nature energy

RESOURCE

https://doi.org/10.1038/s41560-021-00941-3

Check for updates

OPEN

An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles

T. Jesper Jacobsson[1,2]✉, Adam Hultqvist[3], Alberto García-Fernández[4], Aman Anand[5,6], Amran Al-Ashouri[7], Anders Hagfeldt[8], Andrea Crovetto[9], Antonio Abate[10], Antonio Gaetano Ricciardulli[11], Anuja Vijayan[2], Ashish Kulkarni[12], Assaf Y. Anderson[13], Barbara Primera Darwich[14], Bowen Yang[8], Brendan L. Coles[15], Carlo A. R. Perini[16], Carolin Rehermann[1], Daniel Ramirez[17], David Fairen-Jimenez[18], Diego Di Girolamo[19,20], Donglin Jia[21], Elena Avila[18], Emilio J. Juarez-Perez[22], Fanny Baumann[8,23], Florian Mathies[1], G. S. Anaya González[24], Gerrit Boschloo[2], Giuseppe Nasti[19], Gopinath Paramasivam[1,25], Guillermo Martínez-Denegri[26], Hampus Näsström[1], Hannes Michaels[2], Hans Köbler[10], Hua Wu[2], Iacopo Benesperi[2], M. Ibrahim Dar[27], Ilknur Bayrak Pehlivan[28], Isaac E. Gould[29,30], Jacob N. Vagott[16], Janardan Dagar[1], Jeff Kettle[31], Jie Yang[32], Jinzhao Li[1], Joel A. Smith[33,34], Jorge Pascual[10], Jose J. Jerónimo-Rendón[35], Juan Felipe Montoya[17], Juan-Pablo Correa-Baena[16], Junming Qiu[21], Junxin Wang[28,36], Kári Sveinbjörnsson[7], Katrin Hirselandt[1], Krishanu Dey[27], Kyle Frohna[27], Lena Mathies[37], Luigi A. Castriotta[38], Mahmoud. H. Aldamasy[10,39], Manuel Vasquez-Montoya[1,17], Marco A. Ruiz-Preciado[40,41], Marion A. Flatken[10], Mark V. Khenkin[42], Max Grischek[7,43], Mayank Kedia[12,35], Michael Saliba[12,35], Miguel Anaya[27,44], Misha Veldhoen[13], Neha Arora[27], Oleksandra Shargaieva[1], Oliver Maus[1], Onkar S. Game[33], Ori Yudilevich[13], Paul Fassl[40,41], Qisen Zhou[21], Rafael Betancur[17], Rahim Munir[1], Rahul Patidar[15], Samuel D. Stranks[27,44], Shahidul Alam[5,6,45], Shaoni Kar[46], Thomas Unold[9], Tobias Abzieher[41], Tomas Edvinsson[28], Tudur Wyn David[47], Ulrich W. Paetzold[40,41], Waqas Zia[12,35], Weifei Fu[11], Weiwei Zuo[35], Vincent R. F. Schröder[48,49], Wolfgang Tress[50], Xiaoliang Zhang[21], Yu-Hsien Chiang[27], Zafar Iqbal[10], Zhiqiang Xie[51] and Eva Unger[1,23]✉

Large datasets are now ubiquitous as technology enables higher-throughput experiments, but rarely can a research field truly benefit from the research data generated due to inconsistent formatting, undocumented storage or improper dissemination. Here we extract all the meaningful device data from peer-reviewed papers on metal-halide perovskite solar cells published so far and make them available in a database. We collect data from over 42,400 photovoltaic devices with up to 100 parameters per device. We then develop open-source and accessible procedures to analyse the data, providing examples of insights that can be gleaned from the analysis of a large dataset. The database, graphics and analysis tools are made available to the community and will continue to evolve as an open-source initiative. This approach of extensively capturing the progress of an entire field, including sorting, interactive exploration and graphical representation of the data, will be applicable to many fields in materials science, engineering and biosciences.

The halide perovskites have for the last few years been the brightest shining stars on the sky of emerging solar cell materials. They have shown great potential in optoelectronic applications such as tandem solar cells[1–5], LEDs[6,7], lasers[8], photodetectors[9,10], X-ray detectors[11] and for single-junction solar cells the record certified power conversion efficiency (PCE) has reached above 25% (ref. [12]). The halide perovskite semiconductors thus represent a material class with considerable technological relevance where rapid development is occurring. There are, however, remaining problems related to, for example, stability[13–15], scalability[16–19] and

A full list of affiliations appears at the end of the paper.

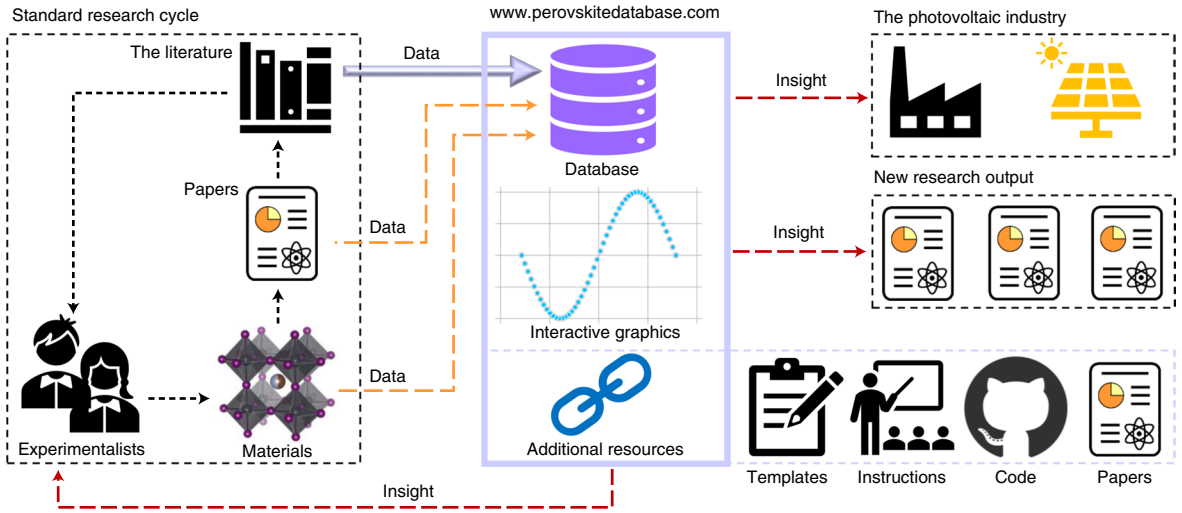

Fig. 1 | Expanding the standard research cycle in experimental material science. An illustration of the standard research cycle and how the Perovskite Database Project can expand it by providing an open database, interactive visualization tools, protocols and a metadata ontology for reporting device data, open-source code for data analysis and so on. Solid data lines refer to data from published papers treated in this project. Dashed data lines refer to raw data from experimentation and analysed full datasets that are natural extensions to be included later. The dashed 'insight' lines represent the use of the expanded research cycle.

reliability[20]; the best material combinations and manufacturing processes are open questions[21,22], and key standards and metrics are still under discussion[23].

