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\textsuperscript{1–5}, LEDs\textsuperscript{6,7}, lasers\textsuperscript{8}, photodetectors\textsuperscript{9,10}, X-ray detectors\textsuperscript{11} and for single-junction solar cells the record certified power conversion efficiency (PCE) has reached above 25\% (ref. \textsuperscript{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\textsuperscript{13–15}, scalability\textsuperscript{16–19} and...
Reliability; the best material combinations and manufacturing processes are open questions, and key standards and metrics are still under discussion.

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 inaccessible 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. The key goals of the project are to: collect 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
The Perovskite Database, a comprehensive resource, contains over 5,500 unique device stacks, allowing users to explore and analyze a wealth of data. The database is designed to facilitate research and innovation in perovskite solar cells by providing detailed information on various aspects of device performance and fabrication.

### 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 photo-voltaic 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$_x$MA$_{1-x}$PbBr$_{1+y}$, and Cs$_x$FA$_x$MA$_{1-x-y}$PbBr$_{1+y}$, 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$_x$-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,
based on methylammonium lead iodide (MAPbI₃), and the ten most common HTLs are used in 85% of all devices, with Spiro-MeOTAD (C₈H₁₅N₅O₇) 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 film⁴⁴. Those could be things such as glove box volume, precise atmospheric composition during fabrication, minor or unintended variations in precursor stoichiometry⁵⁵,⁵⁶, chemical impurities⁵⁷ 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 (Vₜₜ) is given for three common HTLs. For a fair comparison, only MAPbI₃-based devices are included. It turns out that the hole conductor has a notable impact on the Vₜₜ 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 Vₜₜ loss, in line with recent discussions concerning interface recombination⁴⁶, 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 TiO₂ based ETLs in nip-devices with a MAPbI₃ 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 TiO₂ but on an aggregated level the choice of deposition procedure has a fairly small impact and all the depicted deposition procedures
have resulted in a large spread in device performance. Excluding the mesoporous TiO₂ 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 TiO₂ 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 MAPb₀.₅Sn₀.₅I₃ (ref. 39), to above 3 eV for MAPbCl₃ (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\(_{x}\)MA\(_{1-x}\)PbBr\(_y\)I\(_{3-y}\) 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 MAPbI₃ 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 (Fig. 5a). 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 Voc loss, which probably originates from a light-induced partial phase separation in mixed Br/I-perovskites\(^{41}\), sometimes referred to as the Hoke effect\(^{41}\).

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 ≤0.2 cm², 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 cm², 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 that this will be a challenge. There is, however, less than 20% of the devices in the database for which stability data of any kind are available. At the time of writing, the Perovskite Database contains 7,400 cells in the database for which stability data of any kind are available. 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\(^{21}\). The Perovskite Database Project is fully compatible with those ISOS protocols.

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. 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\(^{21}\). 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 \(T_{90}\) value, which is the time it takes for a cell to lose 20% of its initial performance. In Fig. 5c, the \(T_{90}\) versus publication date is given for the nearly 120 devices in the database...
measured under AM 1.5 and MPPT, and where a $T_{so}$ 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 a 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. 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. With the tools here developed, we facilitate
reporting data for all 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 quality, but where it 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, or confusingly written, and errors while transferring data from papers to 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 by doing 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/jesperkenst/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. Snaita, H. J. Perovskites: the emergence of a new era for low-cost, high-efficiency solar cells. J. Mater. Chem. A 1, 3623–3630 (2013).
3. Ballie, 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, 1423–1426 (2014).
9. Domanski, K. et al. Working principles of perovskite photodetectors: analyzing the interplay between photocconductivity and voltage-driven energy-level alignment. Adv. Funct. 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, 1605243 (2017).
11. Kraus, H., Mykhailyk, V. & Saliba, M. Bright and fast scintillation of organolead perovskite MAPbBr3 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. Qu, 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. & Bulovic, 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 Photovoltaic 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., Liat-Kravchilv, 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, 33–49 (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. Gurtarardo, S. et al. AFLOW: a universal 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. Graziulis, 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, E. 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. & Scheller, 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 PbI2 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 FA-PbI2 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. CH3NH3SnPbI3-δ. 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. 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 perovskite photovoltaics. Chem. Sci. 6, 613–617 (2015).

