IE across subdomains presents an instantly useful technology to materials scientists seeking to start DDD projects in unexplored domains, while guiding the development of document-centric

VLMs. We outline an existing, preliminary dataset and the challenges it presents, as well as next steps to develop this dataset into an open, ecolologically valid benchmark for document understanding.

References

Court, C. J. and Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific Data*, 5(1):180111, June 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.111. URL https://www.nature.com/articles/sdata2018111. Publisher: Nature Publishing Group.

Cunningham, W. S., Lei, T., Howard, H. C., Rupert, T. J., and Gianola, D. S. Kinetics of amorphous defect phases measured through ultrafast nanocalorimetry, 2025. URL https://arxiv.org/abs/2506.12179.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., and Jain, A. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418, February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45563-x. URL https://www.nature.com/articles/s41467-024-45563-x. Publisher: Nature Publishing Group.

Gupta, T., Zaki, M., Khatsuriya, D., Hira, K., Krishnan, N. M. A., and Mausam. DiSCoMaT: Distantly supervised composition extraction from tables in materials science articles. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13465–13483, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.753. URL https://aclanthology.org/2023.acl-long.753/.

Horton, M. K., Huck, P., Yang, R. X., Munro, J. M., Dwaraknath, S., Ganose, A. M., Kingsbury, R. S., Wen, M., Shen, J. X., Mathis, T. S., et al. Accelerated data-driven materials science with the materials project. *Nature Materials*, pp. 1–11, 2025.

Itani, S., Zhang, Y., and Zang, J. Large Language Model-Driven Database for Thermoelectric Materials, December 2024. URL http://arxiv.org/abs/2501.00564. arXiv:2501.00564 [cond-mat].

Jensen, Z., Kim, E., Kwon, S., Gani, T. Z. H., Román-Leshkov, Y., Moliner, M., Corma, A., and Olivetti, E. A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Central Science*, 5(5):892–899, May 2019. ISSN 2374-7943. doi: 10.1021/acscentsci.9b00193. URL https://doi.org/10.1021/acscentsci.9b00193. Publisher: American Chemical Society.

Kim, E., Huang, K., Tomala, A., Matthews, S., Strubell, E., Saunders, A., McCallum, A., and Olivetti, E. Machine-learned and codified synthesis parameters of oxide materials. *Scientific Data*, 4(1):170127, September 2017. ISSN 2052-4463. doi: 10.1038/sdata.2017.127. URL https://www.nature.com/articles/sdata2017127. Publisher: Nature Publishing Group.

Lee, D., Chen, W. W., Wang, L., Chan, Y.-C., and Chen, W. Data-driven design for metamaterials and multiscale systems: A review. *Advanced Materials*, 36(8):2305254, 2024. doi: https://doi.org/10.1002/adma.202305254. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adma.202305254.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Lorgouilloux, Y., Dodin, M., Paillaud, J.-L., Caullet, P., Michelin, L., Josien, L., Ersen, O., and Bats, N. IM-16: A new microporous germanosilicate with a novel framework topology containing *d*4*r* and *mtw* composite building units. *Journal of Solid State Chemistry*, 182(3):622–629, March 2009. ISSN 0022-4596. doi: 10.1016/j.jssc.2008.12.002. URL https://www.sciencedirect.com/science/article/pii/S0022459608006348.

Miret, S. and Krishnan, N. M. A. Are LLMs Ready for Real-World Materials Discovery?, September 2024. URL http://arxiv.org/abs/2402.05200. arXiv:2402.05200 [cond-mat].

Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y.-J., and Hiszpanski, A. M. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4):041317, December 2020. ISSN 1931-9401. doi: 10.1063/5.0021106. URL https://doi.org/10.1063/5.0021106.

Si, C., Yang, D., and Hashimoto, T. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T., and Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062, August 2023. ISSN 0002-7863. doi: 10.1021/jacs.3c05819. URL https://doi.org/10.1021/jacs.3c05819. Publisher: American Chemical Society.

A Worked example from Jensen et al. (2019)

Figure 4 represents indices 375-390 from our dataset. We reproduce the first four rows of this table here, and demonstrate how to extract the relevant columns in the first row.

