Curated Materials Data of Hybrid Perovskites: Approaches and Potential Usage

Rayan Chakraborty[1]* and Volker Blum[1,2]*

[1] Thomas Lord Department of Mechanical Engineering and Materials Science, Duke University, Durham, North Carolina 27708, United States

[2] Department of Chemistry, Duke University, Durham, North Carolina 27708, United States

Correspondence:

rayan.chakraborty@duke.edu

volker.blum@duke.edu

ORCID IDs-

Rayan Chakraborty: 0000-0003-2925-4373

Volker Blum: 0000-0001-8660-7230

Abstract. Over the last decade, hybrid perovskite research has evolved to a point where the literature contains an enormous volume of chemical and physical information. However, many essential materials design challenges remain open for researchers to address. The dispersed nature of the large, rapidly growing body of hybrid perovskite materials data poses a barrier to systematic discovery efforts, which can be solved by materials property databases - either by high-throughput or by systematic, accurate human-curated efforts. This opinion article discusses the necessity, challenges, and requirements of building such data libraries. In light of using machine learning (ML) and related tools to solve specific problems, the importance of information related to different material attributes and properties is also highlighted.

Hybrid Halide Perovskites.

Hybrid halide perovskites are a group of semiconductors made of metal halide octahedral frameworks and organic cations. Three-dimensional perovskites share the chemical formulas ABX_3, where A is a small organic cation, B is a divalent metal (e.g., Pb^{2+}, Sn^{2+}, $Cd^{2+,}$ or Cu^{2+}), and X = Cl, Br, or I (e.g., methylammonium lead iodide, $(CH_3NH_3)PbI_3$; Figure 1A) and adopt a lattice framed by octahedra BX_6^{4-},which are connected by sharing corners. The B-site may also be occupied by non-divalent cations (e.g., by equal amounts of Ag^+ and Bi^{3+}), as long as the overall charge balance is kept. For larger organic cations, the three-dimensional connectivity of the inorganic octahedra can be reduced to two, one, or even zero dimensions, with organic cations interspersed. The term "perovskite" still applies to these lower-dimensional crystal structures as long as the characteristic network of corner-sharing octahedra is retained. Hybrid perovskites show excellent structural tunability and remarkable optoelectronic properties, while simultaneously being easy to synthesize. [1,2] This combination has resulted in thousands of research publications in the last two decades, with promising applications in thin-film photovoltaics, superfluorescence, lasing, light-emitting diodes (LEDs), photodetectors, and spintronics. [3-9] Though they have successfully produced research output, their industrial footprint is only just emerging. Challenges of hybrid perovskites, such as poor environmental stability and compositional toxicity of some prominent compounds (e.g., those containing lead), might be to blame. [10] One way to overcome these challenges is to design modified or entirely new materials with well-selected properties. This step is still primarily driven by scientific intuition and requires trial-and-error cycles. The recent acceleration in utilizing artificial intelligence (AI) and machine learning (ML) in solving problems related to chemical and material sciences can make the material discovery step fast and more efficient. [11-15] This approach requires a collection of data that is reliable, accessible, and

machine-readable. The difficulty of applying these methods to solve problems in hybrid perovskite research is the limited availability of such databases.

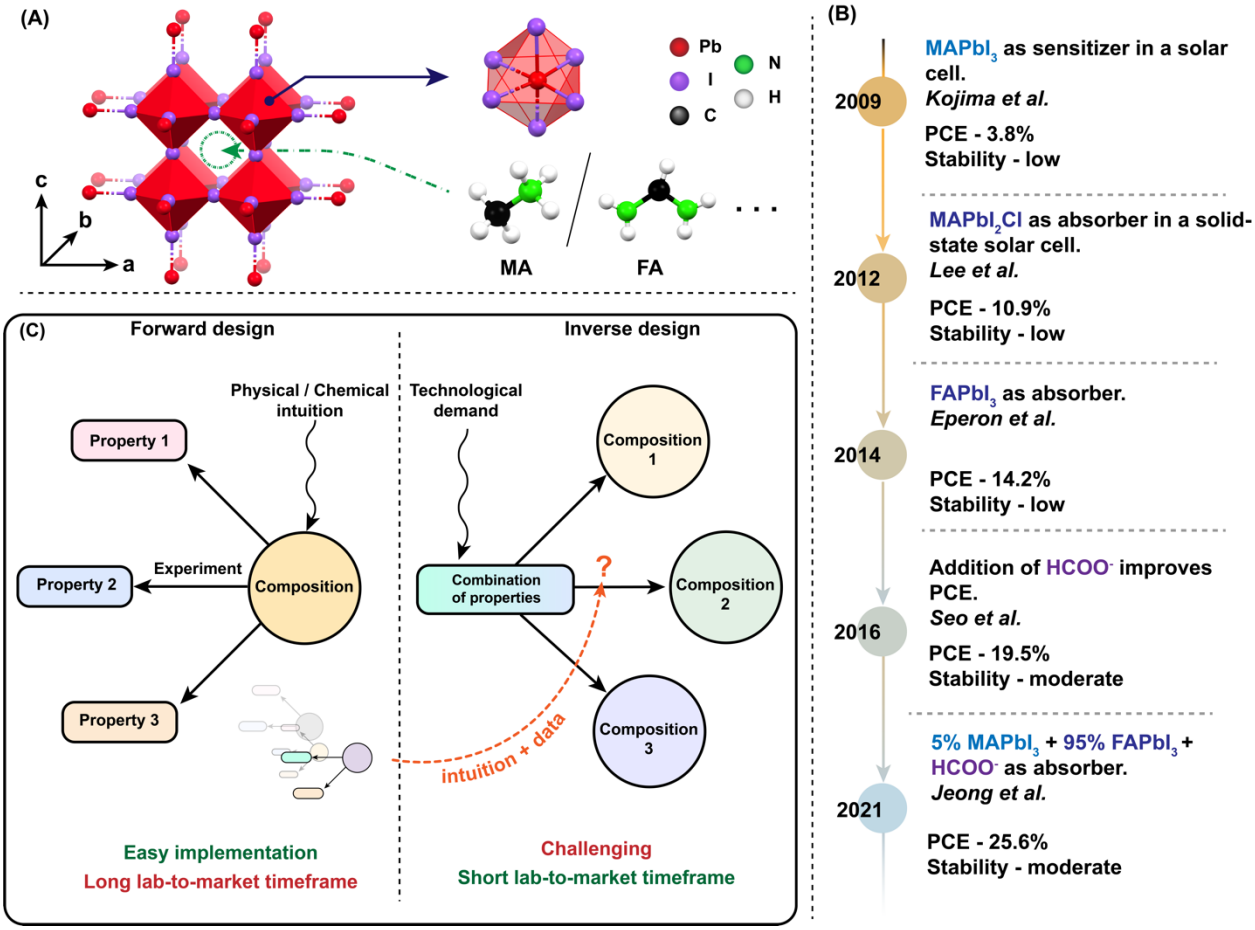

Figure 1. A) Idealized crystal structure of a three-dimensional (3D) perovskite framework with inorganic (Pb-I octahedra); organic cations such as methylammonium (MA) or formamidinium (FA) can fit into the cuboctahedral void. B) Timeline and a few selected developments in hybrid perovskite photovoltaics research. [3, 17-20] C) Schematic of material design approaches.

How can a database address material design challenges?

