The Perovskite Database Project: A Perspectives on Collective Data Sharing

Eva Unger[1,2], T. Jesper Jacobsson[3]*

1. Young Investigator Group Hybrid Materials Formation and Scaling, Helmholtz-Zentrum Berlin für Materialen und Energie GmbH, Berlin, Germany.
2. Chemical Physics and NanoLund, Lund University, Lund, Sweden
3. Institute of Photoelectronic Thin Film Devices and Technology, Key Laboratory of Photoelectronic Thin Film Devices and Technology of Tianjin, College of Electronic Information and Optical Engineering, Nankai University, Tianjin 300350, China

jacobsson.jesper.work@gmail.com

Main text

Since the early 2010s, halide perovskites have evolved from an academic curiosity into a promising solar cell technology. Record solar cell efficiencies have surpassed 25 %,[1, 2] and if the private sector is to be believed, commercial tandem cells may be just around the corner.[3-5] This incredible progress is the result of hard work, over 19000 publications, and probably over a billion US dollars expenditures in research and development. There are, however, still problems to address. Both the stability and the scalability does for example need to improve for perovskite solar cells to reach beyond niche markets.

To successfully confront the remaining challenges, having a good overview by keeping track of the state of knowledge is of utmost importance. The overwhelming number of perovskite papers published every week makes this increasingly difficult, even for veterans in the field. For novices it is even harder to get a grip of the most important innovations and open research questions. The bias towards publication of successful innovations and lack of formats and incentives to make less successful approaches and the resulting data public is also a problem that obscures the overview and leads to unnecessary repetition of failed experiments.

The struggle to keep up to date and a desire for better collective data management was a trigger behind the *Perovskite Database Project* which recently was launched in Nature Energy.[6] We there developed a data protocol based on the FAIR data principles (*i.e.*, Findability, Accessibility, Interoperability, and Reusability)[7, 8] and initiated a communal effort with the ambition to collect data for every perovskite device described in the literature. If this data were to be consistently formatted, openly available, and presented together with interactive tools for data exploration, this would greatly facilitate a good overview of the state of knowledge while also enabling new insights, better experiment design, and an accelerated journey towards commercialisation. In this Energy-focus we will reflect on the *Perovskite Database Project* in a broader perspective and elaborate what can be expected now that the resources are freely available online. We will also share insights into the organisation of this project valuable for future similar initiatives.

We wanted data, but what data to look for and how to format and store it? Experimental material science is not like genetics where the central data structure is a string composed of four different letters. It is instead characterised by small heterogeneous datasets. The amount of characterisation techniques is vast, the number of materials huge, the relevant key performance indicating metrics are application dependent, and properties are affected by both synthesis conditions and sample history. Each sample thus also requires extensive metadata to be properly described. For now, this complexity means that no unified catch-them-all data

ontology for material science yet is in place. The first main challenge of the project was hence to develop and define a list of relevant and suitable metrics to consistently capture the essence of perovskite solar cells.

Based on domain knowledge about the metrics considered important, about what data that usually is reported (and which often is omitted) and based on what we thought we reasonably could capture, we design a protocol for data extraction. The protocols and the details are available via the original publication and focus on materials, deposition procedures, and key performance metrics of single junction cells (fig. 1).[6] The protocol used in the initial data collection campaign contains about 100 attributes and parameters per device. The refined version of the extraction protocol intended for future use currently captures about 400 parameters per device. No data extraction protocol is without flaws and refining and developing them is a constant work in progress. The long-time goal would be to have a data structure that captures metadata with a precision that would enable someone to, solely based on that protocol, reproduce a device with the exact same performance. That is still far away.

A few advice to similar projects. It is important to quickly have a first version of the data extraction protocol that demonstrates the feasibility and usefulness of the project. Then try it out on real data on a smaller scale and be prepared to do modifications. The messy reality is ruthless in revealing flaws and edge cases unthought of. Then it is time for the larger data collection campaign under which larger updates are cumbersome to organise and therefore best is postponed to after the campaign.