In the normal research cycle, researchers read papers, formulate hypotheses, generate data in the laboratory and publish new papers (Fig. 1). With historic data and insights scattered over an inaccessibly large number of papers, this process is not as efficient as it could be. At the time of writing, the keyword 'perovskite solar' does for example find over 19,000 papers in the Web of Science, making it essentially impossible to keep up to date with the literature. The perovskite field could thus be said to have a data management problem at an aggregated level.

Data have always been the foundation of empirical science, but with modern algorithms and artificial intelligence, entirely new opportunities emerge when data are collected in sufficiently large quantities and in a cohesive manner. Big data has become the lifeblood of the tech giants of Silicon Valley, the fuel for artificial intelligence and a cornerstone for the next industrial revolution[24]. The field of materials science is in no way oblivious to this development, and several data initiatives have been initiated, for example the Materials Project[25], Aflow[26], NOMAD[27], the Crystallography Open Database[28], the emerging photovoltaic initiative[29] and the inorganic crystal structure database[30], to mention a few. Despite these efforts, much of the experimental materials science is still struggling to make better use of the data generated[31], and notably so in applied fields where materials are often evaluated primarily by their performance in devices.

A concept of increasing importance is the FAIR data principles, that is, data should be findable, accessible, interoperable and reusable[32,33]. Adhering to those principles can accelerate the development and increase the return on investment as it enables cross-analysis between datasets, data reuse, as well as simplifying the use of artificial intelligence and machine learning. There is also an increased demand from government, funding agencies and journals to disseminate the underlying data accordingly. However, most laboratories are not able to adhere to the FAIR data principles, especially in the applied science fields. There are concurrent reasons behind this, including the lack of suitable data dissemination platforms. However, the largest hurdle is the diversity and complexity of the datasets involved. For instance, sample properties are often influenced by the sample history. Furthermore, they are characterized using a large number of experimental techniques, which vary across disciplines. These small disconnected and heterogeneous datasets also require a substantial amount of metadata to be of use.

In this project, henceforth referred to as the Perovskite Database Project, we have initiated a communal bottom-up effort to transform perovskite research data management. The Perovskite Database Project aims to expand the normal research cycle by collecting all perovskite solar cell data, both past and future, in one place. Apart from making all historical data accessible and providing means to upload new experimental data, interactive graphical data visualization tools have been implemented that enable simple and interactive exploration, analysis and filtering (Fig. 1). This platform will give both academic researchers and the industry an accessible overview of what has been done before, and thereby help in finding relevant knowledge gaps and formulating new scientific questions with the hope of generating new insights, designing better experiments, avoiding known dead ends and accelerating the rate of development. The key goals of the project are to: collect all perovskite solar cell data ever published in one open-access database; develop free interactive web-based tools for simple and interactive exploration, analysis, filtering and visualization of the data; develop procedures and protocols to simplify dissemination and collection of new perovskite data according to the FAIR data principles; release an open-source code base that can be used as a blueprint for similar projects and give a few demonstrations of insights and analysis that can be easily done if all data are consistently formatted and found in one place.

Details of the database

We have manually gone through every paper found in the Web of Science with the search phrase 'perovskite solar' up to the end of February 2020 (that is, over 15,000 papers). In total, we have manually extracted data for over 42,400 devices. While a few devices with extractable data will have slipped through our net, the devices in the database represent almost every device someone has thought is worth the effort to describe in detail in the peer-reviewed literature.

Our original data extraction protocol contained 95 attributes with metadata, process data and performance data. Those can

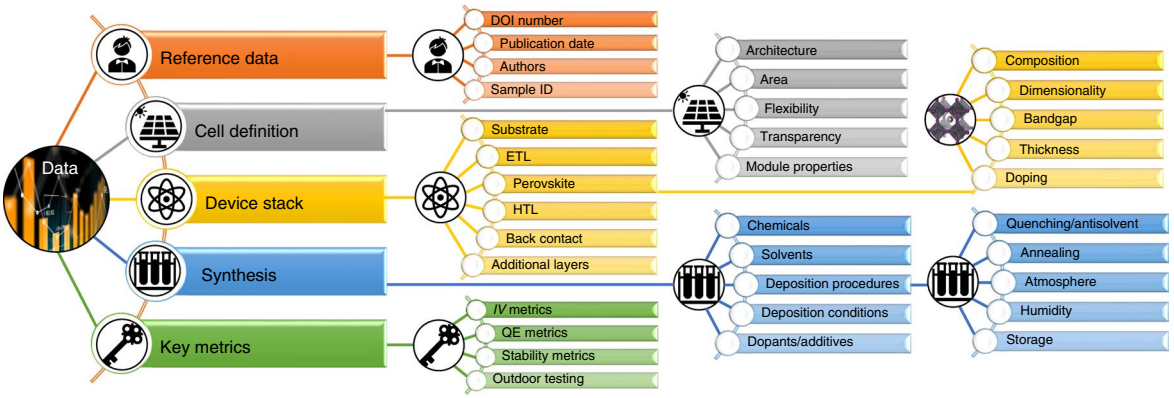

Fig. 2 | Overview of data categories in the Perovskite Database. Overview of the main categories of metadata, process data and performance data in the data extraction protocol. *IV*, current–voltage. QE, quantum efficiency.

be grouped into: reference data; cell-related data; data for every functional layer in the device stack, that is, type of substrate, electron transport layer (ETL), perovskite, hole transport layer (HTL), back contact and so on; synthesis related data for each layer and key metrics related to the performance of the resulting device; that is, current–voltage, quantum efficiency, stability and outdoor performance (Fig. 2). The categories and the formatting guidelines are described in detail in the supporting documentation. For future use, we have developed a more detailed protocol capturing up to 400 parameters per device, which can be found among the resources on the project's webpage.

Once extracted, the data have been consistently formatted according to the instruction in the supporting documentation and is now freely available in the Perovskite Database. To increase the usability of the data, we have developed interactive tools for simple exploration, analysis, filtering and visualization that can be used without programming knowledge. The code base for the project is written in Python and is available at GitHub (https://github.com/Jesperkemist/perovskitedatabase), and everyone is invited to contribute and expand the scope of the project. All the resources are found at the project website (www.perovskitedatabase.com), where they will be updated and maintained for the foreseeable future.

With all the device data consistently formatted and available in one place, a plethora of interesting possibilities opens. What follows is a small selection of analyses, visualizations and insights made possible by the Perovskite Database and the associated toolbox.

Example uses of the Perovskite Database

As a first example, the perovskite solar cell development is illustrated by binning the performance for all available devices and plotting those as a function of publication date (Fig. 3a). This demonstrates the expected trend towards higher-performing devices, as well as offering a sense of the underlying variability by showing the performance distribution, and thereby providing a comprehensive view of the field's progress.

