Renewable and Sustainable Energy Reviews 107 (2019) 554–567

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: www.elsevier.com/locate/rser

Data mining new energy materials from structure databases

Lei Zhang[a,b,*], Zhiqiao Chen[a,b], Jing Su[a,b], Jingfa Li[a,b]

[a] Jiangsu Key Laboratory for Optoelectronic Detection of Atmosphere and Ocean, Nanjing University of Information Science & Technology, Nanjing 210044, China
[b] School of Physics and Optoelectronic Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China

ARTICLE INFO

Keywords:
Solar cell
Energy materials
Structure database
Data mining
Materials design

ABSTRACT

New energy materials that act as clean power sources and data science are developing rapidly in the past decades and the advancement of the two research areas have significantly benefited the development of each other. At the meantime, structural information of materials have been obtained and stored in various structure databases, such as the Cambridge Structure Database (CSD) and the Inorganic Crystal Structure Database (ICSD). Researchers have developed various structure-property relationships of the energy materials, which could be applied to screen the potential suitable materials from structure databases; this has become an efficient route to explore and design new energy materials. In this article, we review recent progresses on the data mining study of new energy materials based on structure databases such as CSD and ICSD, in the context of dye-sensitized solar cells and perovskite solar cells, and also include other energy systems such as water splitting systems, lithium batteries, thermoelectric devices and gas adsorbent materials. The structure descriptors that are more fundamental in the data mining procedure employing the structure-properties relationships are focused; the structural descriptors are complementary to the quantum descriptors and are efficient in the materials design process. We believe that with the successful formulation of more advanced and case-by-case structure-property relationships of energy materials, many new energy materials could be efficiently identified with much lower cost and shorter design period via the data mining process.

1. Introduction

There has been global energy crisis and environmental issues in the past decades due to the overuse of fossil fuels. New energy materials should be identified and developed to provide clean energy [1–5]. For example, efficient solar cell materials should be explored to harness the clean energy from the non-exhaustible solar system to provide the electricity; materials incorporated in water-splitting systems should be screened to provide solar fuels; lithium-based materials could be exploited to be tailored to undergo charge/discharge process reversibly and store electrical energies for portable devices and vehicles; metal-organic frameworks could be discovered to capture the carbon and gas molecules effectively; thermoelectric materials and piezoelectric materials should be identified to covert the thermal and mechanical energy into the electric current. At the current stage, the discovery of these energy materials relies predominantly on the experimental serendipity and the try-and-error experimental process that are inefficient and time-consuming. Nevertheless, many evidences have shown that the discovery process of new energy materials could be greatly accelerated by the data mining process, which have already shown their excellence in pharmaceutical drug discovery, finance, medicine, and marketing [6,7].

A large amount of structure information of new materials have been determined and stored in structure databases, which are mainly prepared by the crystallization processes and solved by the X-ray diffraction techniques and other crystallographic techniques [8]. Most of these materials have their structures accurately determined within the resolution of 0.01 Å and serve as outstanding platforms for the structural analysis and more advanced structure-property analysis. Two outstanding databases of materials structures have emerged: the Cambridge Structure Database (CSD) and the Inorganic Crystal Structure Database (ICSD), which focus on the accurate crystal structures of organic and inorganic materials, respectively [9,10]. These databases formed a good foundation for the data mining process of new energy materials. These databases store structures of materials that have already been synthesized in laboratory and exhibit significant advantages to expedite the process towards achieving promising candidates [11,12].

To help the fundamental understanding of the energy materials and facilitate the following-up materials engineering methods, a series of

* Corresponding author at: Jiangsu Key Laboratory for Optoelectronic Detection of Atmosphere and Ocean, Nanjing University of Information Science & Technology, Nanjing 210044, China.
E-mail address: 002699@nuist.edu.cn (L. Zhang).

https://doi.org/10.1016/j.rser.2019.03.036
Received 21 September 2018; Received in revised form 5 March 2019; Accepted 18 March 2019
1364-0321/

design principles based on materials structures 'structure-property relationships' have been developed. Many researches have been devoted to the formulation of the structure-property relationships of new energy materials to better understand the mechanisms involved. A bonus of the structure-property relationship study is that new energy materials with targeting properties could be efficiently identified from the structure database by proper algorithms that consider the structure-property relationship [13,14]. By simply inputting the algorithms that consider the structure-property relationships using specific codes, potential materials available for experimental validation can be retrieved from structure databases [15–21].

In this manuscript, we focus on recent progresses on data mining new energy materials exploiting the structure–property relationships from structure databases, especially CSD and ICSD, in the context of solar cells, water splitting systems, lithium batteries, thermoelectric devices and gas adsorbents. The structural descriptors are mainly focused, with the quantum descriptors also discussed to reveal the data mining process of the new energy materials. While we emphasize on the fundamental crystal structure databases such as CSD and ICSD, it should be noted that a number of derived databases exist such as those storing the calculated materials properties. The readers could be referred to the existing literatures reviewing the high-throughput calculations and derived databases [14,22–28].

2. Structural databases

2.1. Crystal structure databases

The CSD database presented by the Cambridge Crystallographic Data Centre (CCDC) starts from 1965 to represent the world's largest repository storing organic crystal structures; it provides a foundation for the data mining of organic materials-based energy devices [14,29]. The CSD includes both small organic molecules and metal organic frameworks (MOFs). The structure information is stored in a standard crystallographic information file (CIF) format which includes the crystal structure, the atomic positions, the bond length and the bond angles [30]. The packing mode including the space group and the symmetry elements are also available [31]. CSD now contains over 900,000 entries of accurate 3D structures from X-ray and neutron diffraction analyses [30,32–34]. User-friendly interface software tools are accessible for the detailed structural analysis [35]. The materials stored in CSD have been already synthesized and structurally characterized, but most of them have never been investigated for energy applications [28,36]. Therefore, new materials could be mined directly from the structure database either by a user-ready search interfaces such as *Conquest*, or writing a code with algorithms describing the searching criteria. For example, searches based on CSD have been conducted using ConQuest, performed with short contact analyses and packing diagrams [37].

In contrast with the organic material-based CSD database, the ICSD database contains the crystal structure information of inorganic compounds, with more than 199,000 crystal structure entries in June 2018 [38–40], and ca. 7000 new entries are added per year [16,41]. The ICSD-based data mining process could be coupled with the CSD database, where both organic compounds and inorganic compounds are desired [42].

Other crystal structure database other than CSD and ICSD include the powder diffraction database [43], the Protein Data Bank (PDB) [44–48], and PDB-derived databases such as PDBSite [49,50] and PhosphoSite [51], to name a few. These provide additional structure information on the new materials could be employed to design suitable materials for the targeting applications. In addition, it will not be surprising that with the development of the new devices that requires integration of the energy materials and the biological systems such as skin, the protein-based structure database might be very useful for the development of new energy materials in the future.

In addition to the experimental crystallographic techniques that are employed to obtain the CSD and ICSD databases, the in silico methods have also been used to predict the crystal structures of new materials, since in many cases the crystal structures are impossible to solve experimentally [52–55]. The crystal structure prediction (CSP) method has been widely performed to obtain the hypothetical structures [56–58]. The CSP method relies on the force-field method or the quantum mechanics to identify the proper positions of the neighboring molecules that interact stably with the central molecule to build up the crystal structure. In many cases, a proper space group should be set in advance. The CSP could be accelerated by evolutionary algorithms [53,59–61], R-group enumeration [62], and more advanced analysis [61]. In addition, the existing academic publications storing various structures or property data could form a natural basis for the data mining process [63]. Various techniques have been employed to data mine the existing publications, such as text mining, link mining, citation network analysis, adaptive neuro-fuzzy inference systems, neural networks and multilayer perceptrons [64,65].

2.2. Property databases

The property databases store property information that are closely related to the materials requirements, but are usually computation-intensive to prepare, The Materials Project is a well-known property database and stores the calculated materials properties based on the structures stored in ICSD [66]. The Materials Project database is prepared to identify the "materials genome" that resembles the human genome that could be conceptualized to represent the gene that dictates their properties and applications. The Materials Project contains structural, thermodynamic, electronic, optical and mechanical properties that are calculated by various materials simulation techniques [25,41,45,67–70]. For clarification purpose, Materials Projects and the related calculated databases based on CSD or ICSD are not focused in this study, but are discussed especially in the areas where the structure-property relationships of energy materials are not well-developed at the moment. Similar to the Materials project, other computed properties databases based on CSD or ICSD have been constructed such as Open Quantum Materials Database (OQMD) [38,71], ab-initio electronic transport database [72], MOF databases [11,73,74], JARVIS-DFT database [75], and the Harvard Clean Energy Project [76,77].

3. Data mining techniques

3.1. Structure descriptors and structure-property relationships

The structure-property relationships of energy materials could be coded via proper selections of structure descriptors for the data mining procedure in structure databases. The atomic structures that could be solved by the well-developed crystallographic techniques include the information of the structural parameters specified in terms of bond length, bond angle, conjugation, π…π stacking, intramolecular hydrogen bond (HB), intermolecular HB, molecular weight, packing mode and symmetry that are closely related to the optoelectronic properties of materials [36,62,78–83]. New materials could be mined directly from the structure databases either by a user-ready search interfaces such as *Conquest*, or a code incorporating algorithms describing these searching criteria related to the structure-property relationships.

The bond length is defined as the separation between the two covalently bonded atoms and the number of bonded electrons. The single/double/triple bonds defined by the bond length reveal the strength of interactions between the atoms and the charge transfer characters [84]. Based on the bond length values, the bond length alternation (BLA) values, the harmonic oscillator stabilization energy (HOSE) values, and the quinoidal structures could be determined, while BLA, HOSE and the quinoidal structures are quantitative descriptors to reveal the charge transfer characters [85,86]. The bond angles formed

between multiple atoms across at least two bonds include dihedral/torsion angles where four atoms are involved. The planarity is another structure parameter: and the geometry is planar if all the atoms are embedded on a particular plane [87,88]. Factors such as valency and electronegativity are easily obtained from the structures and have been applied to filter chemically implausible compositions [89].

The hydrogen bond is a moderate bond resulting from an electrostatic attraction between a proton in one molecule and an electronegative atom in the other. The hydrogen bond could be either intramolecular or intermolecular. Similar to the double-helix DNA chain, the hydrogen bond present in materials could greatly influence the structure geometry and stability as well as various properties [90–92]. The π…π stacking is another important structural parameter that dictates the crystal formation of many materials. The π bonds in the aromatic rings leads to intermolecular attractive interactions and result in the arrangements in a sandwich, T-shaped or parallel-manner. Unfortunately, there is a lack of unified theory that accurately quantifies the π…π interactions. Nevertheless, the π…π stacking shows evidence to strongly influence the intermolecular charge transfer and optical properties. The intermolecular interactions including the hydrogen bonds, the π…π stacking, halogen bond, van der Walls forces, etc. from various chemical motifs lead to a number of packing modes of crystals that dictates the related electronic/optical properties such as the emission properties and UV–vis absorption properties. The packing efficiency can be determined by the ratios of the total volume and the volume of the unit cell [93–96].

Certain functional groups are especially important for particular applications (vide infra); therefore a more efficient data mining procedure could include special chemical motifs. For example, the search and detailed analysis of the chemical substructure and ligand conformations have been performed to explore the chemical motif candidates [97]. The donor-π-acceptor motifs are especially important for the design of solar cells and organic light emitting diode (OLED), and these functional groups could be mined to design new energy materials [62].

Symmetry is considered to be critical for the crystal structure formation since the repeating units arrange and stack upon one another in particular ways such as the rotation or translation according to certain symmetry operations. The symmetry is employed to determine the crystal structures during the structural refinement [98–103]. Also, the symmetry of a crystal has been employed to calculate the thermal conductivity and optical properties. The CSD and ICSD databases contain many other structures information that are less studied such as cocrystals and disorder; these structure parameters could be explored in the future [104].

3.2. Quantum descriptors and high-throughput computation

Various quantum descriptors have been proposed to represent physical properties of materials calculated via the Schrödinger equations and related theories such as the band structure and the density of states. The descriptors calculated from the wave function theory include atomic charges, molecular orbital energies including highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), energy levels, frontier orbital densities, superdelocalizabilities, dipole moment, polarity indices, polarizabilities and stability [41,69,105–108]. The quantum descriptors are often considered to be more versatile since they better represent the properties, but they are most often more difficult to obtain and time-consuming compared with the structural descriptors. High throughput calculations, usually based on the first principles calculations, have been employed to calculate the above quantities of the materials stored in database, once their crystal structures are known. The calculations are usually followed by procedures to exclude inappropriate structures in CSD and ICSD such as to eliminate duplicates, disorders and inaccurate structures.

3.3. Other techniques

The computational cost is high for many materials discovery processes and various methods have been employed to speed up the calculation. Genetic algorithms and machine learning approaches have been employed to accelerate the process [62]. The machine learning technique that uses artificial intelligence automatically extracts predictive models from existing materials data. It has been used to extract meaningful chemical trends from training data [109], for example the prediction of solid-state properties with the local spin-density approximation results as a training set [110–112]. Genetics-based machine learning methods have shown scalability capacity which leads to their capability in the large-scale data mining jobs [113,114]. The machine learning models have been used to guide the materials researches such as solvent choice when crystallising a compound [115], soft materials engineering [116], crystal engineering [115], and classifying microstructures [56,117]. High-throughput calculations could be combined with the machine learning procedures for better materials prediction and validation [118]. Artificial neural networks (ANN) have been used for the prediction of the materials properties, on the basis of their dielectric and ionic properties [119,120].

The fast and efficient analysis of the chemical data is highly desirable [16,62,121,122]. Association analysis is used to reveal the patterns of the data [16]. The cluster analysis is used to discover the correlations of closely related groups [78]. Predictive modelling helps build models for targeted objectives as a function of input. Anomaly detection involves the opposite way by identifying data that differ from the normal observation [16]. The classical force field–based methods could be employed for larger systems. Monte Carlo simulations have been used to calculate the adsorption of methane in 650,000 structures [25,123]. Aggregation of experimental data from the published literature have been employed to allow the creation of interactive databases and additional metadata to be visualized involving energy materials [63,124]. Text mining and visualization tools have been applied to search publication databases [125] including papers, books, patents, and literature compilations [126,127]. Many efforts have been made to prepare databases that extract useful information from the pre-existing results in publications [14].

4. Examples of data mining energy materials from structure databases

4.1. Data mining dye-sensitized solar cell (DSSC) materials

The Dye sensitized solar cell is a new-generation solar cell that mimic the solar synthesis in the plant and utilize molecular chromophores that reside stably on semiconductor substrates to capture the solar energy and covert into the electrical power [128–142]. The DSSCs require specific dye structures such as the anchoring group and the D-π-A structure, i.e., the dyes should have an electron donor, a conjugate π bridge and an electron acceptor. The anchoring moiety is usually an cyanoacrylic acid group. The structural aspects including the chemical substitutions, the conjugation, planarity and bond length alternation (BLA) correspond to the charge transfer characters of the DSSC-active dyes [143].