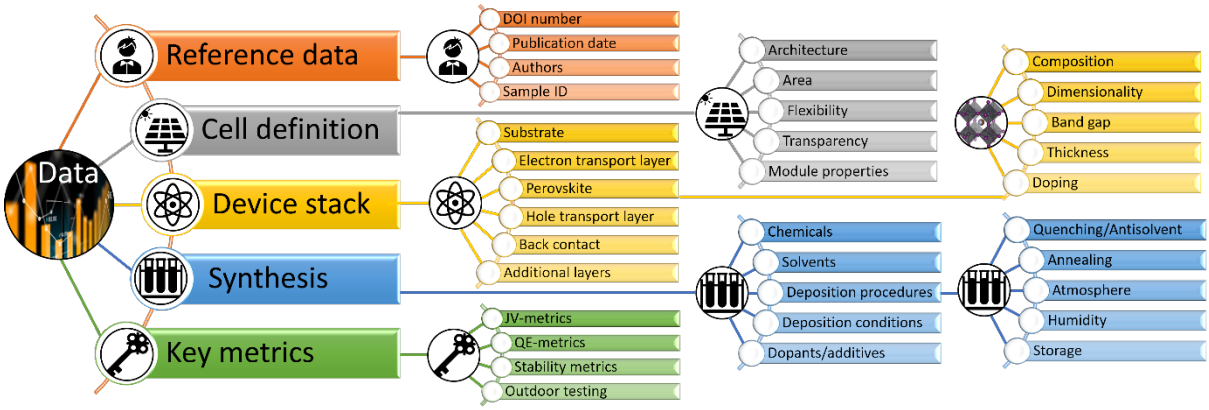

Figure 1. Overview of data categories in the Perovskite Database. Overview of the main categories of metadata, process data, and performance data in the data extraction protocol. JV, current-voltage. QE, quantum efficiency. Figure reproduced from *Nat Energy* **7,** 107–115 (2022).[6]

At the project start in 2019, a search in Web of Science with the search phrase "Perovskite Solar" gave around 12000 entries. When the data extraction campaign ended in 2020, we had gone through close to 16000 papers. All those were not about perovskite solar cells, but most of them were, and around half of them contained original device data.

A first trivial insight was that if we were to succeed in getting all that data, we would need help. Lots of it. That also provided a first sanity checkpoint, *i.e.*, was the idea good enough to convince a sufficient number of people to volunteer their time to do the hard and tedious work of manually data extraction? By utilising combined contact nets and word-of-mouth this went beyond expectation, and we recruited close to 100 volunteers, from master students to professors, that bought into the project's vision.

Distributing volunteer work is a fine balance. Give them too much to do and they may not be able to find the time and drop out. Give them too little and they might feel less invested in the project. We asked every contributor to go through 200 publications and we may have balanced on the border of too much work. Beside undersigned, we estimate that somewhere between 5000 and 10000 volunteer hours have been spent going through papers.

When the data campaign ended, we had extracted data from over 42000 solar cells. This may be a small fraction of all devices ever made, but it represents essentially every device someone has thought is worth the trouble to properly describe in the peer-reviewed literature. It turns out that few publications actually provide the device data described and discussed in their publication in a format that enabled it to be extracted. Very often, device performance metrics are only explicitly included for a few selected devices, *i.e.*, records and references. The devices that provide the statistical justification for the results are unfortunately often only shown as histograms or averages which makes extracting the metrics of the individual devices intractable. By also considering that the reported data probably just is a small fraction of the data that is originally accumulated by individuals in research labs, one can only imagine how much information that has been lost over the years.

With the data collected in The Perovskite Database, a variety of interesting analysis can now be carried out. It is for example possible to extract the evolution of reported power conversion efficiencies, which demonstrate a continuous upwards trend reflecting a collective progress in device engineering (fig. 2). Another option is to explore effects that previously have drowned in the noise from cell-to-cell, batch-to-batch, and lab-to-lab variations, such as if one procedure or material truly is better than another. A few such examples are given in the original publication.