The National Renewable Energy Laboratory (NREL) efficiency chart is probably one of the most reproduced images in the photovoltaic field. It is a highly trustable source as it exclusively relies on externally certified results, but is also limited in scope. The trend in global records illustrated in the NREL chart can easily be reproduced (Fig. 3b), even if some of the data points are different as they are sorted on publication date and include non-certified data. What makes this genuinely interesting is the possibility to filter out the records for any type of cell. With a single mouse click, it is possible to display the performance evolution of, for example, flexible cells, cells based on $CsPbI_3$ or cells fulfilling any combination of constraints (Fig. 3b). With an additional click, the figure can be downloaded and directly incorporated in presentations, applications or in a scientific publication. Clicking on a data point will also redirect the user to the original publication, which is a short-cut when searching for papers on a specific topic of interest.

A typical use case could be someone starting a project on a particular fabrication method, for example, slot-die coating. In the Perovskite Database, one simple command filters out the data for all available devices with slot-die-coated perovskites. Those data can be obtained in tabular form and downloaded with a click that gives an entry point to the key literature for further exploration. Once the relevant subset of data is obtained, it can be separated with respect to any of the dimensions represented in the database. To mention a few examples, these can be the perovskite doping conditions, the use of flexible substrates or, as shown in Fig. 3c, the solvent system used during the deposition of the perovskite. This represents a complex literature search that previously required a substantial amount of non-trivial work, but which can now be accomplished and visualized in a few minutes. With this insight at hand, it is just as easy to go on and explore additional questions, such as what is the importance of the annealing temperature, the choice of hole conductor, the antisolvent or to what extent does the perovskite composition influence the key performance metrics of the device? This illustrates a powerful short-cut towards extracting the historical data relevant for a project, for generating new hypotheses, for finding unexplored areas, for knowledge transfer and for acquiring insights otherwise easily overlooked.

With the aggregated data, it is also possible to visualize trends of how various experimental practices have been developed over the past years. An example is given in Fig. 3d that illustrates how the popularity of a few perovskite compositions, that is, $MAPbI_3$, $FA_xMA_{1-x}PbBr_yI_{3-y}$ and $Cs_zFA_xMA_{1-x-z}PbBr_yI_{3-y}$, have developed over time. That figure embodies both a technical aspect of device optimization, but also the more sociological aspect of how experimental practices and ideas spread through a scientific community.

The data collected in the Perovskite Database demonstrate great flexibility to how a functional perovskite solar cell can be constructed. Among the 42,400 devices found in the database at the time of writing, there are over 5,500 unique device stacks (that is, different combinations of contact materials), not considering the more than 400 different families of perovskite compositions (that is, different combinations of the A, B and C-site ions in the perovskite ABC_3-structure). More than 1,000 of these stacks have champion PCEs above 18%, and more than 300 have demonstrated PCEs above 20%. The multitude of stacks can be broken down into 1,443 unique ETL stacks, 1,957 HTL stacks, 288 back contact configurations and 194 different substrates. Some options are, however, more common than others. Around 60% of all devices are, for example,

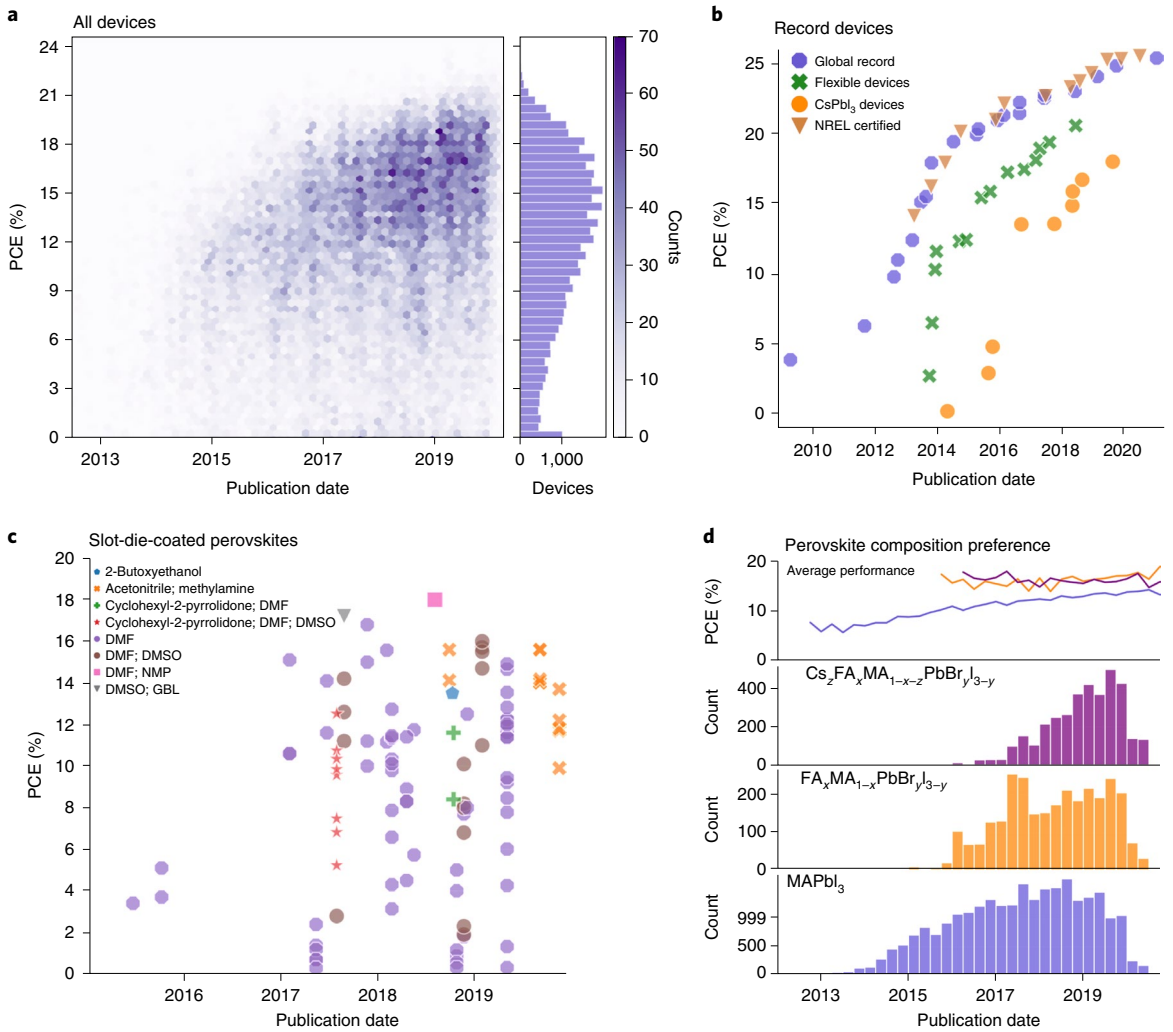

Fig. 3 | Development of perovskite cell efficiencies. Example of analysis from the database. **a**, Hexbin-plot of PCE measured under standard conditions as a function of the publication date for all devices in the database. Efficiency distribution for all devices is shown to the right. **b**, Evolution of record efficiency of all cells, flexible cells and $CsPbI_3$-based cells. Data from the NREL efficiency chart are given as a comparison. **c**, 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. **d**, Average performance and popularity of a $MAPbI_3$, $FA_xMA_{1-x}PbBr_yI_{3-y}$ and $Cs_zFA_xMA_{1-x-z}PbBr_yI_{3-y}$ perovskite compositions as a function of time.

based on methylammonium lead iodide ($MAPbI_3$), and the ten most common HTLs are used in 85% of all devices, with Spiro-MeOTAD ($C_{81}H_{68}N_4O_8$) used in close to half of them.

