



VIPERLAB

FULLY CONNECTED VIRTUAL AND PHYSICAL
PEROVSKITE PHOTOVOLTAICS LAB

D7.5

**Update to D7.2 (metastudy based
on the perovskite database)
taking into account latest data**

**DELIVERABLE
REPORT**



**FULLY CONNECTED VIRTUAL AND
PHYSICAL PEROVSKITE PHOTOVOLTAICS LAB
VIPERLAB**

DELIVERABLE

**D7.5: UPDATE TO D7.2, TAKING INTO ACCOUNT LATEST DATA
ADDED TO THE PEROVSKITE DATABASE**

Project References

Project Acronym	VIPERLAB
Project Title	Fully connected virtual and physical perovskite photovoltaics lab
Project Coordinator	Helmholtz-Zentrum Berlin
Project Start and Duration	1 st June 2021, 42 months

Deliverable References

Deliverable No	D7.5
Type	Report
Dissemination level	Public
Work Package	WP7
Lead beneficiary	HZB
Due date of deliverable	30 th Nov 2024
Actual submission date	28 th Nov 2024

Document history

Version	Status	Date	Beneficiary	Author
1.0	First Draft	27 th Nov 2024	HZB	Eva Unger
1.1	Second Draft	28 th Nov 2024	EPFL	C. Wolff
1.2	Review	28 th Nov 2024	HZB	N. Maticiuc

DISCLAIMER

'Fully connected virtual and physical perovskite photovoltaics lab' VIPERLAB is a Collaborative Project funded by the European Commission under Horizon 2020. Contract: 101006715, Start date of Contract: 01/06/2021; Duration: 42 months.

The authors are solely responsible for this information, and it does not represent the opinion of the European Community. The European Community is not responsible for any use that might be made of the data appearing therein.



Table of content

EXECUTIVE SUMMARY	4
1. INTRODUCTION	5
2. METADATA STUDIES BASED ON THE PEROVSKITE DATABASE	7
2.1 MACHINE LEARNING ON PEROVSKITE PV STABILITY DATA	8
2.2 METADATA STUDY: ARE CHECK-LIST FOR PV DATA PUBLICATIONS SUFFICIENT?.....	9
2.3 METADATA STUDY: BANDGAP TUNING OF HALIDE PEROVSKITES	10
3. THE PEROVSKITE (LITERATURE) DATABASE IN NOMAD	12
4. FIRST RESULTS FROM LLM-BASED DATA MINING	13
5. RESEARCH DATA MANAGEMENT IN NOMAD.....	13
5.1 EXAMPLE: PEROVSKITE PV RESEARCH DATA FROM THE “SLOT DYE COATING BASELINE” AT HZB	14
6. CONCLUSION.....	16
7. REFERENCES	17



EXECUTIVE SUMMARY

One of the main focuses of this task was the broader adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles in the perovskite PV community. A central achievement of has been the establishment of the Perovskite Database, an open-access platform that compiles metadata and performance metrics from over 16,000 publications. This initiative has provided the research community with a valuable resource for analyzing trends, identifying gaps, and fostering collaboration. However, challenges such as the time-intensive nature of manual data extraction, inconsistent reporting practices, and limited community contributions to the database have highlighted the need for more automated and standardized approaches. Emerging tools, such as large language models (LLMs), have shown promise in addressing these challenges by automating data extraction from scientific literature. Based on the database several metastudies were carried out the essence of a few from within the project we highlight in this report.

In this context, we also emphasize the importance of standardized testing protocols, such as ISOS procedures, to ensure the reliability and comparability of aging and stability data. Metadata studies conducted on the Perovskite Database revealed critical insights into device performance, stability, and material optimization, underscoring the need for comprehensive and high-quality datasets to support machine learning (ML) applications and further innovation. In collaboration with the FAIRmat initiative, we developed research data management platforms, based on NOMAD (Oasis), to facilitate the secure and efficient sharing of (experimental) data. These platforms enable researchers to adhere to FAIR principles while maintaining control over sensitive information. The adoption of these tools by research institutions worldwide marks a significant step toward creating a unified and collaborative data-sharing ecosystem for perovskite PV research.

We conclude with actionable recommendations to improve data quality, promote standardized testing, and encourage the adoption of research data management platforms. These efforts aim to accelerate the development and commercialization of perovskite PV technology, ensuring its long-term stability, scalability, and efficiency.



1. INTRODUCTION

The Perovskite Database Project (www.perovskitedatabase.com) was launched as an open-access data initiative that compiles comprehensive metadata and key performance indicators published in the domain of perovskite solar cells. The project is a first step in a larger mission to set up data infrastructure specifically for the photovoltaic R&D community to enable the dissemination and sharing of research data based on FAIR data principles (FAIR standing for Findable, Accessible, Interoperable, and Reusable).

In its initial phase, the project initiated the systematic extraction of research data published in the scientific peer-reviewed literature, which amounted, in 2020, already to 16,000 publications. The data extraction was conducted by experts in the field of perovskite solar cells, mostly Ph.D. students and postdocs, who extracted key meta data and solar cell performance metrics based on a standardized extraction protocol.

