Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

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

Research
Material Science and Engineering—Article

Fifth Paradigm in Science: A Case Study of an Intelligence-Driven Material Design

Can Leng [a,b,c], Zhuo Tang [c,d,*], Yi-Ge Zhou [e], Zean Tian [d], Wei-Qing Huang [f], Jie Liu [a,b], Keqin Li [d,g], Kenli Li [c,d,*]

[a] Science and Technology on Parallel and Distributed Processing Laboratory, National University of Defense Technology, Changsha 410073, China
[b] Laboratory of Software Engineering for Complex Systems, National University of Defense Technology, Changsha 410073, China
[c] National Supercomputing Center in Changsha, Changsha 410082, China
[d] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
[e] Institute of Chemical Biology and Nanomedicine, State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China
[f] Department of Applied Physics, School of Physics and Electronics, Hunan University, Changsha 410082, China
[g] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

ARTICLE INFO

Article history:
Received 8 December 2021
Revised 6 June 2022
Accepted 29 June 2022
Available online xxxx

Keywords:
Catalytic materials
Fifth paradigm
Intelligence-driven
Machine learning
Synergy of interdisciplinary experts

ABSTRACT

Science is entering a new era—the fifth paradigm—that is being heralded as the main character of knowledge integrating into different fields to intelligence-driven work in the computational community based on the omnipresence of machine learning systems. Here, we vividly illuminate the nature of the fifth paradigm by a typical platform case specifically designed for catalytic materials constructed on the Tianhe-1 supercomputer system, aiming to promote the cultivation of the fifth paradigm in other fields. This fifth paradigm platform mainly encompasses automatic model construction (raw data extraction), automatic fingerprint construction (neural network feature selection), and repeated iterations concatenated by the interdisciplinary knowledge ("volcano plot"). Along with the dissection is the performance evaluation of the architecture implemented in iterations. Through the discussion, the intelligence-driven platform of the fifth paradigm can greatly simplify and improve the extremely cumbersome and challenging work in the research, and realize the mutual feedback between numerical calculations and machine learning by compensating for the lack of samples in machine learning and replacing some numerical calculations caused by insufficient computing resources to accelerate the exploration process. It remains a challenging of the synergy of interdisciplinary experts and the dramatic rise in demand for on-the-fly data in data-driven disciplines. We believe that a glimpse of the fifth paradigm platform can pave the way for its application in other fields.

© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The earth-shaking changes in human society are inseparable from the exploration of nature. Such transformative changes have evolved from focusing on natural observations to gradually being realized through various tools and cutting-edge methods [1,2]. In this process, different normative development paradigms covering the overall and interrelated assumptions of various disciplines have been formed [3,4]. Each paradigm shift is caused by the result of changes in the basic assumptions within the ruling theory in a certain era to meet the subsequent requirements, thereby creating a new paradigm [5]. The fifth paradigm has now been characterized as the intelligence-driven and knowledge-centric research paradigm following the paradigm shift from the data-intensive fourth paradigm and is coming on the heels of experimentation, theory, and computer-simulation paradigm shifts from the first to the third paradigms [6–10].

In the fifth paradigm world view, the exploration of the physical universe is not merely projected by the mathematical probable realm of intensive data-driven by intelligence, but the entire research process also involves the undifferentiated conscious process of human expert knowledge. Based on these features, the application of the fifth paradigm can be regarded as a cognitive system or cognitive application [9,10]. Taking the development of materials science as an example, the cognitive system of the fifth

* Corresponding authors.
E-mail addresses: ztang@hnu.edu.cn (Z. Tang), lkl@hnu.edu.cn (K. Li).

https://doi.org/10.1016/j.eng.2022.06.027
2095-8099/© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article as: C. Leng, Z. Tang, Y.-G. Zhou et al., Fifth Paradigm in Science: A Case Study of an Intelligence-Driven Material Design, Engineering, https://doi.org/10.1016/j.eng.2022.06.027

paradigm has evolved from the primitive early paradigm via classic evolutionary spiral processes, in which materials such as metals and ceramics were discovered and used in ancient times before the emergence of Newton's laws and the advent of the theory of relativity. Then, the emergence of relativity and quantum mechanics made it possible to simulate the electronic structure of molecules [11–13]. In recent years, the meteoric rise of artificial intelligence (AI) and machine learning has been transformational to the research of data-driven materials design [14–18]. Therefore, by processing relevant innovative technologies into ever-larger datasets, the hidden properties of new materials, such as metals and ceramics, can be revealed [19–22]. Since then, cognitive materials design has taken the relay baton and formed a new ecosystem through the intellectual collaboration of interdisciplinary experts, thus greatly accelerating the exploration process.