A series of dye structures that are incorporated in DSSC photoanodes have been solved by the X-ray crystallography [144–147]. Ultra-strong X-rays that are produced by the synchrotron have been used to determine the crystal structures of dyes that present weak diffraction signal and previously structurally unattainable [148–152]. The crystal structures of a series of polyoxotitanate clusters have been solved crystallographically to understand the binding modes of dyes on the semiconductor substrates, with their electronic and optical properties determined upon the modification with the dopants and the ligand structures. Apart from the single-crystal X-ray diffraction techniques, the neutron diffraction, the terahertz spectroscopy, the powder

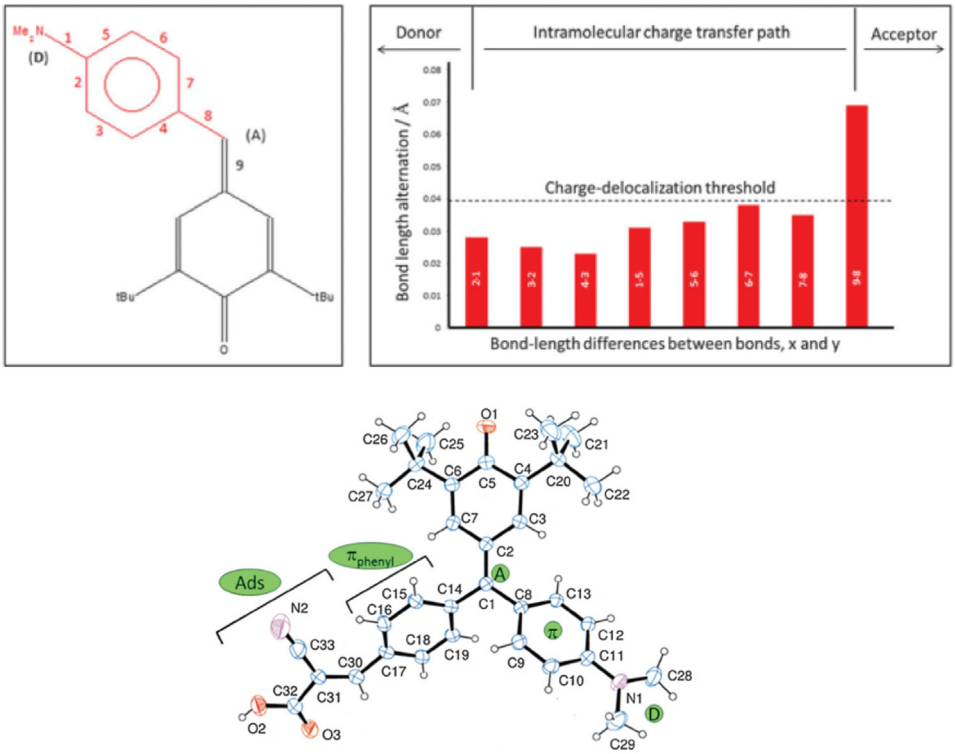

Fig. 1. The bond length analysis involving the bond length alternation (BLA) to filter DSSC-active dye molecules from the CSD database (top) and the new dye identified using the data mining procedure [157]. Reproduced with permission from Ref. [157]. Copyright RSC (2014).

diffraction, NMR, FTIR, as well as other crystallographic and spectroscopic techniques have been involved to understand the structural origin of DSSCs either directly or indirectly [153–155].

Apart from the structure-relationships study on the experimental-proven dyes, the data mining approach has been used to reveal entirely new classes of DSSC dyes, based on the structural analysis of DSSC-active molecular dyes that exhibit certain congruent bond length pattern. Cole et al. employ graph theoretical algorithms and classification tests to discover DSSC-active dyes, using the recursive depth-first, backtracking and graph traversal algorithms (Fig. 1) [156]. The CSD database is crucial since this database contains potential DSSC-active dyes that could be utilized to predict new dyes [157]. In particular, chemical groups and molecular architectures with the D-π-A-phenyl-A backbone, that commonly features in a well-performing DSSC dye have been spotted and translated into the generic design principles for the data mining process in CSD. The crystallographic R1 smaller than 0.07 threshold is used for data mining to ensure the accuracy, with all disordered materials, polymers and ionic salts excluded [157]. The screening process involves the BLA analysis, the anchoring group position, the dipole moments calculation and the Hammett constants calculation [158]. A simulated 'molecular mutation' with the addition of a cyanoacrylate group is performed on the 'top 10' molecular candidate to ensure good electron withdrawing capabilities. Therefore, the screening process relies heavily on the structural descriptors of the DSSC-active dyes.

Apart from the predication of new functional dye molecules, the data mining techniques have also been realized on the photocathode material in a p-type DSSC by Moot et al. [159]. After screening inorganic compounds in ICSD, they identified new photocathode materials such as the lead titanate ($PbTiO_3$), a perovskite material, as the promising photocathode material. The search procedure includes the chemical similarity parameters used in cheminformatics, Tanimoto similarity coefficient, and starts from the known p-type photocathodes including NiO, Co_3O_4, Cu_2O, CuI, $CuAlO_2$, $CuGaO_2$, $NiCo_2O_4$, and $ZnCo_2O_4$ as reference query materials [159].

Since the bond lengths stored in the structure databases are closely related to the intramolecular charge transfer properties in dye molecules, analysis has been carried out on retrieving the bond lengths of different organic dyes to understand the intramolecular charge transfer characteristics [160,161]. For example, the bond length values of the azo group stored in CSD are compared to reveal the efficient charge transfer characters in azo dyes toward the DSSC application [162]. Identifying suitable chemical fragments is important for the DSSC materials design. Donor groups have been screened [163], and DSSCs with novel structures have been obtained [164,165]. Phenothiazine-based dyes have been assembled from fragments in a synthetically tractable manner, with predicted PCEs over 9% [166]. In order to help users analyse the DSSC structure data more efficiently, a DSSC database "DSSCDB" has been generated to afford researchers with information from literatures on the detailed dye structures including triphenylamines, carbazoles, coumarins, phenothiazines, ruthenium and porphyrins [167].

4.2. Data mining perovskite solar cell materials

Perovskite solar cells have their device structures originating from dye-sensitized solar cells, and the perovskite solar cell research could be dated back to 2009 when Miyasaka [168] et al. substitute dye molecules in DSSCs with halide perovskites, and found a power conversion efficiency above 3%. By changing the liquid electrolyte into the solid state, as well as additional optimization of the perovskite crystal formation process, the power conversion efficiencies of perovskite solar cells rocket to 22% [169–200]. Despite of the astonishing development on power conversion efficiency, the origin of the superiority of the halide perovskite materials as well as the detailed mechanisms happening inside the crystal environment or at the interfaces still remains elusive. For example, the halide perovskite structures for solar cell applications have been accurately solved except the disordered cation molecules, and the exact roles of the methylammonium cation are debatable [201,202].

There have been researches carried out to data mine the new halide perovskite materials and similar materials that can replace the traditional lead halide perovskites in solar cells, which suffer from the lead contamination and instability issues [69,159,203–205]. The new materials that are mined from structure database should possess the small effective masses of electrons and holes, the long charge diffusion length, outstanding optical properties and small charge recombination possibilities, all of which could be calculated bases on the ICSD databases. Perovskites could have multiple organic and inorganic constituents (CH_3NH_3, $HC(NH_2)_2$, Cs, Rb, Pb, Sn, I, Br), and various new structures with layered, double, and metal-deficient perovskites, which naturally give flexibility for rational strategies to replace the current experimental trial-and-error-based approaches [203].

The ICSD database has been successfully exploited to identify potential lead-free hybrid perovskites, with the stoichiometry as the starting point to identify possible lead free perovskites. The halide moiety is critical in the functioning of the halide perovskite materials. Using the CSD database, the strength and geometric preferences of non-covalent interactions in the halide polymers have been investigated [206–214]. The structures of the halide polymer with different dimensions could be used for data mining new halide perovskites for solar cells. The 3D halide perovskites are not unique structures for the perovskite solar cells. Replacement of lead with di-, tri- and tetra-valent cations has been performed, and iodosalts, dimers, layered structures, and mixed and milti-dimensional perovskites have been employed (Fig. 2); 12 A^{1+} (e.g., $MA^+/FA^+/Cs^+$), $27M^{2+}$, $35M^{3+}$, and $25M^{4+}$ site cations could be chosen; three halogens (I^-, Br^-, Cl^-) and a mixture of these three elements could be chosen. For example, the divalent element could be Ge, Sn, Bi, Ag and In. The tetravalent metallic cations such as Sn in Cs_2SnI_6 has been tried, while the double perovskites ($A_2{}^+M^+M^{3+}X_6$) uncover the opportunity of including 3^+ cations into the perovskite structures [203]. Taking into consideration of the cationic valence states and volume ratios, possible perovskites combining entries from ICSD databases and chemical intuition are > 24,138 for AMX_3, > 31,290 for $A_3M_2X_9$, > 22,350 for A_2MX_6, and > 9×10^6 for $A_2MM'X_6$ [203].

Efforts have been made to find that ionic radii, tolerance factor, and octahedral factor are the important structural descriptors for the perovskite halides using a support vector machine classification model. With the training model, several novel perovskite structures have been predicted [203,215,216]. Mechanically stable and electronically suitable electrons and holes extracting contacts materials for $CH_3NH_3PbI_3$ have been selected according to the electron affinity, ionisation potential, lattice parameters and crystal structure criteria. Potentially effective new hole and electron transporting layers according to these descriptors include Cu_2O, FeO, SiC, GaN, and ZnTe [205].

4.3. Data mining other solar cell materials

Similar to those in DSSCs and perovskite solar cells, proper descriptors are needed to identify candidate solar cell materials. Starting from the ICSD database, alternative polar chalcogenides have been predicted to substitute the widely studied CIGS material [41]. In the study, the stoichiometry with the AB_2CX_4 chemical formula is targeted, where A = Li, Cu, Ag; B = Zn, Cd; C = Ga, In and X = O, S, Se, Te, leading to a family of 48 members [41]. The criteria could be strong optical absorption coefficients, a proper bandgap value such as ~1.3 eV satisfying the Shockley–Queisser limit, low cost and compatibility with existing technologies. A set of high SLME materials are identified based on ICSD, including the best already known thin-film solar absorbers, such as $CuInSe_2$, $CuGaSe_2$ and $CuInS_2$, while materials different with the 1:1:2 stoichiometry (for example, Cu_7TlS_4, Cu_3TlS_2 and Cu_3TlSe_2) are also identified [217]. Other approaches involving ordinary least squares, sparse partial least squares, elastic net/least absolute shrinkage, selection operator regression methods coupled to rough set and principal component analysis methods have been developed for the band gap prediction of new chalcopyrite compounds [218]. A data mining workflow have been developed and applied to the analysis of solar cell libraries based on titanium and copper oxides [26]. Constituent elements have been focused to find new thin film solar cell materials. Thermodynamic stability, electrical transport, electronic structure, optical and defect properties have been evaluated to predict the Cu_2SnS_3 as the promising photovoltaic materials [219].

High-throughput calculations are most often used to identify new solar cell materials [28,220]. The Materials Project database that provides relaxed structures and related optoelectronic properties are naturally selected as the starting point to identify solar cell materials. For example in the area of the transparent p-type oxide, 3600 quaternary oxides in the Materials Project database are searched and the materials with a small effective mass are obtained according to their band structures [221]. Phosphides and zinc blende boron phosphide BP are found to be promising p-type transparent conducting materials (requiring large band gap for transparency and low hole effective mass for high mobility) by exploiting the weak absorption for indirect optical transitions [222]. The CSD database stores a large amount of organic structures that could be utilized to design organic electronics. Apart from the low carrier mobilities in organic semiconductors, the electronic couplings and intramolecular reorganization energies are identified to be the two main descriptors for charge mobility, and materials with long-range charge percolation pathways are identified [28]. In the ICSD, for every nitride (3008 total entries) there are more than ten oxides (41,529 total entries) [223]. As a consequence, new nitride materials with structural motifs and crystal structures that have never been reported previously i are discovered with good regularity [223]. Similar electronic structure databases and initiatives include the JARVIS-DFT database [75] and the Harvard Clean Energy Project (CEP) toward the new organic solar cell materials [76].

4.4. Data mining water splitting materials

Water splitting materials employing the photocatalyst to harvest the solar energy and produce hydrogen fuel from the water provides an effective alternatives to fossil fuels [176,224–228]. The photocatalysts that split water have been searched from ICSD [229,230]. For example, by analysing the bulk metals and the most stable single- and bi-metal oxides in ICSD and the Materials Project databases, new water splitting materials are proposed [169], based on the descriptors of appropriate band gap, band edges relative to the water redox levels, high mobilities, and chemical stability under light irradiation [169]. 10 oxides and 5 oxynitrides including $AgNbO_3$, $BaSnO_3$, $BaTaO_2N$, $CaTaO_2N$, $SrTaO_2N$, and $LaTiO_2N$ are identified as the candidates for light harvesting materials, which agree with the literature [169]. Oxynitrides and compounds containing d10 cations (Ga^{3+}, In^{3+}, Ge^{4+}, Sn^{4+}, Sb^{5+}, and Bi^{5+}) or d0 cations (Ti^{4+}, Zr^{4+}, Hf^{4+}, V^{5+}, Nb^{5+}, Ta^{5+}, Cr^{6+}, Mo^{6+}, W^{6+}, Sc^{3+}, and Y^{3+}) are targeted in ICSD. To solve the problems of insufficient oxynitrides materials available in the ICSD, some new candidates are proposed to be proper photocatalysts, including three binary nitrides, two ternary oxynitrides and eleven quaternary oxynitrides [229]. The detailed procedure starting from the ICSD structure search and the follow-up electronic property calculations, including the energy level position match and the band gap screen, are summarized in Fig. 3.