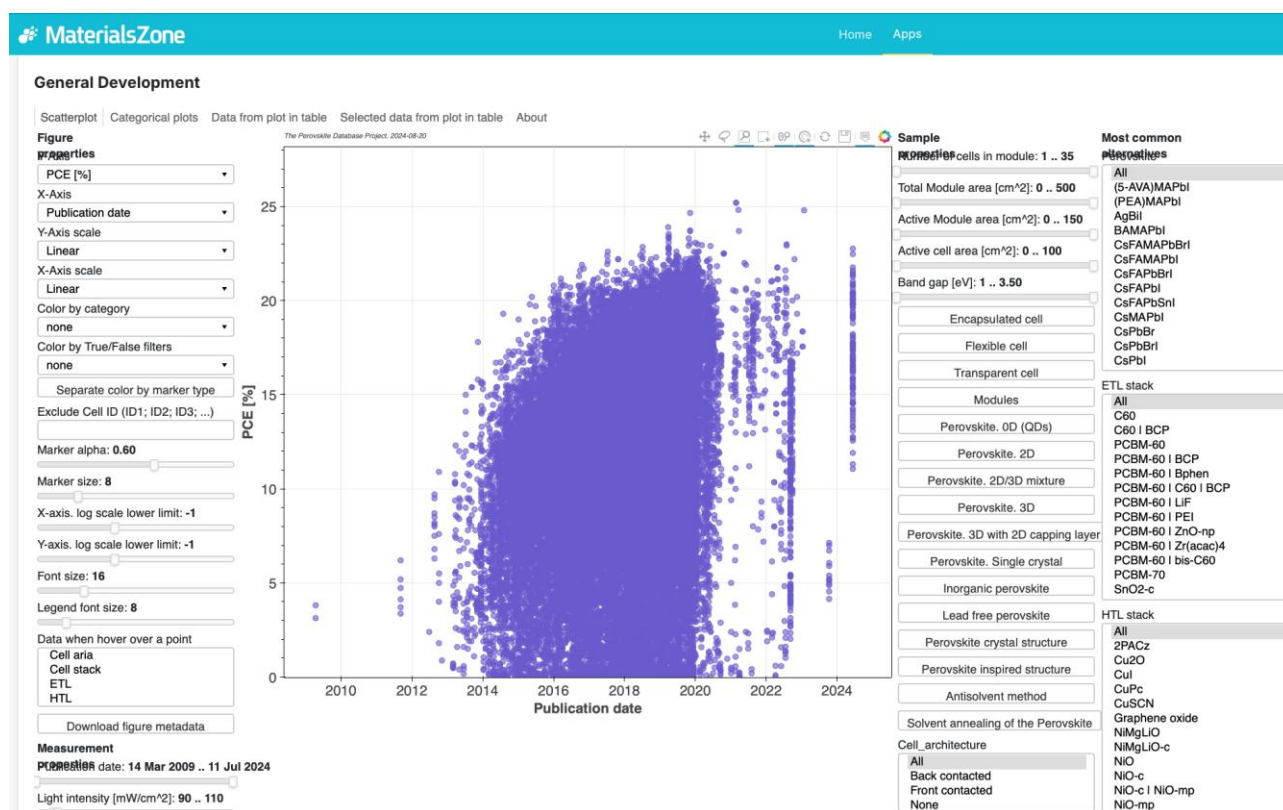


Figure 1: Screenshot of webpage of the App visualizing the complete dataset available on MaterialsZone. **All data collected in the perovskite database that can be viewed, filtered, plotted, modified and downloaded from this web-based source.** Access to the database requires sign-up: Please follow instructions on how to access this dataset on:

www.perovskitedatabase.com



The dataset was launched in the spring of 2022 enabled by the start-up MaterialsZone. Instructions how to access this dataset can be found either on www.perovskitedatabase.com (sign up is required) or following the instructions provided. The launch of the dataset was accompanied with the publication of a peer-reviewed article in Nature Energy, that described the approach and methodology of data extraction, introduced a first ontology for perovskite solar cells, and provided some insight into the evolution of the perovskite photovoltaics (PV) research field captured by the statistical analysis and plots.¹ This initial publication has now been cited more than 230 times and received a lot of attention by the wider research community. About 3 years after the publication and the launch of the database, these are the most important lessons we learned from the project:

- The manual extraction of information from published papers is time-consuming and the accuracy and quality of the dataset strongly depends on the individual extracting the information.
- Coordinating and collecting input from approximately 80 people is not easy. A systematic collection of statistical data comparing the receipt of the extracted data with respect to the set deadline would have been an interesting study for a social or behavioural scientist.
- Despite the implementation of the function/ability for anyone to upload new data to the database, data uploads since the launch of the database are rare; there are no real incitements to motivate researchers to upload their data.
- Recent uploads were the result of a deliberate effort to update the database as part of a review paper on higher bandgap perovskite solar cells (see section 2).
- Emerging tools like large language models (LLMs) provide a viable and automatable strategy to ingest new data published in the scientific literature to the database (see more about this in section 4)
- Ideally, we will establish research data management infrastructures/platforms that enable research scientist to disseminate key performance indicators as well as entire datasets to their research communities enabling also non-published research data to be made available to the community (see more on this in section 5).



2. METADATA STUDIES BASED ON THE PEROVSKITE DATABASE

The dataset first compiled and launched as the Perovskite Database has been utilized by many researchers world-wide. Initially, the dataset provided a valuable source to enable easy and fast visualization of different “state-of-the-art” of different aspect of perovskite PV research. As an example, Figure 2 illustrates a graphic compiled based on the dataset available in the perovskite database to illustrate the historic evolution of reported solar cell performance, the power conversion efficiency (PCE) as a function of absorber bandgap to illustrate the bandgap tunability relevant for multi-junction solar cells, the statistics on published device performance data captured in the T80 value (time to reach 80% of initial performance) and the device performance as a function of active solar cell area illustrating the state-of-the-art in scaling the device technology to larger areas. As perceivable in the data shown below, the data density got very slim after the finalization of the initial data collection campaign ending in mid 2020.