As photoabsorbers in solar cells, solution-processed hybrid perovskites have comparable power conversion efficiencies (PCEs) to those of more expensive crystalline silicon, making them a potential candidate for widespread adoption for renewable energy harvesting. [16] This success can be attributed to an intuitive design process. For example, Figure 1B shows five key reports in this direction. Perovskite solar cells became a research subject after 2009, when Kojima and

colleagues reported 3.8% photoconversion efficiency (PCE), using $MAPbI_3$ (MA: $CH_3NH_3^+$) as the sensitizer in a dye-sensitized solar cell (DSSC). [3] 12 years after, in 2021, Jeong and colleagues achieved 25.6% PCE from a stable perovskite thin-film cell using a combination of $MAPbI_3$ and $FAPbI_3$ (FA: $HC(NH_2)_2^+$) in the absorber layer and formate anions ($HCOO^-$) to passivate the surface defects. [17] Clearly, many modifications have happened in these 12 years, which can be traced back to hundreds of published reports. As three examples, in 2012, Lee and colleagues reported the first solid-state cell with $MAPbI_2Cl$ as the photoabsorber, improving the operational stability and PCE over the previously reported DSSC. [18] Then, in 2014, Eperon and colleagues improved the PCE to 14.2% by using $FAPbI_3$ in the absorber layer. [19] Then, in 2016, Seo and colleagues found that adding formate anions to the surface of perovskite cells improved their stability and performance. [20] These reports are not isolated but are the outcome of a continuous intuitive design process aimed toward gaining control over hybrid perovskites' physical and chemical properties, spanning over a decade and hundreds of research labs around the globe. [21] Tracking these rapid developments and utilizing the newfound knowledge is challenging.

The growth of perovskite solar cell PCE is astonishing; however, the material design challenges are far from over. Though the cells incorporating $FAPbI_3$ or $MAPbI_3$ have achieved appreciable PCE, they can deliver their peak performance for a shorter period than a silicon-based cell. [3,17] The problem is inherently linked to their low formation energies, owing to the ionic bonding between the organic and inorganic components. [22] The ionic bonding makes the crystallization of $FAPbI_3$ realizable at low temperatures; it also makes the material susceptible to degradation in the presence of polar solvents like water, which is abundant in the ambient air. [23] Even in the absence of any solvent, chemical reactions involving halides can form gaseous products that

escape, leaving a degraded material behind. [24] These contradictory desired properties make the intuitive design challenging and require significant modification of the chemical compositions to introduce new functionalities. This situation is not unique to perovskites alone; designing optimum compositions often requires decades of research and is the principal bottleneck in the journey of materials from research labs to the market. [25]

An alternative approach to material design, and perhaps, a solution to shorten the lab-to-market timeframe, is *"inverse design."* [26-28] Figure 1C schematically shows the difference between this process and a traditional (or forward) design approach. As a concept, inverse design is analogous to reverse engineering. The goal is to find materials with a desired combination of properties. This is achieved in two steps – 1) learning correlations between different properties and attributes of materials and 2) combining only the necessary attributes related to a target set of properties to obtain a new material. Though the concept of inverse design is old, modern AI-based methods provide a faster way to realize it. This is because any useful property's relationship with the attributes is usually complex and multi-dimensional, which statistical models are better at deciphering than humans. Importantly, to begin with, the process needs an ample information space where the correlations exist. The excellent structural and compositional tunability of the hybrid perovskites makes the inverse design approach particularly well-suited for the search for new and improved perovskite compositions. However, the hybrid perovskites information space is fragmented and exists in different publications, books, and materials databases. It is inefficient for an AI pipeline, where structured homogenized data is a crucial requirement. A curated hybrid perovskite database will make this space far more accessible.

What are the challenges in building databases?

If a large volume of data can be created from a single source, it is expected to be uniform and well-behaved. Such data are ideal for statistical analyses. High-throughput density functional theory (DFT) simulations have yielded excellent results in building large material databases. [29-32] These databases, in turn, have proven helpful in multiple discoveries in catalysis, phosphors, thermoelectrics, and battery research. [33-36] However, such high-throughput DFT databases are rare for hybrid perovskites, [49] possibly because of the associated high computational cost and manifold structural degrees of freedom, limiting the straightforward application of approximate DFT methods using a restricted number of structure prototypes, unless high-quality experimental crystal structures already exist. The complexities include modeling tens to many hundreds of atoms in the unit cell, an organic-inorganic interface, and strong relativistic modulation of electronic properties. [37-40] Also, for properties related to optoelectronic performances, e.g., electron-hole recombination probability, ion migration, carrier transport, and participation of defect states, the computational costs are significant even for building small-scale simulation datasets.

On the other hand, obtaining experimental materials data in a high-throughput fashion (high throughput experimentation; HTE) is a resource-intensive task. [41,42] In recent years, the use of AI in HTE is gaining attention because they hold great potential for automating the optimization process and decreasing the overall resources required. [43] As a result, multiple exciting developments in materials discovery and processing have been reported. [44-49]. For hybrid perovskites, HTE was recently used to build photoluminescence (PL) datasets, assess thin film quality, and synthesize new perovskite compositions. [50-53] These advancements hold great promise, but they still need to be expanded by the scope and types of data they can produce. It may take a few more years for these methods to get optimized at a scale where complicated experiments for structure or optoelectronic property determination could be reliably carried out.

What are the "important" data for hybrid perovskites?

The alternative to performing high-throughput data acquisition is to compile it from different sources. This approach is, in principle, the most desirable, since it would build on the full breadth of existing work. Conversely, it is bound to be slow because data collected from segregated sources will have different shapes, annotations, and associated errors. To make a collection of such data useful for comparison, rigorous work must be done to bring consistency. This makes it necessary to prioritize certain information over others while building a database. Thankfully, existing knowledge of the materials and the standing challenges can provide clear directions.

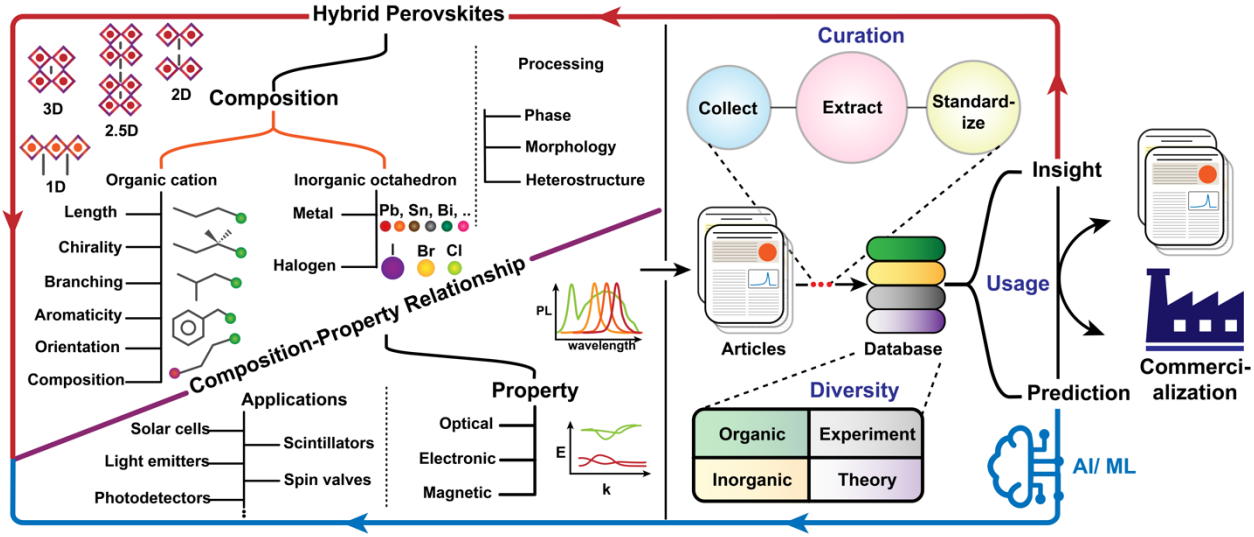

Figure 2. The left panel shows the compositional diversity and a few related critical applications of hybrid perovskites. The right panel shows the steps for building a database and its potential usage.

For example, the observed properties will always have a relationship with the arrangement of the atoms in a crystalline system, i.e., the "atomic structure." Understanding materials' structure-property relationship allows systematic alteration of desired properties. Additionally, in many cases, the composition can also be correlated with observed properties. [54] The composition of a material is easier to obtain and can be used to determine the structure itself. [55] This makes the

structures and compositions of hybrid perovskites the two most important assets for any data collection.