In order to identify potential water splitting materials, calculations have been performed using the Open Quantum Materials Database (OQMD) [204]. 139 materials are identified as potential new candidates for thermochemical water splitting (TWS) application, including $CeCoO_3$ and $BiVO_3$ which are not observed previously, using the large data set of compounds containing stabilities, oxidation states, and ionic sizes [204]. High-throughput screening of materials based on the electronic band gap calculations have been performed, which leads to the discovery of five candidates as the promising new water splitting material [231]. The oxygen evolving reaction (OER) activities of

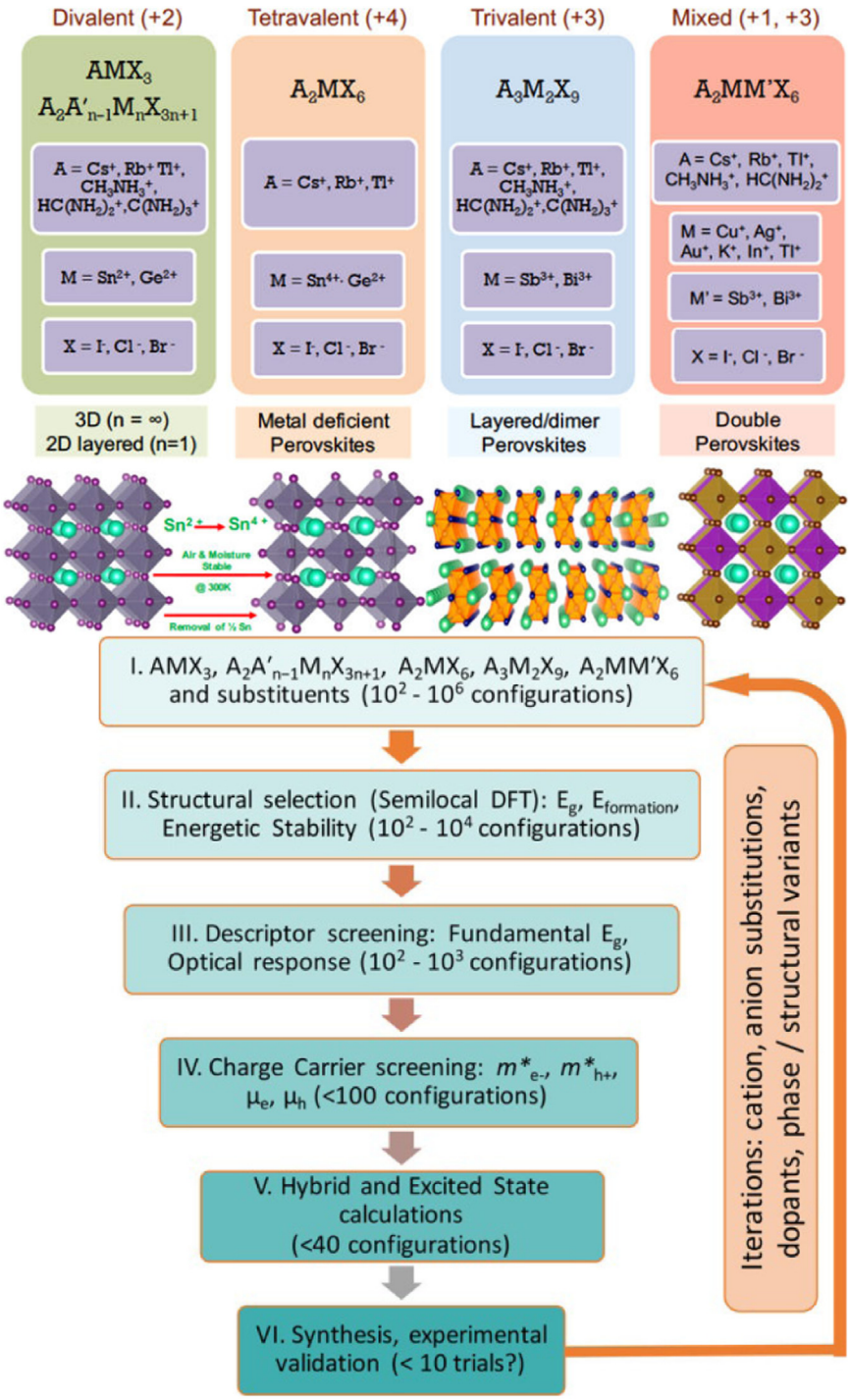

Fig. 2. Perovskite structure with different valence and elements (top) and rational design flowchart starting with the ICSD database and subsequent computational screening and experimental validation (bottom) [203]. Reproduced with permission from Ref. [203]. Copyright ACS (2017).

perovskites are explored according to the descriptors of the metal–oxygen bond strength using factor analysis and linear regression models. Electron occupancy and metal–oxygen covalency are identified as the dominant influences on the OER activity, and multiple descriptors required for better prediction [232]. The associations between composition and catalytic activity are examined for OER for 5429 catalyst compositions in a (Ni–Fe–Co–Ce)O_x library, based on informatics-based clustering of composition property functional relationships [233].

4.5. Data mining organic light emitting diode materials

The data mining process of the OLED molecules based on the structural descriptors are not well-developed compared with the quantum descriptors. Nevertheless, considering the structural similarity in terms of the D-π-A backbone between the OLED materials based on molecules and the DSSC dyes, a similar data mining process borrowing from the DSSC concept could be employed in the future to identify new OLED functional molecules. Compared with the DSSC materials, the OLED materials have larger molecular weight and do not require anchoring groups. The search of OLED molecules could also be based on

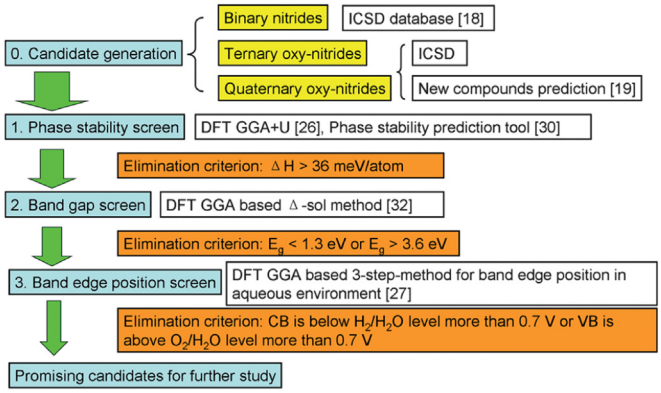

Fig. 3. The searching procedure of water-splitting photocatalysts starting from ICSD [229]. Reproduced with permission from Ref. [229]. Copyright RSC (2013).

completely new fragments. According to the HOMO and LUMO positions and the optical properties, a new set of fragments have been screened [62].

4.6. Data mining thermoelectric materials

The thermoelectric materials generate electricity from thermal energies. Various descriptors have been tried to search the potential thermoelectric materials. Thermoelectric power factors of > 3000 inputs in ICSD are compared [69,234], while a metric to provide a consistent evaluation of the inherent physical properties that dictate the thermoelectric figure of merit zT is proposed [235,236]. After searching for new thermoelectric materials in the ICSD based on the 570 Sb-containing compounds, Zintl compound LiZnSb are suggested as potentially interesting n-type thermoelectric materials [237]. Other descriptors include Lorenz number that lead to different zT enhancement, and the computational search for materials with exceptionally low Lorenz number and high thermoelectric quality factor has been carried out [24]. Based on over 100 publications, databases storing thermoelectric materials have been created [238], which includes 18,000 data points with associated properties measured at several temperatures. The interplay between data mining, informatics, and machine learning approaches allow for a new paradigm in thermoelectric materials development [63,239,240]. A data-driven approach is accomplished by understanding discrete scalar descriptors regarding the crystal and electronic structure and the Curie temperature, and quantitatively exploring the different materials descriptors.

4.7. Data mining lithium battery materials

The Materials Genome Project coupled to ICSD has been data mined to design lithium battery materials where the electrodes consist of inorganic compounds [15,238]. The data-driven approach has proved to be successful to the rational design of Li-ion battery materials [241], since the energy density, power density, discharge capacity, lithiation potential, capacity retention upon cycling and many others could be accurately calculated for the compounds in ICSD [220]. The organic part in the lithium battery such as the electrolyte solvent has also been screened; the preferred electrochemical stability, melting and boiling temperatures, viscosity, dielectric constant, Li-ion conductivity, Li^+ diffusion, electronic conductivity ranges and activation barriers have been calculated in the presence of the organic molecules [62,242,243].

4.8. Data mining gas capture and storage materials

Data mining techniques have been employed to search the gas capture and storage materials, which are predominantly based on

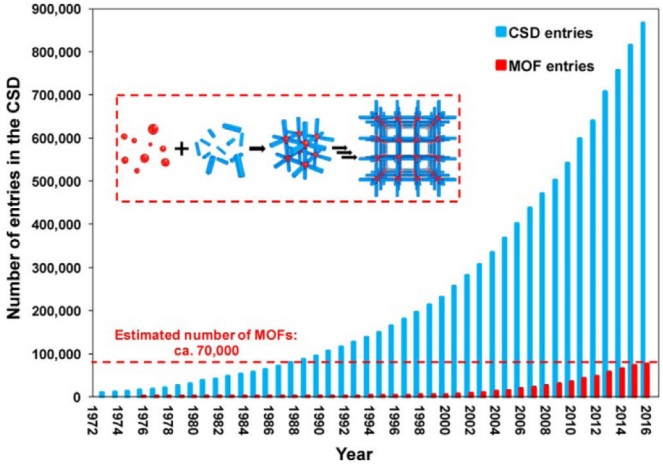

Fig. 4. The increase of MOF entries since 1972 in comparison with the CSD entry. The inset shows the structures of the MOF that is self-assembled by the following moieties: metals (red) and organics (blue) [250]. Reproduced with permission from Ref. [250]. Copyright ACS (2017). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

metal-organic frameworks (MOF) owing to their structural integrity and porosity for molecular adsorption [97,244–248]. These MOFs are mainly for carbon capture and storage, gas separation, as well as hydrogen fuel storage [12,73,74]. The MOFs are essential hybrid systems that are built from the blocks and frameworks consisting of either organic or inorganic moieties. Since CSD stores various synthesis-ready MOF crystal structures and the synthesis of MOF is not straightforward, screening MOFs from the synthesis-ready CSD databases has been found efficient in the design process [11,12,249]. In CSD, the MOF entries have risen up to 6000 in CSD in 2011, and risen to 70,000 in CSD in 2016 (Fig. 4), which means the CSD database is a natural database selection for the carbon capture and storage as well as the hydrogen storage [250].

In order to identify available MOF structures in CSD, researchers have searched for structures with bonds between metallic elements and organic elements, while additional structural screening procedure include the solvent molecules, varying degrees of disorder, missing H atoms and overlapping atoms [22,73,74]. Databases with MOFs free from solvents/disorders and with a pore limiting diameter (PLD) larger than 2.4 Å structures have been provided; the optimal pore size that maximizes net capacity such as 6–10 Å and void fractions of 0.1–0.5, have been set to the optimal criterion for the MOF search [12].

Potential substitutes with possible ligands has been selected according to the chemical substructure searching and ligand conformational analysis [97]. Pores larger than 3.25 Å for CO_2 adsorption are selected based on the possible zeolite-like SiO_2 structures starting from 230 space groups, various unit cell dimensions and densities [251]. Regarding the oxygen uptake, the relationships between structural properties and oxygen adsorption performance at dissimilar pressures have been studied in five dimensions, and UMCM-152 that delivers 22.5% more oxygen has been obtained [249]. NU-12528 is identified from 10,000 hypothetical MOFs and confirmed by experimental synthesis and adsorption measurements [252]. An efficient approach automated assembly of secondary building units (AASBU) has been developed [253], with the building blocks randomly distributed at points, and provides insights for the topological preferences [22]. The methane uptake in over 650,000 materials has been investigated starting from both existing and predicted nanoporous materials [123]. The propylene/propane adsorptive separation have been investigated regarding their selectivity and working capacity, via structural descriptors such as the N_2 surface area, accessible surface area of propane, and PLD. Porous materials that includes elements such as In, Te, Al, and

I, along with the low LCD stipulation has been proposed owing to high crossover distance [254]. In addition, neural networks and grand-canonical Monte Carlo simulations, coupled with Lorentz-Berthelot mixing rules and Lennard-Jones parameters have been employed for the H_2 storage and methane adsorption [12,123]. Apart from CSD, a number of databases for MOFs are available such as Hypothetical MOF Database, MOF-5 Analogues, Reticular Chemistry Structure Resource and porous materials databases including Zeo++, Poreblazer, MOFomics, TOPOS, CoRE MOF and MOFIA [11,22,251].

4.9. Data mining other materials

The data mining technique has generated new insights in various new materials [78,255]. Piezoelectric materials convert mechanical energy into electricity, and the ICSD database has been examined to find previously unrecognized piezoelectric systems from half-Heusler semiconductor compounds, with 987 candidates scanned in terms of the structural, dielectric, insulating and piezoelectric properties [256]. Previously unknown two-dimensional materials have been identified starting from CSD and the first principles calculations [69,257]. A novel data mining algorithm involving the dimensionality of weakly bonded subcomponents has been employed in the Materials Project database, and 325 materials are predicted to exhibit piezoelectric properties [78]. Topological insulators have been data mined by examining the changes in band inversion in the presence or absence of spin-orbit coupling [258]. Stable Dirac-point node are predicted via the combined group theory and data mining approach using the Organic Materials Database, and particular space groups such as *P*212121 is established to be positive to the Dirac nodes construction [259]. Data mining using density of states similarity search in OMDB is performed to identify novel organic High-Tc Superconductors [260]. The relative stability and the relative solubility of cocrystals has been investigated by a data mining force field in CSD [104]. Soft-matter materials is also assisted by the data-driven approaches [261,262].

5. Suggestions and outlook

Although the structure descriptors are sometimes considered to be lower-level quantities than the quantum descriptors, they offer a more direct and faster data mining process than the quantum descriptors to identify new energy materials, while the quantum descriptors often requires more expensive computational resources. The data mining procedures rely heavily on the structural details of energy materials.

Finding a proper algorithm and structural descriptor is critical for the efficient screening. Data mining new energy materials from the CSD and ICSD databases can be accelerated using proper structural descriptors, which describe the crystal structures that could be solved by the well-developed crystallographic methods. The structural parameters such as bond length, bond angle, conjugation, π…π stacking [36,78–82], intramolecular/intermolecular hydrogen bond, molecular weight [62], packing mode and symmetry [83] are closely related to the targeting applications. There can be hundreds of structure descriptors representing certain properties of the materials, and the individual structure descriptors could be artificially combined to form new structure descriptors, owing to the newly developed machine learning techniques. Many experimental researches have given insights on the chemical motifs that could be particularly popular in certain type of devices. For example, the chemical substructure searching, the donor-π-acceptor motif probing, and the ligand conformational analysis are effective to select potential materials. Particular chemical moieties such as amine, triphenyl amine and cyanoacrylic acid group are well-correlated with certain applications. The bond length and angles, the intermolecular interactions, stoichiometry, hydrogen bonds (HBs), intermolecular distance, BLA pattern, quinoidal structures, packing modes, halogen bond and space group should be explored further in relation to particular applications. Many crystal structures of proteins are stored in databases such as the Protein Data Bank (PDB) [44–48] and PDB-derived databases, which might provide additional structure information on the new biological materials, such as the DNA- and protein-based energy devices. It will not be surprising that with the development of the new devices that requires integration of the energy materials into the biological systems such as the skin, the protein-based structure database might be very useful for the development of new energy materials in the future.

Some of the design procedures of energy materials follow an interesting “Lego” type style where suitable chemical subgroups with desired quantities such as the Hammett values are combined in a particular way [263,264]. However, the Hammett values alone only vaguely estimate the electron donating/withdrawing abilities of the materials and many Hammett values of new functional groups are unavailable [86,265]; therefore improvements in the Hammett value determination and more advanced descriptions other than the Hammett values are called for. It would be desirable to develop parameters to better quantify the electron/withdrawing abilities for particular applications.

While most of the structural details have been revealed from the structure-property relationships investigation in the literature, some structural impacts still remain elusive at the moment; for example, the disorder (happening in halide perovskite solar cells) and cocrystals that are different from the crystals of pure components that are considered to be useless previously might be re-investigated [104].

The ever-increasing data mining studies from the structure databases require more crystallographic contributions that are dedicated to obtain the accurate structure information. X-ray crystallography continues to be the undisputed primary technique [266] for picturing molecules and assemblies with the utmost accuracy compared with the in-silico crystal structure prediction, because the structure prediction of even the simplest crystalline solids only starting from the composition is still fundamentally questionable. For example, in many case a possible space group should be set in advance for the structure prediction, which might not be correct and brings out inaccuracies and inconsistencies [69,267,268]. Databases storing materials properties from the Schrödinger equations still depends on the accurately crystal structures inputs in databases, and the crystal structure determination and the structure-property relationships formulation are considered be of paramount importance to realize better design of energy materials in the future.