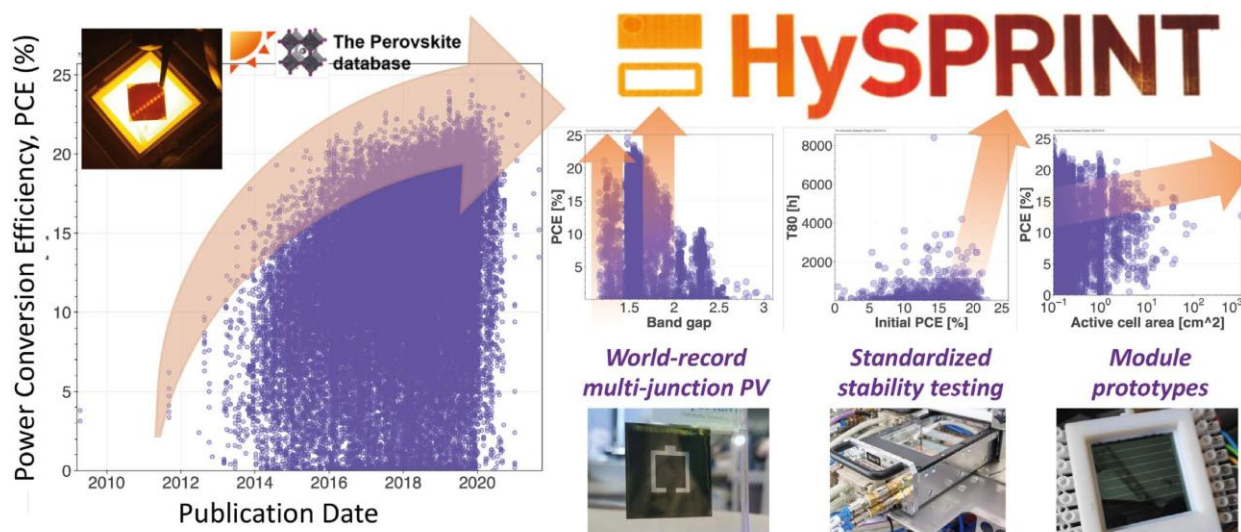


Figure 2: Illustration of the different research goals to bring perovskite PV technology to higher technology readiness level: efficient multi-junction solar cells, stability, and scalability, represented by the three teams of Steve Albrecht, Antonio Abate, and Eva Unger, respectively, in the HySPRINT Innovation Lab. Reproduced with permission from reference ²

The next sub-section will present some examples of meta-data studies carried out on data collected in the Perovskite Database.

2.1 Machine Learning on Perovskite PV Stability Data

Graniero et al.³ attempted to carry out machine learning on the dataset related to stability data collected in the perovskite database. These attempts were of limited success due to data quality issues, particularly missing values and inconsistent reporting. The study highlighted the importance of making more experimental research data directly available instead of reducing the information content to single metrics (e.g., T80) extracted from the actual measurement data. The study also emphasized the issues with using data sets curated to facilitated publication of “progress” in the scientific peer-reviewed literature as the bias towards the best-performance devices does not provide a good basis to train algorithms.

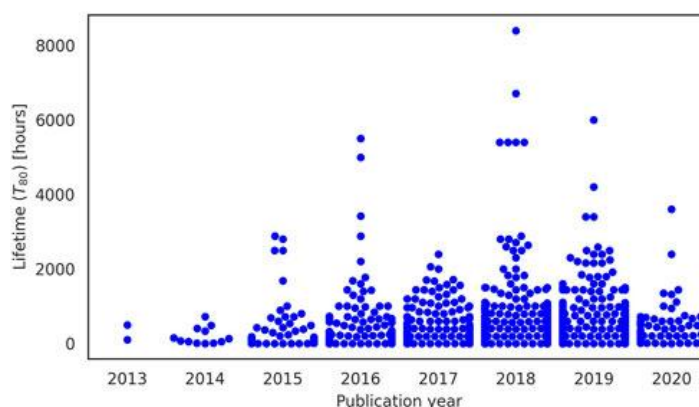


Figure 3: Lifetime T_{80} in hours of every perovskite solar device published between 2013 and February 2020 (for which T_{80} was reported in the Perovskite Database). The vast majority of devices reported, 1,790 out of 1,834, have a short lifetime of at most 2,000 h. Reproduced with permission from reference ³

In conclusion, the study also reiterated the need of the adherence to standardized testing conditions to enable comparison of device degradation data, such as the ISOS protocols⁴ and provided recommendations to a) improve data quality by collecting more complete and standardized aging experimental data, b) develop a universal, fair figure of merit (FOM) for device stability to enhance the reliability and comparability of ML analyses and c) address data gaps and inconsistencies in future studies to enable more robust ML applications.

This study illustrates the limitations in the reuse and statistical analysis of data when the actual experimental dataset is reduced to single metrics such as the T80. We are hoping to solve this shortcoming by setting up research data sharing platforms enabling the dissemination of entire stability test traces (see section 5).



2.2 Metadata Study: Are Check-List for PV data Publications sufficient?

Due to metastabilities and transient effects, determining the power conversion efficiency of halide perovskite solar cells is far from trivial. Based on the data from over 16,000 publications available in the Perovskite Database, Saliba et al. highlighted a consistent discrepancy between two key metrics of device performance: the short circuit current density derived from JV-measurements ($J_{sc,JV}$) and that calculated from integrated external quantum efficiency ($J_{sc,EQE}$). On average, the J_{sc} derived from JV-measurements is 4-5% larger than the J_{sc} determined from integrated EQE data.