A problem faced while developing perovskite solar cells, which is in no way unique for the perovskite field, are cell-to-cell and batch-to-batch variations. Those can be large, thus masking otherwise statistically significant differences. There are also laboratory-to-laboratory variations, and what appears to make a significant difference in one laboratory may not be relevant in another. This is usually ascribed to undescribed, unexplored, unknown or hidden parameters that might influence, for example, the crystallization dynamics of the perovskite film[34]. Those could be things such as glove box volume, precise atmospheric composition during fabrication, minor or unintended variations in precursor stoichiometry[35,36], chemical impurities[37] and so on to mention a few hypotheses. The Perovskite Database can mitigate that problem by combining all the available disseminated device data. That allows for more holistic conclusions about what works, what does not and how reliable and consistent various procedures are. This is illustrated with a few examples below.

In Fig. 4a, the kernel density estimation, that is, the smoothed average, of the open-circuit voltage (V_{oc}) is given for three common HTLs. For a fair comparison, only $MAPbI_3$-based devices are included. It turns out that the hole conductor has a notable impact on the V_{oc} that can be expected on average, which is an example of something that is difficult to verify with a limited number of samples produced in a single laboratory but becomes apparent with such extensive data. The figure also indicates that Spiro-MeOTAD may be associated with a small V_{oc} loss, in line with recent discussions concerning interface recombination[38], and thus not be the best choice of hole conductor from a performance point of view, and the success for Spiro-MeOTAD may be more an effect of a historical coincidence, statistics and it having been heavily optimized rather than it having the highest intrinsic potential. Another example is given in Fig. 4b, which compares deposition procedures for TiO_2 based ETLs in nip-devices with a $MAPbI_3$ perovskite and Spiro-MeOTAD as HTL, which are the most common ETL and HTL stacks. The very best cells have been done using spin-coated mesoporous TiO_2 but on an aggregated level the choice of deposition procedure has a fairly small impact and all the depicted deposition procedures

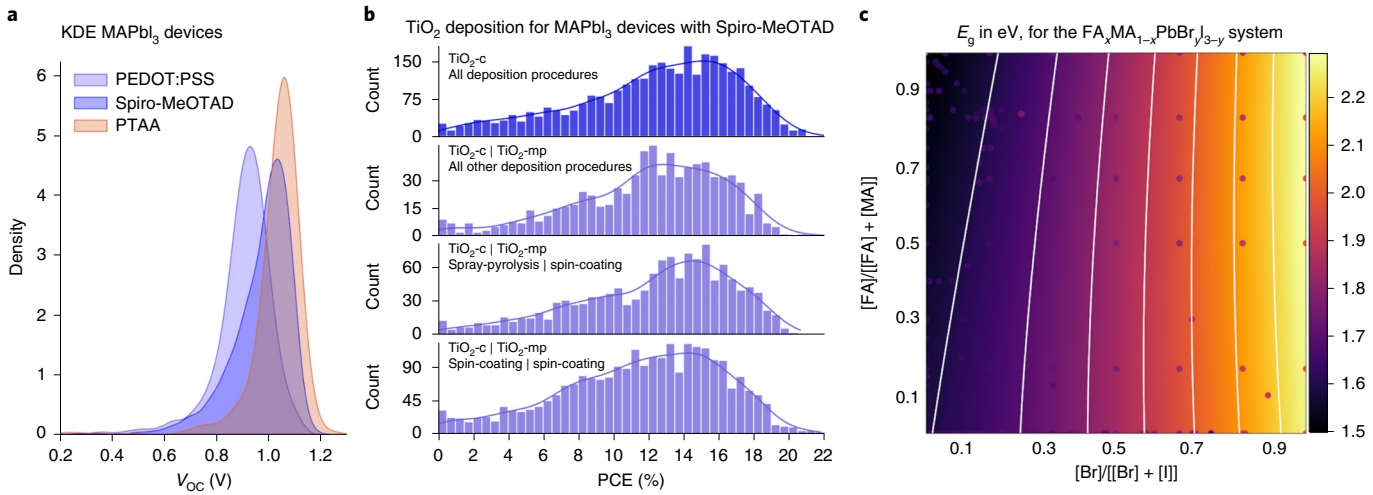

Fig. 4 | Example of analysis from the database. a, The kernel density estimation (KDE) of the V_{oc} for three common HTLs for $MAPbI_3$-based devices. **b**, Performance distributions separated by deposition procedures for the TiO_2-ETL in nip devices with $MAPbI_3$ and Spiro-MeOTAD. The top panel include all cells with a compact TiO_2 layer but without a mesoporous TiO_2. The remaining panels include cells with both compact (-c) and mesoporous (-mp) TiO_2 layers and are separated by the deposition procedure for each layer. The solid lines are the kernel density estimates. **c**, The experimental and fitted bandgap for the $FA_xMA_{1-x}PbBr_yI_{3-y}$ system. The background colour represents the fitted surface, the white lines are isolines and points represent the experimental data. The colour bar represents the bandgap in eV. The colour scheme gives special emphasis to outliers.

have resulted in a large spread in device performance. Excluding the mesoporous TiO_2 layer does not make much of a difference either for the average cell performance, which is interesting given that the very best cells still use a mesoporous TiO_2-layer.

The previous examples illustrate the power of having access to large, diverse, consistently formatted and interoperable datasets. They are also only scratching the surface while raising new questions that invite further explorations by digging deeper into the data. We anticipate this dataset will be an excellent resource for future work in perovskite groups as well as in the broader machine learning and data science communities.

One of the technologically appealing aspects of the metal-halide perovskites is the tunability of the bandgap (E_g), which ranges from below 1.2 eV for $MAPb_{0.5}Sn_{0.5}I_3$ (ref. [39]), to above 3 eV for $MAPbCl_3$ (ref. [40]). One way to use the collected bandgap data is to filter out perovskite compositions in a desired bandgap range. Another is to extrapolate the band gap of previously unexplored compositions, as illustrated in Fig. 4c. Here a second-degree polynomial has been fitted to the bandgap values in the database relating to composition in the $FA_xMA_{1-x}PbBr_yI_{3-y}$ system. Conversely, in such a compositional space, a simple optical measurement could then be used to estimate the perovskite composition. With the analysis code freely available, a fitting procedure such as that in Fig. 4c could easily be done for any compositional range where sufficient data are available and it can be updated whenever new data are made available.

Most devices have been made with perovskites with a bandgap of around 1.55–1.65 eV (Fig. 5a). That is where $MAPbI_3$ is found and it is the most interesting region for perovskite single-junction cells. For tandem integration, the need for optical matching between the subcells means that higher bandgaps are required for the top cell[41]. Unfortunately, from a tandem perspective, there is a drop in performance when the bandgap increases above roughly 1.8 eV, with the trend continuing up to 2.3 eV (Fig. 5a). This is primarily caused by an increased V_{oc} loss, which probably originates from a light-induced partial phase separation in mixed Br/I-perovskites[42], sometimes referred to as the Hoke effect[43].

