ChemComm

HIGHLIGHT

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2023, **59**, 5795

Machine learning integrated photocatalysis: progress and challenges

Luyao Ge, Yuanzhen Ke and Xiaobo Li *

Received 28th February 2023,
Accepted 12th April 2023

DOI: 10.1039/d3cc00989k

rsc.li/chemcomm

Discovering efficient photocatalysts has long been the goal of photocatalysis, which has traditionally been driven by serendipitous or try-and-error strategies. Recent developments in photocatalysis integrated with machine learning techniques promise to accelerate the discovery of photocatalysts, but are also facing significant challenges. In this review, advances in machine learning integrated photocatalysis are first presented from the perspective of three main photocatalytic processes: light harvesting, charge generation and separation, and surface redox reactions. Next, progress in using machine learning to understand complex photoactivity–structure relationships and identify the factors governing activity follows. A future photocatalysis paradigm is then provided with the integration of artificial intelligence, robots and automation. Lastly, we discuss the current challenges in machine learning integrated photocatalysis. This review aims to provide a systematic overview and guidelines to the broad scientific community interested in photocatalysis and artificial intelligence for solar fuel synthesis.

1. Introduction

Nonrenewable energy resources, coal, oil and natural gas, are becoming overexploited as society advances. The exhaustion of nonrenewable energy resources is accompanied by the emission of vast amounts of greenhouse and toxic gases, endangering both the environment and human society. To address the energy crisis and environmental issues, developing clean, renewable energies has become a social priority. Solar energy is one of the most abundant and renewable sources of energy.[1] In 1972, Honda and Fujishima discovered that water could split into H_2 and O_2 on the surface of TiO_2 when exposed to ultraviolet light.[2] This discovery opened the field of photocatalysis, a clean process of converting solar energy into chemical energy in the presence of semiconductor photocatalysts.[3]

Since the 1970s, photocatalysis reactions have been widely developed, including water splitting,[4–6] CO_2 reduction,[7–11] *etc.* Nevertheless, the Solar-to-Chemical conversion efficiency is still less than the required targets for practical application. For example, only very few systems have exhibited Solar-to-Hydrogen (STH) efficiencies exceeding 1%, and most reported systems have a maximum STH of ~0.1%.[12] Photocatalysis is still in an early stage of development in terms of efficiency and requires significant advancements.

Designing and producing a photocatalytic system is a significant challenge. It requires advanced knowledge and synthetic methodologies, assembling photocatalytic units into a device, accomplishing light-induced charge carrier generation/separation/migration, and chemical reaction of charge carriers on the surface (*i.e.*, water oxidation/reduction).[13] Historically, the discovery of many significant breakthroughs in the development of photocatalysts, such as TiO_2, Ta_3N_5, GaN:ZnO, Al:$SrTiO_3$, carbon nitride, *etc.*, was driven by serendipitous or try-and-error methods,[14–17] which are primarily determined by probability: performing a large number of experiments must increase the likelihood of an outcome. However, an exhaustive search costs time and resources, so it is impractical to synthesize and test every semiconductor material.

Machine learning is a crucial area of artificial intelligence that creates models and resolves challenging issues using a data-driven methodology.[18] Through the use of algorithms, the models may "learn on their own," identifying patterns in vast volumes of data and applying those patterns to forecast future samples. Also, it can manage large-scale data systems and resolve multidimensional issues. Machine learning techniques have already been successful in a number of scientific fields, including functional materials,[19–23] biology,[24–26] catalysis,[27–31] batteries[32–34] and organic synthesis.[35,36]

Photocatalysis is by nature a multivariate problem, involving a host of factors spanning multiple length scales, such as bandgaps, thermodynamic driving forces, charge carrier mobility, reaction sites, surface areas and others. Thus, machine learning techniques are acknowledged by the photocatalysis community to supplement the research on this complex system, with the main objective of generating predictive models

Key Laboratory of the Ministry of Education for Advanced Catalysis Materials, Zhejiang Key Laboratory for Reactive Chemistry on Solid Surfaces, Zhejiang Normal University, Jinhua 321004, China. E-mail: xiaobo.li@zjnu.edu.cn

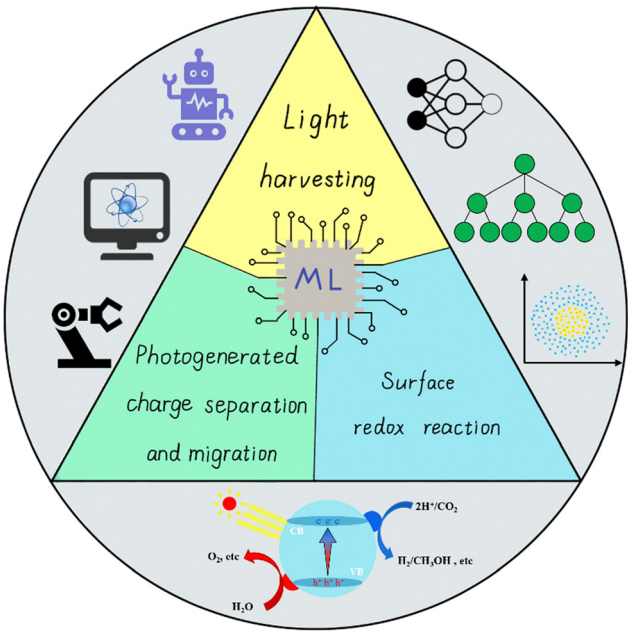

Fig. 1 Schematic diagram of machine learning integrated photocatalysis.

and gaining physical insights: to allow for fast evaluation of potential candidates or to understand the relationship between structure and photoactivity.[31,37]

The advent of machine learning techniques offers the promise of accelerating the discovery of photocatalysts. By learning a given amount of data sets and training them to create machine learning models, machine learning is used to forecast prospective photocatalysts. Moreover, machine learning can be integrated with automation and robots to provide a future research paradigm for photocatalysis research.[38–42]

The basics of photocatalysis have already been described in great detail, and thus, they won't be discussed here. This review aims to present an overview of the machine learning integrated research in photocatalysis (Fig. 1). Note that it does not cover the work of applying machine learning in the photoelectrocatalysis field[43–47] and photocatalytic degradation.[48–51] The advances are first presented from a photocatalysis perspective, focusing on three primary processes: light absorption, charge generation and separation, and surface redox reaction. Then, progress in using machine learning to understand the complex structure–photoactivity relationship and gain insights into the governing factors of activity follows. A future photocatalysis paradigm is then provided with the integration of artificial intelligence, robots and automation. Lastly, the challenges that machine learning integrated photocatalysis face, such as data management and descriptor engineering, are discussed. This review offers a systematic overview and guidelines to the broad scientific community interested in photocatalysis and artificial intelligence for solar fuel synthesis.

2. Machine learning integrated photocatalysis

2.1 Light harvesting

The light harvesting of a photocatalyst determines the theoretical limit of its solar energy utilization efficiency.[52] In the solar spectrum, UV light accounts for 4% and visible light for 50%. Therefore, it is essential to develop photocatalytic materials that absorb visible light. In addition, the conduction band and valence band positions determine the thermodynamic driving force of carriers, respectively.

