AUTOMATED DATA EXTRACTION FROM SOLAR CELL LITERATURE USING LARGE LANGUAGE MODELS

Sherjeel Shabih, HU Berlin

Christoph T. Koch
HU Berlin

José A. Márquez
HU Berlin

Kevin Maik Jablonka FSU Jena, CEEC Jena, JCSM Jena, HIPOLE Jena
mail@kjablonka.com

1 INTRODUCTION

The discovery and development of new materials underpin technological progress in renewable energy (Tabor et al., 2018; Jablonka et al., 2020; Yao et al., 2022). However, the growing complexity and volume of materials research creates an urgent need for systematic data collection and analysis. This challenge is particularly acute in energy materials, where rapid iteration between material design, synthesis, and characterization generates vast amounts of heterogeneous data. Photovoltaics exemplify this challenge. As a key technology for sustainable energy generation (Haegel et al., 2019), solar cells have seen remarkable advances, with perovskite-based devices emerging as one of the fastest-developing technologies to date (Tabor et al., 2018). Their rapid evolution stems from extensive experimentation with device architectures, chemical compositions, and fabrication methods. Understanding the relationships between these variables and device performance is essential for rational design of high-efficiency solar cells. However, the accelerating pace of research makes it increasingly difficult for researchers to track and synthesize findings from published literature. While databases have been created to collect and organize perovskite solar cell data (Jacobsson et al., 2022), manual curation proves both unscalable and error-prone. The largest perovskite solar cell database (Jacobsson et al., 2022), for instance, has remained static since its initial release despite numerous new publications reporting high-performing devices.

Here, we present an automated framework using large language models (LLMs) for extracting structured data from solar cell literature. By systematically validating our approach against expert annotations, we show that LLMs can match or exceed human accuracy in extracting key performance indicators. Our implementation includes thorough validation and normalization protocols, leveraging LLM-as-Judge approaches (Li et al., 2024), ensuring consistent data quality across diverse publication formats. Through comparative analysis of multiple LLM architectures, we identify optimal approaches for the extraction task. Our work advances materials informatics by providing a scalable solution to database maintenance. While demonstrated on perovskite solar cells, this approach offers a template for automated data extraction in other rapidly evolving materials domains. The framework particularly excels at maintaining consistency in extracting numerical parameters and device architectures, addressing key challenges in manual curation across materials science.

2 RELATED WORK

Data Extraction Methods in Materials Science Scientific data extraction has evolved from rule-based approaches to increasingly sophisticated machine learning methods (Schilling-Wilhelmi et al., 2025). Early tools like ChemDataExtractor (Swain & Cole, 2016) and ChemicalTagger (Hawizy et al., 2011) relied on regular expressions and hand-crafted rules, which proved effective but limited in scope and adaptability. The emergence of large language models (LLMs) has transformed this landscape. Recent work demonstrates that LLMs can extract structured scientific data through simple prompting (Patiny & Godin, 2023; Polak & Morgan, 2024) or fine-tuning (Dagdelen et al., 2024), offering greater flexibility and generalization than traditional methods (Jablonka et al., 2023).

Evolution of PV Materials Databases The photovoltaics community has developed several approaches to materials data collection and organization. Initial efforts focused on computational

screening, producing databases of predicted material properties (Stevanović et al., 2012; Castelli et al., 2012). However, the gap between computational predictions and experimental reality motivated the creation of manually curated experimental databases. Notable examples include the Emerging PV Database (Almora et al., 2020) and the Perovskite Database (Jacobsson et al., 2022), which capture detailed device architectures and performance metrics. These resources have enabled new applications, such as fine-tuning LLMs for property prediction (Xie et al., 2023), while highlighting the need for automated curation methods.

3 METHODS

Our experimental framework implements two interconnected pipelines: an extraction pipeline that processes PDF documents using configurable LLM models and an evaluation pipeline that ensures accurate assessment of the extracted data (Figure 1).