When comparing the performance as a function of the perovskite bandgap in more detail, some results are found to be unphysical as they surpass the Shockley–Queisser (SQ) limit, most frequently in terms of a too large short-circuit current. Some of those points can be explained by mislabelled or misreported bandgaps, whereas others may be caused by errors in light source calibration and aperture area. Nevertheless, this illustrates a neglect of basic error checking in historic reports.

Another major challenge towards commercial viability is scalability. Most laboratory cells have an active area $\leq 0.2\,cm^2$, and it is also for these small cells where the highest efficiencies are found. When the cell area increases, there is a downwards trend in maximum performance (Fig. 5b), with a spike at $1\,cm^2$, which is a common cell area used in the first step towards upscaling. The average performance is rather constant with respect to the device area. The reasons for this are unclear, but a possible explanation could be the limited number of cells larger than $5\,cm^2$ reported so far and that upscaling is primarily pursued by groups already producing high-quality small-scale devices.

Long-term stability under operational conditions is a key requirement for any photovoltaic technology, and anyone making perovskite devices, particularly with early methods and recipes, quickly realizes that this will be a challenge. There is, however, less than 20% of the cells in the database for which stability data of any kind are available. At the time of writing, the Perovskite Database contains 7,400 entries with stability data, and 5,500 of those are variations of shelf life in the dark, where devices are stored and remeasured over time. There are around 550 entries with measurements under operational conditions, that is, air mass (AM) 1.5G and maximum power point tracking (MPPT). Historical comparison of stability is complicated both by the scarcity of high-quality data and by a lack of common standards and protocols for measuring and reporting stability data. This is, however, changing due to an active discussion in the field, which recently resulted in a list of International Summit on Organic Photovoltaic Stability (ISOS) consensus protocols related to measuring and reporting of stability data[23]. The Perovskite Database Project is fully compatible with those ISOS protocols.

There is not one single key metric of device stability but several, all with their own merits and limitations. One of the more commonly used is the T_{80} value, which is the time it takes for a cell to lose 20% of its initial performance. In Fig. 5c, the T_{80} versus publication date is given for the nearly 120 devices in the database

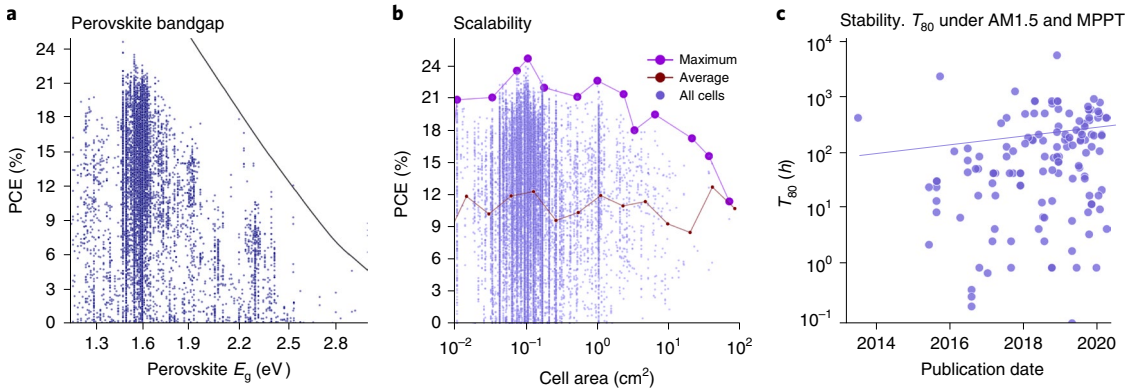

Fig. 5 | Identification of key challenges in the development of perovskite solar cells. Remaining key challenges. **a**, PCE versus E_g for all solar cells in the database. The Shockley–Queisser limit is given as a solid line. **b**, Illustration of perovskite scalability. **c**, T_{80} versus publication date for devices measured under AM 1.5 and MPPT. The solid line represents a linear fit to data.

measured under AM 1.5 and MPPT, and where a T_{80} is stated (that is, less than 0.3% of all cells). There is a general trend towards more devices with higher stabilities as the years progress, even if we still have rather few data points. Given the importance of the problem, we expect a dramatic increase in reporting this type of data in the next few years.

Figure 5 represents a first glimpse of what is found in the Perovskite Database related to the three core technological challenges, namely tandem integration, scalability and stability. All these aspects deserve a much longer analysis, and we expect a multitude of papers to be written based on these open-source resources, both by us and by others. We intend the Perovskite Database to be a living, evolving and scalable project, and we expect future work to expand the scope of the project by adding new data, functionality, analysis, visualizations and open-source code.

Future expansion of the database

The ambition of the Perovskite Database Project is to collect not only historic data but all future device data as well, to create a new standard for disseminating perovskite device data and to build what we can think of as the Wikipedia of perovskite solar cell research. This will require participation from the entire perovskite community, with a mental shift towards a culture where everyone feels that they can, want and will disseminate their device data by uploading it to the Perovskite Database as a complement to traditional publishing.

Uploading new data will take some time and effort. The Perovskite Database Project must therefore deliver a high degree of perceived use, simplicity, visibility, longevity and trustworthiness. In terms of use, we hope the examples in this paper, together with the interactive graphics on the project's website, have demonstrated the power of aggregated datasets adhering to the FAIR data principles, and that this alone provides an incentive to contribute. There are also other benefits to uploading one's own data. Sharing data in this way gives it new life and draws additional attention to the original publication, it is a way to comply with the demands for openness more frequently seen from taxpayers, funding agencies and publishers, and it is a service to the community that helps to accelerate the development of new solar cell technology. Finally, the tools and protocols we provide may help in organizing and improving the local data management and thereby, in the end, simplify planning, analysis and writing.

In terms of simplicity, we have developed intuitive and well-documented data extraction protocols. The backend for data cleaning and validation is written in Python, and the backend for collecting and reporting data is currently in the form of an Excel template. The Excel template is self-explanatory, easy to use, freely available and possible to extend to fit different laboratories' internal needs. By being transparent and freely available, it is possible to build customized data pipelines that directly feed data from laboratory equipment into the template, thereby simplifying data entry even further.

Our vision is that uploading data into databases such as this one will become standard procedure as this will strengthen the associated publication by increasing its visibility and usefulness. We further anticipate involving publishers as important stakeholders in this project. Making experimental data assessible on platforms used by most of the research community will increase the visibility of scientific results. In addition, the accumulation of all device data allows an straightforward assessment as to whether reported device performance metrics are physically possible (for example, that are in the expected performance limits of the Shockley–Queisser limit for single-junction solar cells) or deviate substantially from common trends.

To ensure the project's longevity, we have secured support from the Helmholtz Organization in Germany, which acts as a guarantor ensuring that the web resources, that is, database, webpage and the GitHub account, will be operational and maintained for the coming decade, with an option of possible prolongation.