Bandgap values are mainly obtained from the UV-vis absorption spectrum. Recently a strategy has been developed to extract bandgaps from material observation images without spectral acquisition. Using a set of high throughput instruments, Stein *et al.*[53] constructed a sample space containing 178 994 distinct materials. Variational autoencoders (VAE) were trained on this large experimental dataset using convolutional and deep neural networks, which allowed the prediction of the UV-Vis absorption spectra of the materials from the images (Fig. 2). A material image autoencoder was then developed to enable bandgaps to be extracted from predicted spectra instead of being calculated by *ab initio* methods. Furthermore, the relationship between material images and absorption spectra is used to create

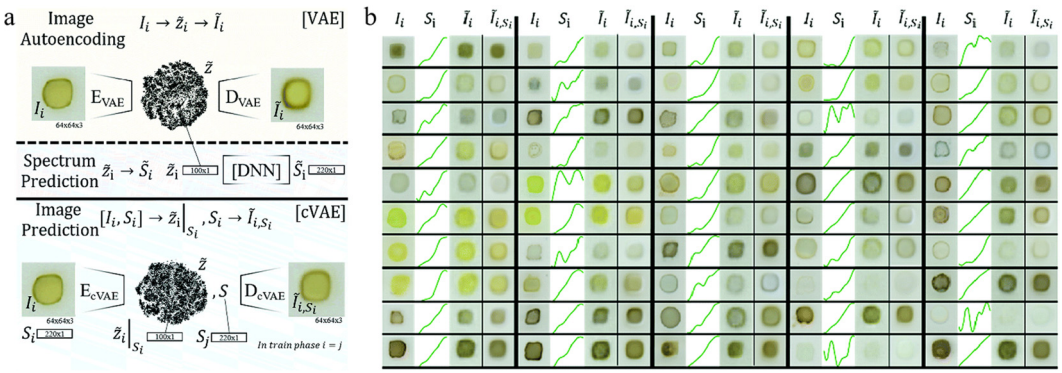

Fig. 2 (a) Schematic visualization of the 3 types of learning models for optical properties of materials. (b) Reconstruction comparison from VAE and cVAE of randomly selected images from the test set. There are 5 large columns of images separated by thick black lines, and within each of these the 4 columns of data are the measured image, measured absorption spectrum, VAE reconstructed image, and cVAE-reconstructed image, respectively. Reprinted with permission from ref. 53.

predictive models for images of materials with specific light absorption properties.

Double perovskite oxide $A_2BB'O_6$ has better stability than ABO_3-type oxide. Xinyang Wan *et al.*[54] used a combination of machine learning and first principles to facilitate rapid screening of double perovskite oxide $A_2BB'O_6$ for overall photocatalytic water splitting (Fig. 3). Around 2500 double perovskite oxides were collected from multiple databases of perovskite materials, and bandgaps were calculated with high throughput calculation. A two-step modeling method, coupling with feature selection, was employed to predict the bandgap. Nearly 8000 candidates with proper bandgaps for water splitting are screened out from 56 894 $A_2BB'O_6$-type double perovskite oxides. Statistical analysis of the results shows that double perovskite oxides containing d^{10} metal ions at the B/B′ position mostly met the bandgap requirements for overall photocatalytic water splitting. The first-principles calculations further found that Sr_2GaSbO_6, Sr_2InSbO_6 and K_2NbTaO_6 have suitable edge positions for overall photocatalytic water splitting.

TiO_2 is a classical photocatalyst but can only absorb UV light. Doping is widely applied to narrow the bandgaps of TiO_2. The lattice parameters and surface area are strongly correlated with bandgap values, which are conventionally simulated and studied through first-principal models, but these models require significant computational resources. Yun Zhang *et al.*[55] collected experimental data of 60 doped-TiO_2 photocatalysts from the literature and developed a Gaussian process regression (GBR) model to predict bandgaps of anatase TiO_2. With the lattice constant as the structural parameter and the surface area as the morphological parameter, the GBR well reveals the relationship between structural and morphological parameters and bandgaps. It was further found that the model could be used for bandgap prediction of undoped or doped-TiO_2 synthesized by different preparation processes.

Khmaissia *et al.*[56] constructed a dataset containing atomic and crystallographic data for ternary chalcopyrite semiconductors, which are compounds that crystallize in the tetragonal form (ABC_2 formula). Two extra descriptors, bond dissociation energy and bond length, were added to the previously-developed machine learning model for the predictions of chalcopyrite bandgaps.[57,58] The original subset of 15 features was reduced to 7 features using the sequential forward feature selection technique, and the prediction accuracy was improved by approximately 40%. Furthermore, the results show that the features associated with the last two elements of chalcopyrite are more relevant to bandgap prediction.

2.2 Photogenerated charge separation and migration

The photogenerated charges must separate and migrate to the surface before they recombine. Thus, charge separation and migration are critical for photocatalytic systems.

Reducing the migration distance of carriers to the surface active site can significantly reduce the chances of charge recombination. Thus, two-dimensional materials are potential candidates as photocatalysts. Kumar *et al.*[59] created a database of 3099 two-dimensional octahedral materials (2DO) with physical properties calculated from first principles. The machine learning model was constructed by considering compositional and chemical hardness features. The SHapley Additive ExPlanations (SHAP) value analysis shows that the predicted highly stable 2DO materials follow the Hard–Soft–Acid–Base (HSAB) principle. A high throughput screening of the database, under criteria such as stability, bandgap and standard redox potentials, yielded 21 potential 2DO materials for overall photocatalytic water splitting (Fig. 4).

Hao Jin *et al.*[60] developed an efficient method for predicting 2D multicomponent photocatalysts using machine learning techniques. Two machine learning models were developed to predict bandgap and band edge positions of 2D photocatalysts, respectively. From more than 4000 2D materials, 75 multicomponent photocatalytic candidates that meet the conditions for photocatalytic water splitting were selected (Fig. 5). It was found that the multinary compounds $A_2P_2X_6$ and ABP_2X_6 with A = Cu/Zn/Ge/Ag/Cd, B = Ga/In/Bi and X = S/Xe have proper bandgap and band edge positions, making them promising photocatalyst candidates.

Doping is another effective strategy for tuning charge separation efficiency in metal oxide-based semiconductors. Despite decades of

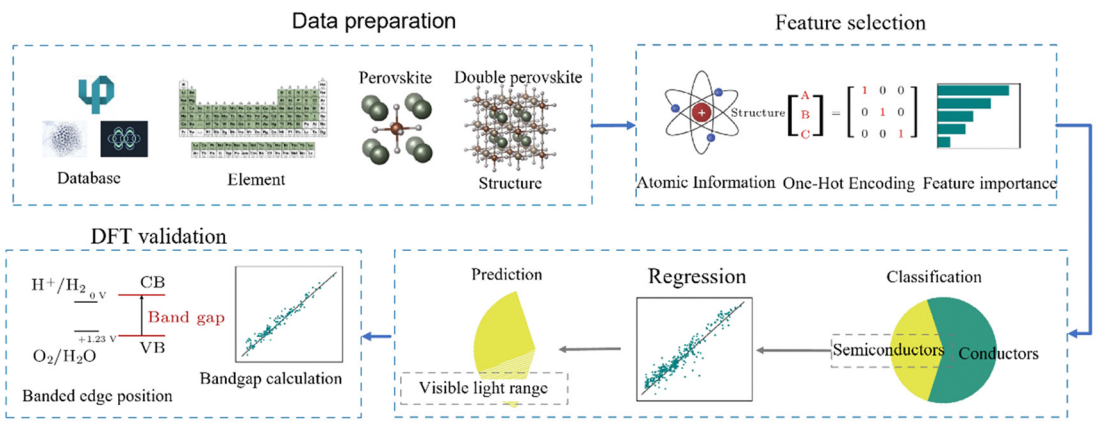

Fig. 3 Multistep machine learning-based screening framework for double perovskite oxides. There are four steps including data collection, feature selection, machine learning process and DFT verification. Reprinted with permission from ref. 54.