Figure 2 (left panel) provides a glimpse of the hybrid perovskite compositional space. For these materials, many additional attributes of the overall composition can be considered, which might be useful for discovering correlations. Though the original term “perovskite” refers to a crystal structure with a three-dimensional connected octahedral framework, the hybrid perovskite composition space has evolved over time to include many related structures in which the corner-sharing octahedral connectivity exists but does not, or not strictly, exhibit three-dimensional connectivity throughout the crystal. [56] Thus, current definitions include any crystalline material with a corner-sharing octahedral network as members of the extended family of hybrid perovskites. This includes structures with different metal and halides, M−X (M = Ge^{2+}, Sn^{2+}, Pb^{2+}, Ag^{+}, Bi^{3+}, $Sb^{3+}$and others; X = Cl^{-}, Br^{-}, I^{-}), and different cations, A (A = MA, FA, Cs^{+}, $Rb^{+,}$ and ammonium cations). The choices of these two sublattices determine the connectivity of the M−X octahedra or so-called "dimensionality" (e.g., 3D, 2D, quasi-2D, 1D, and 0D; D = *dimensional*) of the hybrid structure, specifically, of the interconnections of inorganic component (note that the resulting structures can still be realized as three-dimensional crystals in most cases). Also, other metal halides with octahedral motifs are often frequently grouped among "perovskite-like" materials, but the actual perovskite term is reserved for corner-sharing, not (for example) edge-sharing or face-sharing networks of octahedra.

The dimensionality and the M−X combination determine the electronic properties, such as the band gap, which is essential for semiconducting applications. [57] However, it often gets modified by the properties of the cation A. For the organic sublattice, the variations are near about endless; there exist cations that differ in length (e.g., butylammonium vs. hexadecyl ammonium) [58,59],

in chirality (e.g., R-2-methyl-butylammonium vs. 2-methyl-butylammonium) [7], in branching (e.g., iso-butylammonium vs. cyclobutylammonium) [60,61], in aromaticity (e.g., phenyl-ammonium vs. naphthyl-ammonium) [62,63], in composition (e.g., propylammonium vs. iodopropylammonium) [64,65] and in sometimes much greater complexity of incorporated, organic-semiconductor like conjugated cations. [39,66-68]

These differences also modulate other structural, optical, electrical, thermal, and magnetic responses of the hybrid structure. To name a few, the solubility of the cations determines the formability of a target perovskite phase; the dielectric constants of the sublattices influence excitonic properties; the conjugation in the cation molecule influence optical and transport properties; orientation, arrangement, and dynamics of the cations influence the phase stability of the material.

The properties can also be tuned by varying the synthesis methods. Depending on the experimental conditions, different structural morphologies or phases can be realized without changing the overall composition. For example, crystallite size and shape can be varied to obtain plate-like, cubic, octahedral, or dodecahedral morphologies that differ in stability and optical properties. [69]

Apart from these compositional differences, another essential piece of information is how they were obtained. This is because the experimental measurement of any property generally has an error associated with it, and a single property can be measured using different experimental methods with different associated errors. In many cases, the errors are not explicitly reported and are assumed to be understood by a reader. Additionally, factors such as sample state, temperature, and pressure can significantly impact experimental results.

How to build the database?

The information mentioned above is reported in published articles and requires careful collection, extraction, and standardization before incorporating into a database.

Article collection. With the rise in the use of big data in natural science research, several scientific journal publishers have started to provide application programming interfaces (APIs) to facilitate the automated collection of articles. [70,71] Given a few keywords for a targeted domain, these tools allow the automated download of a large number of articles in a short time.

Data extraction. From the collected articles, information needs to be extracted and repurposed for storage in the database. In contrast to a manual workflow, an automated alternative to this step could be faster, more efficient, and more affordable.

Published articles can be thought to be collections of tables, texts, and figures (see Figure 3A). For example, the presentation of a composition-property relationship can vary significantly with the form of display adopted in the article. There exist ML-based open-source software tools for automatic extraction. For example, natural language processing (NLP) and computer vision (CV) can be used to extract information from texts (paragraphs or tables) and figures, respectively. [72,73] Commercial language models and character recognition tools can also be used for these purposes. [74,75] However, extraction from formats that require context for interpretation, or have a certain degree of abstraction, such as texts or graphs, usually requires some additional supervised domain/subject-aware tuning for these methods to work efficiently.

FAIR principle. Extracted data are little impactful and usable as they are. It usually requires a few additional steps before they can be absorbed into a database and shared. These steps can be designed following the FAIR principle (Figure 3B). [76] FAIR stands for Findable, Accessible, Interoperable, and Reusable and provides a framework for making scientific data accessible and usable by the research community. Unique digital object identifiers (DOIs) can be assigned to

individual datasets to make data findable. This allows the data to be easily located and accessed by others. Data can also be made more accessible by building a user-friendly web interface or providing APIs that allow users to search for and retrieve the information they need. [31,77-79]

Data standardization. The interoperability between different datasets and systems can be achieved by standardizing data formats and metadata. When data is collected from multiple sources, it is common to have different shapes and attributes, even if it carries the same underlying meaning (Figure 3C). For example, the optical properties of a material, such as its absorption or emission spectra, can be expressed in intensity versus photon energy diagrams or intensity versus wavelength diagrams. These are essentially the same information, but they will have different values in one of the axes, making it difficult to compare between different sources.

To address this issue, data standardization involves defining standard formats and conventions for representing scientific data. This allows different datasets to be easily compared and combined and ensures that the data can be easily understood and interpreted by researchers and machines alike.

Date Dissemination. Dissemination of the collected data allows researchers to compare and build upon the work of others. To ensure long-term stability and access, it is important to carefully consider the licensing terms under which the data is shared. This can help ensure that the data is reusable and that the resources required to host, maintain, and curate the database are sustainable.