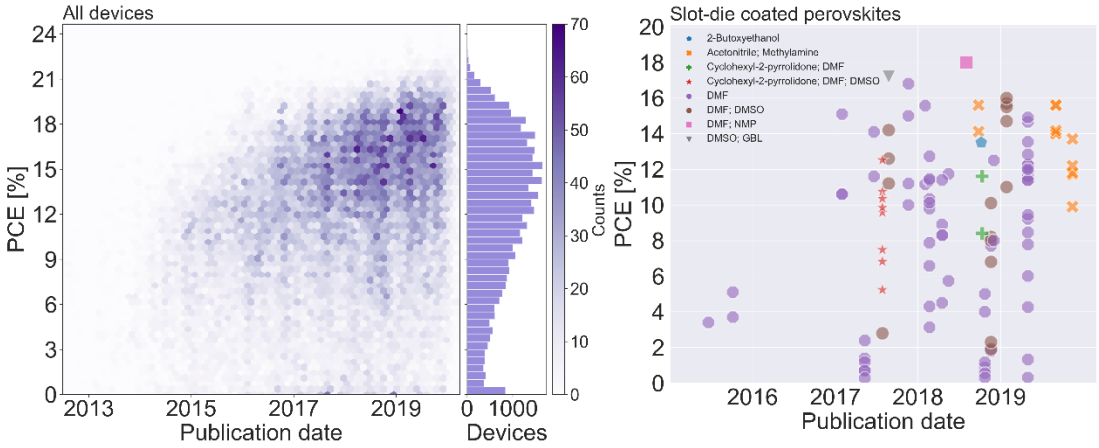

Figure 2. (a) Development of perovskite solar cell efficiencies in the form of a hexbin-plot of PCE measured under standard conditions as a function of the publication date for all devices in the database. Efficiency distribution for all devices is shown to the right. (b) Cell efficiency as a function of the publication date for slot-die-coated perovskites separated by the solvent used for perovskite deposition. DMF, dimethylformamide; DMSO, dimethylsulfoxide; GBL, gamma-butyrolactone; NMP, N-methyl-2-pyrrolidinone. Figures reproduced from *Nat Energy* **7,** 107–115 (2022)[6]

A question we often get is how to ensure that the data is free from errors. The short answer is that errors are unavoidable. The philosophy is that some data is better than no data, and that there is value also in incomplete described devices, even if there are a few errors here and there. After the initial data campaign, we spend a lot of effort correcting errors and inconsistencies, but some will remain. It does for example happen that data in published papers is incorrect, and the database is not able to correct for that. It could also be that data is misinterpreted, which is easily done when papers are ambiguous or confusingly written. Some manual data entry mistakes are also to be expected. We have therefore set up a system where erroneous data points can be updated and we expect some self-correction over time, especially for data points of special interest such as records in subfields. For future data we expect a lower error rate as that primarily will be uploaded by the original authors. We also encourage authors to check their historical data in the database and correct errors they find. Finally, we would like to

encourage the users of the database to not be afraid of the errors that may be, but to embrace them as statistical uncertainty and deal with it as such.

While downloading the latest version of the dataset is easy, it should also be immediately useful and the threshold for data exploration should be as low as possible. We therefore developed interactive graphical tools that can be reached from the project's webpage (https://www.perovskitedatabase.com). Those tools enable exploration of the dataset by providing different options to display, filter, and sorting the data in real time without the need for programming experience. This directly enables insights to be gained and can be used as a quick way to test hypotheses. It also offers the opportunity to facilitate literature research and increase individual research groups visibility by linking all data to its original publications. Making the data accessible in this way is also a strategy to decrease the risk that the project develops into an underutilised and forgotten data dump.

An example of a typical user case is given in fig. 2.b and revolves around slot-die coating. With a simple use of a drop-down menu, all devices made with slot-die coating can be filtered out and plotted against for example publication date to illustrate performance development. With a simple filter, the data can be further narrowed down to only include specific perovskite compositions, and with an additional command the data can be sorted in terms of for example the solvent system used during deposition. Once the filtering is done, a click on a datapoint redirects to the original publication, and with another click the data and the figure can be downloaded and used in whatever way imaginable. This represents a complex literature search that previously would have taken a substantial amount of time and domain knowledge. Now it can be done with a few clicks in less than a minute, thereby really simplifying the kind of overview strived for at the start of the project.