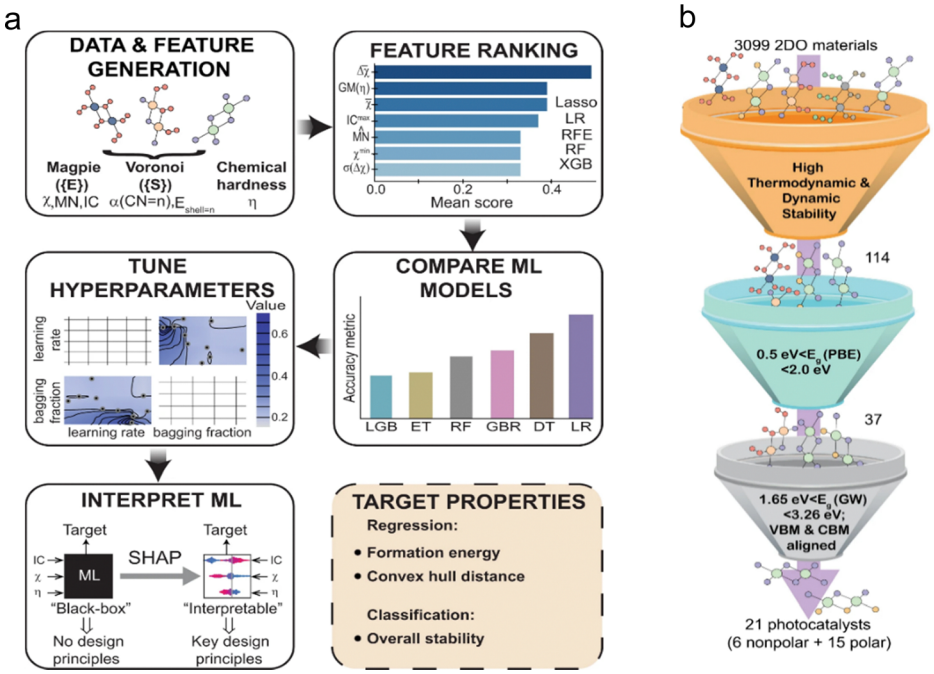

Fig. 4 (a) Schematic of the workflow for machine learning applied in the study. (b) The high-throughput scheme utilized in the current study for screening stable 2DO photocatalysts. Reprinted with permission from ref. 59.

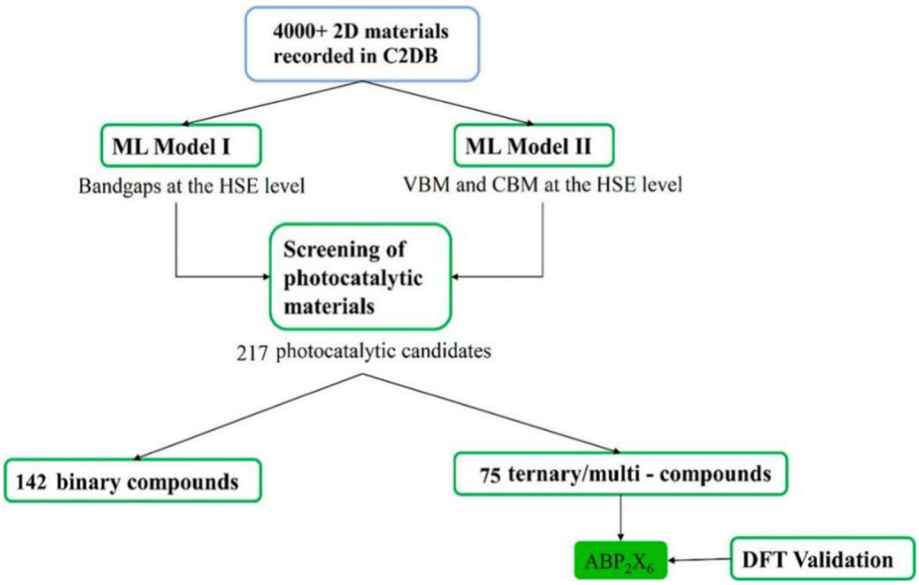

Fig. 5 Schematic diagram to illustrate the search procedure for 2D photocatalytic materials. Reprinted with permission from ref. 60.

extensive research, the dopant selection largely depends on a trial-and-error approach. Zhiliang Wang *et al.*[61] used machine learning techniques to guide the doping of metal oxides in solar-driven photoelectrochemical water splitting (Fig. 6). Using Fe_2O_3 photoanodes as an example, the photochemical properties of Fe_2O_3 with different doping concentrations of 17 doping elements were experimentally determined, and a database containing more than 700 data points was composed. The photocurrent density measured in H_2O_2/NaOH solution was predicted for the charge separation and transfer (CST) of the semiconductor. In terms of $j_{H_2O_2}$, the CST of Zr- and Pt-doped Fe_2O_3 is significantly improved compared to pure Fe_2O_3. Furthermore, the results show that the high dopant concentrations hinder the CST process. The chemical state, ionic radius, and metal–oxygen bond formation enthalpy were found to have the most significant influence on the CST performance by SHAP analysis. Therefore, the dopant selection with high M–O bond formation enthalpy and large ionic radius difference is favorable to promote charge separation. In addition, the dopant

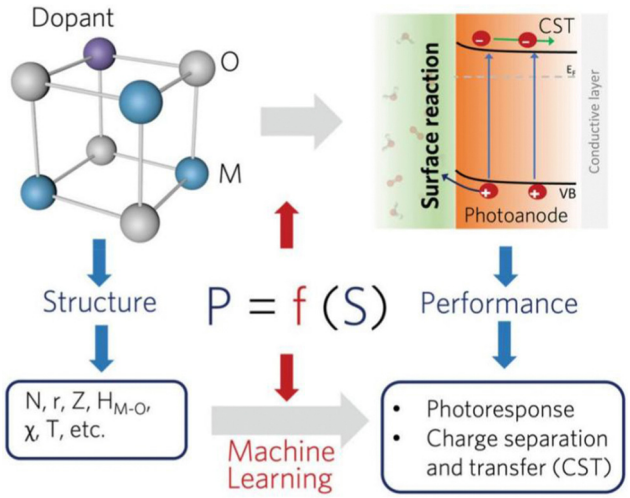

Fig. 6 The schematic illustration of the machine learning guided dopant selection towards an efficient PEC process. Reprinted with permission from ref. 61.

selection criteria derived from machine learning can be extended to CuO-based photoelectrodes, showing the generality of the model.

2.3 Surface redox reaction

The redox reaction on the photocatalyst surface is the final step in the photocatalytic process. However, the surface reaction (*i.e.*, water oxidation) is associated with multiple electron transfers, suffering high overpotential and sluggish kinetics.[62]

Hao Yuan[63] *et al.* used machine learning algorithms to discover new descriptors for predicting the activity of the double-doped system $CsPbBr_3$–$CsPbCl_3$ heterostructures for photocatalytic water splitting. The binding energy between the doped atom and the vacancy was calculated to assess the stability of the doping. Seven metal ions (Ti, V, Cr, Mn, Fe, Co and Ni) were used as dopant metal ions. A total of 49 systems were combined to study the results, showing that some calcium ion systems were unstable and had positive binding energies. Calculations of the HER and OER catalytic process were carried out for the remaining 36 systems, and the best candidate was selected as the $CsPbBr_3$:Ni–$CsPbCl_3$:Co system with a bandgap value of 2.26 eV and a high light absorption coefficient. Using the LASSO method, a descriptor $x = \frac{\sqrt{\frac{X_A}{X_B} + \frac{X_A}{S}}}{\frac{V_A}{X_A} + \frac{X_A}{S}}$ consisting of