At present, the fifth paradigm is in its emergent period and still has a long way to go. Unlike the mature fourth paradigm of data-intensive science, which has exploded rapidly in multiple application domains and has been used in industrial and scientific fields such as self-driving cars, computer vision, and brain modeling [23–27], the intelligence-driven, knowledge-centric fifth paradigm is still in the stage of vigorous development because it needs to break the boundaries of computational and data-intensive paradigms to form a new ecosystem by merging and extending existing technologies. Fortunately, scientists are now on the road to researching and solving these problems. For example, a Spark–message passing interface (MPI) integrated platform proposed by Malitsky et al. [10] can be used to promote the transformation of the fourth paradigm processing pipeline represented by data-intensive applications to the fifth paradigm of knowledge-centric applications. Cognitive computing, such as natural language processing, knowledge representation, and automatic reasoning, is exactly what Zubarev and Pitera [9] suggested that the fifth paradigm should possess. Furthermore, common aspects among diverse computing applications can be inferred in the fifth paradigm by the integration of expert knowledge in different fields and the intensive data from experimental observation and theoretical simulations, steering the development of complementary solutions to meet emerging and future challenges. Therefore, although the task of developing the fifth paradigm is arduous, the prospects of its application are broad.

The strategic transition from data-intensive science toward the fifth paradigm of composite cognitive computing applications is a long-term journey with many unknowns. This paper addresses the fifth paradigm platform by dissecting a framework called generalized adsorption simulations in Python (GASpy) in catalytic materials[†] [28], aiming to bring together human wisdom, algorithms in high-performance scientific computing, and deep-learning approaches for tackling new frontiers of data-driven discovery applications. The remainder of the paper is organized as follows. Section 2 provides a brief overview and a discussion of the fifth paradigm platform. Section 3 further elaborates on the performance evaluation of the platform. Finally, Section 4 concludes with a summary.

2. A platform of the fifth paradigm

In the process of materials research, processing the synergy among experimental data, theoretical models, and machine learning requires experts in different fields to collaboratively analyze and process data, that is, huge human wisdom is needed. Therefore, the intelligence-driven function with knowledge-centric characters that combine each link with versatility and operate in a platform-like manner is particularly important. Here, we introduce a platform of the fifth paradigm used in catalytic materials, as shown in Fig. 1. The platform of the fifth paradigm couples the third and fourth paradigms, and the latter two include the process of the first and second paradigms. Among them, the original data come from experimental observations in the first paradigm and theoretical guidance in the second paradigm, as well as numerical calculations in the third paradigm, which then can be intelligence-driven by machine learning in the fourth paradigm. Combining the knowledge of the work integration of experimental experts and theoretical experts, the materials selected by machine learning can be screened for the second time, and the screening results will be fed back to the numerical simulation of the third paradigm again. The results obtained in the third paradigm can still be driven by the data in the fourth paradigm. The prediction results can then be filtered again through the knowledge integration of experimental and theoretical experts and then fed back to the third paradigm for numerical simulation. These approaches have produced the fifth paradigm platform, which continuously provides samples for machine learning by intelligently controlling the calculation of high-throughput physical models to compensate for the lack of machine-learning samples. Moreover, by using the knowledge integrated into different fields, machine learning can be used to replace part of numerical calculations to solve the problem of the time-consuming massive model due to insufficient computing resources.

The comprehensive work of the fifth paradigm platform stems from the framework designed by Tran and Ulissi [28], for the bimetallic catalysts research in materials science, which uses machine learning to accelerate the numerical calculation based on density functional theory (DFT) that is conducted by a Vienna *ab initio* simulation package (VASP) [29], and can intelligently drive the discovery of high-performance electrocatalysts. The platform can classify the active sites of each stable low-index surface of bimetallic crystals, resulting in hundreds and thousands of possible active sites. At the same time, an alternative model based on artificial neural networks was used to predict the catalytic activity of these sites [30]. The discovered sites with high activity can be further used for future DFT calculations.