For publication quality graphics, the data would normally be downloaded and plotted in the visualisation program of choice. For many purposes it is, however, possible to directly download and use the figures generated from interactive graphics, and a few examples are given in fig. 3.

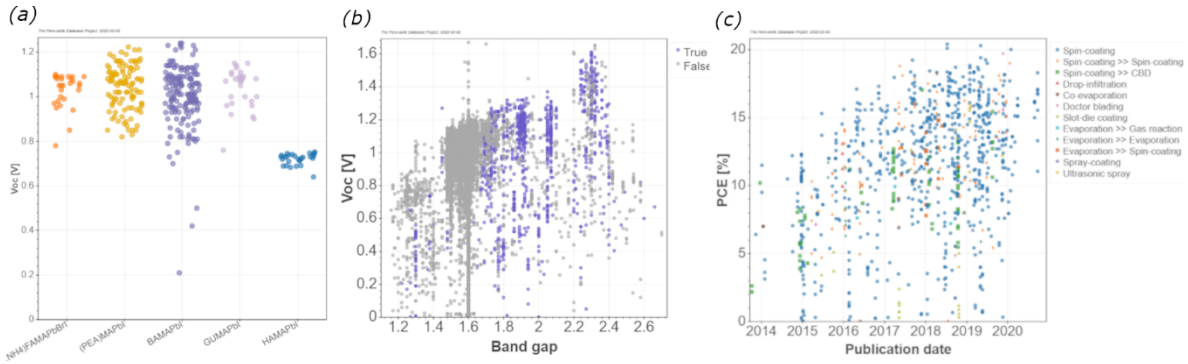

Figure 3. Examples of figures generated within the system without further editing. (a) Swarm plots of the open circuit potential, V_{oc}, for a five 2D perovskites. (b) V_{oc},vs band gap in eV for all devices in the database with inorganic perovskites highlighted in purple. (c) Efficiency as a function of publication data for flexible devices marked by the perovskite deposition procedure.

The release paper discusses a few possible user cases of the interactive graphics and some first insights from analysing the data.[6] Its core value is, however, as an enabler of studies and insights to come and as a secondary dissemination platform to keep track of the progress in this dynamic research area. During the coming year we can expect follow up papers revolving around for example band gap tuning, stability, scalability, module development, tandem cells, and measurement protocol standards. All of which will utilise the perovskite database as an important source of insight. To promote this further we have made it possible to advertise

ongoing initiatives on the webpage to invite researchers in the community to collaborate and contribute

This type of dataset is like an invitation to data scientists to see what machine learning, ML, can provide. At a first glance this could be challenging. Even though the dataset represents all data we could extract from the literature, it is still a small dataset in the context of machine learning. It is also heterogeneous, all samples are not described with the same level of detail, and we know there to be hidden variables obscuring comparisons of results from various labs. Nevertheless, we can expect the first machine learning papers based on this dataset to come out within a year. It will be highly interesting to follow what insights those methods will provide, but also to see how they map out the limitations of the statistical knowledge hidden in the dataset. Discussions around the ML-compatibility of the current dataset will also enable the specification of clearer guidelines and a unified data ontology to expand the relevant meta-dataspace.

Another topic worth exploring is the staggering multitude of device configurations capable of generating a functional perovskite device. At the time of writing, the dataset contains information on solar cells with 194 different substrates, 1957 hole-transport layers, 1443 electron-transport layers, 288 back contact configurations, and over 400 perovskite families (*i.e.*, specific combinations of *A*, *B*, and *C*-site ions in the perovskite ABC_3 structure without taking the ion fractions into account). This represents an astronomically large combinatorial space of possible solar cell stack configurations with already tested layers ($\approx 6*10^{13}$). Obviously, only a vanishingly small fraction of those has been reported, *i.e.,* ≈5500. Nevertheless, over 1000 device configurations have been used to produce cells with an efficiency above 18 %. A small glimpse of the experimental distribution of the combinatorial device stack possibilities is given in the Sankey diagram in fig. 4. One of the things we see there is that despite the multitude of explored device configurations, the stack sequence of the break-through papers in 2012 are still the most explored ones. This is one example where the aggregated data opens for new types of questions, like why does the utilisation of the experimental device stack space looks like it does? What is due to chemistry and device physics, and what is due to chance, serendipity, history, and sociology? We expect to return to this in future works.