Since the excited states and the time-resolved dynamics are particularly important for many energy conversion applications, it could be very useful to obtain the time-resolved structure of energy materials with both atomic and electronic structures determined in the time-domain, such that new mechanisms could be revealed and new materials could be mined from the more accurate structure-properties relationships based on the time-resolved structures. To this end, there are evidences showing that the time-resolved structures of energy materials are achievable via the time-resolved crystallography techniques, which provides the ultra-strong X-rays produced by the synchrotron to determine the crystal structures of new materials with weak signal that are previously unattainable [148–152,269–282]. For example, the time-resolved structures (in nanosecond or picosecond) of several silver- and copper-based organometallic compounds have been reported are reported to be closely related to the optoelectronic properties of the materials especially the UV–vis absorption and the light emission [149], which are critical factors for the solar cells and OLEDs. The extent of the displacement in angstrom of the atoms and the electrons upon the excitation are suggested to be directly associated with the electronic and optical properties, with the larger displacement leading to stronger emission [150,151,270,277,283]. Nevertheless, how the time-resolved structures are related to the energy-related device performance quantitatively remain elusive at the moment, considering the fact that many processes related to the electron transfer and light absorption happen in different time-scales [102,284–287]. In

addition, the energy gap, defects, and the energy level alignment significantly affect the time-domain activities of the energy materials. A formulation of the time resolved structure-property relationship would be thus highly desirable (for instance the differences of the structural changes in femto-, pico-, nano-, micro and milli-second on the optoelectronic properties) to guide the materials design, which emphasize the importance of the accurate determination of materials structures in various timescales using the ultra-strong X-rays or neutron pulses.

6. Conclusions

Data mining energy materials from the structure databases such as CSD and ICSD have been facilitated by the formulation of proper structure-property relationships, and successful algorithms coded with the structural descriptors that consider the structure-property relationship have been rapidly developed to facilitate the data mining process. The structural descriptors could be the chemical moiety, bond length and angles, intermolecular interactions, stoichiometry, valency, elements, ionic substitutions, surface area, pore diameter, HBs, halogen bonds intermolecular distance, BLA pattern, quinoidal structures, stoichiometry, packing modes and space groups, etc. There can be hundreds of structure descriptors representing certain properties of the materials, and the new structure descriptors could be designed via the machine learning techniques to obtain more proper structure descriptors. The data mining study of new energy materials from the crystal structure databases such as CSD and ICSD has been found to be applicable in solar cells, water splitting systems, lithium batteries, thermoelectric devices, piezoelectric materials and gas adsorbent materials. Continuous efforts should be spent on the basic understanding of structure-property relationships of new energy materials, since a bonus of the structure-property relationship study is that new energy materials with targeting properties could be identified quickly and accurately from the structure database by algorithms that include the structure-property relationships. It is believed that with more crystal structures obtained from the X-ray and neutron crystallographic techniques, and a more unified theory of structure-property relationship of energy materials toward particular applications, a Google-like approach based on the data mining process could be realized in the future to design energy materials.

CRediT authorship contribution statement

Lei Zhang: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Zhiqiao Chen:** Data curation, Formal analysis. **Jing Su:** Funding acquisition, Resources. **Jingfa Li:** Funding acquisition, Resources.

Acknowledgment

This work was supported by the Natural Science Foundation of China (No. 51702165), the Jiangsu Provincial Natural Science Foundation (Grant No. BK20160942 and BK20160941) and the Six Talent Summit Project of Jiangsu Province of China (Grant R2016L07). The authors acknowledge computational support from NSCCSZ Shenzhen, China.

Author contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

References

[1] Burlingame Q, Coburn C, Che X, Panda A, Qu Y, Forrest SR. Centimetre-scale electron diffusion in photoactive organic heterostructures. Nature 2018;554:77–80. https://doi.org/10.1038/nature25148.
[2] Pomerantseva E, Gogotsi Y. Two-dimensional heterostructures for energy storage. Nat Energy 2017;2:17089. https://doi.org/10.1038/nenergy.2017.89.
[3] Jeon NJ, Noh JH, Yang WS, Kim YC, Ryu S, Seo J, Il Seok S. Compositional engineering of perovskite materials for high-performance solar cells. Nature 2015;517:476–80. https://doi.org/10.1038/nature14133.
[4] Armand M, Tarascon J-M. Building better batteries. Nature 2008;451:652–7. https://doi.org/10.1038/451652a.
[5] Grätzel M. Photoelectrochemical cells. Nature 2001;414:338–44. https://doi.org/10.1038/35104607.
[6] Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manag 2015;35:137–44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.
[7] Shaw MJ, Subramaniam C, Tan GW, Welge ME. Knowledge management and data mining for marketing. Decis Support Syst 2001;31:127–37. https://doi.org/10.1016/S0167-9236(00)00123-8.
[8] Helliwell JR. Science of crystal structures: highlights in crystallography. Crystallogr Rev 2017;23:231–6. https://doi.org/10.1080/0889311X.2017.1295039.
[9] Glaser R, Nichols GR. Conformational preferences and pathways for enantiomerization and diastereomerization of benzyl alcohol. data mining and ab initio quantum-mechanical study. J Org Chem 2000;65:755–66. https://doi.org/10.1021/jo991423q.
[10] Allen FH. High-throughput crystallography: the challenge of publishing, storing and using the results. Crystallogr Rev 2004;10:3–15. https://doi.org/10.1080/08893110410001664864.
[11] Chung YG, Camp J, Haranczyk M, Sikora BJ, Bury W, Krungleviciute V, Yildirim T, Farha OK, Sholl DS, Snurr RQ. Computation-ready, experimental metal–organic frameworks: a tool to enable high-throughput screening of nanoporous crystals. Chem Mater 2014;26:6185–92. https://doi.org/10.1021/cm502594j.
[12] Thornton AW, Simon CM, Kim J, Kwon O, Deeg KS, Konstas K, Pas SJ, Hill MR, Winkler DA, Haranczyk M, Smit B. Materials genome in action: identifying the performance limits of physical hydrogen storage. Chem Mater 2017;29:2844–54. https://doi.org/10.1021/acs.chemmater.6b04933.
[13] Kajikawa Y, Takeda Y, Matsushima K. Computer-assisted roadmapping: a case study in energy research. Foresight 2010;12:4–15. https://doi.org/10.1108/14636681011035726.
[14] Hill J, Mulholland G, Persson K, Seshadri R, Wolverton C, Meredig B. Materials science with large-scale data and informatics: unlocking new opportunities. MRS Bull 2016;41:399–409. https://doi.org/10.1557/mrs.2016.93.
[15] Ceder G. Opportunities and challenges for first-principles materials design and applications to Li battery materials. MRS Bull 2010;35:693–701. https://doi.org/10.1557/mrs2010.681.
[16] Rajan K. Data mining and inorganic crystallography. In: Hofmann DWM, Kuleshova LN, editors. Data mining in crystallography Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 59–87. https://doi.org/10.1007/978-3-642-04759-6_3.
[17] Apostolakis J. An introduction to data mining. In: Hofmann DWM, Kuleshova LN, editors. Data mining crystallography Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 1–35. https://doi.org/10.1007/978-3-642-04759-6_1.
[18] Ghosh BN, Lahtinen M, Kalenius E, Mal P, Rissanen K. 2,2′:6′,2″-terpyridine trimethylplatinum(IV) Iodide complexes as bifunctional halogen bond acceptors. Cryst Growth Des 2016;16:2527–34. https://doi.org/10.1021/acs.cgd.5b01552.
[19] Strelcov E, Belianinov A, Hsieh YH, Chu YH, Kalinin SV. Constraining data mining with physical models: voltage- and oxygen pressure-dependent transport in multiferroic nanostructures. Nano Lett 2015;15:6650–7. https://doi.org/10.1021/acs.nanolett.5b02472.
[20] Isayev O, Fourches D, Muratov EN, Oses C, Rasch K, Tropsha A, Curtarolo S. Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chem Mater 2015;27:735–43. https://doi.org/10.1021/cm503507h.
[21] Meredig B, Wolverton C. Dissolving the periodic table in cubic zirconia: data mining to discover chemical trends. Chem Mater 2014;26:1985–91. https://doi.org/10.1021/cm403727z.
[22] Colón YJ, Snurr RQ. High-throughput computational screening of metal–organic frameworks. Chem Soc Rev 2014;43:5735–49. https://doi.org/10.1039/C4CS00070F.
[23] Mounet N, Gibertini M, Schwaller P, Campi D, Merkys A, Marrazzo A, Sohier T, Castelli IE, Cepellotti A, Pizzi G, Marzari N. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. Nat Nanotechnol 2018;13:246–52. https://doi.org/10.1038/s41565-017-0035-5.
[24] McKinney RW, Gorai P, Stevanović V, Toberer ES. Search for new thermoelectric materials with low Lorenz number. J Mater Chem A 2017. https://doi.org/10.1039/C7TA04332E.
[25] Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science. APL Mater 2016;4. https://doi.org/10.1063/1.4946894.
[26] Yosipof A, Nahum OE, Anderson AY, Barad HN, Zaban A, Senderowitz H. Data mining and machine learning tools for combinatorial material science of all-oxide photovoltaic cells. Mol Inform 2015;34:367–79. https://doi.org/10.1002/minf.201400174.
[27] Boyd PG, Lee Y, Smit B. Computational development of the nanoporous materials genome. Nat Rev Mater 2017;2. https://doi.org/10.1038/natrevmats.2017.37.
[28] Schober C, Reuter K, Oberhofer H. Virtual screening for high carrier mobility in organic semiconductors. J Phys Chem Lett 2016;7:3973–7. https://doi.org/10.