3.1 DATA EXTRACTION PIPELINE

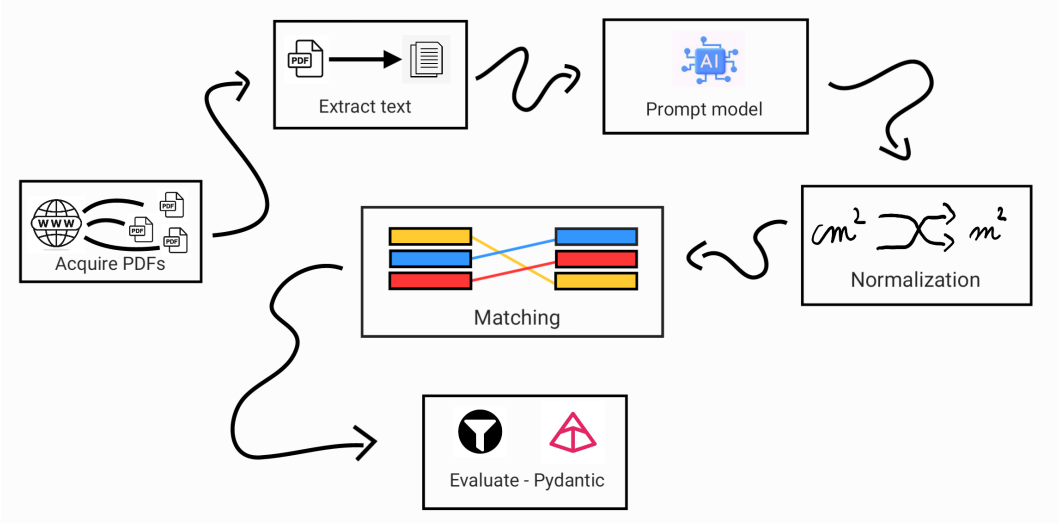

Figure 1: **Overview of the data extraction pipeline.** Our extraction pipeline, Perovscribe, provides bulk processing of research papers. The pipeline handles normalizing and matching extracted cells before evaluation.

The pipeline processes research papers in bulk and stores results by model for NOMAD (Scheidgen et al., 2023) integration. PyMuPDF extracts text from PDFs, which the models (GPT-4o, Gemini 1.5 Pro/Flash, and Claude 3.5 Sonnet at temperature 0) process using the Instructor package (Liu, 2025) to reliably extract only explicitly stated information from papers, with automatic retries for schema validation.

3.2 EVALUATION PIPELINE

The evaluation pipeline incorporates field-specific considerations, including unit normalization across diverse reporting conventions and intelligent cell matching to handle multiple solar cell entries within single papers.

Matching Algorithm and Cost Function Given that extractions can contain multiple solar cells, we aimed to optimally pair extracted cells with the corresponding ground truth entries. To ensure this, we implemented a structured matching process using the Kuhn–Munkres algorithm (Kuhn, 1955; Munkres, 1957). The matching process relies on a cost function that incorporates domain-specific priorities:

$$\text{Cost} = -0.7 \times \text{stack}_{sim(\text{t,e})} - 0.2 \times \text{deposition}_{sim(\text{t,e})} - 0.1 \times \text{all fields}_{sim(\text{t,e})}, \quad (1)$$

where $\text{stack}_{sim(\text{t,e})}$ is the similarity score for the cell stack, $\text{deposition}_{sim(\text{t,e})}$ is the same for deposition data within the layers, and $\text{all fields}_{sim(\text{t,e})}$ for all data extracted. For the cell stack, we take an average Levenshtein ratio for each layer defined as a string. For deposition and all fields, we use a normalized inverted DeepDiff (Dehpour) distance on the Python dictionary objects. This formulation emphasizes structural similarity in the extracted cell stack, which fundamentally defines a solar cell. Deposition details and other metadata play secondary roles. We assessed the robustness of this cost function by conducting an ablation study in which we systematically removed components and modified weight distributions (see Appendix A.2).

3.3 Evaluation Metrics

Our evaluation framework combines multiple complementary metrics to assess extraction quality. The foundation is an overall similarity score computed using DeepDiff (Dehpour), which measures structural differences between extracted data and ground truth.