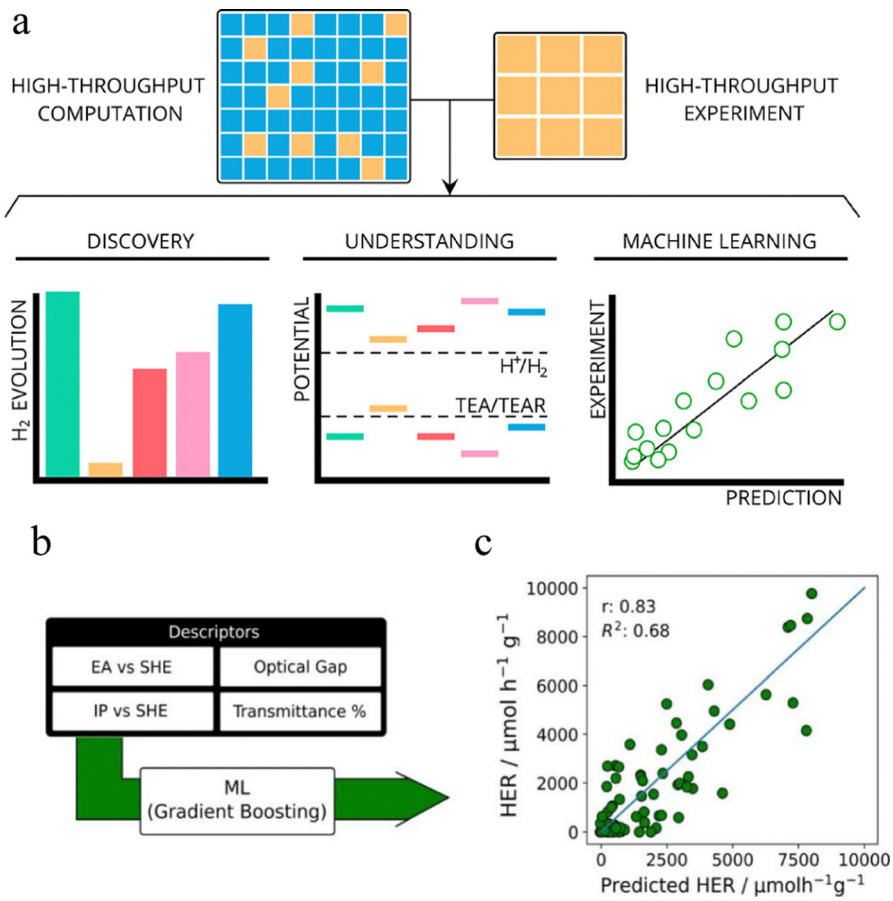

Fig. 7 (a) Schematic of the workflow for machine learning applied in the study. (b) Properties used to train the gradient-boosting model, where ionization potential (IP), electron affinity (E_A), and optical gap are calculated, and transmittance is measured experimentally. (c) Experimentally observed HER vs HER predicted using a gradient-boosted trees machine-learning model. The model is evaluated by leave-one-out cross validation, meaning the data shown are for copolymers not considered during training. Reprinted with permission from ref. 64.

electronegativity (X), valence electron number (V) and surface indicator (S) was discovered. The results show that the descriptor has a linear relationship with the overpotential of the OER, which provides a guide for designing photocatalysts.

2.4 Structure–activity relationship

Developing structure–property relationships for photocatalysts can predict performance or provide insights into photocatalyst design. However, photocatalyst structure–property relationships are multivariate and complex, making extracting patterns and building prediction models challenging. The Cooper group at The University of Liverpool pioneered machine learning to learn the complex structure–property relationships in photocatalysis.[64,65]

Conjugated polymers are an emerging class of hydrogen producing photocatalysts. Yang Bai *et al.*[64] developed a sub-library of over 170 copolymers selected from 6354 copolymers using a combination of high throughput experiments and theoretical calculation. By comparing the experimental and predicted hydrogen evolution rates, the activity of the copolymers was found to be related to the calculated electronic properties (the electron affinity, the ionization potential, and the optical gap) of the polymer and the dispersion of the polymer in the reaction mixture. This relationship was coded into the machine learning model, which was found to explain up to 68% of the variation in the HER between polymers (Fig. 7).

Xiaobo Li *et al.*[65] collected a library of 572 aromatic organic molecules with diverse compositions and structures, obtaining a comparable dataset consisting of 572 photocatalytic hydrogen evolution data points (Fig. 8). Unsupervised learning and supervised classification reveal the structural features and optoelectronic properties that positively impact the activity of these molecular photocatalysts for hydrogen production, which

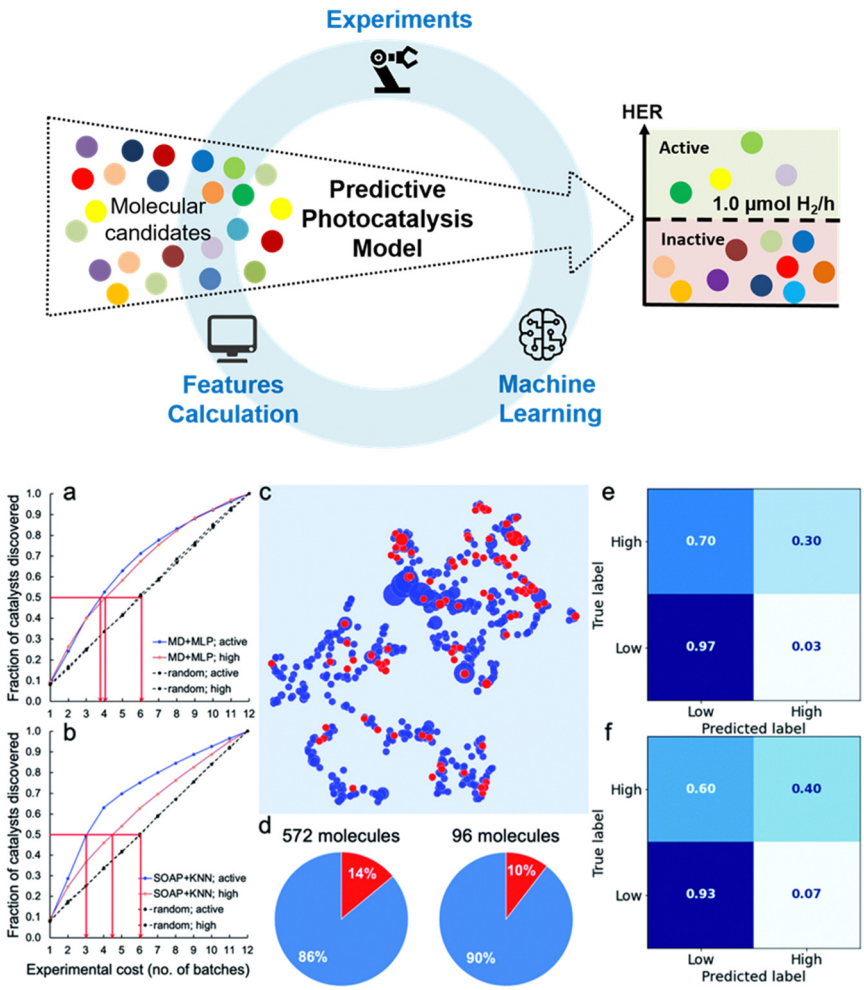

Fig. 8 Schematic of the workflow in the study. Virtual experiments and blind tests. (a) and (b) Virtual experiments comparing an adaptive machine learning approach with random sampling: the 572 molecules were encoded by the molecular descriptors and trained with machine learning P models (a) or encoded by the SOAP descriptors and trained with KNN models (b). (c)–(f) Blind tests of the machine learning models trained on the 572 molecules (referred to as the 572-molecule library) for 96 unseen molecules (referred to as the blind-test set). (c) 2D UMAP embedding of the chemical space (encoded by SOAP) of the 572-molecule library (in blue) and the blind-test set (in red); the symbol size is scaled by the experimentally measure HER. (d) Percentages (in red) of the active samples (HERs $>$ 1.07 μmol h^{-1}) in the 572-molecule library and the blind-test set. (e) and (f) Confusion matrices for the predictions of the blind-test set by models based on the MD + machine learning P protocol (e) or the SOAP + KNN protocol (f), both trained on the 572-molecule library. Reprinted with permission from ref. 65.

also allowed some physical interpretations: for example, the formation of triplet excitons seems to have a beneficial effect. Virtual experiments show that an adaptive machine learning-assisted selection approach outperforms random sampling, significantly reducing the experimental cost of identifying the active photocatalysts in the library. A further evaluation of the trained machine learning advisor on a blind test set of 96 molecules confirmed its potential in assisting the discovery of new molecular photocatalysts.