1021/acs.jpclett.6b01657.
[29] Hofmann DWM. Data mining in organic crystallography. In: Hofmann DWM, Kuleshova LN, editors. Data mining in crystallography Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 89–134. https://doi.org/10.1007/978-3-642-04759-6_4.
[30] Bruno IJ, Groom CR. A crystallographic perspective on sharing data and knowledge. J Comput Aided Mol Des 2014;28:1015–22. https://doi.org/10.1007/s10822-014-9780-9.
[31] Oswald IDH, Motherwell WDS, Parsons S, Pidcock E, Pulham CR. Rationalisation of co-crystal formation through knowledge-mining. Crystallogr Rev 2004;10:57–66. https://doi.org/10.1080/08893110410001664855.
[32] Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr Sect B Struct Sci 2002;58:380–8. https://doi.org/10.1107/S0108768102003890.
[33] Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge structural database. Acta Crystallogr Sect B Struct Sci Cryst Eng Mater 2016;72:171–9. https://doi.org/10.1107/S2052520616003954.
[34] Allen FH. The Cambridge structural database: a quarter of a million crystal structures and rising. Acta Cryst B 2002;58:380–8. https://doi.org/10.1107/S0108768102003890.
[35] Sykes RA, McCabe P, Allen FH, Battle GM, Bruno IJ, Wood PA. New software for statistical analysis of Cambridge structural database data. J Appl Crystallogr 2011;44:882–6. https://doi.org/10.1107/S0021889811014622.
[36] Petrović PV, Janjić GV, Zarić SD. Stacking interactions between square-planar metal complexes with 2,2′-bipyridine ligands. Analysis of crystal structures and quantum chemical calculations. Cryst Growth Des 2014;14:3880–9. https://doi.org/10.1021/cg500447h.
[37] McCrary PD, Chatel G, Alaniz SA, Cojocaru OA, Beasley PA, Flores LA, Kelley SP, Barber PS, Rogers RD. Evaluating ionic liquids as hypergolic fuels: exploring reactivity from molecular structure. Energy Fuels 2014;28:3460–73. https://doi.org/10.1021/ef500264z.
[38] Bergerhoff G, Hundt R, Sievers R, Brown ID. The inorganic crystal structure data base. J Chem Inf Model 1983;23:66–9. https://doi.org/10.1021/ci00038a003.
[39] Inorganic Crystal Structure Database, 〈http://icsd.fiz-karlsruhe.de/icsd/〉, (n.d.).
[40] Buchsbaum C, Hähler-Schlimm S, Rehme S. Data bases, the base for data mining. In: Hofmann DWM, Kuleshova LN, editors. Data mining in crystallography Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 37–58. https://doi.org/10.1007/978-3-642-04759-6_2.
[41] Bennett JW. Discovery and design of functional materials: integration of database searching and first principles calculations. Phys Procedia 2012;34:14–23. https://doi.org/10.1016/j.phpro.2012.05.003.
[42] Cumby J, Attfield JP. Ellipsoidal analysis of coordination polyhedra. Nat Commun 2017;8:14235. https://doi.org/10.1038/ncomms14235.
[43] Faber J, PDF-4 ICDD'S New. Organic database: search indexes, full pattern analysis and data mining. Crystallogr Rev 2004;10:97–107. https://doi.org/10.1080/08893110410001664873.
[44] van Beusekom B, Perrakis A, Joosten RP. Data mining of macromolecular structures. In: Carugo O, Eisenhaber F, editors. Data mining techniques fo life sciences New York, NY: Springer New York; 2016. p. 107–38. https://doi.org/10.1007/978-1-4939-3572-7_6.
[45] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank, 1999. Int tables crystallogr Chester, England: International Union of Crystallography; 2006. p. 675–84. https://doi.org/10.1107/97809553602060000722.
[46] Idicula-Thomas S, Balaji PV. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. Protein Sci 2005;14:582–92. https://doi.org/10.1110/ps.041009005.
[47] Cheng H, Sen TZ, Jernigan RL, Kloczkowski A. Data mining for protein secondary structure prediction. In: Hofmann DWM, Kuleshova LN, editors. Data mining in crystallography Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 135–67. https://doi.org/10.1007/978-3-642-04759-6_5.
[48] Zheng H, Chruszcz M, Lasota P, Lebioda L, Minor W. Data mining of metal ion environments present in protein structures. J Inorg Biochem 2008;102:1765–76. https://doi.org/10.1016/j.jinorgbio.2008.05.006.
[49] Ivanisenko VA. PDBSite: a database of the 3D structure of protein functional sites. Nucleic Acids Res 2004;33:D183–7. https://doi.org/10.1093/nar/gki105.
[50] Zheng H, Porebski PJ, Grabowski M, Cooper DR, Minor W. Databases, repositories, and other data resources in structural biology. In: Wlodawer A, Dauter Z, Jaskolski M, editors. Protein crystallogr methods protoc New York, New York, NY: Springer; 2017. p. 643–65. https://doi.org/10.1007/978-1-4939-7000-1_27.
[51] Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics 2004;4:1551–61. https://doi.org/10.1002/pmic.200300772.
[52] Catlow CRA. Crystallography, materials and computation. IUCrJ 2014;1:200–1. https://doi.org/10.1107/S2052252514014122.
[53] Oganov AR, Lyakhov AO, Valle M. How evolutionary crystal structure prediction works—and why. Acc Chem Res 2011;44:227–37. https://doi.org/10.1021/ar1001318.
[54] Curtarolo S, Morgan D, Persson K, Rodgers J, Ceder G. Predicting crystal structures with data mining of quantum calculations. Phys Rev Lett 2003;91:1–4. https://doi.org/10.1103/PhysRevLett.91.135503.
[55] Manuel Serra J, Allen Baumes L, Moliner M, Serna P, Corma A. Zeolite synthesis modelling with support vector machines: a combinatorial approach. Comb Chem High Throughput Screen 2007;10:13–24. https://doi.org/10.2174/138620707779802779.
[56] Cerqueira TFT, Sarmiento-Pérez R, Amsler M, Nogueira F, Botti S, Marques MAL. Materials design on-the-fly. J Chem Theory Comput 2015;11:3955–60. https://doi.org/10.1021/acs.jctc.5b00212.
[57] Hautier G. Data mining approaches to high-throughput crystal structure and compound prediction. In: Atahan-Evrenk S, Aspuru-Guzik A, editors. Predict calc cryst struct methods appl Cham: Springer International Publishing; 2014. p. 139–79. https://doi.org/10.1007/128_2013_486.
[58] Friedrichs OD, Dress AWM, Huson DH, Klinowski J, Mackay AL. Systematic enumeration of crystalline networks. Nature 1999;400:644–7. https://doi.org/10.1038/23210.
[59] Ryan K, Lengyel J, Shatruk M. Crystal structure prediction via deep learning. J Am Chem Soc 2018. https://doi.org/10.1021/jacs.8b03913. [jacs.8b03913].
[60] Zhang X, Stevanović V, D'Avezac M, Lany S, Zunger A. Prediction of A_2BX_4 metal-chalcogenide compounds via first-principles thermodynamics. Phys Rev B 2012;86:014109. https://doi.org/10.1103/PhysRevB.86.014109.
[61] Nauha E, Bernstein J. "Predicting" Crystal Forms of Pharmaceuticals Using Hydrogen Bond Propensities: two Test Cases. Cryst Growth Des 2014;14:4364–70. https://doi.org/10.1021/cg500478p.
[62] Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J, Aspuru-Guzik A. What is high-throughput virtual screening? A perspective from organic materials discovery. Annu Rev Mater Res 2015;45:195–216. https://doi.org/10.1146/annurev-matsci-070214-020823.
[63] Chen W, Pöhls J-H, Hautier G, Broberg D, Bajaj S, Aydemir U, Gibbs ZM, Zhu H, Asta M, Snyder GJ, Meredig B, White MA, Persson K, Jain A. Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment. J Mater Chem C 2016;4:4414–26. https://doi.org/10.1039/C5TC04339E.
[64] Kajikawa Y, Yoshikawa J, Takeda Y, Matsushima K. Tracking emerging technologies in energy research: toward a roadmap for sustainable energy. Technol Forecast Soc Change 2008;75:771–82. https://doi.org/10.1016/j.techfore.2007.05.005.
[65] Colak I, Sagiroglu S, Yesilbudak M. Data mining and wind power prediction: a literature review. Renew Energy 2012;46:241–7. https://doi.org/10.1016/j.renene.2012.02.015.
[66] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1:011002. https://doi.org/10.1063/1.4812323.
[67] Sun W, Dacek ST, Ong SP, Hautier G, Jain A, Richards WD, Gamst AC, Persson KA, Ceder G. The thermodynamic scale of inorganic crystalline metastability. Sci Adv 2016;2. https://doi.org/10.1126/sciadv.1600225. [e1600225–e1600225].
[68] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1. https://doi.org/10.1063/1.4812323.
[69] Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O. The high-throughput highway to computational materials design. Nat Mater 2013;12:191–201. https://doi.org/10.1038/nmat3568.
[70] Materials Genome Initiative for Global Competitiveness,OSTP. 2011.
[71] Belsky A, Hellenbrandt M, Karen VL, Luksch P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. Acta Crystallogr Sect B Struct Sci 2002;58:364–9. https://doi.org/10.1107/S0108768102006948.
[72] Ricci F, Chen W, Aydemir U, Snyder GJ, Rignanese G-M, Jain A, Hautier G. An ab initio electronic transport database for inorganic materials. Sci Data 2017;4:170085. https://doi.org/10.1038/sdata.2017.85.
[73] Watanabe T, Sholl DS. Accelerating applications of metal–organic frameworks for gas adsorption and separation by computational screening of materials. Langmuir 2012;28:14114–28. https://doi.org/10.1021/la301915s.
[74] Goldsmith J, Wong-Foy AG, Cafarella MJ, Siegel DJ. Theoretical limits of hydrogen storage in metal–organic frameworks: opportunities and trade-offs. Chem Mater 2013;25:3373–82. https://doi.org/10.1021/cm401978e.
[75] Choudhary K, Zhang Q, Reid ACE, Chowdhury S, Van Nguyen N, Trautt Z, Newrock MW, Congo FY, Tavazza F. Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms. Sci Data 2018:1–31〈http://arxiv.org/abs/1804.01032〉.
[76] Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera RS, Gold-Parker A, Vogt L, Brockway AM, Aspuru-Guzik A. The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J Phys Chem Lett 2011;2:2241–51. https://doi.org/10.1021/jz200866s.
[77] Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. MatErials Design And Discovery With High-throughput Density Functional Theory: thE Open Quantum Materials Database (OQMD). JOM 2013;65:1501–9. https://doi.org/10.1007/s11837-013-0755-4.
[78] Cheon G, Duerloo KAN, Sendek AD, Porter C, Chen Y, Reed EJ. Data mining for new two- and one-dimensional weakly bonded solids and lattice-commensurate heterostructures. Nano Lett 2017;17:1915–23. https://doi.org/10.1021/acs.nanolett.6b05229.
[79] Gadde S, Batchelor EK, Kaifer AE. Controlling the formation of cyanine dye H- and J-aggregates with cucurbituril hosts in the presence of anionic polyelectrolytes. Chem - Eur J 2009;15:6025–31. https://doi.org/10.1002/chem.200802546.
[80] Upreti S, Ramanan A. Structure-directing role of hydrogen-bonded of phenylenediammonium cations: supramolecular assemblies of octamolybdate-based organic-inorganic hybrids. Cryst Growth Des 2005;5:1837–43. https://doi.org/10.1021/cg050100m.

[81] Planells M, Pellejà L, Clifford JN, Pastore M, De Angelis F, López N, Marder SR, Palomares E. Energy levels, charge injection, charge recombination and dye regeneration dynamics for donor–acceptor π-conjugated organic dyes in mesoscopic TiO_2 sensitized solar cells. Energy Environ Sci 2011;4:1820. https://doi.org/10.1039/c1ee01060c.

[82] Kahr B, Ward MD. Flexible molecular materials: stressed out crystals. Nat Chem 2017;10:4–6. https://doi.org/10.1038/nchem.2913.

[83] Fayos J. Molecular crystal prediction approach by molecular similarity: data mining on molecular aggregation predictors and crystal descriptors. Cryst Growth Des 2009;9:3142–53. https://doi.org/10.1021/cg801122m.

[84] Cole JM. Single-crystal X-ray diffraction studies of photo-induced molecular species. Chem Soc Rev 2004;33:501–13. https://doi.org/10.1039/b205339j.

[85] Liu X, Cole JM, Waddell PG, Lin T, Mckechnie S, Thomson JJ. Molecular origins of optoelectronic properties in coumarins 343, 314T, 445, and 522B. J Phys Chem C 2013;117:14130–41. https://doi.org/10.1021/jp400614e.

[86] Liu X, Cole JM, Waddell PG, Lin T-C, McKechnie S. Molecular origins of optoelectronic properties in coumarins 343, 314T, 445, and 522B. J Phys Chem C 2013;117:14130–41. https://doi.org/10.1021/jp400614e.

[87] Albert IDL, Marks TJ, Ratner MA. Large molecular hyperpolarizabilities in "push–pull" porphyrins. molecular planarity and auxiliary donor–acceptor effects. Chem Mater 1998;10:753–62. https://doi.org/10.1021/cm970478a.

[88] Zhang L, Cole JM. Dye aggregation in dye-sensitized solar cells. J Mater Chem A 2017;5:19541–59. https://doi.org/10.1039/C7TA05632J.

[89] Davies DW, Butler KT, Jackson AJ, Morris A, Frost JM, Skelton JM, Walsh A. Computational screening of all stoichiometric inorganic materials. Chem 2016;1:617–27. https://doi.org/10.1016/j.chempr.2016.09.010.

[90] Nagy PI, Erhardt PW. Ab initio study of hydrogen-bond formation between cyclic ethers and selected amino acid side chains. J Phys Chem A 2006;110:13923–32. https://doi.org/10.1021/jp061113t.

[91] Lee JH, Lee J-H, Kong E-H, Jang HM. The nature of hydrogen-bonding interaction in the prototypic hybrid halide perovskite, tetragonal $CH_3NH_3PbI_3$. Sci Rep 2016;6:21687. https://doi.org/10.1038/srep21687.

[92] Cheon K-S, Park YS, Kazmaier PM, Buncel E. Studies of azo–hydrazone tautomerism and H-bonding in azo-functionalized dendrimers and model compounds. Dye Pigment 2002;53:3–14. https://doi.org/10.1016/S0143-7208(01)00096-1.

[93] Görner H, Chibisov AK, Slavnova TD. Kinetics of J-aggregation of cyanine dyes in the presence of gelatin. J Phys Chem B 2006;110:3917–23. https://doi.org/10.1021/jp055876c.

[94] Peyratout C, Daehne L. Aggregation of thiacyanine derivatives on polyelectrolytes. Phys Chem Chem Phys 2002;4:3032–9. https://doi.org/10.1039/b111581b.

[95] Nibbering ETJ, Fidder H, Pines E. Ultrafast chemistry: using time-resolved vibrational spectroscopy for interrogation of structural dynamics. Annu Rev Phys Chem 2005;56:337–67. https://doi.org/10.1146/annurev.physchem.56.092503.141314.

[96] Diebold U. The surface science of titanium dioxide. Surf Sci Rep 2003;48:53–229. https://doi.org/10.1016/S0167-5729(02)00100-0.

[97] Witman M, Ling S, Anderson S, Tong L, Stylianou KC, Slater B, Smit B, Haranczyk M. In silico design and screening of hypothetical MOF-74 analogs and their experimental synthesis. Chem Sci 2016;7:6263–72. https://doi.org/10.1039/C6SC01477A.

[98] Upreti S, Ramanan A. Role of hydrogen-bonded interactions in the crystal packing of phenylenediammonium phosphomolybdates. Cryst Growth Des 2006;6:2066–71. https://doi.org/10.1021/cg0601610.

[99] Zhao Y, Zhu K. Organic–inorganic hybrid lead halide perovskites for optoelectronic and electronic applications. Chem Soc Rev 2016;45:655–89. https://doi.org/10.1039/C4CS00458B.

[100] Zhu L, Yi Y, Li Y, Kim E-G, Coropceanu V, Brédas J-L. Prediction of remarkable ambipolar charge-transport characteristics in organic mixed-stack charge-transfer crystals. J Am Chem Soc 2012;134:2340–7. https://doi.org/10.1021/ja210284s.

[101] Nguyen KL, Friscić T, Day GM, Gladden LF, Jones W. Terahertz time-domain spectroscopy and the quantitative monitoring of mechanochemical cocrystal formation. Nat Mater 2007;6:206–9. https://doi.org/10.1038/nmat1848.

[102] Neukirch AJ, Hyeon-Deuk K, Prezhdo OV. Time-domain ab initio modeling of excitation dynamics in quantum dots. Coord Chem Rev 2014;263–264:161–81. https://doi.org/10.1016/j.ccr.2013.08.035.

[103] Cortecchia D, Soci C, Cametti M, Petrozza A, Mart-Rujas J. Crystal engineering of a two-dimensional lead-free perovskite with functional organic cations by second-sphere coordination. Chempluschem 2016:1–6. https://doi.org/10.1002/cplu.201600603.

[104] Kuleshova LN, Hofmann DWM, Boese R. Lattice energy calculation – a quick tool for screening of cocrystals and estimation of relative solubility. Case flavonoids, Chem Phys Lett 2013;564:26–32. https://doi.org/10.1016/j.cplett.2013.02.008.

[105] Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 1996;96:1027–44. https://doi.org/10.1021/cr950202r.

[106] Zhang X, Yu L, Zakutayev A, Zunger A. Sorting stable versus unstable hypothetical compounds: the case of multi-functional ABX half-heusler filled tetrahedral structures. Adv Funct Mater 2012;22:1425–35. https://doi.org/10.1002/adfm.201102546.

[107] Zakutayev A, Zhang X, Nagaraja A, Yu L, Lany S, Mason TO, Ginley DS, Zunger A. Theoretical prediction and experimental realization of new stable inorganic materials using the inverse design approach. J Am Chem Soc 2013;135:10048–54. https://doi.org/10.1021/ja311599g.

[108] Hautier G, Jain A, Ong SP. From the computer to the laboratory: materials discovery and design using first-principles calculations. J Mater Sci 2012;47:7317–40. https://doi.org/10.1007/s10853-012-6424-0.

[109] Janet JP, Gani TZH, Steeves AH, Ioannidis EI, Kulik HJ. Leveraging cheminformatics strategies for inorganic discovery: application to redox potential design. Ind Eng Chem Res 2017;56:4898–910. https://doi.org/10.1021/acs.iecr.7b00808.

[110] Schütt KT, Glawe H, Brockherde F, Sanna A, Müller KR, Gross EKU. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. Phys Rev B 2014;89:205118. https://doi.org/10.1103/PhysRevB.89.205118.

[111] Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, Thompson A, Zhang K, Choudhary A, Wolverton C. Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys Rev B - Condens Matter Mater Phys 2014;89:1–7. https://doi.org/10.1103/PhysRevB.89.094104.

[112] Saad Y, Gao D, Ngo T, Bobbitt S, Chelikowsky JR, Andreoni W. Data mining for materials: computational experiments with AB compounds. Phys Rev B - Condens Matter Mater Phys 2012;85:1–10. https://doi.org/10.1103/PhysRevB.85.104104.

[113] Bacardit J, Llorà X. Large-scale data mining using genetics-based machine learning, Wiley Interdiscip. Rev Data Min Knowl Discov 2013;3:37–61. https://doi.org/10.1002/widm.1078.

[114] Balachandran PV, Theiler J, Rondinelli JM, Lookman T. Materials prediction via classification learning. Sci Rep 2015;5:13285. https://doi.org/10.1038/srep13285.

[115] Pillong M, Marx C, Piechon P, Wicker JGP, Cooper RI, Wagner T. A publicly available crystallisation data set and its application in machine learning. CrystEngComm 2017;19:3737–45. https://doi.org/10.1039/C7CE00738H.

[116] Ferguson AL. Machine learning and data science in soft materials engineering. J Phys Condens Matter 2018;30:043002. https://doi.org/10.1088/1361-648X/aa98bd.

[117] Gola J, Britz D, Staudt T, Winter M, Schneider AS, Ludovici M, Mücklich F. Advanced microstructure classification by data mining methods. Comput Mater Sci 2018;148:324–35. https://doi.org/10.1016/j.commatsci.2018.03.004.