At the device level, we track the total number of solar cells identified in both ground truth and extraction, along with the number of successfully matched devices. Device recall quantifies the proportion of ground truth cells that were correctly extracted. We supplement this with structural accuracy measures that assess how well the extraction preserves device architecture, focusing on stack composition and layer-wise similarities.

For numerical and categorical fields, we compute precision and recall using predefined tolerances to accommodate acceptable variations in reported values. These tolerances are particularly important for experimental measurements like efficiency and current density, where minor variations should not be penalized. We aggregate these scores at both the individual field level and across entire devices to provide granular and holistic performance measures.

LLM-as-Judge For cases in which our manually created scoring methods do not work, we employ an LLM-as-Judge approach for scoring. For cases in which our manually created scoring methods do not work, we use LLMs to evaluate the extraction performance (Li et al., 2024). For this, we use `claude-3-5-haiku-20241022` with a user prompt that asks to compare two fields and to return a boolean indicating if the fields match. This methodology proves particularly valuable for chemical formulas where equivalent representations exist. For instance, our evaluation recognizes that "MAPbI3" and "CH3NH3PbI3" represent identical compounds, with the former using a common abbreviation for methylammonium. Similarly, the judge correctly identifies chemical equivalence between variants like "MAPbI3-xClx" and "CH3NH3PbI3" or between "(PEA)2MA4Pb5I16" and "PEA2MA4Pb5I16". The LLM-based evaluation also handles naming variations in material layers, successfully matching "TiO2-c" with "TiO2", "Spiro-MeOTAD" with "Spiro-OMeTAD", and "Perovskite with PCBM/PEG network" with "CH3NH3PbI3 w/ PCBM/PEG", while correctly rejecting non-equivalent matches like different atmosphere conditions or conflicting treatment durations such as "10 min TiCl4 treatment" versus "TiCl4 treatment for 30 min".

3.4 Data Collection

To evaluate our extraction framework, we assembled a dataset of 68 solar cell research papers, including 54 from Jacobsson et al. (2021) and 14 additional publications manually selected to cover edge cases such as tandem architectures, self-assembled monolayers (SAMs), and stability studies. We established a validation protocol involving 13 domain experts in solar cell research, with each paper assigned to three independent reviewers to ensure robust cross-validation.

To optimize the validation process, we first performed automated extractions using Claude-3.5-Sonnet, providing experts with baseline structured data through the NOMAD platform Scheidgen et al. (2023). Experts were instructed to thoroughly verify and correct these preliminary extractions over a four-week period, identifying any errors or omissions in the model's output. This process yielded 63 validated papers from 11 experts.

Inter-annotator agreement analysis revealed limited consensus among expert extractions (see Figure 4). Only 15 papers achieved agreement between two or more experts on fundamental parameters (device count and four primary performance metrics), with complete consensus among all assigned reviewers in four cases. These 15 validated extractions, representing contributions from all participating experts, were established as the ground truth dataset for evaluating our extraction framework's performance.

4 Results and Discussion

The overall precision was consistently high across models, with all achieving scores above 0.88. Claude 3.5 Sonnet led with precision of 0.938, followed by GPT-4o at 0.904, while both Gemini variants performed similarly (Flash: 0.887, Pro: 0.886). The small performance spread of approximately 0.05 between models indicates that current LLM architectures have achieved robust capabilities in technical data extraction. Bootstrapped performance analysis revealed more nuanced differences between models (Figure 2). Claude 3.5 Sonnet achieved perfect precision on multiple papers and showed statistically significant improvements over GPT-4o.

A detailed performance breakdown across multiple metrics revealed important trade-offs (Figure 3). While Claude 3.5 Sonnet excelled in precision, it scored lowest in recall among the tested models. We also observed substantial variations in how models handled complex extractions. Notably, Gemini 1.5 Flash required frequent LLM Judge calls, indicating that its outputs often fell outside our standard evaluation pipeline. In contrast, Claude 3.5 Sonnet needed minimal judge interventions, suggesting more standardized output formatting.