2.1. Automatic model construction and verification

The ability of raw data extraction to be driven by intelligence is reflected in automatic model construction and verification in the

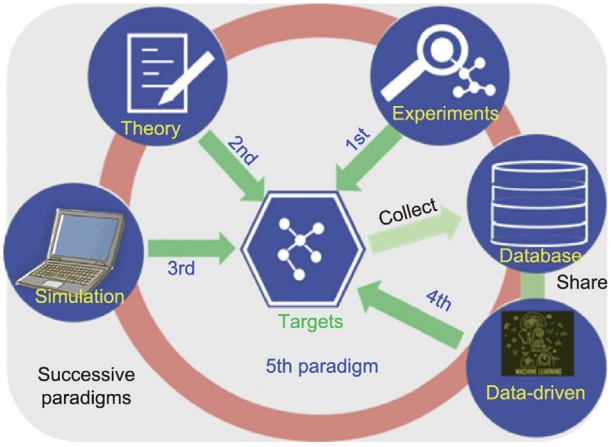


Fig. 1. The paradigms in science. The evolution of the scientific paradigm has been developed from the simple 1st paradigm to the complex 5th paradigm. The core of the 5th paradigm is knowledge-centric and intelligence-driven including the successive paradigms from 1st, 2nd, 3rd to 4th marked by the experiments, theory, simulation, and data-driven process, respectively.

[†] https://github.com/ulissigroup/GASpy

fifth paradigm platform. The ever-larger structures with and without the adsorbates can be automatically constructed and verified by DFT calculations. Since the adsorption of surface species is an indispensable process in heterogeneous catalysis, constructing many structures in experiments and DFT calculations can be time-intensive before determining the catalytic activity by evaluating the adsorption energy. Therefore, automated model construction and verification are essential to solving the problem.

As shown in Fig. 2, the entire task calculation includes the preparation process of raw data for standard simulation and then the numerical calculation. All the raw data used for the theoretical simulation come from the Material Project website, which can be realized by the module of gas/bulk generation through the Generate_Gas/Generate_Bulk function, and they can be processed into a list form with the items of user information, task location, calculation status, and other attributes, as well as be stored in the database by the update_atom_collection function with the collection creation named "Firework," "Atoms," "Catalog," and "Adsorption."

The relaxation calculation of the task in Fig. 2(a) can then be generated by the FireWorks workflow manager for submission in