[118] Cerqueira TFT, Lin S, Amsler M, Goedecker S, Botti S, Marques MAL. Identification of novel Cu, Ag, and Au ternary oxides from global structural prediction. Chem Mater 2015;27:4562–73. https://doi.org/10.1021/acs.chemmater.5b00716.

[119] Scott DJ, Coveney PV, Kilner JA, Rossiny JCH, Alford NMN. Prediction of the functional properties of ceramic materials from composition using artificial neural networks. J Eur Ceram Soc 2007;27:4425–35. https://doi.org/10.1016/j.jeurceramsoc.2007.02.212.

[120] Rajan K. Materials Informatics: the materials "gene" and big data. Annu Rev Mater Res 2015;45:153–69. https://doi.org/10.1146/annurev-matsci-070214-021132.

[121] Suh C, Rajan K. Combinatorial design of semiconductor chemistry for bandgap engineering: "virtual" combinatorial experimentation. Appl Surf Sci 2004;223:148–58. https://doi.org/10.1016/S0169-4332(03)00918-8.

[122] Suh C, Rajan K. Invited review: data mining and informatics for crystal chemistry: establishing measurement techniques for mapping structure–property relationships. Mater Sci Technol 2009;25:466–71. https://doi.org/10.1179/174328409X430483.

[123] Simon CM, Kim J, Gomez-Gualdron DA, Camp JS, Chung YG, Martin RL, Mercado R, Deem MW, Gunter D, Haranczyk M, Sholl DS, Snurr RQ, Smit B. The materials genome in action: identifying the performance limits for methane storage. Energy Environ Sci 2015;8:1190–9. https://doi.org/10.1039/C4EE03515A.

[124] Tseng Y-H, Lin C-J, Lin Y-I. Text mining techniques for patent analysis. Inf Process Manag 2007;43:1216–47. https://doi.org/10.1016/j.ipm.2006.11.011.

[125] Porter AL, Guo Y, Chiavatta D. Tech mining: text mining and visualization tools, as applied to nanoenhanced solar cells. Wiley Interdiscip Rev Data Min Knowl Discov 2011;1:172–81. https://doi.org/10.1002/widm.7.

[126] Guo Y, Xu C, Huang L, Porter A. Empirically informing a technology delivery system model for an emerging technology: illustrated for dye-sensitized solar cells. R&D Manag 2012;42:133–49. https://doi.org/10.1111/j.1467-9310.2012.00674.x.

[127] Swain MC, Cole JM. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. J Chem Inf Model 2016;56:1894–904. https://doi.org/10.1021/acs.jcim.6b00207.

[128] O'Regan B, Grätzel M. A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. Nature 1991;353:737–40. https://doi.org/10.1038/353737a0.

[129] Hagfeldt A, Boschloo G, Sun L, Kloo L, Pettersson H. Dye-sensitized solar cells. Chem Rev 2010;110:6595–663. https://doi.org/10.1021/cr900356p.

[130] Cai N, Moon S-J, Cevey-Ha L, Moehl T, Humphry-Baker R, Wang P, Zakeeruddin SM, Grätzel M. An organic D-π-A dye for record efficiency solid-state sensitized heterojunction solar cells. Nano Lett 2011;11:1452–6. https://doi.org/10.1021/nl104034e.

[131] Wu Y, Marszalek M, Zakeeruddin SM, Zhang Q, Tian H, Grätzel M, Zhu W. High-conversion-efficiency organic dye-sensitized solar cells: molecular engineering on D–A–π-A featured organic indoline dyes. Energy Environ Sci 2012;5:8261. https://doi.org/10.1039/c2ee22108j.

[132] Haid S, Marszalek M, Mishra A, Wielopolski M, Teuscher J, Moser J-E, Humphry-Baker R, Zakeeruddin SM, Grätzel M, Bäuerle P. Significant improvement of dye-sensitized solar cell performance by small structural modification in π-conjugated donor-acceptor dyes. Adv Funct Mater 2012;22:1291–302. https://doi.org/10.1002/adfm.201102519.

[133] Zhang L, Liu X, Rao W, Li J. Multilayer dye aggregation at dye/TiO_2 interface via π…π stacking and hydrogen bond and its impact on solar cell performance: a DFT analysis. Sci Rep 2016;6:35893. https://doi.org/10.1038/srep35893.

[134] Zhang Y, Porter AL, Hu Z, Guo Y, Newman NC. "Term clumping" for technical intelligence: a case study on dye-sensitized solar cells. Technol Forecast Soc Change 2014;85:26–39. https://doi.org/10.1016/j.techfore.2013.12.019.

[135] Grätzel M. Dye-sensitized solar cells. J Photochem Photobiol C Photochem Rev

2003;4:145–53. https://doi.org/10.1016/S1389-5567(03)00026-1.
[136] Nazeeruddin MK, De Angelis F, Fantacci S, Selloni A, Viscardi G, Liska P, Ito S, Takeru B, Grätzel M. Combined experimental and DFT-TDDFT computational study of photoelectrochemical cell ruthenium sensitizers. J Am Chem Soc 2005;127:16835–47. https://doi.org/10.1021/ja052467l.
[137] Hamann TW, Jensen RA, Martinson ABF, Van Ryswyk H, Hupp JT. Advancing beyond current generation dye-sensitized solar cells. Energy Environ Sci 2008;1:66. https://doi.org/10.1039/b809672d.
[138] Ning Z, Fu Y, Tian H. Improvement of dye-sensitized solar cells: what we know and what we need to know. Energy Environ Sci 2010;3:1170. https://doi.org/10.1039/c003841e.
[139] Nedelcu M, Guldin S, Orilall MC, Lee J, Hüttner S, Crossland EJW, Warren SC, Ducati C, Laity PR, Eder D, Wiesner U, Steiner U, Snaith HJ. Monolithic route to efficient dye-sensitized solar cells employing diblock copolymers for mesoporous TiO2. J Mater Chem 2010;20:1261. https://doi.org/10.1039/b920077k.
[140] Grätzel M. Recent advances in sensitized mesoscopic solar cells. Acc Chem Res 2009;42:1788–98. https://doi.org/10.1021/ar900141y.
[141] Labat F, Le Bahers T, Ciofini I, Adamo C. First-principles modeling of dye-sensitized solar cells: challenges and perspectives. Acc Chem Res 2012;45:1268–77. https://doi.org/10.1021/ar200327w.
[142] Mathew S, Yella A, Gao P, Humphry-Baker R, Curchod BFE, Ashari-Astani N, Tavernelli I, Rothlisberger U, Nazeeruddin MK, Grätzel M. Dye-sensitized solar cells with 13% efficiency achieved through the molecular engineering of porphyrin sensitizers. Nat Chem 2014;6:242–7. https://doi.org/10.1038/nchem.1861.
[143] Justin Thomas KR, Hsu Y-C, Lin JT, Lee K-M, Ho K-C, Lai C-H, Cheng Y-M, Chou P-T. 2,3-disubstituted thiophene-based organic dyes for solar cells. Chem Mater 2008;20:1830–40. https://doi.org/10.1021/cm702631r.
[144] Zhang L, Cole JM, Waddell PG, Low KS, Liu X. Relating electron donor and carboxylic acid anchoring substitution effects in azo dyes to dye-sensitized solar cell performance. ACS Sustain Chem Eng 2013;1:1440–52. https://doi.org/10.1021/sc400183t.
[145] Cole JM, Low KS, Gong Y. Discovery of black dye crystal structure polymorphs: implications for dye conformational variation in dye-sensitized solar cells. ACS Appl Mater Interfaces 2015;7:27646–53. https://doi.org/10.1021/acsami.5b07364.
[146] Cole JM, Blood-Forsythe MA, Lin TC, Pattison P, Gong Y, Vázquez-Mayagoitia Á, Waddell PG, Zhang L, Koumura N, Mori S. Discovery of S···C≡N intramolecular bonding in a thiophenylcyanoacrylate-based dye: realizing charge transfer pathways and dye···TiO_2 anchoring characteristics for dye-sensitized solar cells. ACS Appl Mater Interfaces 2017;9:25952–61. https://doi.org/10.1021/acsami.7b03522.
[147] Monat JE, Rodriguez JH, Mccusker JK. Ground- and excited-state electronic structures of the solar cell sensitiser bis(4,4′-dicarboxylato-2,2′-bipyridine)bis (isothiocyanato)ruthenium(II). J Phys Chem A 2002;2:7399–406.
[148] Coppens P. The new photocrystallography. Angew Chem Int Ed 2009;48:4280–1. https://doi.org/10.1002/anie.200900910.
[149] Coppens P, Zhang L, Thomas R, Chen Y, Jarzembska K, Kaminski R, Trzop E, Fournier B. Can we deconvolute electron density changes from the dominant influence of the atomic rearrangement on molecular excitation in time-resolved diffraction studies? Phys Scr 2016;91:023003. https://doi.org/10.1088/0031-8949/91/2/023003.
[150] Benedict JB, Makal A, Sokolow JD, Trzop E, Scheins S, Henning R, Graber T, Coppens P. Time-resolved Laue diffraction of excited species at atomic resolution: 100 ps single-pulse diffraction of the excited state of the organometallic complex Rh2(μ-PNP)2(PNP)2·BPh4. Chem Commun (Camb) 2011;47:1704–6. https://doi.org/10.1039/c0cc04997b.
[151] Coppens P. Molecular excited-state structure by time-resolved pump–probe X-ray diffraction. What is new and what are the prospects for further progress? J Phys Chem Lett 2011;2:616–21. https://doi.org/10.1021/jz200050x.
[152] Jarzembska KN, Chen Y, Nasca JN, Trzop E, Watson DF, Coppens P. Relating structure and photoelectrochemical properties: electron injection by structurally and theoretically characterized transition metal-doped phenanthroline–polyoxotitanate nanoparticles. Phys Chem Chem Phys 2014;16:15792. https://doi.org/10.1039/C4CP02509A.
[153] Patrocínio AOT, Paniago EB, Paniago RM, Iha NYM. XPS characterization of sensitized n-TiO_2 thin films for dye-sensitized solar cell applications. Appl Surf Sci 2008;254:1874–9. https://doi.org/10.1016/j.apsusc.2007.07.185.
[154] McCree-Grey PJ, Cole J, Evans JM, McCree-Grey J, Cole JM, Evans PJ. Preferred molecular orientation of Coumarin 343 on TiO_2 surfaces: application to dye-sensitized solar cells. ACS Appl Mater Interfaces 2015;7:16404–9. https://doi.org/10.1021/acsami.5b03572.
[155] Zhang L, Cole JM. Anchoring groups for dye-sensitized solar cells. ACS Appl Mater Interfaces 2015;7:3427–55. https://doi.org/10.1021/am507334m.
[156] Cole JM, Simos TE, Psihoyios G, Tsitouras C, Anastassi Z. Systematic prediction of dyes for dye sensitized solar cells: data-mining via molecular charge-transfer algorithms. AIP Conf. Proc. 2011:999–1002. https://doi.org/10.1063/1.3637778.
[157] Cole JM, Low KS, Ozoe H, Stathi P, Kitamura C, Kurata H, Rudolf P, Kawase T. Data mining with molecular design rules identifies new class of dyes for dye-sensitised solar cells. Phys Chem Chem Phys 2014;16:26684–90. https://doi.org/10.1039/C4CP02645D.
[158] Hammett LP. The effect of structure upon the reactions of organic compounds. Benzene derivatives. J Am Chem Soc 1937;59:96–103. https://doi.org/10.1021/ja01280a022.
[159] Moot T, Isayev O, Call RW, McCullough SM, Zemaitis M, Lopez R, Cahoon JF, Tropsha A. Material informatics driven design and experimental validation of lead titanate as an aqueous solar photocathode. Mater Discov 2016;6:9–16. https://doi.org/10.1016/j.md.2017.04.001.
[160] Low N, kian Sing, Cole Jacqueline M, Zhou Xiaolan, Yufa. Rationalising the molecular origins of Ru and research papers. Acta Crystallogr Sect B 2009:1–12.
[161] Low KS, Cole JM, Zhou X, Yufa N. Rationalizing the molecular origins of Ru- and Fe-based dyes for dye-sensitized solar cells. Acta Crystallogr B 2012;68:137–49. https://doi.org/10.1107/S0108768112009263.
[162] Zhang L, Cole JM, Liu X. Tuning solvatochromism of azo dyes with intramolecular hydrogen bonding in solution and on titanium dioxide nanoparticles. J Phys Chem C 2013;117:26316–23. https://doi.org/10.1021/jp4088783.
[163] Yang Z, Shao C, Cao D. Screening donor groups of organic dyes for dye-sensitized solar cells. RSC Adv 2015;5:22892–8. https://doi.org/10.1039/C4RA17261B.
[164] Yang Z, Liu Y, Liu C, Lin C, Shao C. TDDFT screening auxiliary withdrawing group and design the novel D-A-π-A organic dyes based on indoline dye for highly efficient dye-sensitized solar cells. Spectrochim Acta Part A Mol Biomol Spectrosc 2016;167:127–33. https://doi.org/10.1016/j.saa.2016.05.041.
[165] Yang Z, Liu C, Shao C, Zeng X, Cao D. Screening π -conjugated bridges of organic dyes for dye-sensitized solar cells with panchromatic visible light harvesting. Nanotechnology 2016;27:265701. https://doi.org/10.1088/0957-4484/27/26/265701.
[166] Venkatraman V, Foscato M, Jensen VR, Alsberg BK. Evolutionary de novo design of phenothiazine derivatives for dye-sensitized solar cells. J Mater Chem A 2015;3:9851–60. https://doi.org/10.1039/C5TA00625B.
[167] Venkatraman V, Raju R, Oikonomopoulos SP, Alsberg BK. The dye-sensitized solar cell database. J Chemin 2018;10:18. https://doi.org/10.1186/s13321-018-0272-0.
[168] Kojima A, Teshima K, Shirai Y, Miyasaka T. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. J Am Chem Soc 2009;131:6050–1. https://doi.org/10.1021/ja809598r.
[169] Castelli IE, Olsen T, Datta S, Landis DD, Dahl S, Thygesen KS, Jacobsen KW. Computational screening of perovskite metal oxides for optimal solar light capture. Energy Environ Sci 2012;5:5814–9. https://doi.org/10.1039/C1EE02717D.
[170] Cao DH, Stoumpos CC, Farha OK, Hupp JT, Kanatzidis MG. 2D Homologous Perovskites as Light-Absorbing Materials for Solar Cell Applications. J Am Chem Soc 2015;137:7843–50. https://doi.org/10.1021/jacs.5b03796.
[171] Haruyama J, Sodeyama K, Han L, Tateyama Y. Surface properties of $CH_3NH_3PbI_3$ for perovskite solar cells. Acc Chem Res 2016;49:554–61. https://doi.org/10.1021/acs.accounts.5b00452.
[172] Battaglia C, Cuevas A, De Wolf S, Manuscript A. High-efficiency crystalline silicon solar cells: status and perspectives. Energy Environ Sci 2016;9. https://doi.org/10.1039/C5EE03380B.
[173] Yuan Y, Huang J. Ion migration in organometal trihalide perovskite and its impact on photovoltaic efficiency and stability. Acc Chem Res 2016;49:286–93. https://doi.org/10.1021/acs.accounts.5b00420.
[174] Kim H-S, Lee J-W, Yantara N, Boix PP, a Kulkarni S, Mhaisalkar S, Grätzel M, Park N-G. High efficiency solid-state sensitized solar cell-based on submicrometer rutile TiO2 nanorod and CH3NH3PbI3 perovskite sensitizer. Nano Lett 2013;13:2412–7. https://doi.org/10.1021/nl400286w.
[175] Wu Z, Raga SR, Juarez-Perez EJ, Yao X, Jiang Y, Ono LK, Ning Z, Tian H, Qi Y. Improved efficiency and stability of perovskite solar cells induced by C□O functionalized hydrophobic ammonium-based additives. Adv Mater 2018;30:1703670. https://doi.org/10.1002/adma.201703670.
[176] Luo J, Im J-H, Mayer MT, Schreier M, Nazeeruddin MK, Park N-G, Tilley SD, Fan HJ, Gratzel M. Water photolysis at 12.3% efficiency via perovskite photovoltaics and Earth-abundant catalysts. Science 2014;345:1593–6. https://doi.org/10.1126/science.1258307.
[177] O'Regan BC, Barnes PRF, Li X, Law C, Palomares E, Marin-Beloqui JM. Optoelectronic studies of methylammonium lead iodide perovskite solar cells with mesoporous TiO_2: separation of electronic and chemical charge storage, understanding two recombination lifetimes, and the evolution of band offsets during J-V hyst. J Am Chem Soc 2015;137:5087–99. https://doi.org/10.1021/jacs.5b00761.
[178] Choi H, Jeong J, Kim HB, Kim S, Walker B, Kim GH, Kim JY. Cesium-doped methylammonium lead iodide perovskite light absorber for hybrid solar cells. Nano Energy 2014;7:80–5. https://doi.org/10.1016/j.nanoen.2014.04.017.
[179] Hu Q, Wu J, Jiang C, Liu T, Que X, Zhu R, Gong Q. Engineering of electron-selective contact for perovskite solar cells with efficiency exceeding 15%. ACS Nano 2014;8:10161–7. https://doi.org/10.1021/nn5029828.
[180] Chen Q, De Marco N, (Michael) Yang Y, Song T-B, Chen C-C, Zhao H, Hong Z, Zhou H, Yang Y. Under the spotlight: the organic–inorganic hybrid halide perovskite for optoelectronic applications. Nano Today 2015;10:355–96. https://doi.org/10.1016/j.nantod.2015.04.009.
[181] Boix PP, Agarwala S, Koh TM, Mathews N, Mhaisalkar SG. Perovskite solar cells: beyond methylammonium lead iodide. J Phys Chem Lett 2015;6:898–907. https://doi.org/10.1021/jz502547f.
[182] Malinkiewicz O, Yella A, Lee YH, Espallargas GMM, Graetzel M, Nazeeruddin MK, Bolink HJ. Perovskite solar cells employing organic charge-transport layers. Nat Photonics 2014;8:128–32. https://doi.org/10.1038/nphoton.2013.341.
[183] Yang S, Wang Y, Liu P, Cheng Y-B, Zhao HJ, Yang HG. Functionalization of perovskite thin films with moisture-tolerant molecules. Nat Energy 2016;1:15016. https://doi.org/10.1038/nenergy.2015.16.
[184] McMeekin DP, Sadoughi G, Rehman W, Eperon GE, Saliba M, Horantner MT, Haghighirad A, Sakai N, Korte L, Rech B, Johnston MB, Herz LM, Snaith HJ. A mixed-cation lead mixed-halide perovskite absorber for tandem solar cells. Science 2016;351:151–5. https://doi.org/10.1126/science.aad5845.
[185] Kim H-S, Mora-Sero I, Gonzalez-Pedro V, Fabregat-Santiago F, Juarez-Perez EJ, Park N-G, Bisquert J. Mechanism of carrier accumulation in perovskite thin-