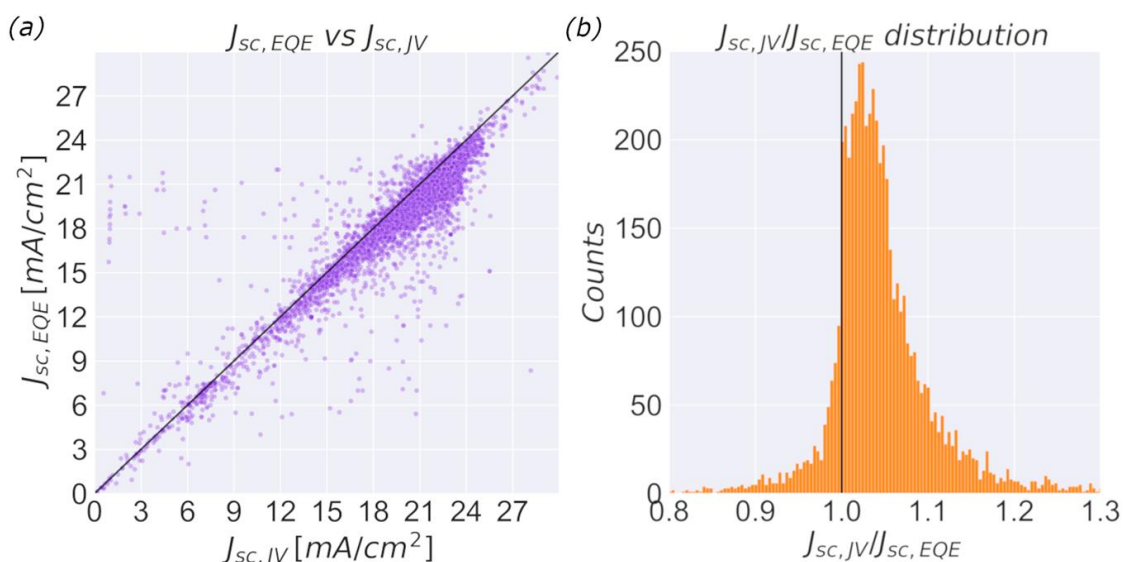


Figure 4: Short Circuit Current Density, J_{sc} , determined from Current-Voltage (JV) as well as integrated External Quantum Efficiency (EQE) measurements reported in the peer-reviewed literature based on data available in The Perovskite Database. Plot reproduced with permission from reference⁵

Potential causes for the discrepancy include differences in measurement conditions, such as light intensity or transient effects during JV sweeps. The findings raise concerns about the reliability of reporting practices for metrics reflecting the performance of perovskite solar cells. As dynamic phenomena during JV-measurements affect the reliability of the performance measurements on perovskite PV, the study recommends greater emphasis on maximum power point tracking (MPP) as a performance measure, as it better reflects operational conditions. The study underscores the importance of standardized testing protocols and highlights the value of large, communal data platforms like the Perovskite Database for uncovering systematic trends and overlooked phenomena in PSCs.



2.3 Metadata Study: Bandgap Tuning of Halide Perovskites

VIPERLAB report D7.2 (non-public) presented a first version of collaborative efforts led by the VIPERLAB partner EPFL to extract new data published since 2020 from the literature to update the Perovskite Database. This was carried out as part of a review paper, published end of 2023⁶, on the topic of bandgap tuning for higher bandgap perovskite absorbers. The statistical analysis of then over 45 000 experimental datapoints provided a clear picture of the general trends in the research community in terms of the most commonly researched materials, illustrated in Figure 5. The histogram (a) illustrates that the most significant research effort focussed on halide perovskites with a bandgap of 1.6 eV, which is the archetypical methylammonium lead iodide, MAPbI₃, which was the material first utilized as absorbers in solar cells. The 2D heat map (b) illustrates how the exploration of halide perovskite semiconductors in a broader bandgap ranged expanded since 2012, with some early examples featuring the higher bandgap MAPbBr₃ and since about 2016, there is also the exploration of lower bandgap Sn/Pb perovskite absorbers perceivable from the plot.