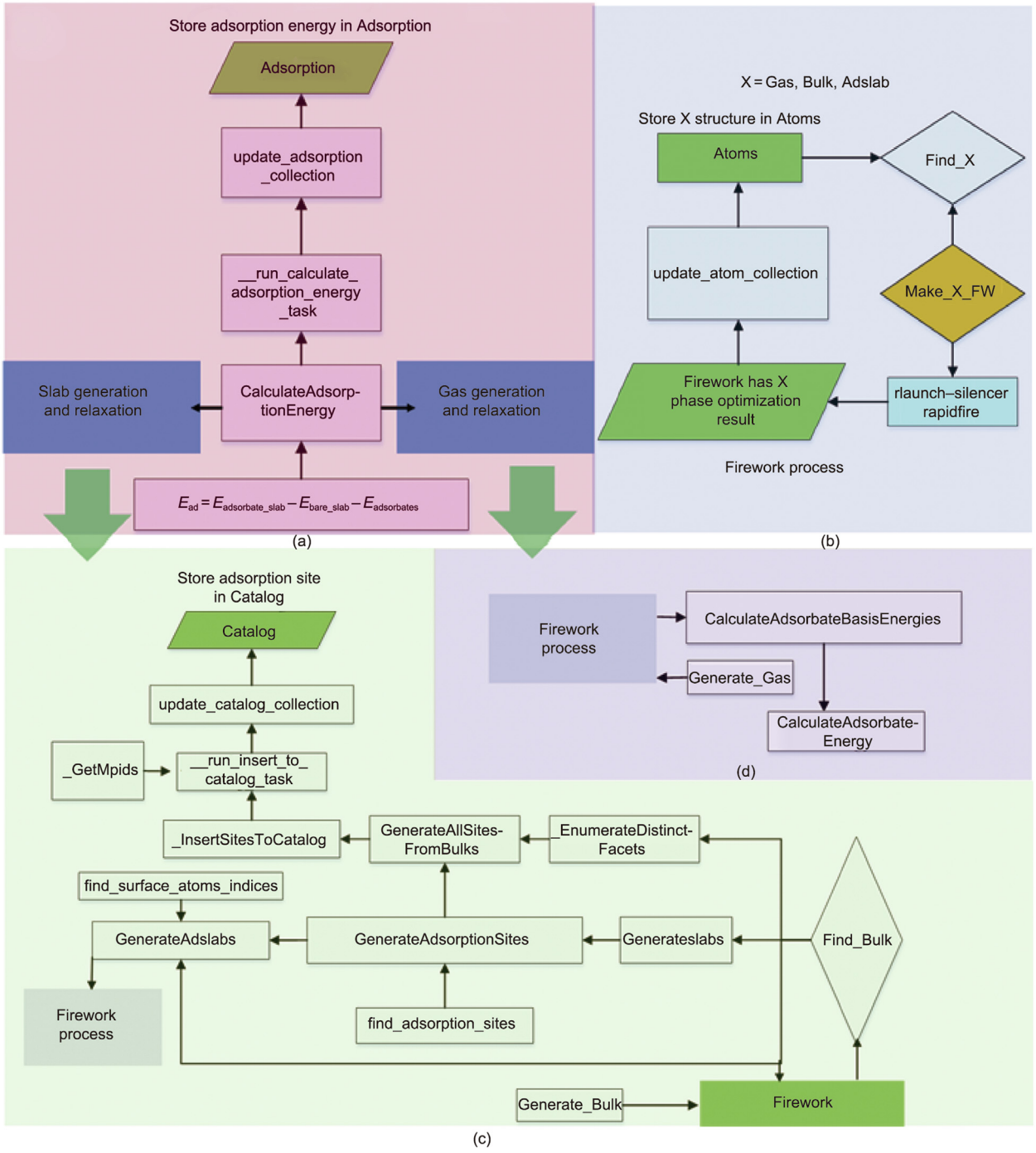


Fig. 2. The framework of this fifth paradigm case. The intelligence-driven of raw data extraction in the framework of GASpy is realized through modules of atomic operation, generation, and calculation. (a) The function of the module is to automatically calculate the adsorption energy from the gas and slab phase in the fifth paradigm platform. (b) This module is used to automatically create high-throughput tasks for the optimization of gas, bulk, and adslab with/without adsorbates through Firework. (c, d) The modules represent the (c) slab generation and (d) gas generation and structural relaxation described in part (a).

Fig. 2(b). The attribute of the results in FireWorks contains "gas-phase optimization" as the list format for gas relaxation, as well as the "unit cell optimization" for bulk optimization (bulk_relaxation). The attribute "status" is the calculation status of "COMPLETED," "RUNNING," "READY," and other statuses, such as "FIZZLED," among others, and is judged by the Find_Bulk/Find_Gas function to either store the completed calculation process in the Atoms collection or generate a FireWorks task workflow waiting for calculation that has not started yet.

If the status determined by Find_Bulk/Find_Gas is "COMPLETED," on one hand, the calculated result will be stored in the database. On the other hand, the irreducible crystal face index enumeration (realized by the EnumerateDistinctFacets function) can be carried out by obtaining the optimized crystal structure from the Atoms collection, followed by crystal slab cutting to generate slabs (realized by the Generateslabs function) according to the given Miller index, and then, all adsorption sites on the slab (realized by the GenerateAdsorptionSites function) are found by the extending primitive units (the function of Atom_operates), enumerating crystal slabs, and adding adsorbents, as shown in Figs. 2(c) and (d). For all the adsorption sites on all bulk materials, the GenerateAllSitesFromBulks function, composed of the EnumerateDistinctFacets function and GenerateAdsorptionSites function, can enumerate the irreducible Miller index in each slab and generate all the adsorption sites. All such information is written into the Catalog collection by the function update_catalog_collection.