absorber solar cells. Nat Commun 2013;4:2242. https://doi.org/10.1038/ncomms3242.
[186] Stranks SD, Eperon GE, Grancini G, Menelaou C, Alcocer MJP, Leijtens T, Herz LM, Petrozza A, Snaith HJ. Electron-hole diffusion lengths exceeding 1 micrometer in an organometal trihalide perovskite absorber. Science 2013;342:341–4. https://doi.org/10.1126/science.1243982.
[187] Lin Q, Armin A, Nagiri RCR, Burn PL, Meredith P. Electro-optics of perovskite solar cells. Nat Photonics 2014;9:106–12. https://doi.org/10.1038/nphoton.2014.284.
[188] Li W, Fan J, Li J, Mai Y, Wang L. Controllable grain morphology of perovskite absorber film by molecular self-assembly toward efficient solar cell exceeding 17%. J Am Chem Soc 2015;137:10399–405. https://doi.org/10.1021/jacs.5b06444.
[189] Grätzel M. The light and shade of perovskite solar cells. Nat Mater 2014;13:838–42. https://doi.org/10.1038/nmat4065.
[190] Eames C, Frost JM, Barnes PRF, O'Regan BC, Walsh A, Islam MS. Ionic transport in hybrid lead iodide perovskite solar cells. Nat Commun 2015;6:7497. https://doi.org/10.1038/ncomms8497.
[191] Lee MM, Teuscher J, Miyasaka T, Murakami TN, Snaith HJ. Efficient hybrid solar cells based on meso-superstructured organometal halide perovskites. Science 2012;338:643–7. https://doi.org/10.1126/science.1228604. [doi].
[192] Docampo P, Ball JM, Darwich M, Eperon GE, Snaith HJ. Efficient organometal trihalide perovskite planar-heterojunction solar cells on flexible polymer substrates. Nat Commun 2013;4:2761. https://doi.org/10.1038/ncomms3761.
[193] Lee MM, Teuscher J, Miyasaka T, Murakami TN, Snaith HJ. Efficient hybrid solar cells based on meso-superstructured organometal halide perovskites. Science 2012;338:643–7. https://doi.org/10.1126/science.1228604.
[194] Yang WS, Noh JH, Jeon NJ, Kim YC, Ryu S, Seo J, Seok SI. High-performance photovoltaic perovskite layers fabricated through intramolecular exchange. Science 2015;348:1234–7. https://doi.org/10.1126/science.aaa9272.
[195] Hanusch FC, Wiesenmayer E, Mankel E, Binek A, Angloher P, Fraunhofer C, Giesbrecht N, Feckl JM, Jaegermann W, Johrendt D, Bein T, Docampo P. Efficient planar heterojunction perovskite solar cells based on formamidiniumlead bromide. J Phys Chem Lett 2014;5:2791–5. https://doi.org/10.1021/jz501237m.
[196] Chueh C-C, Li C-Z, Jen AK-Y. Recent progress and perspective in solution-processed Interfacial materials for efficient and stable polymer and organometal perovskite solar cells. Energy Environ Sci 2015;8:1160–89. https://doi.org/10.1039/C4EE03824J.
[197] Park BW, Philippe B, Zhang X, Rensmo H, Boschloo G, Johansson EMJ. Bismuth based hybrid perovskites $A_3Bi_2I_9$ (A: methylammonium or cesium) for solar cell application. Adv Mater 2015;27:6806–13. https://doi.org/10.1002/adma.201501978.
[198] Zhou HP, Chen Q, Li G, Luo S, Song T-bB, Duan H-SS, Hong ZR, You JB, Liu YS, Yang Y. Interface engineering of highly efficient perovskite solar cells. Science 2014;345:542–6. https://doi.org/10.1126/science.1254050.
[199] Boix PP, Agarwala S, Koh TM, Mathews N, Mhaisalkar SG. Perovskite solar cells: beyond methylammonium lead iodide. J Phys Chem Lett 2015;6:898–907. https://doi.org/10.1021/jz502547f.
[200] Graetzel M, Park N-G. Organometal halide perovskite photovoltaics: a diamond in the rough. Nano 2014;9:1–7. https://doi.org/10.1142/S1793292014400025.
[201] Su J, Lu D, Zhang L, Wang D, Bai Y, Wang W. Growth and properties of $CH3NH_3PbI_3$ single crystal. Cryst Res Technol 2017;52:1700171. https://doi.org/10.1002/crat.201700171.
[202] Su J, Wang WF, Lei Y, Zhang L, Xu LH, Wang D, Lu D, Bai Y. On the growth of CH3NH3PbI3-xClx single crystal and characterization. Phys B Condens Matter 2018;537:7–11. https://doi.org/10.1016/j.physb.2018.01.057.
[203] Chakraborty S, Xie W, Mathews N, Sherburne M, Ahuja R, Asta M, Mhaisalkar SG. Rational design: a high-throughput computational screening and experimental validation methodology for lead-free and emergent hybrid perovskites. ACS Energy Lett 2017;2:837–45. https://doi.org/10.1021/acsenergylett.7b00035.
[204] Emery AA, Saal JE, Kirklin S, Hegde VI, Wolverton C. High-throughput computational screening of perovskites for thermochemical water splitting applications. Chem Mater 2016;28:5621–34. https://doi.org/10.1021/acs.chemmater.6b01182.
[205] Butler KT, Kumagai Y, Oba F, Walsh A. Screening procedure for structurally and electronically matched contact layers for high-performance solar cells: hybrid perovskites. J Mater Chem C 2016;4:1149–58. https://doi.org/10.1039/C5TC04091D.
[206] Janiak C. A critical account on π–π stacking in metal complexes with aromatic nitrogen-containing ligands. J Chem Soc Dalton Trans 2000:3885–96. https://doi.org/10.1039/b003010o.
[207] Pyykkö P. Strong closed-shell interactions in inorganic chemistry. Chem Rev 1997;97:597–636. https://doi.org/10.1021/cr940396v.
[208] Malathy Sony SM, Ponnuswamy MN. Nature of π-interactions in nitrogen-containing heterocyclic systems: a structural database analysis. Cryst Growth Des 2006;6:736–42. https://doi.org/10.1021/cg050514f.
[209] Aullón G, Bellamy D, Guy Orpen A, Brammer L, Bruton Eric A. Metal-bound chlorine often accepts hydrogen bonds. Chem Commun 1998:653–4. https://doi.org/10.1039/a709014e.
[210] Nangia A. Database research in crystal engineering. CrystEngComm 2002;4:93. https://doi.org/10.1039/b201206e.
[211] Desiraju GR. The C–H···O hydrogen bond: structural implications and supramolecular design. Acc Chem Res 1996;29:441–9. https://doi.org/10.1021/ar950135n.
[212] Steiner T. The hydrogen bond in the solid state. Angew Chem Int Ed 2002;41:48–76. https://doi.org/10.1002/1521-3773(20020104)41:1<48::AID-ANIE48>3.0.CO;2-U.
[213] Roesky HW, Andruh M. The interplay of coordinative, hydrogen bonding and π–π stacking interactions in sustaining supramolecular solid-state architectures. Coord Chem Rev 2003;236:91–119. https://doi.org/10.1016/S0010-8545(02)00218-7.
[214] Slabbert C, Rademeyer M. One-dimensional halide-bridged polymers of metal cations with mono-heterocyclic donor ligands or cations: a review correlating chemical composition, connectivity and chain conformation. Coord Chem Rev 2015;288:18–49. https://doi.org/10.1016/j.ccr.2015.01.004.
[215] Pilania G, Balachandran PV, Kim C, Lookman T. Finding new perovskite halides via machine learning. Front Mater 2016;3. https://doi.org/10.3389/fmats.2016.00019.
[216] Castelli IE, Jacobsen KW. Designing rules and probabilistic weighting for fast materials discovery in the Perovskite structure. Model Simul Mater Sci Eng 2014;22:055007. https://doi.org/10.1088/0965-0393/22/5/055007.
[217] Yu L, Zunger A. Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials. Phys Rev Lett 2012;108:1–5. https://doi.org/10.1103/PhysRevLett.108.068701.
[218] Dey P, Bible J, Datta S, Broderick S, Jasinski J, Sunkara M, Menon M, Rajan K. Informatics-aided bandgap engineering for solar materials. Comput Mater Sci 2014;83:185–95. https://doi.org/10.1016/j.commatsci.2013.10.016.
[219] Zawadzki P, Baranowski LL, Peng H, Toberer ES, Ginley DS, Tumas W, Zakutayev A, Lany S. Evaluation of photovoltaic materials within the Cu-Sn-S family. Appl Phys Lett 2013;103:253902. https://doi.org/10.1063/1.4851896.
[220] Evans R, Dore J, Van Voorthuysen E, Zhu J, Green MA. Data mining photovoltaic cell manufacturing data. 2014 IEEE 40th Photovolt Spec Conf PVSC 2014;2014:2699–704. https://doi.org/10.1109/PVSC.2014.6925486.
[221] Bhatia A, Hautier G, Nilgianskul T, Miglio A, Sun J, Kim HJ, Kim KH, Chen S, Rignanese G-M, Gonze X, Suntivich J. High-mobility bismuth-based transparent p-type oxide from high-throughput material screening. Chem Mater 2016;28:30–4. https://doi.org/10.1021/acs.chemmater.5b03794.
[222] Varley JB, Miglio A, Ha VA, Van Setten MJ, Rignanese GM, Hautier G. High-throughput design of non-oxide p-type transparent conducting materials: data mining, search strategy, and identification of boron phosphide. Chem Mater 2017;29:2568–73. https://doi.org/10.1021/acs.chemmater.6b04663.
[223] Zakutayev A. Design of nitride semiconductors for solar energy conversion. J Mater Chem A 2016;4:6742–54. https://doi.org/10.1039/C5TA09446A.
[224] Zhang L, Cui P, Yang H, Chen J, Xiao F, Guo Y, Liu Y, Zhang W, Huo F, Liu B. Metal-organic frameworks as promising photosensitizers for photoelectrochemical water splitting. Adv Sci 2015. https://doi.org/10.1002/advs.201500243.
[225] McNamara WR, Milot RL, Song H, Snoeberger III RC, Batista VS, Schmuttenmaer CA, Brudvig GW, Crabtree RH. Water-stable, hydroxamate anchors for functionalization of TiO_2 surfaces with ultrafast interfacial electron transfer. Energy Environ Sci 2010;3:917. https://doi.org/10.1039/c001065k.
[226] Toma FM. Disentangling interfacial energetics. Nat Energy 2018;3:6–7. https://doi.org/10.1038/s41560-017-0066-z.
[227] Kim H, Park J, Park I, Jin K, Jerng SE, Kim SH, Nam KT, Kang K. Efficient water oxidation catalyst. Nat Commun 2015;6:1–11. https://doi.org/10.1038/ncomms9253.
[228] Han B, Stoerzinger KA, Tileli V, Gamalski AD, Stach EA, Shao-Horn Y. Nanoscale structural oscillations in perovskite oxides induced by oxygen evolution. Nat Mater 2017;16:121–6. https://doi.org/10.1038/nmat4764.
[229] Cai B, Xing Y, Yang Z, Zhang W, Qiu J. High performance hybrid solar cells sensitized by organolead halide perovskites. Energy Environ Sci 2013;6:1480. https://doi.org/10.1039/c3ee40343b.
[230] Peng Q, Xiong R, Sa B, Zhou J, Wen C, Wu B, Anpo M, Sun Z. Computational mining of photocatalysts for water splitting hydrogen production: two-dimensional InSe-family monolayers. Catal Sci Technol 2017;7:2744–52. https://doi.org/10.1039/C7CY00090A.
[231] Castelli IE, Hüser F, Pandey M, Li H, Thygesen KS, Seger B, Jain A, Persson KA, Ceder G, Jacobsen KW. New light-harvesting materials using accurate and efficient bandgap calculations. Adv Energy Mater 2015;5:1400915. https://doi.org/10.1002/aenm.201400915.
[232] Hong WT, Welsch RE, Shao-Horn Y. Descriptors of oxygen-evolution activity for oxides: a statistical evaluation. J Phys Chem C 2016;120:78–86. https://doi.org/10.1021/acs.jpcc.5b10071.
[233] Suram SK, Haber JA, Jin J, Gregoire JM. Generating information-rich high-throughput experimental materials genomes using functional clustering via multitree genetic programming and information theory. ACS Comb Sci 2015;17:224–33. https://doi.org/10.1021/co5001579.
[234] Wang S, Wang Z, Setyawan W, Mingo N, Curtarolo S. Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations. Phys Rev X 2011;1:1–8. https://doi.org/10.1103/PhysRevX.1.021012.
[235] Yan J, Gorai P, Ortiz B, Miller S, Barnett SA, Mason T, Stevanović V, Toberer ES. Material descriptors for predicting thermoelectric performance. Energy Environ Sci 2015;8:983–94. https://doi.org/10.1039/C4EE03157A.
[236] Gorai P, Gao D, Ortiz B, Miller S, Barnett SA, Mason T, Lv Q, Stevanović V, Toberer ES, Design Lab TE. A virtual laboratory for thermoelectric material design. Comput Mater Sci 2016;112:368–76. https://doi.org/10.1016/j.commatsci.2015.11.006.
[237] Madsen GKH. Automated search for new thermoelectric materials: the case of LiZnSb. J Am Chem Soc 2006;128:12140–6. https://doi.org/10.1021/ja062526a.
[238] Gaultois MW, Sparks TD, Borg CKH, Seshadri R, Bonificio WD, Clarke DR. Data-driven review of thermoelectric materials: performance and resource considerations. Chem Mater 2013;25:2911–20. https://doi.org/10.1021/cm400893e.
[239] Sparks TD, Gaultois MW, Oliynyk A, Brgoch J, Meredig B. Data mining our way to the next generation of thermoelectrics. Scr Mater 2016;111:10–5. https://doi.org/10.1016/j.scriptamat.2015.04.026.
[240] Balachandran PV, Broderick SR, Rajan K. Identifying the "inorganic gene" for high-temperature piezoelectric perovskites through statistical learning. Proc R Soc