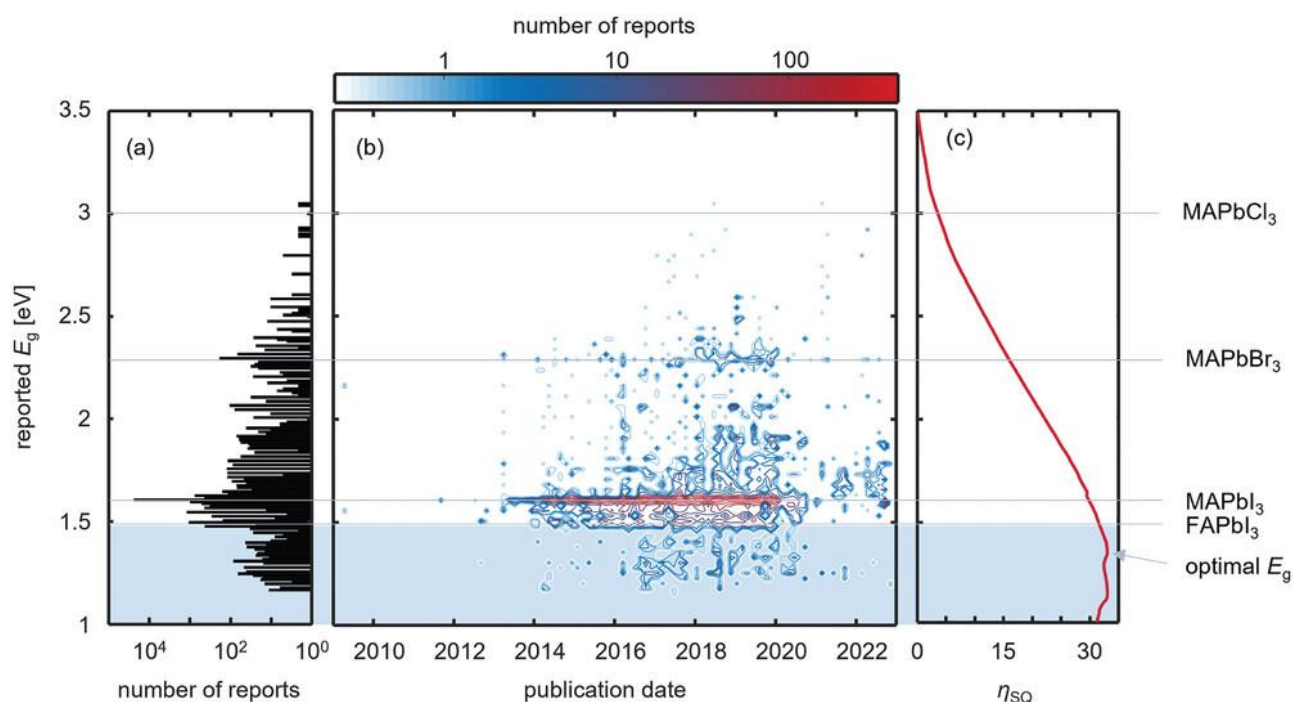


Figure 5: a) Histogram of perovskite E_g distribution. The bin size is 0.02 eV. b) Evolution of the reported E_g in the literature. 1.6 eV perovskites dominate the reports throughout the years, and even in the 2022, most research is conducted on formulations close to pure MAPbI₃ (1.6 eV) and FAPbI₃ (1.5 eV). c) Comparison with the PCE in the SQ-limit shows that the most exploited E_g region is above the optimal E_g region for single-junction solar cells. Reproduced from reference ⁶



In comparison, the exploration of materials in other bandgap ranges were limited. In particular for triple junction solar cells, top cell absorber with a bandgap of around 2 eV will become of high strategic relevance. In this bandgap range, there is very limited data available, illustrated in Figure 6, indicating limited research activities, and the current generation of materials exhibit the proclivity to phase-segregate under illumination leading to severe performance losses, as first discussed by Hoke et al.⁷ The findings of the meta-data analysis highlight intrinsic and optimization-related challenges, such as suboptimal band alignment, poor material quality, and chemical instability, which impact device performance.

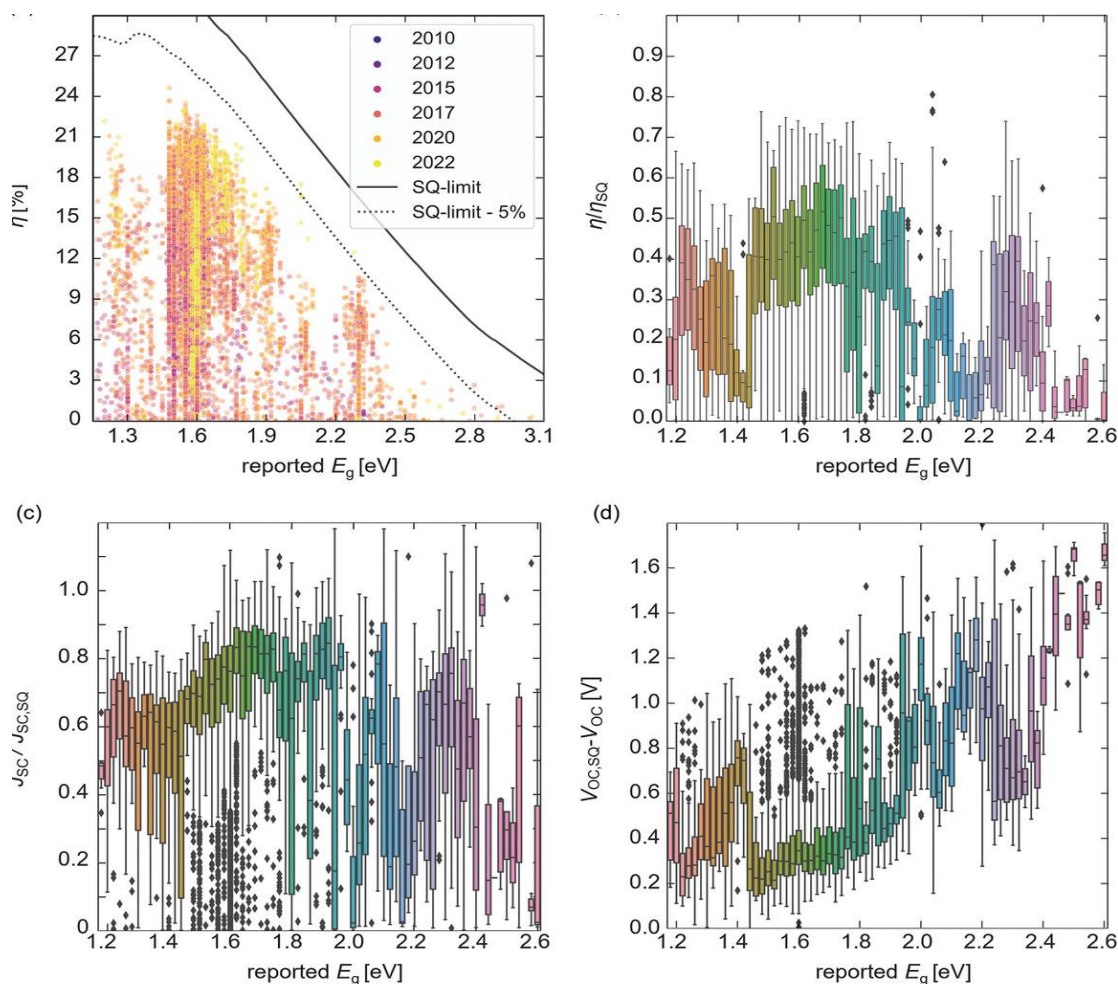


Figure 6: a) Device efficiency as a function of the reported perovskite E_g for all cells in The Perovskite Database together with the SQ-limit (solid) and the SQ-limit minus 5% (broken). The colour of the dot represents the publication date. Box plots of (b) the cell efficiency vs. perovskite E_g , (c) the short circuit current, J_{SC} , and (d) the VOC loss, that is, V_{OC} , $SQ - V_{OC}$, as a function of the E_g . Both the PCE and J_{SC} are normalised with the values given by the SQ-limit, that is, η/η_{SQ} and $J_{SC}/J_{SC,SQ}$. The bin size is 0.02 eV, and the leftmost bin is at 1.18 eV. The end of the boxes represents the 25 and the 75 percentiles. The whiskers are placed at an interquartile