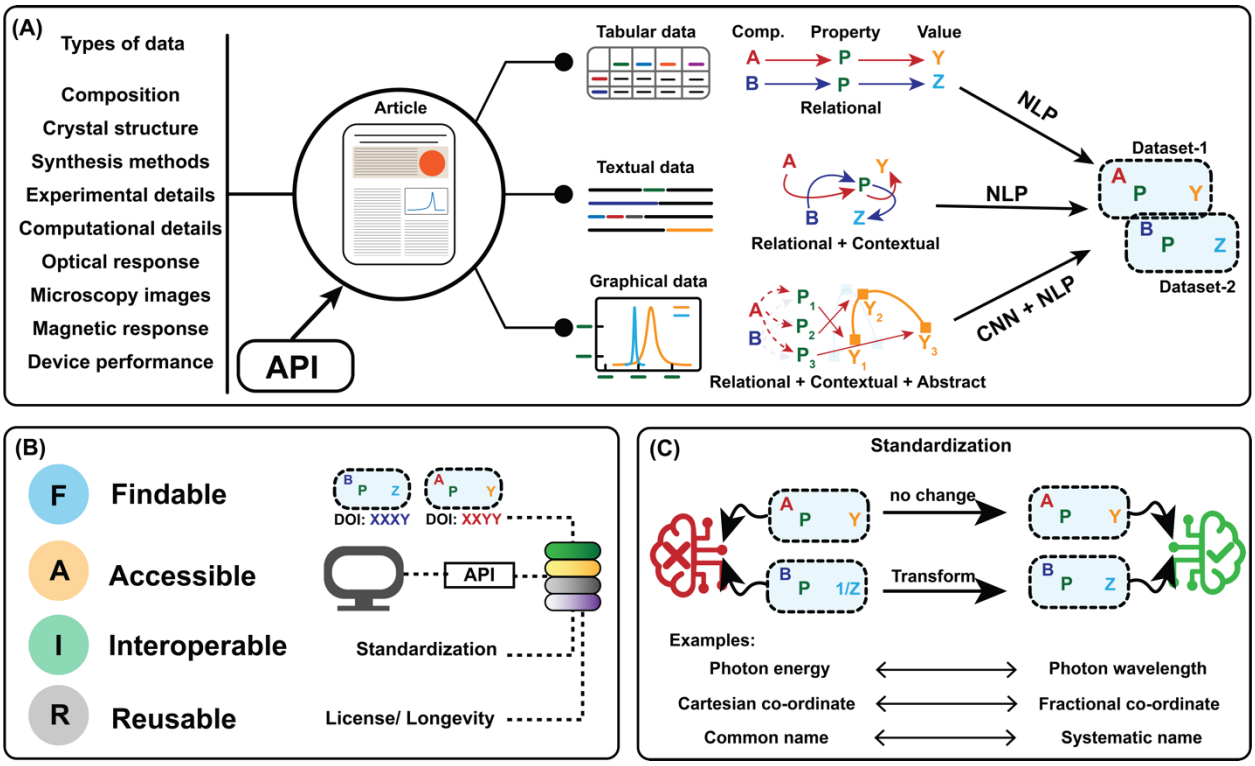

Figure 3. (A) Collection and extraction steps for different data types. Abbreviations used: API - Application Programming Interface; NLP - Natural Language Processing; CNN - Convolutional Neural Network. Schematics of (B) the FAIR principle and (C) Data standardization.

How to use databases efficiently?

In the last few years, there have been several efforts to find composition-property relationships in hybrid perovskites using ML. Table 1 lists a few examples where labeled data was used for this purpose. Such methods fall under the broad class of supervised learning (Figure 4A-top panel). A set of systems and measured properties are supplied to an ML model as input. A successful training process replicates the property values within an acceptable error limit. The model can then be used to predict properties of a system that could not be acquired by other means. A more detailed discussion of different ML models typically used in material sciences can be found in previously published reviews. [13,80]

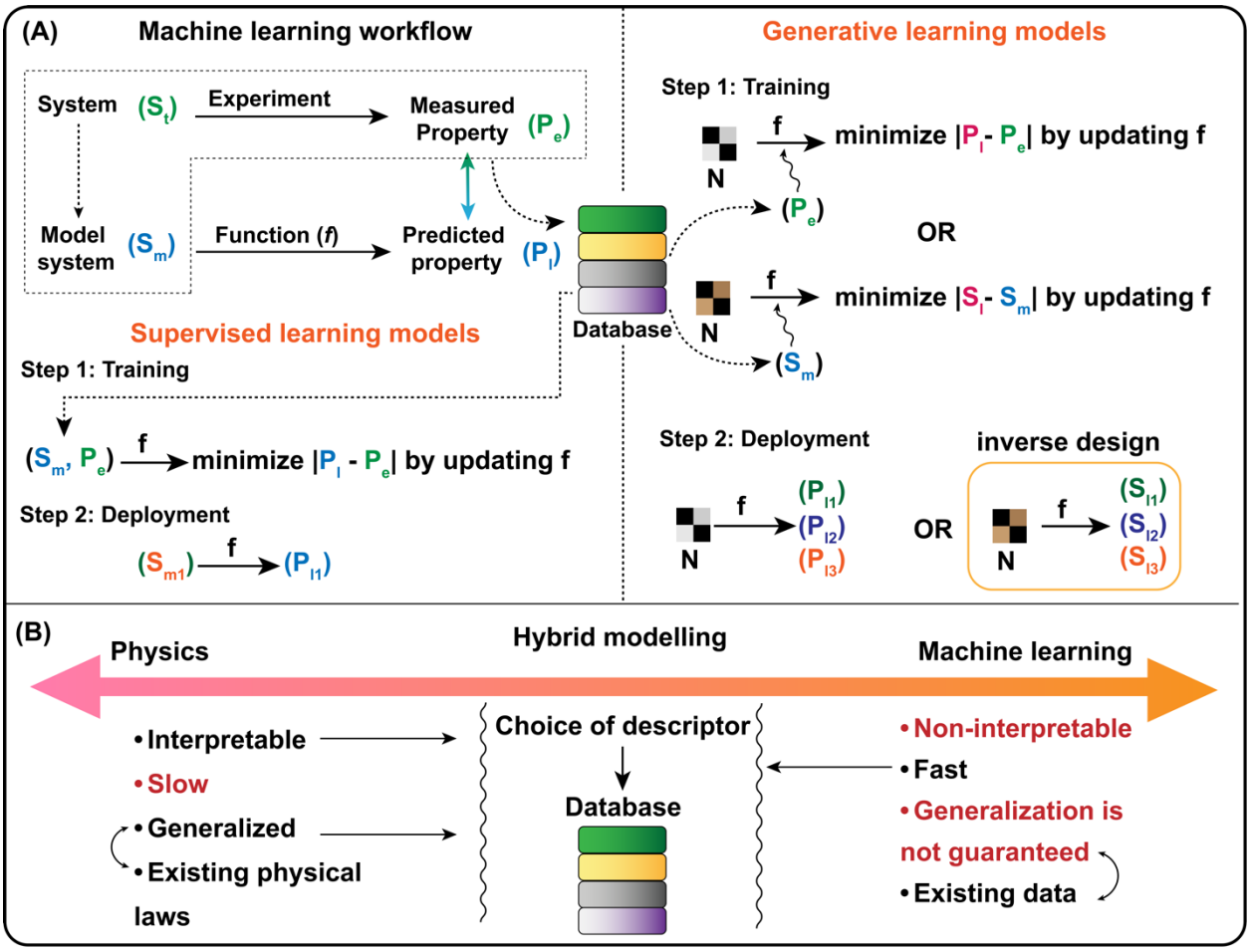

Figure 4. (A) Schematic showing general workflows for typical machine learning (ML, e.g., supervised models; bottom left panel, and generative models; right panel) and how databases can fit into it. The subscripts t, e, m, and, l stand for true, experimental, modelled, and learned. (B) The strategy of choosing descriptors for ML models.

Similar to computational methods like DFT, supervised learning requires a description of the system for initialization. Descriptors are numerical values that represent the systems. While for theoretical computations, some descriptors are well known and standardized, e.g., atomic structure or electron density, for ML methods, there is no standard descriptor, varying instead depending on the objective. Cases of simple descriptors for a given property are rare, although they exist. [7] More generally, multiple descriptors are pursued to describe the systems using experimental or

computational data. However, choosing appropriate descriptors that balance the trade-off between the model not being generalized and not being fast enough to get trained is crucial.

Figure 4B outlines a strategy for choosing descriptors. Generally, a hybrid approach that combines physical intuition and available data can be more beneficial than selecting all available information. [81] This is because though any observed property is expected to depend on several material attributes, the attributes that can be linked to the property through physical laws will have a larger impact. The availability of such descriptors in a database is expected to speed up the workflow.

For example, for quasi-2D perovskites (i.e., layered perovskites with inorganic layers that are thicker than a monolayer), the number of inorganic layers dictates the band gap and emission energy. [82] The distortions in the inorganic layer, induced by the organic cations, fine-tune the emission energy further. However, it is difficult to predict which combination of cations would result in which color output. Wang et al. attempted to tackle this problem using a probabilistic model with experimental composition-emission energy data. [83] Interestingly, they could also predict a set of compositions suitable for deep-blue emitters, which are relevant for blue-emitting diodes.

The presence of Pb in compositions like $MAPbI_3$ is regarded as a drawback for its toxicity. Consequently, replacing Pb with other less-toxic elements without compromising the optoelectronic properties is a standing challenge. [84] Though elements like Sn, Ge, Ag, Bi, In, etc. have shown some promise as possible substituents, the examples of such compounds are very few compared to the Pb-based ones. While it is well known that the ionic radii determine the formation of perovskite-like structures, empirical descriptors like the Goldsmith tolerance factor have a low success rate. Li et al. found a correlation between the ionic radii and formation energy

from a computational dataset generated by high-throughput DFT. [85] The trained classification model promises to provide better guidance for future experiments on discovering new lead-free compositions than the empirical parameters.