Furthermore, for each slab in which the adsorption site has been found, the adsorbates will be added to the adsorption sites by the GenerateAdslabs function to generate a "slab + adsorbate optimization" calculation model (adslab_relaxation); in addition, the adsorbates can also be eliminated by the GenerateAdslabs function to generate a "bare slab optimization" calculation model (bare_slab_relaxation). These calculation models can then be submitted for calculation through the FireWorks workflow manager.

When completed, all the calculated results will be stored in the database collections by the function update_atom_collection. The Find_Adslab function will determine whether a relaxation task should be started by finding if there is a corresponding calculated result in the Atoms collection. For the adsorption energy E_{ad} calculation, the CalculateAdsorptionEnergy function is used to extract the gas energy $E_{adsorbates}$, adsorbate_slab energy $E_{adsorbate\_slab}$, and bare_slab energy E_{bare_slab} from the Atoms collection: $E_{ad} = E_{adsorbate\_slab} - E_{bare\_slab} - E_{adsorbates}$. The E_{ad} and the associated initial and final structure information can then be added to the Adsorption collection by the update_adsorption_collection function, where the neural network feature selection that will be discussed next can be extracted as the input of machine learning. Thus, the process of intelligence-driven model construction and verification is realized.

2.2. Automated fingerprint construction

The intelligence-driven quality of a neural network feature selection is reflected in the automatic fingerprint construction in the fifth paradigm platform. In this framework, the automatically constructed fingerprint is converted from all the atomic structures of each material adsorption model into a graphical representation of the numerical input of a convolutional neural network (CNN) [31]. In the atomic structure information, three types of features are considered, as shown in Fig. 3, namely atomic feature (F_{N1}), neighbor feature (F_{N2}), and connection distance (F_{N3}). The basic atomic properties in atomic feature characteristics are atomic number, electronegativity, coordination number/covalent radius, group, period, the valence electron, first ionization energy, electron affinity, block, and atomic volume. The basic neighborhood feature properties are composed of the coordination number between adjacent atoms near the adsorption site calculated by the Voronoi polyhedron algorithm [32]. The connection distances are the distances from the adsorbate to all atoms. The target fingerprint is the adsorption energy (E_{adN}).

The process of automatic fingerprint construction includes the process of extracting the final structures and adsorption energy by DFT calculation, fingerprint generation, and the process of machine learning, as well as the learning problem stating. The fingerprint constructed in GASpy comes from the original model without DFT calculation and the DFT calculation result. After DFT calculation, the initial targets E_{adN} are obtained, as shown in Fig. 3(a), and then, these DFT relaxation structures are used to