Yuzhi Xu *et al.*[66] used machine learning techniques to achieve hydrogen evolution prediction of alternating conjugated copolymers. 157 organic conjugated polymers with existing HER data were collected from the literature, and their electronic property composition descriptors were calculated by DFT and used to train the model (Fig. 9). Two types of multidimension fragmentation descriptors were developed, of which the structure-based multi-dimension fragmentation descriptor helped to achieve high accuracy in electronic property prediction. In addition, a machine learning model trained on an electronic property-based multi-dimension fragmentation descriptor was developed to predict the HER with a measured accuracy = 0.91. Lastly, the machine learning technique was combined with high-throughput computing to discover a new copolymer material with high photocatalytic properties using a virtual generator.

Conjugated polyelectrolytes (CPEs) are versatile organic materials with diverse applications. Yangyang Wan *et al.*[67] constructed a first-principles database of CPEs by combining machine learning with high-throughput first-principles calculations to establish structure–property relationships for CPE materials. It is shown that the HOMO/LUMO front orbitals and bandgap of CPE materials are related to the electrostatic interactions between the ionic group and the counter ion on the CPE backbone. Machine learning reveals that bandgap depends primarily on the backbone, especially in relation to $HOMO_D$ and $LUMO_A$.

Elemental doping of graphite-phase carbon nitride can significantly increase its photocatalytic activity. A machine learning model was developed by Liqing Yan *et al.*[68] for exploring the effect of elemental doping on the rate of photocatalytic hydrogen production. The database was built from published research papers on photocatalytic H_2 generation using D-g-C_3N_4 as a photocatalyst. H_2 evolution rate was selected as the output, while experimental conditions were used as inputs for machine learning model training. The synthesis parameters of the material, the properties of the material and the conditions of H_2 production were used as features to fit the rate of H_2 production. Using the SHAP technique, it was found that the synthesis conditions of the material (type of dopant, type of precursor, method of synthesis) were the main factors affecting the rate of hydrogen production. The data were grouped according to the type of non-metal doped elements and it was found that the machine learning predictions matched the experimental results, except for the O elements. Among them, P element doping resulted in the highest average hydrogen production rate and the largest SHAP value for D-g-C_3N_4, making it the best candidate for D-g-C_3N_4 doping (Fig. 10).

Copper-based semiconductors, *i.e.*, Cu_2O, have excellent catalytic performance in photocatalytic CO_2 reduction reactions,

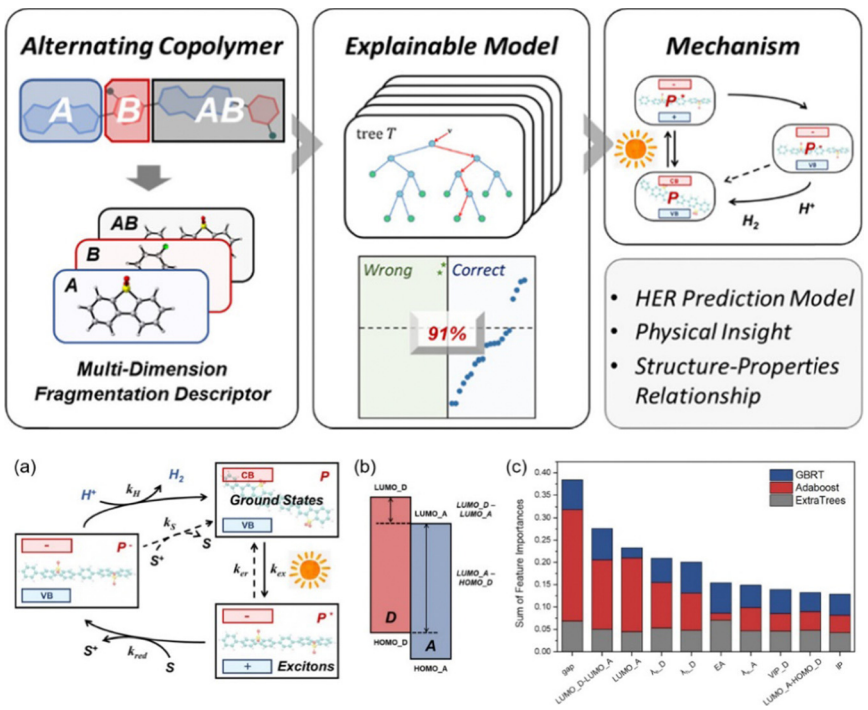

Fig. 9 Schematic of the workflow in the study. (a) Diagram representing the photocatalytic HER (*S* response to the sacrificial reagent). (b) Schematic of the energy level of the fragmentations in A–B alternating copolymers and selected energy level difference. (c) Sum of top 10 important descriptors in EPMDFD selected by three DT-based classifier models. Reprinted with permission from ref. 66.

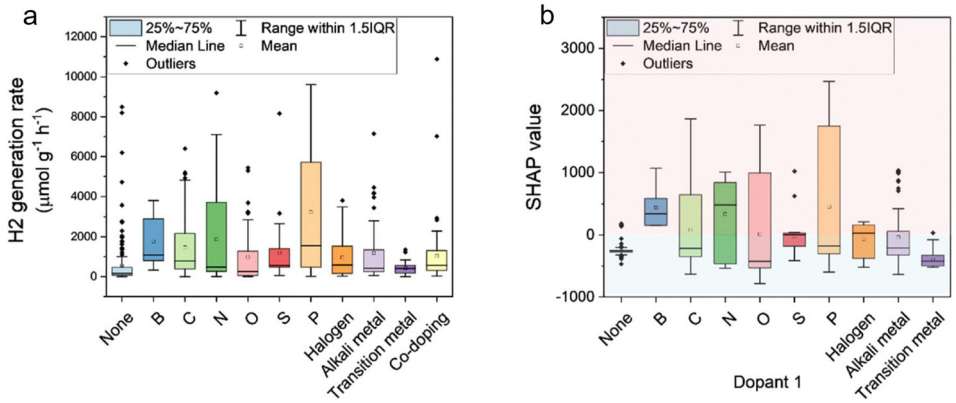

Fig. 10 Experimental H_2 production rate (a) and SHAP value (b) grouped by doping element. Reprinted with permission from ref. 68.

but our understanding of the intrinsic link between structure and properties is unclear. A data-driven approach to predicting the rate of photocatalytic reduction of CO_2 to methanol conversion on Cu_2O was modelled by Voleti *et al.*[69] to gain insight into the structure–operation–property relationship. 505 data points were extracted from 68 papers by data mining to develop machine learning models. Five machine learning models were tested for statistical performance, and GBRT (tree-based models) was found to be the best model for predicting the rate of methanol production from CO_2 conversion. Furthermore, it was found that the active metal component and the light source were the experimental conditions that contributed most to the predicted rate model.

Mageed[70] compared a variety of machine learning models for photocatalytic hydrogen production from ethanol over copper oxide nanoparticles. Among them, LMNN (Levenberg–Marquardt neural networks) had the highest R value of 0.998. The importance of the input parameters was analyzed using the Garson algorithm, with irradiation time and CuO content having the most significant effect on hydrogen production.