These patterns highlight an important consideration for deployment: while all tested models can extract data with high precision, their varying recall scores and output formats suggest different optimal use cases. Models requiring fewer judge calls might be preferable for large-scale automated extraction, while those with higher recall might better suit applications where completeness is critical. If we look at detailed scores for the primary solar cell metrics, we see that no one model outperforms the others.

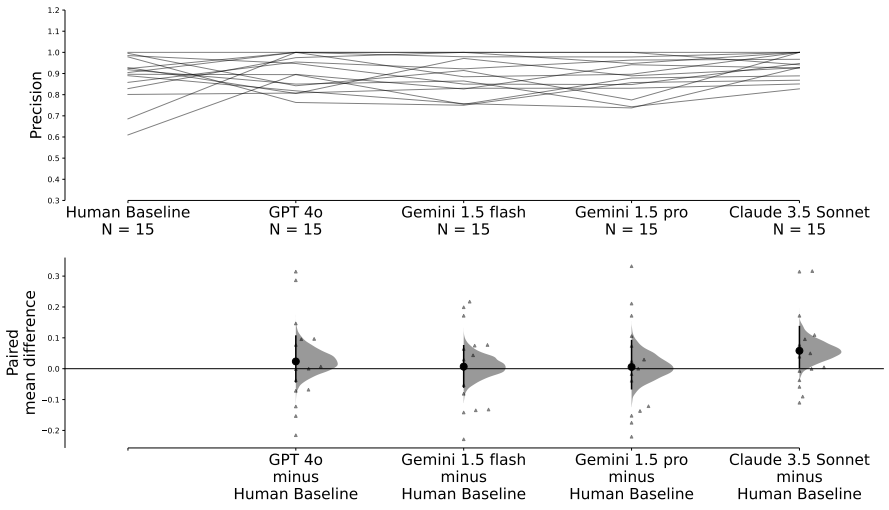

Figure 2: **Bootstrapped performance differences between models.** For this analysis, we use DABEST (Ho et al., 2019) to bootstrap effect sizes between the overall precision scores of the models. The top plot shows the precision scores for different papers in our test set, where each paper is represented by one line. We can see that for multiple papers Claude 3.5 Sonnet achieves perfect precision. The bottom plot shows bootstrapped effect sizes relative to a random Human Baseline. We find that Claude 3.5 Sonnet significantly outperforms GPT 4o in overall precision.

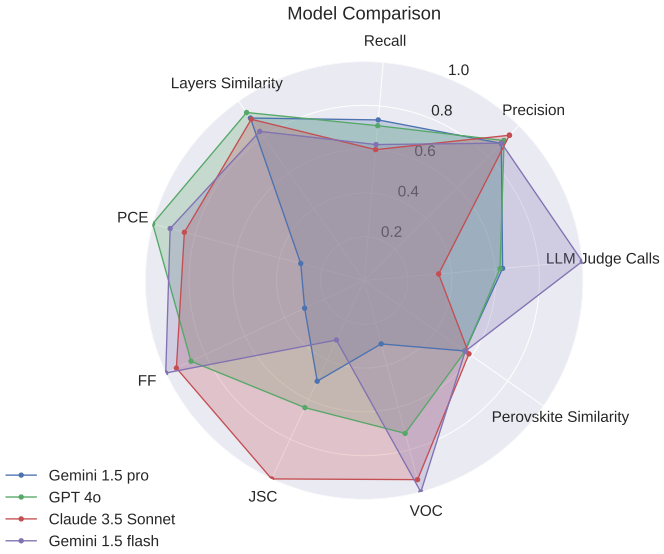

Figure 3: **Performance breakdown for different models.** In all cases, the best performance is on the outer rim. We show different models in color. The precision and recall scores are averaged over all fields. Layers similarity is a DeepDiff score on the detailed characteristics of the device stack (i.e., including processing conditions). While we observe that Claude 3.5 Sonnet scores best in precision, it scores worst in recall. We also observe large differences in the number of calls to the LLM Judge. While Claude 3.5 Sonnet requires the least calls, Gemini 1.5 flash requires the most — indicating that many of the results Gemini 1.5 flash produces cannot be handled by our manually developed scoring functions but instead require calls to the LLM Judge. While inspecting the main solar cell performance metrics, PCE, FF, J_{sc}, and V_{oc}, we can see that even though there is no clear winner, Claude 3.5 Sonnet does best overall.