A Math Phys Eng Sci 2011;467:2271–90. https://doi.org/10.1098/rspa.2010.0543.
[241] Jain A, Hautier G, Ong SP, Persson K. New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. J Mater Res 2016;31:977–94. https://doi.org/10.1557/jmr.2016.80.
[242] Deng Z, Zhu Z, Chu I-H, Ong SP. Data-driven first-principles methods for the study and design of alkali superionic conductors. Chem Mater 2017;29:281–8. https://doi.org/10.1021/acs.chemmater.6b02648.
[243] Chen H, Hautier G, Jain A, Moore C, Kang B, Doe R, Wu L, Zhu Y, Tang Y, Ceder G. Carbonophosphates: a new family of cathode materials for Li-ion batteries identified computationally. Chem Mater 2012;24:2009–16. https://doi.org/10.1021/cm203243x.
[244] Wilmer CE, Leaf M, Lee CY, Farha OK, Hauser BG, Hupp JT, Snurr RQ. Large-scale screening of hypothetical metal–organic frameworks. Nat Chem 2012;4:83–9. https://doi.org/10.1038/nchem.1192.
[245] Moghadam PZ, Fairen-Jimenez D, Snurr RQ. Efficient identification of hydrophobic MOFs: application in the capture of toxic industrial chemicals. J Mater Chem A 2016;4:529–36. https://doi.org/10.1039/C5TA06472D.
[246] Wilmer CE, Farha OK, Bae Y-S, Hupp JT, Snurr RQ. Structure–property relationships of porous materials for carbon dioxide separation and capture. Energy Environ Sci 2012;5:9849. https://doi.org/10.1039/c2ee23201d.
[247] Banerjee D, Simon CM, Plonka AM, Motkuri RK, Liu J, Chen X, Smit B, Parise JB, Haranczyk M, Thallapally PK. Metal–organic framework with optimally selective xenon adsorption and separation. Nat Commun 2016;7. https://doi.org/10.1038/ncomms11831. [ncomms11831].
[248] Jain A, Seyed-Reihani S-A, Fischer CC, Couling DJ, Ceder G, Green WH. Ab initio screening of metal sorbents for elemental mercury capture in syngas streams. Chem Eng Sci 2010;65:3025–33. https://doi.org/10.1016/j.ces.2010.01.024.
[249] Moghadam PZ, Islamoglu T, Goswami S, Exley J, Fantham M, Kaminski CF, Snurr RQ, Farha OK, Fairen-Jimenez D. Computer-aided discovery of a metal–organic framework with superior oxygen uptake. Nat Commun 2018;9:1378. https://doi.org/10.1038/s41467-018-03892-8.
[250] Moghadam PZ, Li A, Wiggin SB, Tao A, Maloney AGP, Wood PA, Ward SC, Fairen-Jimenez D. Development of a cambridge structural database subset: a collection of metal-organic frameworks for past, present, and future. Chem Mater 2017;29:2618–25. https://doi.org/10.1021/acs.chemmater.7b00441.
[251] Lin LC, Berger AH, Martin RL, Kim J, Swisher JA, Jariwala K, Rycroft CH, Bhown AS, Deem MW, Haranczyk M, Smit B. In silico screening of carbon-capture materials. Nat Mater 2012;11:633–41. https://doi.org/10.1038/nmat3336.
[252] DeCoste JB, Weston MH, Fuller PE, Tovar TM, Peterson GW, LeVan MD, Farha OK. Metal-organic frameworks for oxygen storage. Angew Chem Int Ed 2014;53:14092–5. https://doi.org/10.1002/anie.201408464.
[253] Mellot Draznieks C, Newsam JM, Gorman AM, Freeman CM, Férey G. De novo prediction of inorganic structures developed through automated assembly of secondary building units (AASBU method). Angew Chem Int Ed 2000;39:2270–5. https://doi.org/10.1002/1521-3773(20000703)39:13<2270::AID-ANIE2270>3.0.CO;2-A.
[254] Yeo BC, Kim D, Kim H, Han SS. High-throughput screening to investigate the relationship between the selectivity and working capacity of porous materials for propylene/propane adsorptive separation. J Phys Chem C 2016;120:24224–30. https://doi.org/10.1021/acs.jpcc.6b08177.
[255] Nockemann P, Van Deun R, Thijs B, Huys D, Vanecht E, Van Hecke K, Van Meervelt L, Binnemans K. Uranyl complexes of carboxyl-functionalized ionic liquids. Inorg Chem 2010;49:3351–60. https://doi.org/10.1021/ic902406h.
[256] Roy A, Bennett JW, Rabe KM, Vanderbilt D. Half-heusler semiconductors as piezoelectrics. Phys Rev Lett 2012;109:1–5. https://doi.org/10.1103/PhysRevLett.109.037602.
[257] Rasmussen FA, Thygesen KS. Computational 2d materials database: electronic structure of transition-metal dichalcogenides and oxides. J Phys Chem C 2015;119:13169–83. https://doi.org/10.1021/acs.jpcc.5b02950.
[258] Klintenberg M, Haraldsen JT, Balatsky AV. Computational search for strong topological insulators: an exercise in data mining and electronic structure. Appl Phys Res 2014;6. https://doi.org/10.5539/apr.v6n4p31.
[259] Borysov SS, Geilhufe RM, Balatsky AV. Organic materials database: an open-access online database for data mining. PLoS One 2017;12:e0171501. https://doi.org/10.1371/journal.pone.0171501.
[260] Geilhufe RM, Borysov SS, Kalpakchi D, Balatsky AV. Towards novel organic high-T_c superconductors: data mining using density of states similarity search. Phys Rev Mater 2018;2:024802. https://doi.org/10.1103/PhysRevMaterials.2.024802.
[261] Bereau T, Andrienko D, Kremer K. Research update: computational materials discovery in soft matter. APL Mater 2016;4. https://doi.org/10.1063/1.4943287.
[262] Krishnamohan Sharma CV. Crystal engineering – where do we go from here? Cryst Growth Des 2002;2:465–74. https://doi.org/10.1021/cg0200356.
[263] Rozes L, Sanchez C. Titanium oxo-clusters: precursors for a Lego-like construction of nanostructured hybrid materials. Chem Soc Rev 2011;40:1006. https://doi.org/10.1039/c0cs00137f.
[264] Cole JM. Organic materials for second-harmonic generation: advances in relating structure to function. Philos Trans R Soc Lond A 2003;361:2751–70. https://doi.org/10.1098/rsta.2003.1271.
[265] Rizzo F, Polo F, Bottaro G, Fantacci S, Antonello S, Armelao L, Quici S, Maran F. From blue to green: fine-tuning of photoluminescence and electrochemiluminescence in bifunctional organic dyes. J Am Chem Soc 2017;139:2060–9. https://doi.org/10.1021/jacs.6b12247.
[266] Deml AM, Holder AM, O'Hayre RP, Musgrave CB, Stevanović V. Intrinsic material properties dictating oxygen vacancy formation energetics in metal oxides. J Phys Chem Lett 2015;6:1948–53. https://doi.org/10.1021/acs.jpclett.5b00710.
[267] Geilhufe RM, Borysov SS, Bouhon A, Balatsky AV. Data mining for three-dimensional organic dirac materials: focus on space group. Sci Rep 2017;7:1–7. https://doi.org/10.1038/s41598-017-07374-7.
[268] Chi W-J, Li Z-S. The theoretical investigation on the 4-(4-phenyl-4-α-naphthyl-butadieny)-triphenylamine derivatives as hole transporting materials of perovskites-type solar cells. Phys Chem Chem Phys 2015;17:5991–8. https://doi.org/10.1039/C4CP05096G.
[269] Coppens P, Fomitchev DV, Carducci MD, Culp K. Crystallography of molecular excited states. Transition-metal nitrosyl complexes and the study of transient species. J Chem Soc Dalton Trans 1998:865–72. https://doi.org/10.1039/a708604k.
[270] Zheng SL, Wang Y, Yu Z, Lin Q, Coppens P. Direct observation of a photoinduced nonstabilized nitrile imine structure in the solid state. J Am Chem Soc 2009;131:18036–7. https://doi.org/10.1021/ja9094523.
[271] Bowes KF, Cole JM, Husheer SLG, Raithby PR, Savarese TL, Sparkes HA, Teat SJ, Warren JE. Photocrystallographic structure determination of a new geometric isomer of $[Ru(NH_3)_4(H_2O)(\eta^1\text{-OSO})][MeC_6H_4SO_3]_2$. Chem Commun 2006:2448–50. https://doi.org/10.1039/b604039j.
[272] Kovalevsky AY, a Bagley K, Coppens P. The first photocrystallographic evidence for light-induced metastable linkage isomers of ruthenium sulfur dioxide complexes. J Am Chem Soc 2002;124:9241–8〈http://www.ncbi.nlm.nih.gov/pubmed/12149030〉.
[273] Fournier B, Coppens P. Measuring picosecond excited-state lifetimes at synchrotron sources. J Synchrotron Radiat 2012;19:497–502. https://doi.org/10.1107/S0909049512010710.
[274] Bitterwolf TE. Photochemical nitrosyl linkage isomerism/metastable states. Coord Chem Rev 2006;250:1196–207. https://doi.org/10.1016/j.ccr.2005.12.016.
[275] Cole JM. Applications of photocrystallography: a future perspective. Z Krist 2008;223:363–9. https://doi.org/10.1524/zkri.2008.0036.
[276] Coppens P, Sokolow J, Trzop E, Makal A, Chen Y. On the biexponential decay of the photoluminescence of the two crystallographically-independent molecules in crystals of $[Cu(I)(phen)(PPh_3)_2][BF_4]$. J Phys Chem Lett 2013;4:579–82. https://doi.org/10.1021/jz400013b.
[277] Coppens P, Novozhilova I, Kovalevsky A. Photoinduced linkage isomers of transition-metal nitrosyl compounds and related complexes. Chem Rev 2002;102:861–84. https://doi.org/10.1021/cr000031c.
[278] Cole JM. A new form of analytical chemistry: distinguishing the molecular structure of photo-induced states from ground-states. Analyst 2011;136:448–55. https://doi.org/10.1039/c0an00584c.
[279] Phillips AE. In situ X-ray characterisation of photoinduced structural changes in solid ruthenium-sulfur dioxide complexes; 2010.
[280] Makal A, Benedict J, Trzop E, Sokolow J, Fournier B, Chen Y, Kalinowski JA, Graber T, Henning R, Coppens P. Restricted photochemistry in the molecular solid state: structural changes on photoexcitation of Cu(I) phenanthroline metal-to-ligand charge transfer (MLCT) complexes by time-resolved diffraction. J Phys Chem A 2012;116:3359–65. https://doi.org/10.1021/jp300313s.
[281] Cole JM. Photocrystallography. Acta Crystallogr A 2008;64:259–71. https://doi.org/10.1107/S0108767307065324.
[282] Kovalevsky AY, Bagley KA, Cole JM, Coppens P. Light-induced metastable linkage isomers of ruthenium sulfur dioxide complexes. Inorg Chem 2003;42:140–7. https://doi.org/10.1021/ic025997g.
[283] Negre CFA, Young KJ, Oviedo MB, Allen LJ, Sánchez CG, Jarzembska KN, Benedict JB, Crabtree RH, Coppens P, Brudvig GW, Batista VS. Photoelectrochemical hole injection revealed in polyoxotitanate nanocrystals functionalized with organic adsorbates. J Am Chem Soc 2014;136:16420–9. https://doi.org/10.1021/ja509270f.
[284] Baxter JB, Richter C, Schmuttenmaer CA. Ultrafast carrier dynamics in nanostructures for solar fuels. Annu Rev Phys Chem 2014:423–47. https://doi.org/10.1146/annurev-physchem-040513-103742.
[285] Correa Baena JP, Steier L, Tress W, Saliba M, Neutzner S, Matsui T, Giordano F, Jacobsson TJ, Srimath Kandada AR, Zakeeruddin SM, Petrozza A, Abate A, Nazeeruddin MK, Grätzel M, Hagfeldt A. Highly efficient planar perovskite solar cells through band alignment engineering. Energy Environ Sci 2015;8:2928–34. https://doi.org/10.1039/C5EE02608C.
[286] Gao C, Wang J, Xu H, Xiong Y. Coordination chemistry in the design of heterogeneous photocatalysts. Chem Soc Rev 2017;46:2799–823. https://doi.org/10.1039/c6cs00727a.
[287] Gumerova NI, Rompel A. Synthesis, structures and applications of electron-rich polyoxometalates. Nat Rev Chem 2018;2. https://doi.org/10.1038/s41570-018-0112.