Another key aspect related to trustworthiness is the open-source nature of the project, which means transparency, to which users could suggest improvements and provide additional functionality, and it enables easy restart in case of disruption.

The database could also easily be expanded to include data relevant to, for example, LEDs, lasers, scintillators and so on, and we actively encourage initiatives in that direction.

A key problem addressed in this project is the challenge of keeping track of the field's progress when data are inconsistently formatted and scattered over an inaccessible large number of papers. A related problem is data loss, or the iceberg problem[44,45]. In a typical project, there may be hundreds and sometimes thousands of devices made before the paper is written. Despite this, the average number of devices for which we could extract data was fewer than six per publication with original device data. A common pattern is that one parameter is changed in few steps, and for each of those steps data for the best device could be found. Some of the data for the missing devices are presented as statistical averages, even if the data for the individual devices cannot be extracted from the papers. Data for other devices are, for various reasons, never disseminated and are essentially lost forever. Data for most of the best devices are probably disseminated, but there is a wealth of information hidden in the data now lost[44,45]. With the tools here developed, we facilitate

reporting data for also those kinds of device in future reports, which could mitigate the bias for not disseminating data for failed experiments and less successful devices.

Conclusions

In this Perovskite Database Project, we have created an open-access database for perovskite solar cell device data and visualization tools for interactive data exploration, and we have populated the database with data for over 42,000 devices described in the peer-reviewed literature up until spring 2020. We also demonstrate the capabilities of the database and the associated tools by giving a few examples of insights that can be gleaned from the analysis of this large dataset in terms of, for example, record development, tandem integration, stability and scalability. We hope that this project will prompt better data management in the perovskite field as well as a culture of data sharing, as well as inspiring other experimental fields to do the same. We could then get data with a more fine-grained data mesh and make those data available for most devices ever made, not just a few highlighted in papers as has been the case historically. In a few years, we could then have data for millions of devices, which will enable us to finally take greater advantage of machine learning and other artificial intelligence-based methods to accelerate development even further.

Methods

The search phrase 'perovskite solar' in the Web of Science generated over 15,000 entries by the end of February 2020. Not all of those publications relate to metal-halide perovskites and photovoltaic applications, but most do. Similarly, a few relevant papers will be missed in this search. From here, our collective team has manually gone through every paper and extracted data for all the described devices.

Of the publications we went through, we found original experimental device data in close to half of them, that is, around 7,400. Among the remaining papers, we found reviews, theoretical investigations and studies focused on material properties, as well as some non-photovoltaic-/perovskite-related publications. In total, we have manually extracted data for over 42,400 devices. The total time consumption to do this is in the range of 5,000–10,000 man hours.

On the basis of our collective experience of perovskite device development and optimization, the total number of devices ever made is probably at least two orders of magnitude larger, but for data for most of those devices cannot be extracted from the publications. In fact, data for most devices are only available as average values, in scatterplots or not disseminated at all.

One database entry per device has been the default procedure, but if only averaged data were found, we entered that as belonging to one cell but specified the number of devices the averaged is based on. Another guiding principle has been that, while preferably having all possible data for a device, having some data is better than having none. We have thus not discarded data based on poor or limited device descriptions in the scientific publications. We also considered a best estimate of a perovskite composition, for example, to be worth more than stating the information as unknown, which for example could be the case for solvent-based ion exchange procedure where the ionic fractions in the perovskite cannot be derived from the composition of the precursor solutions, but where it can be inferred from optical or X-ray diffraction data.

All data contain errors. That is unavoidable. Some sources of errors include: the data stated in the original papers are erroneous due to several possible reasons; misinterpretation of data, which is easily done when papers are ambiguous or confusingly written, and errors while transferring data from papers to the database. We have therefore set up a system for reporting dubious data points, and we thereby expect some self-correction over time, especially for data points of special interest such as records in subfields. To reduce the errors, we went through the extracted data to check for errors, misunderstandings, confusing entries and inconsistent formatting. For future data, where we expect authors to upload their own data, we expect a lower error rate than for the historical dataset. It is, nevertheless, advisable to double check outliers, especially when the applied search filters generate small datasets, so as not to draw erroneous conclusions. We also encourage authors, who know their own data best, to double check their devices in the database.

Every data point in the database is linked to the DOI number of the original publication. Every data point is thus effectively cited in the database, and for everyone who uses the data found there it is straightforward to use this DOI linkage to both find and cite the original sources of the data used.

Data availability

The project has a dedicated website, www.perovskitedatabase.com that provide access to all resources. Among those are: the Perovskite Database, interactive graphics exploring the database, instructions for what is found in the database, templates and instructions for uploading new data, links to all works related to the project and so on.

Code availability

Codes reproducing all analyses in this paper are available in the following GitHub repository at https://github.com/Jesperkemist/perovskitedatabase.

Received: 28 February 2021; Accepted: 19 October 2021;
Published online: 13 December 2021