For lower-dimensional hybrid perovskites, it is known that the cation impacts the dimensionality of the inorganic sublattice. However, it is usually difficult to predict the outcome without carrying out the experiment and resolving the crystal structure. This is because organic cations are usually flexible and can orient themselves in many different ways during crystallization, resulting in different dimensionalities which may not be helpful for a specific application. Lyu et al. used a dataset of organic cations to train a probabilistic model to predict whether a specific cation will form a 2D structure. [86] Using a quantitative equation structured on the key features of the model, the authors could experimentally synthesize four new hybrid perovskite compositions with targeted dimensionality.

The relative humidity is known to be detrimental to the PCE of thin-film devices. The PL intensity, another optoelectronic response, is relatively easier to measure than PCE. Howard et al. found a correlation between PCE and relative humidity using a PL-relative humidity dataset. [87] The model could then be used to predict which films have higher moisture resistivity ahead of time and should be used for optoelectronic devices. Then, using an X-ray diffraction dataset, Oviedo et al. could predict dimensionality and space groups of perovskites from powder X-ray diffraction patterns. [88]

The PCE of solar cells depends on the band gap of the absorber material. A smaller band gap absorbs a higher fraction of the solar spectrum, resulting in higher PCE. Alloying at A, B, and X sites for ABX_3 composition has proven to be an effective strategy for optimizing the bandgap of 3D perovskites. However, the nonlinear behavior of the bandgap with changing composition

makes the ideal composition for a required bandgap difficult to predict. Li et al. used reported experimental data to train a neural network to predict compositions with an ideal band gap for a solar cell. [89] The same technique could then be used to predict the band gap of unknown compositions and, finally, to optimize other device parameters, such as the transport layers, for obtaining higher PCE. Also, similar attempts to optimize material composition for an ideal band gap were carried out by Marchenko et al. and Lu et al. for different types of hybrid perovskites. [90,91]

Among all these exciting advances, there remain many unexplored properties that are technologically important. A few of them are listed in Table 2. For example, broadband emission originating from self-trapped excitons is relevant for developing white-light-emitting LEDs. It is known that structural distortions influence the formation and stabilization of such trapped excitons. [92] However, designing specific materials with high quantum yields is challenging. Similarly, optimizing dopant concentrations for tuning electrical or emissive properties, [93] designing compositions for specific exciton binding energies, [94] high PL or electroluminescent quantum yields, [95] higher non-linear susceptibilities, [96], etc. remain yet to be explored. Since the physical phenomena are well understood, choosing descriptors is expected to take little effort.

Additionally, the application of generative models (Figure 4A-right panel), such as generative adversarial networks or diffusion models, has been limited in the hybrid perovskite research domain. [97-100] In these approaches, the goal is to generate an observation from an approximated information space rather than replicate the exact observation as seen in an experiment. Unlike supervised methods, they do not require labeled datasets. They might help in data augmentation, which can expand initial datasets and help in making a supervised follow-up model more generalized. These approaches can also enable the inverse design of hybrid perovskite

compositions for many other exciting applications. However, experimental verification of such prediction results will also need to occur. While models trained on small datasets can provide examples for proof-of-concept, generalization and accurate prediction of experimental outcomes would likely require much larger datasets that are relatively rare to come by.

Table 1. Use of data for predicting properties of hybrid perovskites. Abbreviations used: exp – experimental; sim – simulated; aug – augmented.

Problem	**Property**	**Compositional space**	**key feature**	**Number of datapoints in training set**	**Best performing model**	**Reference**
Search for blue emitter	Emission energy	Quasi-2D $L_2A_{n-1}Pb_nX_{3n+1}$ with $n \leq 5$	Cation composition	106 (exp.)	Random forest	[100]
Search for stable Pb-free perovskite	Decomposition energy	3D $A_2BB'X_6$	Composition	354 (sim.)	Kernel ridge regression	[102]
Cation selection strategy	Dimensionality	2D, 1D, 0D	Size, shape, aromaticity, number of H-bond donors	86 (exp.)	Decision tree	[103]
Stability of perovskite film	PL intensity	3D $MAPbI_3$ and $MAPbBr_3$	Relative humidity	5 (2 exp. + 3 aug.)	Recurrent neural network	[104]

Rapid structural analysis	Space group	3D, 2D, 0D with B = Pb, Sn, Sb, Bi, Ag, Cu	PXRD pattern	4279 (115 exp. + 164 sim. + 4000 aug.)	Convolutional neural network	[105]
Search for photovoltaic absorber	Photo conversion efficiency	3D ABX_3	Composition, Band gap, device structure	333 (exp.)	Deep neural network	[106]
	Band gap	Quasi-2D/2D	Number of layers, M—X—M angle	515 (exp.)	Gradient boosting	[107]
		3D ABX_3	Tolerance factor, octahedral factor, ionic charges, electron affinity, orbital radii	212 (sim.)	Gradient boosting	[108]

Table 2. A list of important properties relevant to the applications of hybrid perovskites and physically intuitive descriptors.

Unexplored property	**Potential descriptors for supervised learning**
Broadband emission	Cation structure, dimensionality, distortion factors, composition
Exciton binding energy	Composition, band gap, dimensionality, dielectric constants
PL quantum yield (PLQY)	Emission energy, dimensionality, morphology, film roughness
External quantum efficiency (EQE)	PLQY, film morphology, device structure, composition
Phase transition temperature	Composition, dimensionality, distortion factors, decomposition energy
Photon upconversion	Excitation energy, space group, composition

Where to find curated hybrid perovskite data?

There exist at least two general databases that are focused on hybrid perovskite data. The schematic in Figure 5A lists the different information available in these databases. For example, the perovskite database provides extensive solar cell device data collection. [21] The database hosts all important device-related parameters for more than 20,000 reported devices. The database is ideal for finding a correlation between PCE and other device parameters.

On the other hand, the HybriD^3 database (developed in our group) hosts structural, optical, and electrical characterization data of different hybrid perovskites. [101] The database currently hosts over 1500 datasets on 550 hybrid materials which are manually curated, and verified. The most

important limiting factor is the human curation of materials data, but simultaneously, the anticipated reliability of validated data would be ideal for learning composition-property relationships. The data collected in HybriD^3 database is also incorporated into the much broader, curated SpringerMaterials database, significantly enhancing its reach.

In addition, multiple specialized databases provide collection properties that might be relevant to hybrid perovskites. Some of these are listed in Table 3. For example, the Cambridge Structural Database (CSD), Inorganic Crystal Structure Database (ICSD), and Crystallography Open Database (COD) contain XRD-derived crystal structures of most of the reported hybrid perovskite compositions. [102-104] Importantly, CSD can be used with a Python-based API, and a subscription service for filtering and extracting crystal structure from the database. The organic material database (OMDB) contains data on the properties of many organic amines that are the precursor to the cations. [105] The SpringerMaterials and the Novel Materials Discovery (NOMAD) database provides experimental and computational data on many different classes of materials. [101, 106] Then, databases like Automatic FLOW library (AFLOWLIB), Open Quantum Materials Database (OQMD), and Materials project contain computational datasets of different semiconductor materials. [29-31]

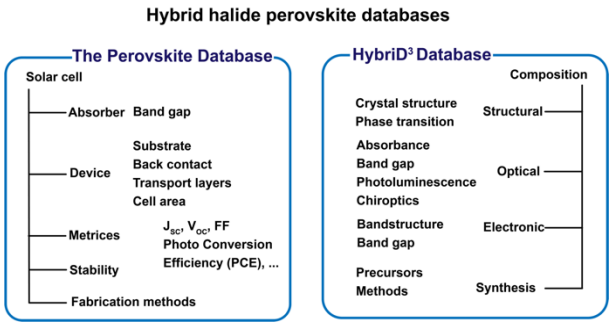

Figure 5. Types of data in existing curated hybrid halide perovskite databases – the perovskite database and HybriD^3 database. [21,101]

Table 3. Types of data relevant to hybrid perovskite research and databases

Type of data	Database name	Reference
Experimental crystal structure	Cambridge Structural Database (CSD)	[102]
	Inorganic Crystal Structure Database (ICSD)	[103]
	Crystallography Open Database (COD)	[104]
Organic amines' property	Organic Materials Database (OMDB)	[105]
Experimental and computational data on semiconductors and solids	SpringerMaterials database	[101]
	Novel Materials Discovery (NOMAD) database	[106]
Computational data on electronic/thermodynamic properties of semiconductors	The Materials Project database	[31]
	Open Quantum Materials Database (OQMD)	[30]
	Automatic FLOW LIBrary (AFLOWLIB)	[29]

Concluding Remarks.