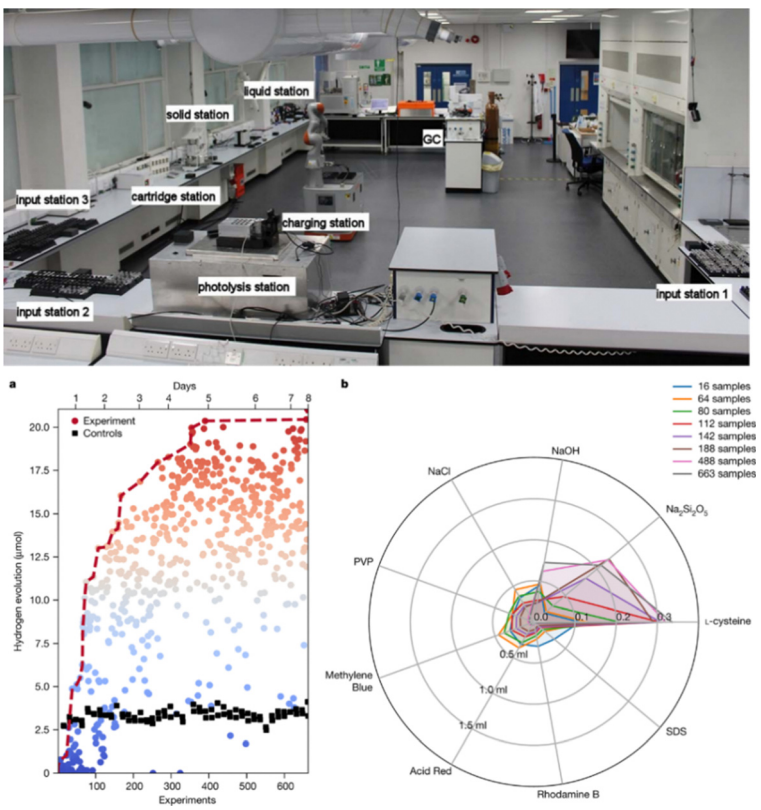

Fig. 11 Laboratory space used for the autonomous experiments. (a) Plot showing hydrogen evolution achieved per experiment in an autonomous search that extended over 8 days. (b) Radar plot showing the evolution of the average sampling of the search space in milliliters; the scale denotes the fraction of maximum solution volume dispensed. Reprinted with permission from ref. 38.

3. Autonomous discovery in photocatalysis

Chemistry experiments are tedious, involving repetitive solid measuring, liquid dispersing *etc.* Embracing automation in the chemistry lab can significantly reduce the repetitive work. Apart from saving time, reliability, productivity and safety in the chemistry lab could also be potentially improved. The combination of automation with AI presents new opportunities in chemical research. In photocatalysis, Cooper *et al.*[38] exemplify this with the autonomous discovery of improved photocatalysts for hydrogen generation (Fig. 11). The robotic chemist is able to transport the vials between different functional stations to complete tasks, including solid and liquid component dispersion, capping under certain atmospheres, photolysis and gas chromatography measurement. With built-in Bayesian optimization into a mobile robotic workflow, the robotic chemist can autonomously navigate the ten-dimensional space, finding the optimized reaction conditions for hydrogen evolution. The robot operates them in essentially the same way that a human researcher would. And more modules could be added depending on the research tasks.

Later, groups in USTC built an all-round AI-Chemist that included a machine reading module to capture existing chemical knowledge by automatically reading massive chemical literature, a mobile robot module to produce experimental data by executing various chemical experiments, and a computational brain module to generate physics/theory-based predictive models by carrying out theoretical calculations (Fig. 12).[39] The competence of the AI-Chemist has been scrutinized by three different chemical tasks, including photocatalytic degradation of rhodamine B (RhB). The equipped computational brain could bias searches towards components that are more likely to yield the desired property. This will be important for search spaces with even larger numbers of components where purely combinatorial approaches may become inefficient.

4. Challenges and outlook

Machine learning has emerged to make it more efficient to discover high-performance photocatalysts and to understand the complex structure–activity relationship. It allows the researcher to consider a much broader chemical space than we have contemplated so far and points out important parameters to consider while designing photocatalytic systems. However, several questions remain to be addressed. How big of a dataset is required for machine learning to recognize the structure–property pattern? This is somehow up to the size and diversity of the chemical space defined. For chemical space with constraints on structure and functions, as exemplified in the polymer photocatalysts case,[64] the size of the dataset may not need to be substantial to achieve a moderate performance of machine learning, but the generality of the model to other kinds of photocatalyst would be limited. In the case of molecular photocatalysts,[65] apart from aromaticity and availability, no other prior knowledge about the desirable properties of the candidate photocatalysts was applied in the library selection, thus minimizing prior chemical knowledge from skewing the structure–activity correlation. The generality of the generated model performance is expected to be better. However, as a result of this broad selection approach, the model performance is deemed to be modest due to the limited size and unbalanced data structure (low activity data dominates) in the dataset. Apparently, it needs more extensive and larger standardized datasets to have a chance of learning the underlying structure–property rules. For photocatalytic overall water splitting systems, the developed photocatalysts are limited

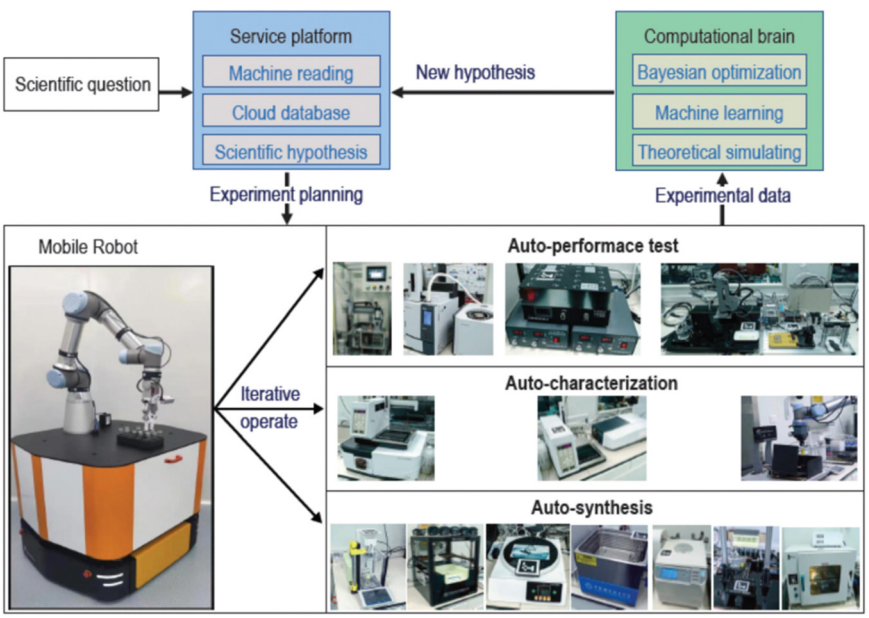

Fig. 12 The workflow of the AI-Chemist and the functions of each module. Reprinted with permission from ref. 39.

and sparse in many different types of materials, so collecting extensive and qualified data would be a big challenge. To complicate things further, small changes in experimental parameters such as temperature, pressure and light duration or wavelength can cause large fluctuations in catalytic performance: as such, notwithstanding the surge of interest and publications in this area, the lack of experimental standardisation between laboratories makes it challenging to implement data-mining approaches across a large number of different studies. We can envisage that this may only be realized with a collective effort from the whole community. Somehow, sharing the samples and data in the community is necessary.