5 Limitations

Some limitations affect this work's scope and conclusions. A fundamental challenge stems from the significant variance in our human-labeled dataset (Figure 4). Expert annotators showed considerable disagreement in their extractions, highlighting the inherent complexity of standardizing solar cell data extraction. This variance complicates both the training and evaluation of automated approaches.

In addition, the current framework processes only textual content. Solar cell publications often present critical performance data and device architectures through visual elements, which our system cannot interpret. Additionally, the analysis covers only the main manuscript text, excluding supplementary information that frequently contains detailed experimental procedures and additional characterization data.

While our validation approach involved multiple expert reviewers per paper, the dataset remains modest at 15 papers. Expanding this dataset would strengthen our conclusions, but the resource-intensive nature of expert validation presents a practical barrier. The substantial time and expertise required for thorough paper review makes large-scale validation prohibitively expensive.

6 Conclusions

Data extraction from scientific literature is a foundational challenge in materials science. For simple cases, rule-based approaches and regular expressions can reliably extract structured information. Traditional natural language processing methods work well for standardized formats and clearly defined parameters. For small-scale efforts, manual curation by experts remains viable. However, there are vast amounts of complex, heterogeneous scientific data for which these approaches prove inadequate.

For critical research domains like perovskite solar cells, researchers have relied on manual extraction to build comprehensive databases, a process that becomes increasingly unsustainable as the field grows. Here, we have shown that large language models can automate this process while maintaining high accuracy. Our results demonstrate that these models not only match human performance but can provide consistency in cases where expert annotators disagree.

This success in handling the complexities of solar cell literature — with its intricate device architectures, varied reporting formats, and technical nuance — reveals the transformative potential of LLMs in scientific knowledge management. While the path to fully automated scientific data extraction still faces challenges, our results suggest that the collective knowledge embedded in the scientific literature can now be systematically accessed and structured at scale. This capability promises to accelerate not just solar cell research but any field where the wealth of published knowledge currently lies trapped in unstructured text, waiting to be unlocked.

REFERENCES

Osbel Almora, Derya Baran, Guillermo C. Bazan, Christian Berger, Carlos I. Cabrera, Kylie R. Catchpole, Sule Erten-Ela, Fei Guo, Jens Hauch, Anita W. Y. Ho-Baillie, T. Jesper Jacobsson, Rene A. J. Janssen, Thomas Kirchartz, Nikos Kopidakis, Yongfang Li, Maria A. Loi, Richard R. Lunt, Xavier Mathew, Michael D. McGehee, Jie Min, David B. Mitzi, Mohammad K. Nazeeruddin, Jenny Nelson, Ana F. Nogueira, Ulrich W. Paetzold, Nam-Gyu Park, Barry P. Rand, Uwe Rau, Henry J. Snaith, Eva Unger, Lídice Vaillant-Roca, Hin-Lap Yip, and Christoph J. Brabec. Device performance of emerging photovoltaic materials (version 1). *Advanced Energy Materials*, 11(11), December 2020. ISSN 1614-6840. doi: 10.1002/aenm.202002774. URL http://dx.doi.org/10.1002/aenm.202002774.

Ivano E Castelli, Thomas Olsen, Soumendu Datta, David D Landis, Søren Dahl, Kristian S Thygesen, and Karsten W Jacobsen. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy & Environmental Science*, 5(2):5814–5819, 2012.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1), February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45563-x. URL http://dx.doi.org/10.1038/s41467-024-45563-x.

Sep Dehpour. DeepDiff. URL https://github.com/seperman/deepdiff.