Hybrid perovskite research is leading to immense amounts of high-quality materials research data and is poised to benefit from ML methods for materials discovery. The availability of optimized, curated databases that collect and disseminate this data will greatly enhance the value of hybrid perovskite research in the future. The recent progress of specialized tools for information extraction and database generation is promising. However, the lack of standardized rules for data

reporting makes automated extraction challenging. During the period of writing this article, the general large language processing models such as chatGPT and its successors became widely known and usable through OpenAI's web interface and API. It will be interesting to see if such a general model connecting massive amounts of information will be capable of subsuming the difficult problem of automated, high-quality materials data collection and dissemination (see Outstanding Questions). In the authors' current experience, the level of detail captured in scientific publications as yet still eludes the full grasp of generalized AI, but combining general AI with detailed curation may eventually and drastically help speed up data accessibility in the hybrid perovskite community and beyond.

Acknowledgments

The authors acknowledge support from SpringerMaterials through a funded project. Part of this research was supported by the National Science Foundation under award number 1729297.

References:

1. Weber, D. (1978) $CH_3NH_3PbX_3$, Ein Pb(II)-System Mit Kubischer Perowskitstruktur / $CH_3NH_3PbX_3$, a Pb(II)-System with Cubic Perovskite Structure. *Z. Naturforsch. B* 33, 1443-1445.
2. Saparov, B. and Mitzi, D.B. (2016) Organic-Inorganic Perovskites: Structural Versatility for Functional Materials Design. *Chem. Rev.* 116, 4558-4596.
3. Kojima, A. *et al.* (2009) Organometal Halide Perovskites as Visible-Light Sensitizers for Photovoltaic Cells. *J. Am. Chem. Soc.* 131, 6050-6051.

4. Lu, H. *et al.* (2020) Highly Distorted Chiral Two-Dimensional Tin Iodide Perovskites for Spin Polarized Charge Transport. *J. Am. Chem. Soc.* 142, 13030-13040.
5. Lei, L. *et al.* (2020) Efficient Energy Funneling in Quasi-2D Perovskites: From Light Emission to Lasing. *Adv. Mater.* 32, e1906571.
6. Kim, Y.H. *et al.* (2021) Chiral-Induced Spin Selectivity Enables a Room-Temperature Spin Light-Emitting Diode. *Science* 371, 1129-1133.
7. Jana, M.K. *et al.* (2021) Structural Descriptor for Enhanced Spin-Splitting in 2D Hybrid Perovskites. *Nat. Commun.* 12, 4982.
8. Biliroglu, M. *et al.* (2022) Room-Temperature Superfluorescence in Hybrid Perovskites and Its Origins. *Nat. Photon.* 16, 324-329.
9. Moon, J. *et al.* (2023) Metal-Halide Perovskite Lasers: Cavity Formation and Emission Characteristics. *Adv. Mater.*, e2211284.
10. Williams, S.T. *et al.* (2016) Current Challenges and Prospective Research for Upscaling Hybrid Perovskite Photovoltaics. *J. Phys. Chem. Lett.* 7, 811-819.
11. Kalinin, S.V. *et al.* (2015) Big-Deep-Smart Data in Imaging for Guiding Materials Design. *Nat. Mater.* 14, 973-980.
12. Agrawal, A. and Choudhary, A. (2016) Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Mater.* 4, 053208.
13. Cole, J.M. (2021) How the Shape of Chemical Data Can Enable Data-Driven Materials Discovery. *Trends Chem.* 3, 111-119.
14. Baird, S.G. *et al.* (2022) DiSCoVeR: A Materials Discovery Screening Tool for High Performance, Unique Chemical Compositions. *Digital Discovery* 1, 226-240.
15. Yao, Z. *et al.* (2023) Machine Learning for a Sustainable Energy Future. *Nat. Rev. Mater.* 8, 202-215.
16. Miyasaka, T. and Jena, A.K. (2021) Overview of Hybrid Perovskite Solar Cells. In *Hybrid Perovskite Solar Cells*, pp. 29-64,
17. Jeong, J. *et al.* (2021) Pseudo-Halide Anion Engineering for α-$FAPbI_3$ Perovskite Solar Cells. *Nature* 592, 381-385.
18. Lee, M.M. *et al.* (2012) Efficient Hybrid Solar Cells Based on Meso-Superstructured Organometal Halide Perovskites. *Science* 338, 643-647.

19. Eperon, G.E. *et al.* (2014) Formamidinium Lead Trihalide: A Broadly Tunable Perovskite for Efficient Planar Heterojunction Solar Cells. *Energy Environ. Sci.* 7, 982-988.
20. Seo, J.Y. *et al.* (2016) Ionic Liquid Control Crystal Growth to Enhance Planar Perovskite Solar Cells Efficiency. *Adv. Energy Mater.* 6, 1600767.
21. Jacobsson, T.J. *et al.* (2021) An Open-Access Database and Analysis Tool for Perovskite Solar Cells Based on the Fair Data Principles. *Nat. Energy* 7, 107-115.
22. Park, B.W. and Seok, S.I. (2019) Intrinsic Instability of Inorganic-Organic Hybrid Halide Perovskite Materials. *Adv. Mater.* 31, e1805337.
23. Leguy, A.M.A. *et al.* (2015) Reversible Hydration of $CH_3NH_3PbI_3$ in Films, Single Crystals, and Solar Cells. *Chem. Mater.* 27, 3397-3407.
24. Ciccioli, A. and Latini, A. (2018) Thermodynamics and the Intrinsic Stability of Lead Halide Perovskites $CH_3NH_3PbX_3$. *J. Phys. Chem. Lett.* 9, 3756-3765.
25. White, A. (2012) The Materials Genome Initiative: One Year On. *MRS Bulletin* 37, 715-716.
26. Ren, Z. *et al.* (2022) An Invertible Crystallographic Representation for General Inverse Design of Inorganic Crystals with Targeted Properties. *Matter* 5, 314-335.
27. Zunger, A. (2018) Inverse Design in Search of Materials with Target Functionalities. *Nat. Rev. Chem.* 2, 0121.
28. Lu, Z. (2021) Computational Discovery of Energy Materials in the Era of Big Data and Machine Learning: A Critical Review. *Materials Reports: Energy* 1, 100047.
29. Curtarolo, S. *et al.* (2012) Aflowlib.Org: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* 58, 227-235.
30. Saal, J.E. *et al.* (2013) Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *Jom* 65, 1501-1509.
31. Jain, A. *et al.* (2013) Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* 1, 011002.
32. Choudhary, K. *et al.* (2020) The Joint Automated Repository for Various Integrated Simulations (Jarvis) for Data-Driven Materials Design. *Npj Comput. Mater.* 6, 173.
33. Gorai, P. *et al.* (2017) Computationally Guided Discovery of Thermoelectric Materials. *Nat. Rev. Mater.* 2, 17053.
34. Yan, Q. *et al.* (2017) Solar Fuels Photoanode Materials Discovery by Integrating High-Throughput Theory and Experiment. *Proc. Natl. Acad. Sci. U. S. A.* 114, 3040-3043.