44. Heidorn, P. B. Shedding light on the dark data in the long tail of science. Nature 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.materialszone.net) 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 Scholarships, 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. 756692, ERC Project ‘HYPERION’), (grant no. 764047, EU Project ‘ESPResSo’ and grant no. 850937), GCRF/EP/SRC SUNRISE (EP/P030287/1), German Federal Ministry for Education and Research (BMBF), HyPerFORME, NanoMatFutur (grant no. 03XP0091), PEROSED (ZT-0024), Helmholtz Energy Materials Foundry, The Helmholtz Innovation Laboratory HySPRINT. BMBF (grant no. 05S05040, 05S0557A), HyPerCells graduate school, Helmholtz Association, Helmholtz International Research School (Hi-SCORE), the Erasmus programme (CDT-PV, grant no. EP/K01551X/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), SNASPhoTs (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 Colombina Scientific Programme in the framework of the call Ecosistema Científico (Contract no. FP4842-218-2018), the committee for the development of research (CODI) of the Universidad de Antioquia (grant no. 2017-16008), 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.T., B.L.C., C.A.R.P., C.R., D.F.-J., D.D.G., D.J., E.A., E.J.J.-P., F.B., E.M., G.S.A.G., G.B., G.N., G.P., G.M., D., H.N., H.M., H.K., H.W., I.B., I.M.D., I.R.P., I.G., I.N.Y., J.D., I.K., J.Y., J.L., J.A.S., J.P., J.I.-R., J., J.-P., J.F., J.Q., J.W., J.S., K.H., K.D., K.F., L.M., L.A.C., M.H.A., V.M.-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., T.W.D., U.W.P., W.Z., W.F., W.F., W.R.F.S., W.T., X.Y.-C., Y.-H.C., Z.I., Z.X. and E.U. all contributed to the laborsious 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. 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.T.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.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021
Swansea, UK. 16School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. 17Center for Research, Innovation and Development of Materials-CIDEMAT, Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia. 18Adsorption and Advanced Materials Lab, Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK. 19Department of Chemical Materials and Production Engineering, University of Naples Federico II, Naples, Italy. 20Department of Chemistry, University of Rome La Sapienza, Rome, Italy. 21School of Materials Science and Engineering, Beihang University, Beijing, China. 22Aragon Agency for Research and Development (ARAI), Instituto de Nanociencia y Materiales de Aragón (INMA) CSIC-Universidad de Zaragoza, Zaragoza, Spain. 23Chemical Physics and NanoLund, Lund University, Lund, Sweden. 24Benemérita Universidad Autónoma de Puebla, CIDS-ICUAP, San Claudio, Puebla, México. 25Faculty IV—Electrical Engineering and Computer Science, TU Berlin, Berlin, Germany. 26ICFO—Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, Castelldefels, Spain. 27 Cavendish Laboratory, University of Cambridge, Cambridge, UK. 28Department of Materials Science and Engineering, Solid State Physics, Uppsala University, Uppsala, Sweden. 29Materials Science and Engineering, University of Colorado, Boulder, CO, USA. 30National Renewable Energy Laboratory, Golden, CO, USA. 31James Watt School of Engineering, University of Glasgow, Glasgow, UK. 32Department of Physics, Chemistry and Biology (IFM), Linköping University, Linköping, Sweden. 33Department of Physics and Astronomy, University of Sheffield, Sheffield, UK. 34Department of Physics, Clarendon Laboratory, Oxford University, Oxford, UK. 35Institute for Photovoltaics (IPV), University of Stuttgart, Stuttgart, Germany. 36Chemistry Research Laboratory, Department of Chemistry, University of Oxford, Oxford, UK. 37Interdisziplinäres Zentrum für Materialwissenschaften, Martin-Luther Universität, Halle, Germany. 38Centre for Hybrid and Organic Solar Energy, Electronic Engineering Department, University of Rome Tor Vergata, Rome, Italy. 39Egyptian Petroleum Research Institute, Nasr City, Egypt. 40Institute of Microstructure Technology, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. 41Light Technology Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany. 42Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany. 43Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK. 44Physical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), KAUST Solar Center, Thuwal, Saudi Arabia. 45Interdisciplinary Graduate School, Energy Research Institute @ Nanyang Technological University (ERI@N), Singapore, Singapore. 46School of Computer Science and Electronic Engineering, Bangor University, Bangor, UK. 47Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin, Germany. 48Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany. 49Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK. 50Physical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), KAUST Solar Center, Thuwal, Saudi Arabia. 51Interdisciplinary Graduate School, Energy Research Institute @ Nanyang Technological University (ERI@N), Singapore, Singapore. 52School of Computer Science and Electronic Engineering, Bangor University, Bangor, UK. 53Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Berlin, Germany. 54Department of Chemistry, Department of Physics, Humboldt-Universität zu Berlin, IRIS Adlershof, Berlin, Germany. 55Novel Semiconductor Devices Group, Institute for Computational Physics, Zurich University of Applied Sciences, Winterthur, Switzerland. 56Department of Physics, University of York, York, UK. 57e-mail: jacobsson.jesper.work@gmail.com; eva.unger@helmholtz-berlin.de