Nancy M. Haegel, Harry Atwater, Teresa Barnes, Christian Breyer, Anthony Burrell, Yet-Ming Chiang, Stefaan De Wolf, Bernhard Dimmler, David Feldman, Stefan Glunz, Jan Christoph Goldschmidt, David Hochschild, Ruben Inzunza, Izumi Kaizuka, Ben Kroposki, Sarah Kurtz, Sylvere Leu, Robert Margolis, Koji Matsubara, Axel Metz, Wyatt K. Metzger, Mahesh Morjaria, Shigeru Niki, Stefan Nowak, Ian Marius Peters, Simon Philipps, Thomas Reindl, Andre Richter, Doug Rose, Keiichiro Sakurai, Rutger Schlatmann, Masahiro Shikano, Wim Sinke, Ron Sinton, B.J. Stanbery, Marko Topic, William Tumas, Yuzuru Ueda, Jao van de Lagemaat, Pierre Verlinden, Matthias Vetter, Emily Warren, Mary Werner, Masafumi Yamaguchi, and Andreas W. Bett. Terawatt-scale photovoltaics: Transform global energy. *Science*, 364 (6443):836–838, May 2019. ISSN 1095-9203. doi: 10.1126/science.aaw1845. URL http://dx.doi.org/10.1126/science.aaw1845.

Lezan Hawizy, David M Jessop, Nico Adams, and Peter Murray-Rust. Chemicaltagger: A tool for semantic text-mining in chemistry. *Journal of cheminformatics*, 3:1–13, 2011.

Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. Moving beyond p values: data analysis with estimation graphics. *Nature Methods*, 16(7):565–566, June 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0470-3. URL http://dx.doi.org/10.1038/s41592-019-0470-3.

Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-data science in porous materials: Materials genomics and machine learning. *Chemical Reviews*, 120(16):8066–8129, June 2020. ISSN 1520-6890. doi: 10.1021/acs.chemrev.0c00004. URL http://dx.doi.org/10.1021/acs.chemrev.0c00004.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M. Bran, Stefan Bringuier, L. Catherine Brinson, Kamal Choudhary, Defne Circi, Sam Cox, Wibe A. de Jong, Matthew L. Evans, Nicolas Gastellu, Jerome Genzling, María Victoria Gil, Ankur K. Gupta, Zhi Hong, Alishba Imran, Sabine Kruschwitz, Anne Labarre, Jakub Lála, Tao Liu, Steven Ma, Sauradeep Majumdar, Garrett W. Merz, Nicolas Moitessier, Elias Moubarak, Beatriz Mouriño, Brenden Pelkie, Michael Pieler, Mayk Caldas Ramos, Bojana Ranković, Samuel G. Rodriques, Jacob N. Sanders, Philippe Schwaller, Marcus Schwarting, Jiale Shi, Berend Smit, Ben E. Smith, Joren Van Herck, Christoph Völker, Logan Ward, Sean Warren, Benjamin Weiser, Sylvester Zhang, Xiaoqi Zhang, Ghezal Ahmad Zia, Aristana Scourtas, K. J. Schmidt, Ian Foster, Andrew D. White, and Ben Blaiszik. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023. ISSN 2635-098X. doi: 10.1039/d3dd00113j. URL http://dx.doi.org/10.1039/D3DD00113J.