range of 1.5, which means that for a normal distributed dataset, 99.3 % of points should be within that range. Reproduced from reference ⁶

3. THE PEROVSKITE (LITERATURE) DATABASE IN NOMAD

As part of the activities of VIPERLAB, we carried out substantial efforts to develop a research data dissemination and sharing platform in collaboration with the German National Research Data (NFDI) Project FAIRmat (<https://www.fairmat-nfdi.eu/fairmat/>). Visualization tools are evidently slightly different but this is in principle the same dataset as initially collected and uploaded to MaterialsZone.

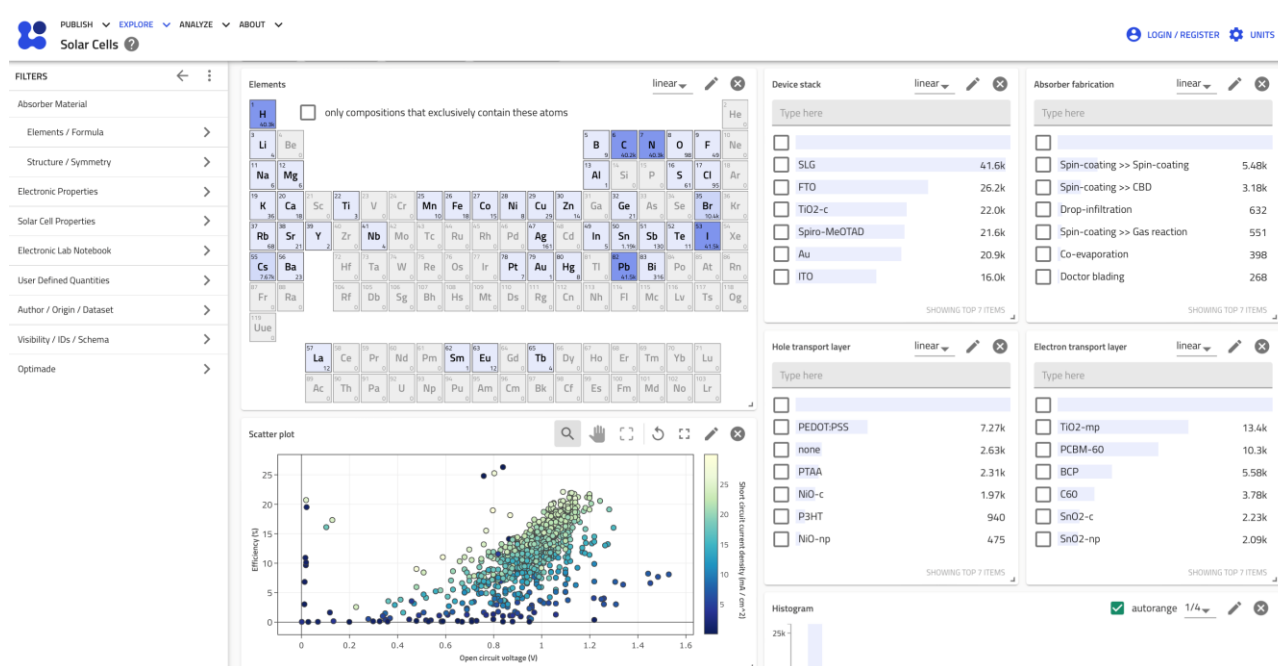


Figure 7: Screenshot of the Perovskite Database dataset on the NOMAD research data management platform facilitated in collaboration with the FAIRmat project.

4. FIRST RESULTS FROM LLM-BASED DATA MINING

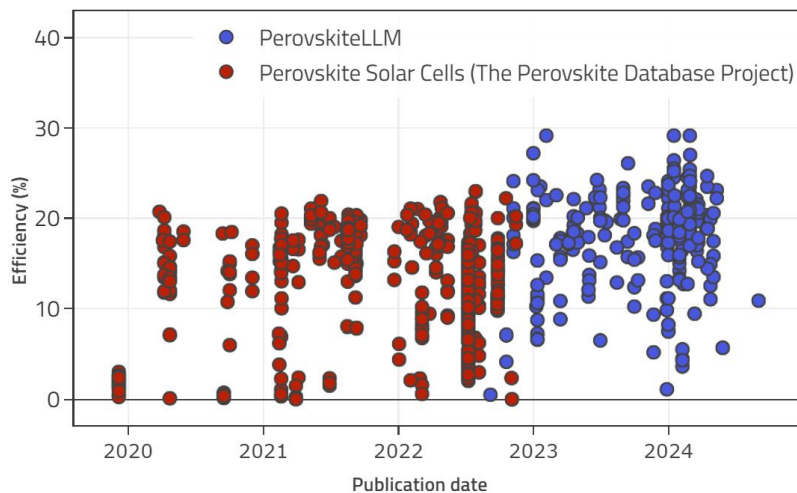


Figure 8: First results from LLM-based literature data extraction using the dataset from the Perovskite Database Project as a training dataset.