Another question is, ''what kind of knowledge patterns can be gained from machine learning?''. It is well recognized that the nature of catalysis is complex, as many of the catalytic processes involve multiple steps and dynamic active sites; many experimental researchers in the field of photocatalysis (or heterogeneous catalysis) are skeptical about whether current machine learning techniques can extract knowledge patterns while the mechanism of the (photo)catalysis is not yet clear. Note that in homogeneous-based catalysis with defined active sites, machine learning is more adopted with success.[71–77] In the photovoltaic area, where more standardization of the test system is adopted, machine learning has also shown potential.[78–81] Therefore, to extract meaningful knowledge patterns from photocatalysis with machine learning, besides the data requirements stated above, looking for descriptors describing the primary steps of photocatalysis is essential too. For example, photo-induced charge transfer from a photocatalyst to catalytic surface sites is key in ensuring photocatalytic efficiency. However, descriptors describing the charge-transfer, surface redox reaction on cocatalysts, *etc.*, are associated with solid states or defects, which are challenging to measure experimentally or computationally. Recently, Can Li *et al.* demonstrated that quasi-ballistic inter-facet electron transfer and spatially selective trapping are the dominant processes that facilitate efficient charge separation in photocatalysis.[82] To improve the performance of machine learning, descriptors related to anisotropic facets and defect structures could be considered.

As machine learning techniques continue to advance, models developed are increasingly difficult to interpret and are often used as black box models. How to interpret the models is still a tough issue. Understanding the logic behind prediction and suggestion made by the model can provide insights for the design of next-generation photocatalysts. The more interpretable the models are, the easier for the research community to adopt the approach and trust the model. Also, it is worth noting that overfitting is a common problem in machine learning: the model performs well on training data but can't generalize with unseen data. Techniques, such as train/test split, feature selection, and cross-validation, are recommended to prevent overfitting.

The exploration of photocatalysts is to search for photocatalysts with desirable properties. Analogous to the natural photosynthesis system, the long search for photocatalysts will most likely be a complex system containing components whereby each contributes to the required functions such as light absorption, charge transfer and surface redox reactions. The complex and multivariate relationship between photocatalysts and performance points to the fact that the desirable properties must be discrete and intersect with each other. As such, there is great value in considering machine learning to deconvolute the multidimensional dataset to accelerate the search of the photocatalyst.[83] The examples picked up in this review demonstrate that the exploration of photocatalysts led by machine learning is a promising new approach to accelerate the discovery of photocatalysts.

Perhaps the most significant barrier to adopting this strategy in solar fuel synthesis is the collection of big standardized data. This challenge pushes researchers in photocatalysis to rethink the value of data generated in the laboratory and the use of data. For example, the ''negative'' data, deliberately skewed from publication and buried in the paper notebook, are valuable for machine learning,[84] and the essential nature of reporting standardized data.[85] Another barrier is engineering descriptors describing the primary process of photocatalysis.

In conclusion, this does not mean the expert knowledge of chemists has been less critical in developing photocatalysts. It needs chemists to define the chemical space and rules of exploration by choosing search algorithms and approaches. The use of machine learning is to allow the chemist to go beyond their bias and allow them to consider a much bigger chemical space, go into the unknown for the exploration of photocatalysts, and have a systematical view into the complex structure–activity relationship. By integrating machine learning with automation and robots, autonomous discovery of photocatalysts in the future could be envisioned. Indeed, we can hope that, along with machine learning, it will catalyze several future photocatalyst discoveries.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors acknowledge the funding from Zhejiang Normal University and the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (No. 2022R01007).

References

1 N. Kannan and D. Vakeesan, *Renewable Sustainable Energy Rev.*, 2016, **62**, 1092–1105.
2 A. Fujishima and K. Honda, *Nature*, 1972, **238**, 37–38.
3 A. Kudo and Y. Miseki, *Chem. Soc. Rev.*, 2009, **38**, 253–278.
4 Q. Wang and K. Domen, *Chem. Rev.*, 2020, **120**, 919–985.
5 K. A. Hajj and K. A. Whitehead, *Nat. Rev. Mater.*, 2017, **2**, 17056.
6 Z. Wang, C. Li and K. Domen, *Chem. Soc. Rev.*, 2019, **48**, 2109–2125.
7 J. Ran, M. Jaroniec and S.-Z. Qiao, *Adv. Mater.*, 2018, **30**, 1704649.
8 J. Low, B. Cheng and J. Yu, *Appl. Surf. Sci.*, 2017, **392**, 658–686.
9 J. Fu, B. Zhu, C. Jiang, B. Cheng, W. You and J. Yu, *Small*, 2017, **13**, 1603938.
10 O. Ola and M. M. Maroto-Valer, *J. Photochem. Photobiol., C*, 2015, **24**, 16–42.