T. Jesper Jacobsson, Adam Hultqvist, Alberto García-Fernández, Aman Anand, Amran Al-Ashouri, Anders Hagfeldt, Andrea Crovetto, Antonio Abate, Antonio Gaetano Ricciardulli, Anuja Vijayan, Ashish Kulkarni, Assaf Y. Anderson, Barbara Primera Darwich, Bowen Yang, Brendan L. Coles, Carlo A. R. Perini, Carolin Rehermann, Daniel Ramirez, David Fairen-Jimenez, Diego Di Girolamo, Donglin Jia, Elena Avila, Emilio J. Juarez-Perez, Fanny Baumann, Florian Mathies, G. S. Anaya González, Gerrit Boschloo, Giuseppe Nasti, Gopinath Paramasivam, Guillermo Martínez-Denegri, Hampus Näsström, Hannes Michaels, Hans Köbler, Hua Wu, Iacopo Benesperi, M. Ibrahim Dar, Ilknur Bayrak Pehlivan, Isaac E. Gould, Jacob N. Vagott, Janardan Dagar, Jeff Kettle, Jie Yang, Jinzhao Li, Joel A. Smith, Jorge Pascual, Jose J. Jerónimo-Rendón, Juan Felipe Montoya, Juan-Pablo Correa-Baena, Junming Qiu, Junxin Wang, Kári Sveinbjörnsson, Katrin Hirselandt, Krishanu Dey, Kyle Frohna, Lena Mathies, Luigi A. Castriotta, Mahmoud. H. Aldamasy, Manuel Vasquez-Montoya, Marco A. Ruiz-Preciado, Marion A. Flatken, Mark V. Khenkin, Max Grischek, Mayank Kedia, Michael Saliba, Miguel Anaya, Misha Veldhoen, Neha Arora, Oleksandra Shargaieva, Oliver Maus, Onkar S. Game, Ori Yudilevich, Paul Fassl, Qisen Zhou, Rafael Betancur, Rahim Munir, Rahul Patidar, Samuel D. Stranks, Shahidul Alam, Shaoni Kar, Thomas Unold, Tobias Abzieher, Tomas Edvinsson, Tudur Wyn David, Ulrich W. Paetzold, Waqas Zia, Weifei Fu, Weiwei Zuo, Vincent R. F. Schröder, Wolfgang Tress, Xiaoliang Zhang, Yu-Hsien Chiang, Zafar Iqbal, Zhiqiang Xie, and Eva Unger. An open-access database and analysis tool for perovskite solar cells based on the fair data principles. *Nature Energy*, 7(1):107–115, December 2021. ISSN 2058-7546. doi: 10.1038/s41560-021-00941-3. URL http://dx.doi.org/10.1038/s41560-021-00941-3.

T Jesper Jacobsson, Adam Hultqvist, Alberto García-Fernández, Aman Anand, Amran Al-Ashouri, Anders Hagfeldt, Andrea Crovetto, Antonio Abate, Antonio Gaetano Ricciardulli, Anuja Vijayan, et al. An open-access database and analysis tool for perovskite solar cells based on the fair data principles. *Nature Energy*, 7(1):107–115, 2022.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL https://arxiv.org/abs/2412.05579.

Jason Liu. Instructor, 2025. URL https://github.com/instructor-ai/instructor.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

Luc Patiny and Guillaume Godin. Automatic extraction of fair data from publications using llm. *ChemRxiv preprint*, December 2023. doi: 10.26434/chemrxiv-2023-05v1b-v2. URL http://dx.doi.org/10.26434/chemrxiv-2023-05v1b-v2.

Maciej P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1), February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-45914-8. URL http://dx.doi.org/10.1038/s41467-024-45914-8.

Markus Scheidgen, Lauri Himanen, Alvin Noe Ladines, David Sikter, Mohammad Nakhaee, Ádám Fekete, Theodore Chang, Amir Golparvar, José A. Márquez, Sandor Brockhauser, Sebastian Brückner, Luca M. Ghiringhelli, Felix Dietrich, Daniel Lehmberg, Thea Denell, Andrea Albino, Hampus Näsström, Sherjeel Shabih, Florian Dobener, Markus Kühbach, Rubel Mozumder, Joseph F. Rudzinski, Nathan Daelman, José M. Pizarro, Martin Kuban, Cuauhtemoc Salazar, Pavel Ondračka, Hans-Joachim Bungartz, and Claudia Draxl. Nomad: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8 (90):5388, 2023. doi: 10.21105/joss.05388. URL https://doi.org/10.21105/joss.05388.

Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T. Koch, José A. Márquez, and Kevin Maik Jablonka. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews*, 2025. ISSN 1460-4744. doi: 10.1039/d4cs00913d. URL http://dx.doi.org/10.1039/D4CS00913D.

Vladan Stevanović, Stephan Lany, Xiuwen Zhang, and Alex Zunger. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Physical Review B*, 85(11), March 2012. ISSN 1550-235X. doi: 10.1103/physrevb.85.115104. URL http://dx.doi.org/10.1103/PhysRevB.85.115104.