The lack of data entries after the initial campaign points to a critical shortcoming in the approach. If the datasets are not being entered by the community, alternative approaches have to be found. We discussed several options including consulting with editors from some of the higher-ranking journals on whether providing the metadata alongside the submission must be made compulsory, yet the feedback was inconclusive. With the emergence of the extremely powerful new generations of LLMs an alternative procedure could be realized, where a LLM is fed with the existing database as a learning set and based thereon learn to read out new publications automatically. Reaching beyond VIPERLAB first attempts were carried out, with the results shown in Figure 8. We're currently evaluating the accuracy of the initially extracted data.

5. RESEARCH DATA MANAGEMENT IN NOMAD

The NOMAD platform offers extensive capabilities for the systematic management of research data. At the HySPRINT lab of HZB, we are currently customizing the capabilities of the NOMAD research data management platform to suit the particular workflow of our research laboratory. For this, we have set up a so-called NOMAD Oasis, which is a local installation of NOMAD that users or user groups can tailor to their specific needs. The local installation of NOMAD allows users to make use of any functionality available by the open



NOMAD platform while keeping all research data protected from unauthorized access or use. This system ensures the researchers full control over their research data while providing a seamless pathway to share information selectively with collaborators or disseminate openly to the wider research community. NOMAD offers hence the opportunity to research data dissemination fully adhering to FAIR data principles while also enabling research communities to create research data management platforms tailored to their specific need to facilitate collaboration and sharing. The activities have been pursued in close collaboration with other PV researchers working in the domain of perovskite photovoltaics in other Helmholtz Centers in Germany, namely the KIT and Forschungszentrum Jülich with the Helmholtz Institute Erlangen-Nürnberg, illustrated in Figure 9.

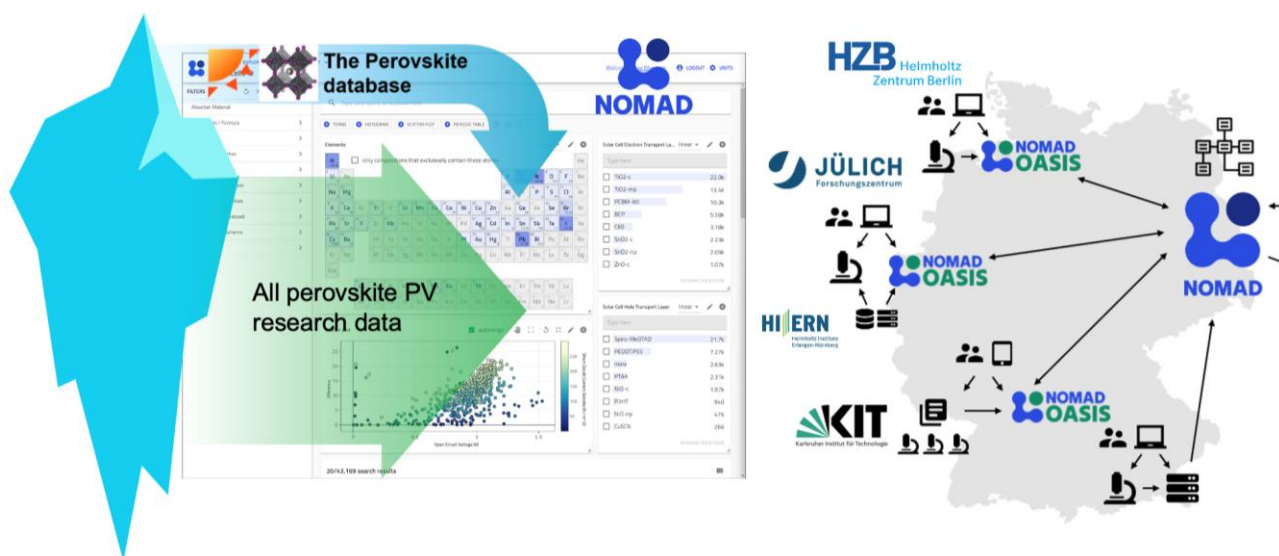


Figure 9: (left) Illustration of the migration of the “Perovskite Literature Database” to NOMAD as well as the effort to build a research data management platform underneath, that allows the handling of all relevant research data generated in the development of solar cells. (right) Illustration of the collaborative effort between Helmholtz Centers in Germany creating common research data sharing platform through NOMAD and local NOMAD OASIS installations.

5.1 Example: Perovskite PV research data from the “Slot dye coating baseline” at HZB

To provide a better sense of the current workflow implemented in the HySPRINT laboratory of HZB to handle research data generated for perovskite solar cells, the following figure illustrates the main principles. For the time being, we found that for lab users the collection

of all experimental variables of a specific solar cell batch made in the laboratory was most conveniently handled using customized excel tables. This is to be considered a temporary solution as our experience from the used needs are now being utilized to further customize the graphical user interface of NOMAD to reflect the workflow of sample manufacturing in our laboratory. Switching to all data of the so-called “baseline” for slot-die coated perovskite solar cells being handled in NOMAD is creating a lot of added value for our lab users as it becomes so much more convenient to track the evolution of solar cell batches, identify common challenges, transparent exchange of information through common research data management between team members, and direct data visualization.