References

1. Al-Ashouri, A. et al. Monolithic perovskite/silicon tandem solar cell with >29% efficiency by enhanced hole extraction. *Science* **370**, 1300–1309 (2020).
2. Snaith, H. J. Perovskites: the emergence of a new era for low-cost, high-efficiency solar cells. *J. Phys. Chem. Lett.* **4**, 3623–3630 (2013).
3. Bailie, C. D. et al. Semi-transparent perovskite solar cells for tandems with silicon and CIGS. *Energy Environ. Sci.* **8**, 956–963 (2015).
4. Albrecht, S. et al. Monolithic perovskite/silicon-heterojunction tandem solar cells processed at low temperature. *Energy Environ. Sci.* **9**, 81–88 (2016).
5. Jošt, M., Kegelmann, L., Korte, L. & Albrecht, S. Monolithic perovskite tandem solar cells: a review of the present status and advanced characterization methods toward 30% efficiency. *Adv. Energy Mater.* **10**, 1904102 (2020).
6. Tan, Z.-K. et al. Bright light-emitting diodes based on organometal halide perovskite. *Nat. Nanotechnol.* **9**, 687–692 (2014).
7. Van Le, Q., Jang, H. W. & Kim, S. Y. Recent advances toward high-efficiency halide perovskite light-emitting diodes: review and perspective. *Small Methods* **2**, 1700419 (2018).
8. Deschler, F. et al. High photoluminescence efficiency and optically pumped lasing in solution-processed mixed halide perovskite semiconductors. *J. Phys. Chem. Lett.* **5**, 1421–1426 (2014).
9. Domanski, K. et al. Working principles of perovskite photodetectors: analyzing the interplay between photoconductivity and voltage-driven energy-level alignment. *Adv. Func. Mater.* **25**, 6936–6947 (2015).
10. Ahmadi, M., Wu, T. & Hu, B. A review on organic–inorganic halide perovskite photodetectors: device engineering and fundamental physics. *Adv. Mater.* **29**, 1605242 (2017).
11. Kraus, H., Mykhaylyk, V. & Saliba, M. Bright and fast scintillation of organolead perovskite $MAPbBr_3$ at low temperatures. *Mater. Horiz.* **6**, 1740–1747 (2019).
12. Green, M. A. et al. Solar cell efficiency tables (version 56). *Prog. Photovolt. Res. Appl.* **28**, 629–638 (2020).
13. Wali, Q. et al. Advances in stability of perovskite solar cells. *Org. Electron.* **78**, 105590 (2020).
14. Krishnan, U., Kaur, M., Kumar, M. & Kumar, A. Factors affecting the stability of perovskite solar cells: a comprehensive review. *J. Photon. Energy* **9**, 021001 (2019).
15. Howard, J. M., Tennyson, E. M., Neves, B. R. & Leite, M. S. Machine learning for perovskites' reap-rest-recovery cycle. *Joule* **3**, 325–337 (2019).
16. Park, N.-G. & Zhu, K. Scalable fabrication and coating methods for perovskite solar cells and solar modules. *Nat. Rev. Mater.* **5**, 333–350 (2020).
17. Qiu, L., He, S., Ono, L. K., Liu, S. & Qi, Y. Scalable fabrication of metal halide perovskite solar cells and modules. *ACS Energy Lett.* **4**, 2147–2167 (2019).
18. Swartwout, R., Hoerantner, M. T. & Bulović, V. Scalable deposition methods for large-area production of perovskite thin films. *Energy Environ. Mater.* **2**, 119–145 (2019).
19. Matteocci, F., Castriotta, L. A. & Palma, A. L. in *Photoenergy and Thin Film Materials* (ed. Yang, X.-Y.) 121–155 (Wiley, 2019).
20. Li, N., Niu, X., Chen, Q. & Zhou, H. Towards commercialization: the operational stability of perovskite solar cells. *Chem. Soc. Rev.* **49**, 8235–8286 (2020).
21. Howard, I. A. et al. Coated and printed perovskites for photovoltaic applications. *Adv. Mater.* **31**, 1806702 (2019).
22. Mathies, F., List-Kratochvil, E. J. & Unger, E. L. Advances in inkjet-printed metal halide perovskite photovoltaic and optoelectronic devices. *Energy Technol.* **8**, 1900991 (2020).
23. 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).
24. Schwab, K. & Davis, N. *Shaping the Future of the Fourth Industrial Revolution* (Currency, 2018).
25. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
26. Curtarolo, S. et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comp. Mater. Sci.* **58**, 218–226 (2012).
27. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).

28. Gražulis, S. et al. Crystallography Open Database–an open-access collection of crystal structures. *J. Appl. Crystallogr.* **42**, 726–729 (2009).
29. Almora, O. et al. Device performance of emerging photovoltaic materials (version 1). *Adv. Energy. Mater.* **11**, 2002774 (2020).
30. Bergerhoff, G., Brown, I. D. & Allen, F. *Crystallographic Databases* (International Union of Crystallography (1987).
31. Empty rhetoric over data sharing slows science. *Nature* **546**, 327 (2017).
32. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
33. Draxl, C. & Scheffler, M. NOMAD: the FAIR concept for big data-driven materials science. *MRS Bull.* **43**, 676–682 (2018).
34. Zeng, L. et al. Controlling crystallization dynamics of photovoltaic perovskite layers on larger-area coatings. *Energy Environ. Sci.* **13**, 4666–4690 (2020).
35. Jacobsson, T. J. et al. Unreacted PbI_2 as a double-edged sword for enhancing the performance of perovskite solar cells. *J. Am. Chem. Soc.* **138**, 10331–10343 (2016).
36. Fassl, P. et al. Fractional deviations in precursor stoichiometry dictate the properties, performance and stability of perovskite photovoltaic devices. *Energy Environ. Sci.* **11**, 3380–3391 (2018).
37. Zhang, Y. et al. Achieving reproducible and high-efficiency (>21%) perovskite solar cells with a presynthesized $FAPbI_3$ powder. *ACS Energy Lett.* **5**, 360–366 (2019).
38. Gharibzadeh, S. et al. Record open-circuit voltage wide-bandgap perovskite solar cells utilizing 2D/3D perovskite heterostructure. *Adv. Energy Mater.* **9**, 1803699 (2019).
39. Ogomi, Y. et al. $CH_3NH_3Sn_xPb_{(1-x)}I_3$ Perovskite solar cells covering up to 1,060 nm. *J. Phys. Chem. Lett.* **5**, 1004–1011 (2014).
40. Liu, D., Yang, C. & Lunt, R. R. Halide perovskites for selective ultraviolet-harvesting transparent photovoltaics. *Joule* **2**, 1827–1837 (2018).
41. Jacobsson, T. J. et al. 2-Terminal CIGS-perovskite tandem cells: a layer by layer exploration. *Sol. Energy* **207**, 270–288 (2020).
42. Jacobsson, T. J. et al. Exploration of the compositional space for mixed lead halogen perovskites for high efficiency solar cells. *Energy Environ. Sci.* **9**, 1706–1724 (2016).
43. Hoke, E. T. et al. Reversible photo-induced trap formation in mixed-halide hybrid perovskites for photovoltaics. *Chem. Sci.* **6**, 613–617 (2015).
44. Heidorn, P. B. Shedding light on the dark data in the long tail of science. *Libr. Trends* **57**, 280–299 (2008).
45. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).

Acknowledgements

The core funding of the project has been received from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 787289. We acknowledge MaterialsZone (https://www.materials.zone/) for technical assistance and for hosting the project's cloud resources. We acknowledge Helmholtz-Zentrum Berlin für Materialien und Energie for guaranteeing economic and technical support for keeping the project online for the next decade. We acknowledge the following sources for individual funding. Cambridge India Ramanujan Scholarship, China Scholarship Council, Deutscher Akademischer Austauschdienst (DAAD), EPSRC (grant no. EP/S009213/1), European Union's Horizon 2020 research and innovation programme (grant no. 764787, EU Project 'MAESTRO'), (grant no. 756962, ERC Project 'HYPERION'), (grant no. 764047, EU Project 'ESPResSo' and grant no. 850937), GCRF/EPSRC SUNRISE (EP/P032591/1), German Federal Ministry for Education and Research (BMBF), HyPerFORME, NanoMatFutur (grant no. 03XP0091). PEROSEED (ZT-0024), Helmholtz Energy Materials Foundry, The Helmholtz Innovation Laboratory HySPRINT. BMBF (grant nos. 03SF0540, 03SF0557A), HyPerCells graduate school, Helmholtz Association, Helmholtz International Research School (HI-SCORE), the Erasmus programme (CDT-PV, grant no. EP/L01551X/1), the European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie grant agreement nos. 841386, 795079 and 840751), Royal Society University Research Fellowship (grant no. UF150033). SNaPSHoTs (BMBF), SPARC II, German Research Foundation (DFG, grant no. SPP2196), The National Natural Science Foundation of China (grant no. 51872014), the Recruitment Programme of Global Experts, Fundamental Research Funds for the Central Universities and the '111' project (grant no. B17002), the US Department of Energy's Office of Energy Efficiency and Renewable Energy under Solar Energy Technologies Office (SETO) agreement no. DE-EE0008551, the Colombia Scientific Programme in the framework of the call Ecosistema Cientifico (Contract no. FP44842-218-2018), the committee for the development of research (CODI) of the Universidad de Antioquia (grant no. 2017-16000), Spanish MINECO (Severo Ochoa programme, grant no. SEV-2015-0522), the Swedish research council (VR, grant no. 2019-05591) and the Swedish Energy Agency (grant no. 2020-005194).