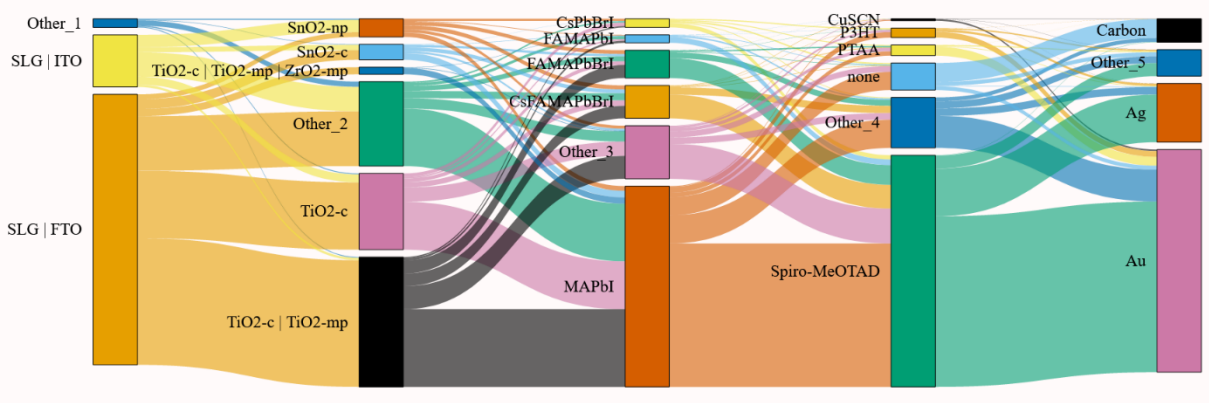

Figure 4. A Sankey diagram of stack configuration for perovskite cells with nip-architecture. The node layers from left to right are: the substrate, the electron transport layer, the perovskite, the hole transport layer, and finally the back contact. Abbreviations in the figure: c for compact, np for nanoparticles, and mp for mesoporous. Vertical bars *i.e.* "|" represent a boundary between two layers in a stack.

While Perovskite PV was the domain of this project, the core themes were Open Science, data sharing and dissemination platforms, big data, and FAIR data principles. Those are themes growing in popularity, in part due to the global tech giants showing how to find both value and knowledge in large datasets. There are now both push and pull factors driving research in that

direction. In the research community, there is an increasing demand from funding agencies, governments, and publishers for making research data publicly available to maximize the return on investment. Developments in data science, machine learning and artificial intelligence also provide a strong pull by the promise of new insights if just more structured data was available. Shared datasets seem to inevitably be the direction we are heading towards.

With that said, most experimental groups do not yet have adequate routines for sharing their data. In part, this is due to lack of good, user friendly, trustable, and widely recognised data sharing platforms that cater to the needs of specific research communities. In part it is due to old habits and an incentive structure that primarily rewards reporting state-of-the-art device performance in the form of papers and that puts little value in disseminating data for all devices made. There is also the unavoidable and undeniable fact that sharing data comes at a cost. It requires thought, time, and effort to format data to agreed standards. It is also necessary to associate the raw data with extensive metadata for making the data valuable and adhering to FAIR data principles. That can be a daunting task for labs where the competence and the digital infrastructure for automating the process not yet is in place. If the researcher does not perceive that the immediate benefit outweighs the effort, there is a high risk that much data never is shared, which essentially mirrors the tragedy of the commons.

For a community that believes in the value of peer-review, it should reasonably be possible to foster a culture of sharing data for the common good. It will, however, require a push from both publishers and funding agencies, that the pull factors in terms of direct perceived utility get stronger, and that the barrier for data sharing decreases. The *Perovskite Database Project* is an attempt to facilitate data sharing by providing clear guidelines and a platform for simple data uploading. It also aims at increasing the pull factors by improving the utility of the uploaded data and increasing the visibility of an individual's research output. This is done with the help of the interactive graphics that not only provide a simple means for data exploration, but also makes the original papers more visible by linking every data point to the original report.