If present, the silicon content is always the basis of normalization, and so receives a value of 1 in the `Si` column. This therefore leads us to normalize the germanium value, in the ratio of Si:Ge 0.4:0.6, to 0.667. This paper uses neither aluminum nor boron, leading to 0 values for both of those. Water and HF content are similarly normalized by dividing by 0.6.

In the table in Figure 4, the R column is interpreted as the OSDA, even though this is not specified in the paper. This is a common substitution, alongside others, such as using "T" as the basis for normalization. We therefore use the values in the R column for the SDA value.

Text found elsewhere on the page provides additional information that must be incorporated. Synthesis paragraph 2.1 implies that the OSDA is also the source of OH ions: "and 3-ethyl-1-methyl-3H-imidazol-1-ium bromide (98%, Solvionic), which was transformed into its OH form by ion exchange in water." The time and temperature (170°C for 14 days) are from the same paragraph; 14 days must be normalized to 336 hours.

The name of the OSDA is specified in the table caption. The names of the products are extracted into column S, but must be expanded using the table footnotes to indicate that "Arg" is argutite, and "Q" is quartz.

This table demonstrates several of the challenges in this dataset, from table understanding, to resolving in-table references, having conventional knowledge, and using contextual text that is not explicitly part of the table being considered or extracted.

Si	Ge	Al	OH	H_2O	HF	SDA	B	Time	Temp	SDA Type	Extracted
1	0.667	0	0.8335	33.34	0	0.8335	0	336	170	3-ethyl-1-meth...	TON+MFI+argutite
1	0.667	0	1.667	33.34	0	1.667	0	336	170	3-ethyl-1-meth...	MFI+unknown
1	0	0	0.5	8	0.5	0.5	0	336	170	3-ethyl-1-meth...	Amorphous
1	0.25	0	0.625	10	0.625	0.625	0	336	170	3-ethyl-1-meth...	IM-16+unknown

Table 1: Sample rows from our dataset, filtered from Jensen et al. (2019). This table represents the first four rows of the table seen in Figure 4

B Example Charts

These example charts are reproduced from Cunningham et al. (2025).

Table 1
Selection of the most representative synthesis of zeolite IM-16 with 3-ethyl-1-methyl-3*H*-imidazol-1-ium as OSDA.

Sample	Molar gel composition (T = Si+Ge)				Material
	H_2O/T	R/T	HF/T	Si/Ge	
1[a]	20	0.5	0	0.6:0.4	**TON**+**MFI**+Arg[c]
2[a]	20	1	0	0.6:0.4	**MFI**+ε?[d]
3[a]	8	0.5	0.5	1:0	Amorphous
4[a]	8	0.5	0.5	0.8:0.2	IM−16+ε?[d]
5[a]	8	0.5	0.5	0.6:0.4	IM−16+ε?[d]
6[a]	8	0.5	0.5	0.4:0.6	Q[c]+IM−16
7[a]	8	0.5	0.5	0.2:0.8	Q[c]
8[a]	8	0.6	0.4	0.8:0.2	IM−16+ε?[d]
9[a]	20	1	0.5	0.6:0.4	IM−16+ε?[d]
10[a]	3	0.3	0.3	0.8:0.2	IM−16+ε?[d]
11[a]	20	1	1	0.8:0.2	IM−16+ε?[d]
12[a]	20	1	1	0.6:0.4	IM−16+ε?[d]
13[a]	20	1	1	0.5:0.5	IM−16+Q[e]
14[b]	20	1	1	0.8:0.2	IM−16+**MFI**
15[b]	20	1	1	0.6:0.4	IM−16+ε?[d]
16[b]	20	1	1	0.5:0.5	IM−16

Silica sources:
[a] TEOS (tetraethylorthosilicate).
[b] Aerosil 200.
[c] Argutite.
[d] ε?: small quantity of one or more unknown impurities.
[e] Quartz.

Figure 1: Example table from the dataset, reproduced from Lorgouilloux et al. (2009, Table 1). This table demonstrates several of the challenges with table extraction in this dataset, including: (1) Generic table layout understanding; (2) Processing information related to tables, such as captions and footnotes; (3) Understanding and resolving in-document substitutions; and (4) Numerical reasoning to normalize ratios. Note that table understanding, being partially in text, remains easier than chart understanding.