Matthew C Swain and Jacqueline M Cole. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904, 2016.

Daniel P. Tabor, Loïc M. Roch, Semion K. Saikin, Christoph Kreisbeck, Dennis Sheberla, Joseph H. Montoya, Shyam Dwaraknath, Muratahan Aykol, Carlos Ortiz, Hermann Tribukait, Carlos Amador-Bedolla, Christoph J. Brabec, Benji Maruyama, Kristin A. Persson, and Alán Aspuru-Guzik. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials*, 3(5):5–20, April 2018. ISSN 2058-8437. doi: 10.1038/s41578-018-0005-z. URL http://dx.doi.org/10.1038/s41578-018-0005-z.

Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, and Bram Hoex. Large language models as master key: Unlocking the secrets of materials science with gpt, 2023.

Zhenpeng Yao, Yanwei Lum, Andrew Johnston, Luis Martin Mejia-Mendoza, Xin Zhou, Yonggang Wen, Alán Aspuru-Guzik, Edward H. Sargent, and Zhi Wei Seh. Machine learning for a sustainable energy future. *Nature Reviews Materials*, 8(3):202–215, October 2022. ISSN 2058-8437. doi: 10.1038/s41578-022-00490-5. URL http://dx.doi.org/10.1038/s41578-022-00490-5.

A Appendix

A.1 Data collection

An overview of the interrater agreement is shown in Figure 4.

A.2 Matching function sensitivity analysis

We performed the following modifications:

- Removing deposition information had no significant effect, as deposition data was sparsely extracted by both experts and models.
- Adjusting the weight distribution (e.g., shifting to 0.5 stacks, 0.4 depositions, 0.1 all fields or 0.5 stacks, 0.1 depositions, 0.4 all fields) resulted in negligible differences due to the lack of extracted deposition information.
- Sensitivity analysis confirmed that our approach is stable; small perturbations in weight values did not significantly alter the results.

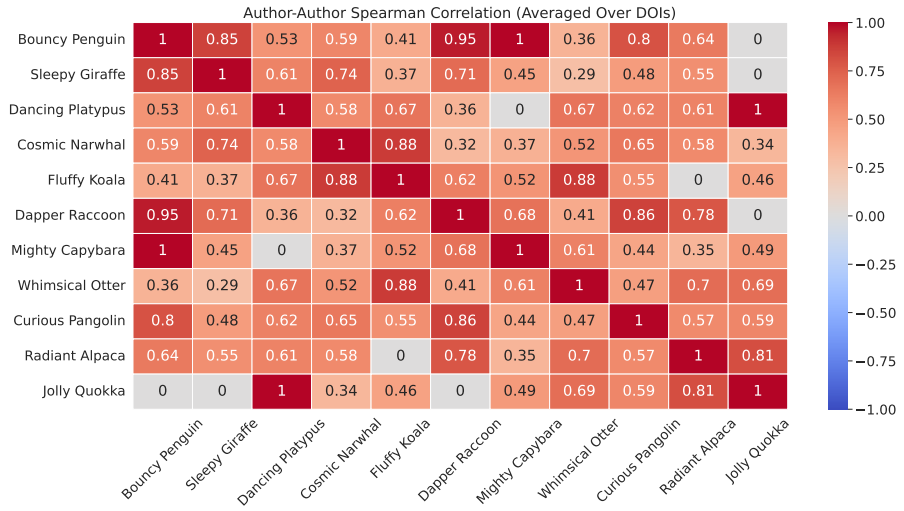

Figure 4: **Interrater agreement between human labelers.** Each row and column represents a human labeler. The color coding shows the Spearman rank correlation coefficient, which, by definition, is perfect for all diagonal elements. However, we observe that the correlation for the off-diagonal elements is often rather weak.

These findings reinforce our decision to prioritize cell stack similarity as the defining criterion for matching, as meaningful variations in solar cell composition are typically reflected in the stack structure rather than minor deposition or metadata differences.