The publication of the release article is meant to be the beginning and our intention has always been to open the platform for community-driven expansion and improvements. This is now where you, dear reader of this, can and shall feel invited to become involved and here are some simple ideas, how:

- Visit the project website: www.perovskitedatabase.com, sign up and get acquainted with the data and tools available
- Check, whether all your published device data is already online; if yes, please check whether the data was entered correctly and please supply any additional information that was not yet included for your dataset; if not, download the data extraction protocol, fill in (also including statistical data if you please) and upload to the database
- Every time you read a perovskite PV paper not yet included in the database, feel free to extract the relevant information and upload to the database or send the authors of the work an e-mail inviting them to do so (thus helping to spreading the word)
- Consider making some or all of your historic device data available that has been disseminated through peer-reviewed publication

Our hope is that the perovskite community will embrace this as a valuable resource and upload not only a few selected high performing devices, which essentially is what is extractable from most papers, but also the full datasets for all devices made, both good and bad. If that happens, we may in a few years from now have data for millions of devices which will open for even more interesting possibilities to extract value and insights from the data.

We will not initiate a new data extraction campaign but believe it to be more beneficial at this point to involve the entire research community in updating and utilizing the database. We will, however, do our best to continue developing the Perovskite Database, which also includes

linking or integrating the database and platform with other data infrastructures. To enable the community to profit from the effort so far, the data, the extraction protocols, and the source code for data formatting, data editing, the database structure, and for the interactive graphics is shared on GitHub. That enables anyone with a good idea to build further from what we have done, or to use it as inspiration to build something better, either for perovskites or for other material systems and applications.

With projects like this, one never knows where they will lead, and there will be reasons to return in a few years' time to evaluate the progress and see what can be learned. Will the project succeed in its ambition to evolve into a standard for data dissemination in the perovskite field, or will it be superseded by something different and better building on top of it? Regardless, we are convinced that the general direction of sharing data in this way is an inevitable part of the future and something that will accelerate the development of functional devices.

Data availability

The Perovskite Database Project was first described in Nature Energy (https://doi.org/10.1038/s41560-021-00941-3). The associated resources, like the full dataset, can be reach from the project's webpage https://www.perovskitedatabase.com. The dataset will periodically also be mirrored/uploaded and made available via Zenodo (DOI: 10.5281/zenodo.5837035, https://zenodo.org/record/5837035#.YfIIFv5ByUk)

Code availability

The code base for the perovskite database project is found in the Github repository, https://github.com/Jesperkemist/perovskitedatabase.

Acknowledgements

Ministry of science and technology in China via the National Key Research and Development Program of China (Grant No. 2021YFF0500500). The GRECO project funded by the European Union's Horizon 2020 research and innovation programme (grant no. 787289). This project will now be continued in the framework of the VIPERLAB project.

Competing interests.

The author declares no competing interests.

References

1. M. A. Green, E. D. Dunlop, J. Hohl-Ebinger, M. Yoshita, N. Kopidakis and X. Hao, *Progress in Photovoltaics: Research and Applications*, 2021, **29**, 657-667.
2. H. Min, D. Y. Lee, J. Kim, G. Kim, K. S. Lee, J. Kim, M. J. Paik, Y. K. Kim, K. S. Kim and M. G. Kim, *Nature*, 2021, **598**, 444-450.
3. J. Gifford, *Journal*, 2021, **July 23**.
4. M. ASA, *Journal*, 2021, **December 16**.
5. J. F. Weaver, *Journal*, 2021, **July 2**.
6. T. J. Jacobsson, A. Hultqvist, A. García-Fernández, A. Anand, A. Al-Ashouri, A. Hagfeldt, A. Crovetto, A. Abate, A. G. Ricciardulli and A. Vijayan, *Nature Energy*, 2021, 1-9.
7. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos and P. E. Bourne, *Scientific data*, 2016, **3**, 1-9.
8. C. Draxl and M. Scheffler, *Mrs Bulletin*, 2018, **43**, 676-682.