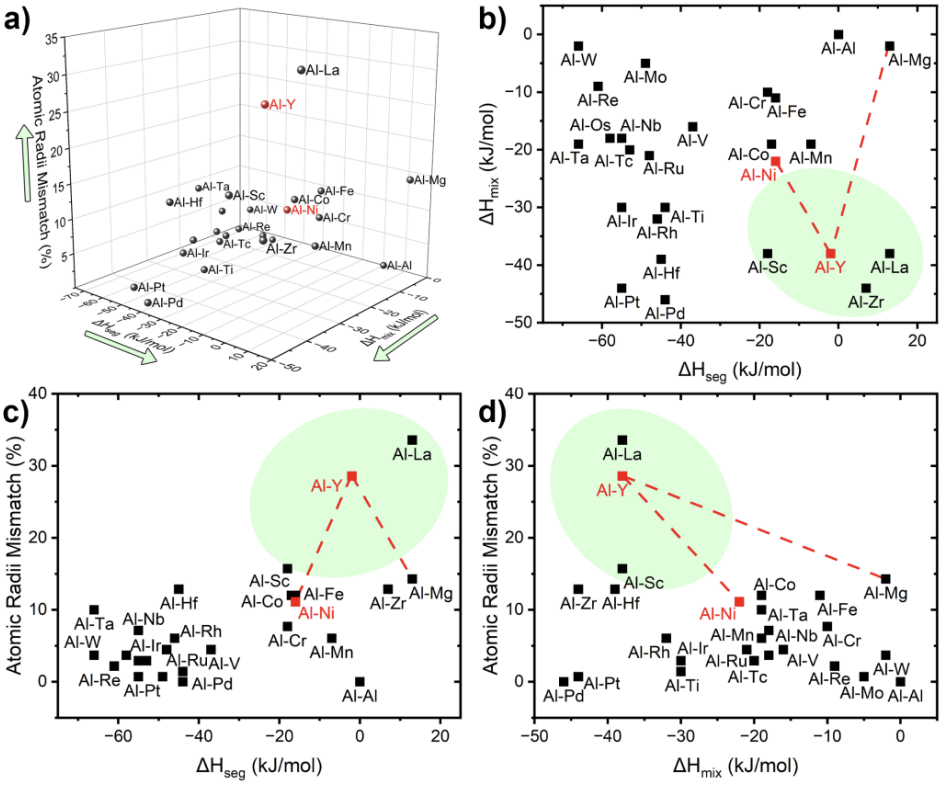

Figure 2: (a) 3D plot of binary Al alloys as a function of atomic radius mismatch, ΔH_{seg}, and ΔH_{mix} from Refs. [57, 59]. Green arrows indicate the directions for increased propensities for amorphous defect phase formation. The binary alloys explored in this study, Al-Ni and Al-Y, are colored in red. (b-d) 2D sections of (a) to facilitate comprehension with regions of optimal parameters for amorphous defect phase formation colored in green. Red dashed lines are drawn between the binary systems to illustrate the two ternary systems explored in this study, Al-Ni-Y and Al-Mg-Y.

Figure 2: An example chart from a materials science paper. This example features a 3-D visualization, which should likely not be parsed, along with a chart that makes precise use of color and line to convey information.