Author contributions

T.J.J. and E.U. designed the project. T.J.J. coordinated the project, wrote much of the code for the interactive graphics and wrote the first draft of the paper. M.V., A.Y.A. and O.Y. worked on coding. T.J.J., A.H., A.G.-F., A. Anand, A.A-A., A.H., A.C., A. Abate, A.G.R., A.V., A.K., B.P.D., B.Y., B.L.C., C.A.R.P., C.R., D.R., D.F.-J., D.D.G., D.J., E.A., E.J.J.-P., F.B., F.M., G.S.A.G., G.B., G.N., G.P., G.M.-D., H.N., H.M., H.K., H.W., I.B., M.I.D., I.B.P., I.E.G., J.N.V., J.D., J.K., J.Y., J.L., J.A.S., J.P., J.J.J.-R., J.F.M., J.-P.C-B., J.Q., J.W., K.S., K.H., K.D., K.F., L.M., L.A.C., M.H.A., M.V.-M., M.A.R.-P., M.A.F., M.V.K., M.G., M.K., M.S., M.A., N.A., O.S., O.M., O.S.G., P.F., Q.Z., R.B., R.M., R.P., S.S., S.A., S.K., T.U., T.A., T.E., T.W.D., U.W.P., W.Z., W.F., W.Z., V.R.F.S., W.T., X.Z., Y.-H.C., Z.I., Z.X. and E.U. all contributed to the laborious task of going through the literature, extracting the data found there and formatting consistently. All authors have participated in preparing the final draft of the paper.

Funding

Open access funding provided by Helmholtz-Zentrum Berlin für Materialien und Energie GmbH.

Competing interests

MaterialsZone is a web platform used for managing, standardizing, sharing and analysing data in the field of Materials Science, and is aimed at researchers in both academia and industry. In this project, MaterialsZone worked in collaboration with Helmholtz-Zentrum Berlin to make the Perovskite Database Project easily accessible to anyone interested in these data in an open and convenient manner. The people who participated in this project from MaterialsZone are: A.Y.A (Chief Executive Officer of the company, and PhD in Materials Science), O.Y. (Chief Technology Officer of the company and PhD in Mathematics) and M.V. (Senior Developer and Data Scientist, and PhD in Experimental Physics). The remaining authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T. Jesper Jacobsson or Eva Unger.

Peer review information *Nature Energy* thanks Chris Deline, Sang Il Seok, Marina Leite and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

[1]Young Investigator Group Hybrid Materials Formation and Scaling, Helmholtz-Zentrum Berlin für Materialen und Energie GmbH, Berlin, Germany. [2]Department of Chemistry, Uppsala University, Uppsala, Sweden. [3]Department of Materials Science and Engineering, Solar Cell Technology, Uppsala University, Uppsala, Sweden. [4]Division of Applied Physical Chemistry, Department of Chemistry, KTH Royal Institute of Technology, Stockholm, Sweden. [5]Center for Energy and Environmental Chemistry Jena, Friedrich Schiller University Jena, Jena, Germany. [6]Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Jena, Germany. [7]Young Investigator Group Perovskite Tandem Solar Cells, Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin, Germany. [8]Laboratory of Photomolecular Science, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [9]Department of Structure and Dynamics of Energy Materials, Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin, Germany. [10]Department Novel Materials and Interfaces for Photovoltaic Solar Cells, Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin, Germany. [11]Institute of Materials Science, Technische Universität Darmstadt, Darmstadt, Germany. [12]IEK5-Photovoltaics, Forschungszentrum Jülich, Jülich, Germany. [13]Materials Zone, Tel Aviv-Yafo, Israel. [14]Laboratory for Molecular Engineering of Optoelectronic Nanomaterials, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [15]SPECIFIC, College of Engineering, Swansea University,

Swansea, UK. [16]School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [17]Center for Research, Innovation and Development of Materials-CIDEMAT, Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia. [18]Adsorption and Advanced Materials Lab, Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK. [19]Department of Chemical Materials and Production Engineering, University of Naples Federico II, Naples, Italy. [20]Department of Chemistry, University of Rome La Sapienza, Rome, Italy. [21]School of Materials Science and Engineering, Beihang University, Beijing, China. [22]Aragon Agency for Research and Development (ARAID), Instituto de Nanociencia y Materiales de Aragón (INMA) CSIC-Universidad de Zaragoza, Zaragoza, Spain. [23]Chemical Physics and NanoLund, Lund University, Lund, Sweden. [24]Benemérita Universidad Autónoma de Puebla, CIDS-ICUAP, San Claudio, Puebla, México. [25]Faculty IV—Electrical Engineering and Computer Science, TU Berlin, Berlin, Germany. [26]ICFO—Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, Castelldefels, Spain. [27] Cavendish Laboratory, University of Cambridge, Cambridge, UK. [28]Department of Materials Science and Engineering, Solid State Physics, Uppsala University, Uppsala, Sweden. [29]Materials Science and Engineering, University of Colorado, Boulder, CO, USA. [30]National Renewable Energy Laboratory, Golden, CO, USA. [31]James Watt School of Engineering, University of Glasgow, Glasgow, UK. [32]Department of Physics, Chemistry and Biology (IFM), Linkoping University, Linkoping, Sweden. [33]Department of Physics and Astronomy, University of Sheffield, Sheffield, UK. [34]Department of Physics, Clarendon Laboratory, Oxford University, Oxford, UK. [35]Institute for Photovoltaics (IPV), University of Stuttgart, Stuttgart, Germany. [36]Chemistry Research Laboratory, Department of Chemistry, University of Oxford, Oxford, UK. [37]Interdisziplinäres Zentrum für Materialwissenschaften, Martin-Luther Universität, Halle, Germany. [38]Centre for Hybrid and Organic Solar Energy, Electronic Engineering Department, University of Rome Tor Vergata, Rome, Italy. [39]Egyptian Petroleum Research Institute, Nasr City, Egypt. [40]Institute of Microstructure Technology, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. [41]Light Technology Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany. [42]PVcomB, Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin, Germany. [43]Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany. [44]Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK. [45]Physical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), KAUST Solar Center, Thuwal, Saudi Arabia. [46]Interdisciplinary Graduate School, Energy Research Institute @ Nanyang Technological University (ERI@N), Singapore, Singapore. [47]School of Computer Science and Electronic Engineering, Bangor University, Bangor, UK. [48]Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin, Germany. [49]Department of Chemistry, Department of Physics, Humboldt-Universität zu Berlin, IRIS Adlershof, Berlin, Germany. [50]Novel Semiconductor Devices Group, Institute for Computational Physics, Zurich University of Applied Sciences, Winterthur, Switzerland. [51]Department of Physics, University of York, York, UK. ✉e-mail: jacobsson.jesper.work@gmail.com; eva.unger@helmholtz-berlin.de