35. Li, S.X. *et al.* (2019) Data-Driven Discovery of Full-Visible-Spectrum Phosphor. *Chem. Mater.* 31, 6286-6294.

36. Shen, J.X. *et al.* (2020) A Charge-Density-Based General Cation Insertion Algorithm for Generating New Li-ion Cathode Materials. *Npj Comput. Mater.* 6, 161.

37. Even, J. *et al.* (2014) Understanding Quantum Confinement of Charge Carriers in Layered 2d Hybrid Perovskites. *Chemphyschem* 15, 3733-3741.

38. Whalley, L.D. *et al.* (2017) Perspective: Theory and Simulation of Hybrid Halide Perovskites. *J Chem Phys* 146, 220901.

39. Liu, C. *et al.* (2018) Tunable Semiconductors: Control over Carrier States and Excitations in Layered Hybrid Organic-Inorganic Perovskites. *Phys. Rev. Lett.* 121, 146401.

40. Jana, M.K. *et al.* (2020) Organic-to-Inorganic Structural Chirality Transfer in a 2d Hybrid Perovskite and Impact on Rashba-Dresselhaus Spin-Orbit Coupling. *Nat. Commun.* 11, 4699.

41. Shevlin, M. (2017) Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* 8, 601-607.

42. Akinc, A. *et al.* (2003) Parallel Synthesis and Biophysical Characterization of a Degradable Polymer Library for Gene Delivery. *J. Am. Chem. Soc.* 125, 5316-5323.

43. Eyke, N.S. *et al.* (2021) Toward Machine Learning-Enhanced High-Throughput Experimentation. *Trends Chem.* 3, 120-132.

44. Szymanski, N.J. *et al.* (2021) Toward Autonomous Design and Synthesis of Novel Inorganic Materials. *Mater Horiz* 8, 2169-2198.

45. Xie, Y. *et al.* (2021) Accelerate Synthesis of Metal-Organic Frameworks by a Robotic Platform and Bayesian Optimization. *ACS Appl. Mater. Interfaces* 13, 53485-53491.

46. Jenewein, K.J. *et al.* (2022) Automated High-Throughput Activity and Stability Screening of Electrocatalysts. *Chem Catalysis* 2, 2778-2794.

47. Bateni, F. *et al.* (2022) Autonomous Nanocrystal Doping by Self-Driving Fluidic Micro-Processors. *Adv. Intell. Syst.* 4, 2200017.

48. Moradi, S. *et al.* (2022) High-Throughput Synthesis of Thin Films for the Discovery of Energy Materials: A Perspective. *ACS Mater. Au* 2, 516-524.

49. Raccuglia, P. *et al.* (2016) Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* 533, 73-76.

50. Sun, S. *et al.* (2019) Accelerated Development of Perovskite-Inspired Materials Via High-Throughput Synthesis and Machine-Learning Diagnosis. *Joule* 3, 1437-1451.

51. Anwar, H. *et al.* (2022) High-Throughput Evaluation of Emission and Structure in Reduced-Dimensional Perovskites. *ACS Cent. Sci*. 8, 571-580.

52. Moradi, S. *et al.* (2022) High-Throughput Exploration of Halide Perovskite Compositionally-Graded Films and Degradation Mechanisms. *Commun. Mater.* 3, 13.

53. Epps, R.W. *et al.* (2020) Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. *Adv. Mater.* 32, e2001626.

54. Goodall, R.E.A. and Lee, A.A. (2020) Predicting Materials Properties without Crystal Structure: Deep Representation Learning from Stoichiometry. *Nat. Commun.* 11, 6280.

55. Jumper, J. *et al.* (2021) Highly Accurate Protein Structure Prediction with Alphafold. *Nature* 596, 583-589.

56. Mercier, N. (2019) Hybrid Halide Perovskites: Discussions on Terminology and Materials. *Angew. Chem. Int. Ed.* 58, 17912.

57. Pandey, M. *et al.* (2016) Band Gap Tuning and Defect Tolerance of Atomically Thin Two-Dimensional Organic-Inorganic Halide Perovskites. *J. Phys. Chem. Lett.* 7, 4346-4352.

58. Mitzi, D.B. (1996) Synthesis, Crystal Structure, and Optical and Thermal Properties of (C4h9nh3)2mi4 (M = Ge, Sn, Pb). *Chem. Mater.* 8, 791-800.

59. Billing, D.G. and Lemmerer, A. (2007) Synthesis, Characterization and Phase Transitions in the Inorganic-Organic Layered Perovskite-Type Hybrids $[(C_nH_{2n+1}NH_3)_2PbI_4]$, n = 4, 5 and 6. *Acta. Crystallogr. B* 63, 735-747.

60. Hoffman, J.M. *et al.* (2020) Long Periodic Ripple in a 2D Hybrid Halide Perovskite Structure Using Branched Organic Spacers. *Chem. Sci.* 11, 12139-12148.

61. Billing, D.G. and Lemmerer, A. (2007) Inorganic-Organic Hybrid Materials Incorporating Primary Cyclic Ammonium Cations: The Lead Iodide Series. *Crystengcomm* 9, 236-244.

62. Papavassiliou, G.C. *et al.* (1999) Preparation and Characterization of $[C_6H_5CH_2NH_3]_2PbI_4$, $[C_6H_5CH_2CH_2SC(NH_2)_2]_3PbI_5$ and $[C_{10}H_7CH_2NH_3]PbI_3$ Organic-Inorganic Hybrid Compounds. *Z. Naturforsch. B* 54, 1405-1409.

63. Du, K.Z. *et al.* (2017) Two-Dimensional Lead(Ii) Halide-Based Hybrid Perovskites Templated by Acene Alkylamines: Crystal Structures, Optical Properties, and Piezoelectricity. *Inorg. Chem.* 56, 9291-9302.

64. Hoffman, J.M. *et al.* (2019) From 2D to 1D Electronic Dimensionality in Halide Perovskites with Stepped and Flat Layers Using Propylammonium as a Spacer. *J. Am. Chem. Soc.* 141, 10661-10676.

65. Chakraborty, R. *et al.* (2021) Iodine–Iodine Interactions Suppressing Phase Transitions of 2d Layered Hybrid $(I\text{-}(CH_2)_n\text{-}NH_3)_2PbI_4$ (n = 2–6) Perovskites. *Chem. Mater.* 34, 288-296.

66. Mitzi, D.B. *et al.* (1999) Design, Structure, and Optical Properties of Organic-Inorganic Perovskites Containing an Oligothiophene Chromophore. *Inorg. Chem.* 38, 6246-6256.

67. Dunlap-Shohl, W.A. *et al.* (2019) Tunable Internal Quantum Well Alignment in Rationally Designed Oligomer-Based Perovskite Films Deposited by Resonant Infrared Matrix-Assisted Pulsed Laser Evaporation. *Mater. Horiz.* 6, 1707-1716.

68. Gao, Y. *et al.* (2019) Molecular Engineering of Organic-Inorganic Hybrid Perovskites Quantum Wells. *Nat Chem* 11, 1151-1157.

69. Dey, A. *et al.* (2021) State of the Art and Prospects for Halide Perovskite Nanocrystals. *ACS Nano* 15, 10775-10981.

70. https://www.springernature.com/gp/researchers/text-and-data-mining, Accessed on: 24th December, 2022.

71. https://dev.elsevier.com, Accessed on: 24th December, 2022.

72. Swain, M.C. and Cole, J.M. (2016) Chemdataextractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model* 56, 1894-1904.

73. Baibakova, V. *et al.* (2022) Optical Emissivity Dataset of Multi-Material Heterogeneous Designs Generated with Automated Figure Extraction. *Sci. Data* 9, 589.

74. https://developer.adobe.com/document-services/apis/pdf-services/, Accessed on: 24th December, 2022.