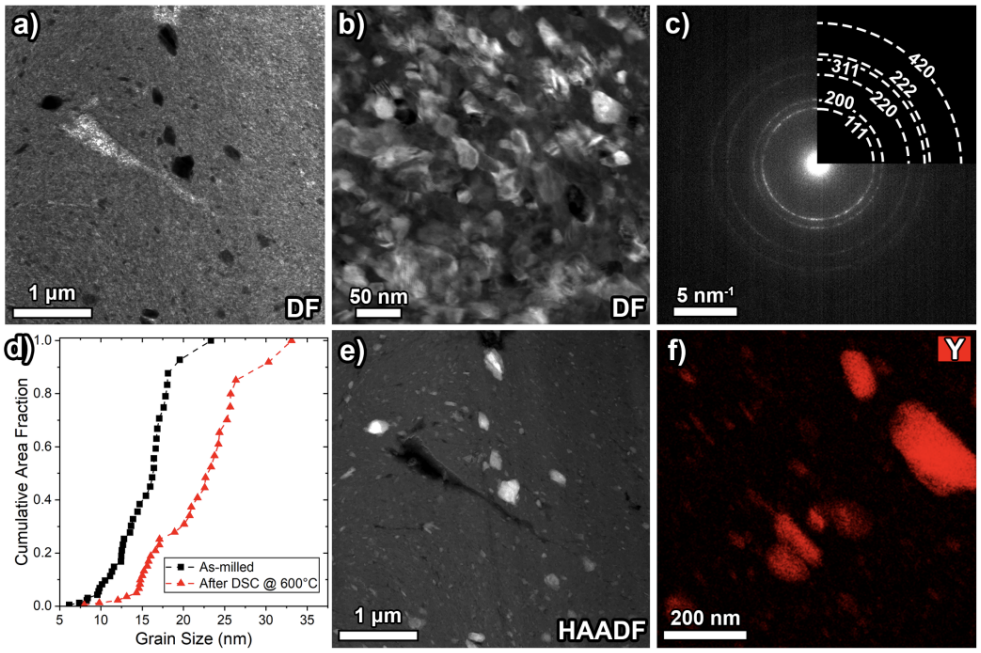

Figure 7: (a) STEM-DF micrograph of Al-Y following a single ultrafast DSC annealing run at 3,000 °C/s to 600 °C followed by a -3,000 °C/s quench. (b) A higher magnification STEM-DF micrograph showing a clear nanocrystalline structure. (c) Selected area diffraction pattern of the nanocrystalline Al-Y in (a), with matching diffraction rings corresponding to FCC Al. (d) Cumulative grain size distributions for the as-milled Al-Y from **Figure 6** (black) and post-annealed Al-Y (red). (e) Corresponding STEM-HAADF micrograph for the region in (a). (f) A higher magnification complementary STEM-EDS map of Y (red).

Figure 3: A second example chart from a materials science paper. This example features the side-by-side presentation of data visualization alongside the output of instruments. Ideally, all of these modalities contribute to the information extraction we propose.

C Baseline Results on Preliminary Datasets

These results come from running a GPT-4o based pipeline on a preliminary version of our dataset.

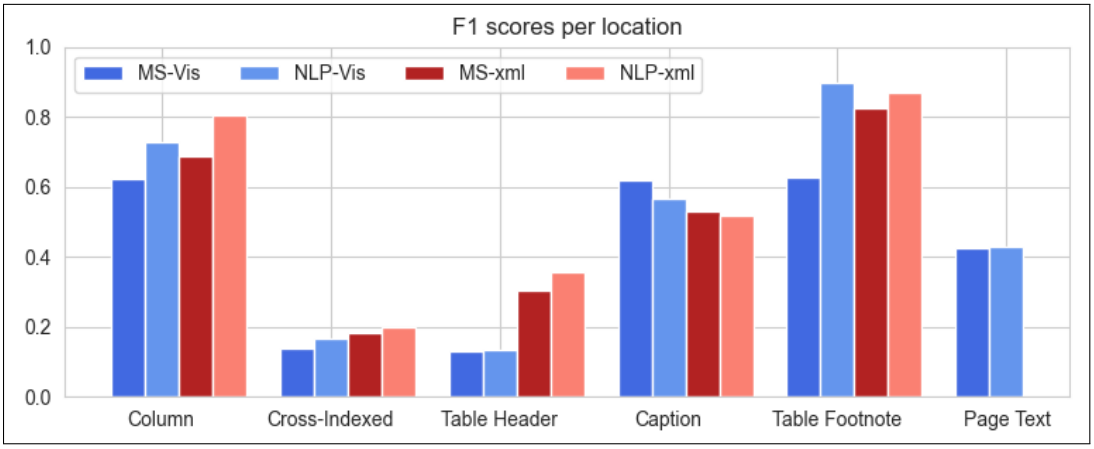

Figure 4: Results from GPT-4o on the preliminary version of our dataset. Groups indicate where information is found in a paper; "Column" and "Cross-Indexed" indicate two common table formats. MS- and NLP- indicate prompts from authors who were more familiar with materials science and NLP, respectively, and -Vis and -XML indicate the PDF and XML settings. These results imply a significant variance based on prompt construction, information modality, and information presentation; charts are not present in the original dataset at all.