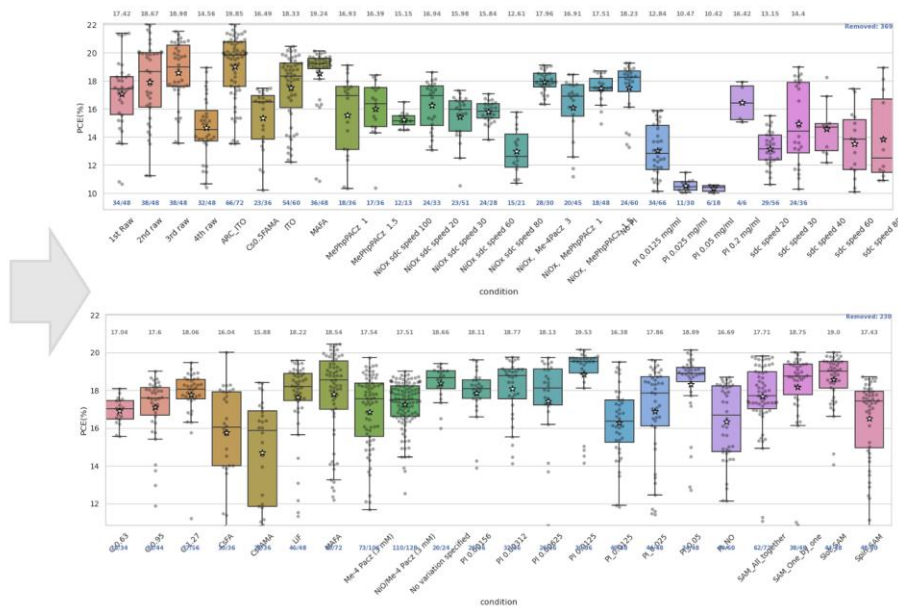
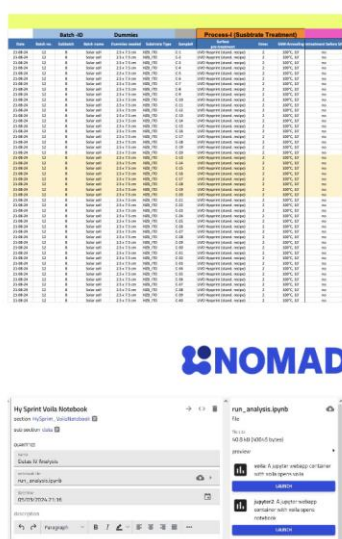


Figure 10: (left) Example of Metadata and experimental data feed-in to the NOMAD OASIS database of HySPRINT-HZB (right) overview of all perovskite solar cells fabricated in the slot-die coating baseline of HySPRINT-HZB.

The data schema customized for our laboratory are now being transferred and further adapted by other research teams world-wide. In Helmholtz, we are in direct exchange with the colleagues from KIT and HI-ERN/Jülich. From the VIPERLAB partners, there are now first initiatives to implement NOMAD OASIS’ at UNITOV and TNO and we will support partners and collaborators strategically to help them set up their own data infrastructure making all of our data schema available to them. This effort will be continued through EU projects which just started where HZB is one of the partner organizations (e.g. SolMates,



PERSEUS, LUMINOSITY). We are also strategically collaborating with PV researchers at NREL to enable them to adopt NOMAD for their own needs.

In collaboration with the FAIRmat project, we are making educational resources available, offer workshops, and will, to the best of our ability, offer on-site assistance to enable other laboratories to join the NOMAD community.

6. CONCLUSION

The VIPERLAB project has successfully addressed critical challenges in the field of perovskite photovoltaics by fostering collaboration, advancing data management practices, and promoting standardized testing protocols. The establishment of the Perovskite Database has provided the research community with a powerful tool for analysing trends and identifying opportunities for innovation. However, we also highlight the limitations of manual data extraction and the need for more automated solutions, such as large language models, to ensure the sustainability and growth of the database.

The adoption of FAIR data principles and the development of platforms like NOMAD and/or the local version NOMAD Oasis represent a significant leap forward in research data management. These tools not only enhance the accessibility and interoperability of data but also empower researchers to share and analyse information more effectively. By addressing data quality issues, promoting standardized aging procedures, and encouraging the use of advanced data-sharing platforms, VIPERLAB has laid a strong foundation for the future of perovskite PV research, with first integrations in the Helmholtz centers in Germany (KIT HI-ERN, Jülich, HZB), pioneered by HZB, and hopefully laying the groundwork for adoption by other researchers in Europe and worldwide.

As the project concludes, these findings and recommendations will serve as a roadmap for the continued development of perovskite PV technology. By fostering a culture of collaboration, transparency, and innovation, VIPERLAB has positioned the research community to overcome existing challenges and unlock the full potential of perovskite photovoltaics as a sustainable and scalable energy solution.



7. REFERENCES

1. Jacobsson, T. J. *et al.* An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nat. Energy* **7**, 107–115 (2021).
2. Unger, E. HySPRINT: Synergetic Halide Perovskites Material and Device Development for Future Photovoltaics. *Bunsen-Mag. Jahrgang* **24**, 96–98.
3. Graniero, P. *et al.* The challenge of studying perovskite solar cells' stability with machine learning. *Front. Energy Res.* **11**, 1118654 (2023).
4. Khenkin, M. V. *et al.* Consensus statement for stability assessment and reporting for perovskite photovoltaics based on ISOS procedures. *Nat. Energy* **5**, 35–49 (2020).
5. Saliba, M., Unger, E., Etgar, L., Luo, J. & Jacobsson, T. J. A systematic discrepancy between the short circuit current and the integrated quantum efficiency in perovskite solar cells. *Nat. Commun.* **14**, 5445 (2023).
6. Suchan, K. *et al.* Rationalizing Performance Losses of Wide Bandgap Perovskite Solar Cells Evident in Data from the *Perovskite Database*. *Adv. Energy Mater.* **14**, 2303420 (2024).
7. Hoke, E. T. *et al.* Reversible photo-induced trap formation in mixed-halide hybrid perovskites for photovoltaics. *Chem. Sci.* **6**, 613–617 (2015).