75. Polak, M.P. and Morgan, D. (2023) Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering -- Example of Chatgpt. *arXiv:2303.05352.*

76. Wilkinson, M.D. *et al.* (2016) The Fair Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* 3, 160018.

77. Laasner, R. *et al.* (2020) $MatD^3$: A Database and Online Presentation Package for Research Data Supporting Materials Discovery, Design, and Dissemination. *J. Open Source Softw.* 5, 1945.

78. https://materials.springer.com, Accessed on: 24th December, 2022.
79. Andersen, C.W. *et al.* (2021) OPTIMADE, an API for Exchanging Materials Data. *Sci. Data* 8, 217.
80. Butler, K.T. *et al.* (2018) Machine Learning for Molecular and Materials Science. *Nature* 559, 547-555.
81. Ouyang, R.H. *et al.* (2018) SISSO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Mater.* 2, 083802.
82. Stoumpos, C.C. *et al.* (2016) Ruddlesden–Popper Hybrid Lead Iodide Perovskite 2D Homologous Semiconductors. *Chem. Mater.* 28, 2852-2867.
83. Wang, W. *et al.* (2022) Predicting the Photon Energy of Quasi-2D Lead Halide Perovskites from the Precursor Composition through Machine Learning. *Nanoscale Adv.* 4, 1632-1638.
84. Ke, W. and Kanatzidis, M.G. (2019) Prospects for Low-Toxicity Lead-Free Perovskite Solar Cells. *Nat. Commun.* 10, 965.
85. Li, Z.Z. *et al.* (2019) Thermodynamic Stability Landscape of Halide Double Perovskites Via High-Throughput Computing and Machine Learning. *Adv. Funct. Mater.* 29, 1807280.
86. Lyu, R. *et al.* (2021) Predictive Design Model for Low-Dimensional Organic-Inorganic Halide Perovskites Assisted by Machine Learning. *J. Am. Chem. Soc.* 143, 12766-12776.
87. Howard, J.M. *et al.* (2022) Quantitative Predictions of Moisture-Driven Photoemission Dynamics in Metal Halide Perovskites Via Machine Learning. *J. Phys. Chem. Lett.* 13, 2254-2263.
88. Oviedo, F. *et al.* (2019) Fast and Interpretable Classification of Small X-Ray Diffraction Datasets Using Data Augmentation and Deep Neural Networks. *Npj Comput. Mater.* 5, 60.
89. Li, J.X. *et al.* (2019) Predictions and Strategies Learned from Machine Learning to Develop High-Performing Perovskite Solar Cells. *Adv. Energy Mater.* 9, 1901891.
90. Marchenko, E.I. *et al.* (2020) Database of Two-Dimensional Hybrid Perovskite Materials: Open-Access Collection of Crystal Structures, Band Gaps, and Atomic Partial Charges Predicted by Machine Learning. *Chem. Mater.* 32, 7383-7388.
91. Lu, S. *et al.* (2018) Accelerated Discovery of Stable Lead-Free Hybrid Organic-Inorganic Perovskites Via Machine Learning. *Nat. Commun.* 9, 3405.

92. Dohner, E.R. *et al.* (2014) Intrinsic White-Light Emission from Layered Hybrid Perovskites. *J. Am. Chem. Soc.* 136, 13154-13157.

93. Euvrard, J. *et al.* (2021) Electrical Doping in Halide Perovskites. *Nat. Rev. Mater.* 6, 531-549.

94. Dyksik, M. *et al.* (2021) Tuning the Excitonic Properties of the 2D $(PEA)_2(MA)_{n-1}Pb_nI_{3n+1}$ Perovskite Family Via Quantum Confinement. *J. Phys. Chem. Lett.* 12, 1638-1643.

95. Zhong, J.X. *et al.* (2020) The Rise of Textured Perovskite Morphology: Revolutionizing the Pathway toward High-Performance Optoelectronic Devices. *Adv. Energy Mater.* 10, 1902256.

96. Shen, W.L. *et al.* (2021) Nonlinear Optics in Lead Halide Perovskites: Mechanisms and Applications. *ACS Photonics* 8, 113-124.

97. Kim, S. *et al.* (2020) Generative Adversarial Networks for Crystal Structure Prediction. *ACS Cent. Sci.* 6, 1412-1420.

98. Gomez-Bombarelli, R. *et al.* (2018) Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* 4, 268-276.

99. Xu, M. *et al.* (2022) GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. arXiv.2203.02923

100. Dan, Y.B. *et al.* (2020) Generative Adversarial Networks (Gan) Based Efficient Sampling of Chemical Composition Space for Inverse Design of Inorganic Materials. *Npj Comput. Mater.* 6, 84.

101. https://materials.hybrid3.duke.edu, Accessed on: 12th April, 2023.

102. Groom, C.R. *et al.* (2016) The Cambridge Structural Database. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* 72, 171-179.

103. Bergerhoff, G. *et al.* (2002) The Inorganic Crystal Structure Data Base. *J. Chem. Inf. Comput.* 23, 66-69.

104. Grazulis, S. *et al.* (2009) Crystallography Open Database - an Open-Access Collection of Crystal Structures. *J. Appl. Crystallogr.* 42, 726-729.

105. Borysov, S.S. *et al.* (2017) Organic Materials Database: An Open-Access Online Database for Data Mining. *PLoS One* 12, e0171501.

106. Scheffler, M. *et al.* (2022) FAIR data enabling new horizons for materials research. *Nature* 604, 635.

Glossary.

Photo conversion efficiency (PCE): PCE is a measure of a photovoltaic cell's ability to convert light into electricity. It's calculated by dividing the electrical power output (in watts) by the incident light power (in watts/m^2). When expressed as a percentage, the formula is PCE = P_{out}/P_{in} * 100%.

Dye-Sensitized Solar Cell (DSSC): DSSCs are a type of thin-film solar cell comprising a porous layer of titanium dioxide nanoparticles coated with a photosensitive dye. They offer an efficient, low-cost alternative to traditional silicon solar cells.

Density functional theory (DFT): DFT is a quantum mechanical theory that is typically used to investigate the electronic properties of many-body systems, particularly atoms, molecules, and crystals, using computational calculations. It uses functionals, often approximated, of the electron density.

Machine learning (ML): ML is a subset of artificial intelligence that provides systems the ability to learn and improve from experience without being explicitly programmed. It involves the use of algorithms and statistical models to perform tasks by relying on patterns and inference instead of explicit instructions.

Photoluminescence quantum yield (PLQY): PLQY is a measure of the efficiency of luminescence (emission of light) after the optical excitation of a material. It's calculated as the ratio of the number of photons emitted to the number of photons absorbed. PLQY = (Photons emitted) / (Photons absorbed).

External quantum efficiency (EQE): EQE is a measure of the effectiveness of a device in converting incident photons into electrons (or current). It is expressed as the ratio of the number of charge carriers collected to the number of incident photons, often expressed as a percentage. EQE = (Charge carriers collected)/(Incident photons) * 100%.

Application program interface (API): An API is a set of rules and protocols for building and interacting with software applications. It defines methods of communication between various software components, enabling different software systems to interact with each other. In essence, it's a contract between different software components on how to interact.

Natural language processing (NLP): NLP is a field of artificial intelligence that focuses on the interaction between humans and computers using natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of human language in a valuable way. It involves techniques to convert human language into data that a computer can understand and process.

Convolutional Neural Network (CNN): CNN is a class of deep neural networks often used in image and video processing. They are designed to automatically and adaptively learn spatial hierarchies of features from the input data, utilizing layers of filters that scan input data for recognizable and often hierarchical features.

Data Augmentation in Machine Learning: Data augmentation is a strategy in machine learning that increases the diversity of data available for training models, without actually collecting new data. Techniques can include transformations like rotations or flips for image data, or synonym replacement and sentence shuffling in text data. This can help improve model robustness and performance.