11 J. Yu, J. Low, W. Xiao, P. Zhou and M. Jaroniec, *J. Am. Chem. Soc.*, 2014, **136**, 8839–8842.
12 Z.-S. Wang, H. Kawauchi, T. Kashima and H. Arakawa, *Coord. Chem. Rev.*, 2004, **248**, 1381–1389.
13 K. Takanabe, *ACS Catal.*, 2017, **7**, 8006–8022.
14 T. Takata, J. Jiang, Y. Sakata, M. Nakabayashi, N. Shibata, V. Nandal, K. Seki, T. Hisatomi and K. Domen, *Nature*, 2020, **581**, 411–414.
15 X. Wang, K. Maeda, A. Thomas, K. Takanabe, G. Xin, J. M. Carlsson, K. Domen and M. Antonietti, *Nat. Mater.*, 2009, **8**, 76–80.
16 M. Hara, G. Hitoki, T. Takata, J. N. Kondo, H. Kobayashi and K. Domen, *Catal. Today*, 2003, **78**, 555–560.
17 K. Maeda, K. Teramura, D. Lu, T. Takata, N. Saito, Y. Inoue and K. Domen, *Nature*, 2006, **440**, 295.
18 Z. Ghahramani, *Nature*, 2015, **521**, 452–459.
19 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
20 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, *Nat. Mater.*, 2016, **15**, 1120–1127.
21 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D. G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, *Nat. Mater.*, 2016, **15**, 1120–1127.
22 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. J. N. Walsh, *Nature*, 2018, **559**, 547–555.
23 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Comput. Mater.*, 2017, **3**, 54.
24 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, *Nat. Rev. Drug Discovery*, 2019, **18**, 463–477.
25 M. W. Libbrecht and W. S. Noble, *Nat. Rev. Genet.*, 2015, **16**, 321–332.
26 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.
27 M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C. T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C. S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S. C. Lo, A. Ip, Z. Ulissi and E. H. Sargent, *Nature*, 2020, **581**, 178–183.
28 J. R. Kitchin, *Nat. Catal.*, 2018, **1**, 230–232.
29 P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen and T. Bligaard, *ChemCatChem*, 2019, **11**, 3581–3601.
30 S. Ma and Z.-P. Liu, *ACS Catal.*, 2020, **10**, 13213–13226.
31 H. Mai, T. C. Le, D. Chen, D. A. Winkler and R. A. Caruso, *Chem. Rev.*, 2022, **122**, 13478–13515.
32 M. F. Ng, J. Zhao, Q. Yan, G. J. Conduit and Z. W. Seh, *Nat. Mach.*, 2020, **2**, 161–170.
33 P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y. H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon and W. C. Chueh, *Nature*, 2020, **578**, 397–402.
34 E. Chemali, P. J. Kollmeyer, M. Preindl and A. Emadi, *J. Power Sources*, 2018, **400**, 242–255.
35 J. M. Granda, L. Donina, V. Dragone, D. L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.
36 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.
37 H. Masood, C. Y. Toe, W. Y. Teoh, V. Sethu and R. J. A. C. Amal, *ACS Catal.*, 2019, **9**, 11774–11787.
38 B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237–241.
39 Q. Zhu, F. Zhang, Y. Huang, H. Xiao, L. Zhao, X. Zhang, T. Song, X. Tang, X. Li, G. He, B. Chong, J. Zhou, Y. Zhang, B. Zhang, J. Cao, M. Luo, S. Wang, G. Ye, W. Zhang, X. Chen, S. Cong, D. Zhou, H. Li, J. Li, G. Zou, W. Shang, J. Jiang and Y. Luo, *Natl. Sci. Rev.*, 2022, **9**, 190.
40 C. W. Coley, N. S. Eyke and K. F. Jensen, *Angew. Chem., Int. Ed.*, 2020, **59**, 22858–22893.
41 C. W. Coley, N. S. Eyke and K. F. Jensen, *Angew. Chem., Int. Ed.*, 2020, **59**, 23414–23436.
42 L. Buglioni, F. Raymenants, A. Slattery, S. D. A. Zondag and T. Noël, *Chem. Rev.*, 2022, **122**, 2752–2906.
43 D. Guevarra, L. Zhou, M. H. Richter, A. Shinde, D. Chen, C. P. Gomes and J. M. Gregoire, *npj Comput. Mater.*, 2022, **8**, 57.
44 B. Sa, R. Hu, Z. Zheng, R. Xiong, Y. Zhang, C. Wen, J. Zhou and Z. Sun, *Chem. Mater.*, 2022, **34**, 6687–6701.
45 K. P. Sokol and V. Andrei, *Nat. Rev. Mater.*, 2022, **7**, 251–253.
46 L. Zhang, W. Hu, M. He, K. Xu and Z. Pan, *J. Phys. Chem. C*, 2022, **126**, 6482–6490.
47 B. Oral, E. Can and R. Yildirim, *Int. J. Hydrogen Energy*, 2022, **47**, 19633–19654.
48 K. H. Ng, Y. S. Gan, C. K. Cheng, K.-H. Liu and S.-T. Liong, *Environ. Pollut.*, 2020, **267**, 115500.
49 A. H. Navidpour, A. Hosseinzadeh, Z. Huang, D. Li and J. L. J. C. R. Zhou, *Chem. Rev.*, 2022, **392**, 1–26.
50 Z. H. Jaffari, A. Abbas, S.-M. Lam, S. Park, K. Chon, E.-S. Kim and K. H. Cho, *J. Hazard. Mater.*, 2023, **442**, 130031.
51 Z. Jiang, J. Hu, M. Tong, A. C. Samia, H. Zhang and X. Yu, *Catalysts*, 2021, **11**, 1107.
52 S. Rühle, *Sol. Energy*, 2016, **130**, 139–147.
53 H. S. Stein, D. Guevarra, P. F. Newhouse, E. Soedarmadji and J. M. Gregoire, *Chem. Sci.*, 2019, **10**, 47–55.
54 X. Wan, Y. Zhang, S. Lu and Y. Wu, *Acta Phys.-Chim. Sin.*, 2022, **71**, 177101.
55 Y. Zhang and X. Xu, *ACS Omega*, 2020, **5**, 15344–15352.
56 F. Khmaissia, H. Frigui, M. Sunkara, J. Jasinski, A. M. Garcia, T. Pace and M. Menon, *Comput. Mater. Sci.*, 2018, **147**, 304–315.
57 Y. Zeng, S. J. Chua and P. Wu, *Chem. Mater.*, 2002, **14**, 2989–2998.
58 C. Suh, A. Rajagopalan, X. Li and K. J. S. Rajan, *Mater. Res. Soc.*, 2003, **1999**, 333–342.
59 R. Kumar and A. K. Singh, *npj Comput. Mater.*, 2021, **7**, 197.
60 H. Jin, X. Tan, T. Wang, Y. Yu and Y. Wei, *J. Phys. Chem. Lett.*, 2022, **13**, 7228–7235.
61 Z. Wang, Y. Gu, L. Zheng, J. Hou, H. Zheng, S. Sun and L. Wang, *Adv. Mater.*, 2022, **34**, 2106776.
62 C. Ding, J. Shi, Z. Wang and C. Li, *ACS Catal.*, 2017, **7**, 675–688.
63 H. Yuan, Y. Min and L. Xu, *J. Phys. Chem. Lett.*, 2021, **12**, 822–828.
64 Y. Bai, L. Wilbraham, B. J. Slater, M. A. Zwijnenburg, R. S. Sprick and A. I. Cooper, *J. Am. Chem. Soc.*, 2019, **141**, 9063–9071.
65 X. Li, P. M. Maffettone, Y. Che, T. Liu, L. Chen and A. I. Cooper, *Chem. Sci.*, 2021, **12**, 10742–10754.
66 Y. Xu, C. W. Ju, B. Li, Q. S. Ma, Z. Chen, L. Zhang and J. Chen, *ACS Appl. Mater. Interfaces*, 2021, **13**, 34033–34042.
67 Y. Wan, F. Ramirez, X. Zhang, T.-Q. Nguyen, G. C. Bazan and G. Lu, *npj Comput. Mater.*, 2021, **7**, 69.
68 L. Yan, S. Zhong, T. Igou, H. Gao, J. Li and Y. Chen, *Int. J. Hydrogen Energy*, 2022, **47**, 34075–34089.
69 L. D. Voleti, R. Kamesh and K. Y. Rani, *Mater. Today*, 2023, **72**, 494–499.
70 A. K. Mageed, *Biomass Convers. Biorefin.*, 2023, **13**, 3319–3327.
71 J. T. Margraf, H. Jung, C. Scheurer and K. Reuter, *Nat. Catal.*, 2023, **6**, 112–121.
72 T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa and K.-I. Shimizu, *ACS Catal.*, 2020, **10**, 2260–2297.
73 S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan and R. B. Sunoj, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 1339–1345.
74 K. Jorner, A. Tomberg, C. Bauer and C. Sköld, *Nat. Rev. Chem.*, 2021, **5**, 240–255.
75 A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, 5631.
76 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89–96.
77 L.-C. Xu, J. Frey, X. Hou, S.-Q. Zhang, Y.-Y. Li, J. C. A. Oliveira, S.-W. Li, L. Ackermann and X. Hong, *Nat. Synth.*, 2023, **2**, 321–330.
78 W. Sun, Y. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li and K. Sun, *Sci. Adv.*, 2019, **5**, 4275.
79 X. Du, L. Lüer, T. Heumueller, J. Wagner, C. Berger, T. Osterrieder, J. Wortmann, S. Langner, U. Vongsaysy, M. Bertrand, N. Li, T. Stubhan, J. Hauch and C. J. Brabec, *Joule*, 2021, **5**, 495–506.

80 B. Cao, L. A. Adutwum, A. O. Oliynyk, E. J. Luber, B. C. Olsen, A. Mar and J. M. Buriak, *ACS Nano*, 2018, **12**, 7434–7444.
81 T. W. David, H. Anizelli, T. J. Jacobsson, C. Gray, W. Teahan and J. Kettle, *Nano Energy*, 2020, **78**, 105342.
82 R. Chen, Z. Ren, Y. Liang, G. Zhang, T. Dittrich, R. Liu, Y. Liu, Y. Zhao, S. Pang, H. An, C. Ni, P. Zhou, K. Han, F. Fan and C. Li, *Nature*, 2022, **610**, 296–301.
83 P. S. Gromski, A. B. Henson, J. M. Granda and L. Cronin, *Nat. Rev. Chem.*, 2019, **3**, 119–128.
84 P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.
85 Z. Wang, T. Hisatomi, R. Li, K. Sayama, G. Liu, K. Domen, C. Li and L. Wang, *Joule*, 2021, **5**, 344–359.