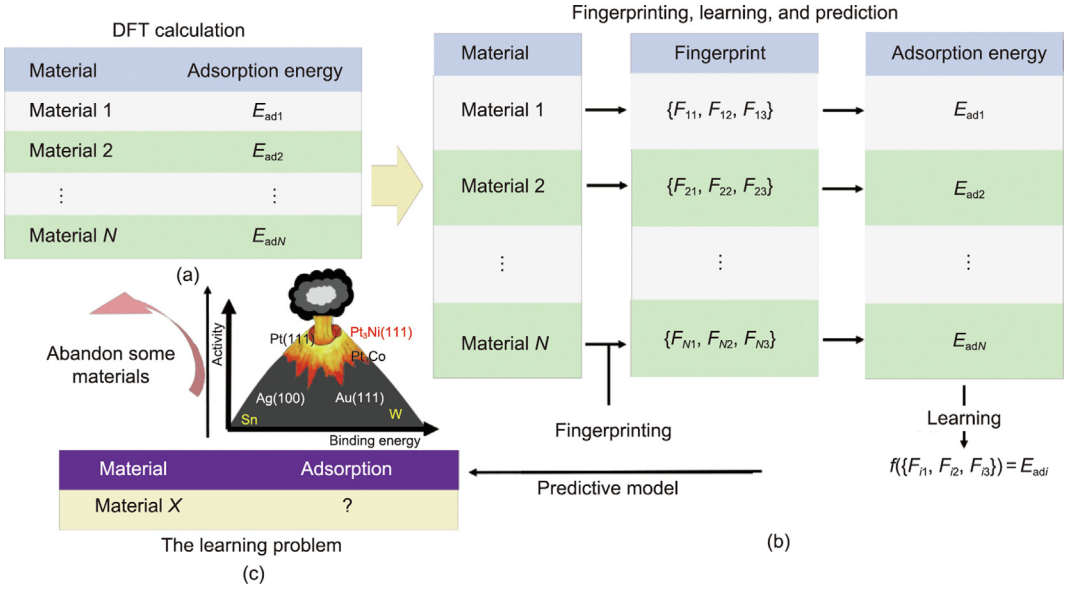


Fig. 3. The intelligence-driven of neural network feature selection in the fifth paradigm platform. It is realized by the automatic fingerprint construction in the framework of GASpy. (a) The DFT calculation is schematically viewed as an example dataset (*N* is the number of training examples); (b) the automatic fingerprint construction is achieved by a predictive model through the fingerprinting and learning steps process; (c) the learning problem is stated, followed by abandoning some materials from the learning results through the scaling relationship, and carrying out further DFT calculation screening.

extract fingerprints {F_{N1}, F_{N2}, F_{N3}} for learning and prediction, as shown in Fig. 3(b). These features will be used as a cross-validation dataset in machine learning, and then, the function f will be found by the learning process for the next prediction. In the prediction process, the fingerprints are obtained from the initial structure without any DFT calculation and are used to predict the adsorption energy of material X, as shown in Fig. 3(c), and then, the DFT calculation candidates required for the next cycle are screened through the learning problem. This learning problem is determined by the famous scaling relationship [33,34], as shown in Fig. 3. The scaling relationship is the adsorption energy–catalytic activity (also known as the binding energy–catalytic activity) curve, like a volcano, which rises first and then declines, also known as the "volcano plot."

The data on adsorption energy and catalytic activity in the scaling relationship come from the work of many attempts by theoretical and experimental scientists and are further used by AI experts to screen the results of machine learning. Hence, the knowledge-centric collaboration of these interdisciplinary experts formed this fifth paradigm platform. With the help of the knowledge-centric module, the predicted materials described in Fig. 3(c) will be further exploited, which means that some materials with predicted adsorption energies that do not match the "volcano plot" will be discarded, and only those predicted materials that match the "volcano plot" can be further quantified by DFT calculation. In the next cycle, the exploited candidates will be calculated again by DFT, and the dataset will be increased through exploration. As the types of materials calculated by DFT increase, the number of datasets also increases. The automated exploration and exploitation process enables the constantly updated number of fingerprints.

2.3. The theoretical model for both DFT calculation and machine learning

In the fifth paradigm platform, the Kohn–Sham theory and a method that integrates the CNN and Gaussian process (GP) [31,35–37] are the core theoretical models for both DFT and machine learning processes. Thus, we briefly introduce the details of these theoretical models.

2.3.1. The theoretical model for DFT calculation

In the process of numerical calculation, namely the DFT calculation, the adsorption energy calculation process mainly involves the optimization process of each slab through the continuous adjustment of the atomic and electronic structure to achieve the most energy-stable structural state, which can be achieved by approximately solving the many-body Schrödinger equation based on quantum mechanics, and solving the Kohn–Sham equation DFT is one of the main methods for this approximate solution.

The Kohn–Sham equation is

$$E[n(r)] = T[n(r)] + \int v(r)n(r)\mathrm{d}^3r + E_{\mathrm{xc}}[n(r)] + \frac{\mathrm{e}^2}{2}\iint \frac{n(r)n(r')}{|r-r'|}\mathrm{d}^3r\mathrm{d}^3r' \tag{1}$$

$$n(r) = \sum_{i=1}^{K} |\psi_i(r)|^2 \tag{2}$$

$$T[n(r)] = \sum \psi_i^*(\boldsymbol{r})\left(-\frac{\hbar^2}{2m}\nabla^2\right)\psi_i(\boldsymbol{r})\mathrm{d}^3r \tag{3}$$

$$E_{\mathrm{xc}}[n(\boldsymbol{r})] = \int n(\boldsymbol{r})\varepsilon_{\mathrm{xc}}[n(\boldsymbol{r})]\mathrm{d}^3r \tag{4}$$

Given a system that contains K ions, namely K occupied orbitals in three-dimensional coordinate space r, $\psi_i(r)$ refers to the wave function of ion i with its coordinate in r, while its conjugate wave function is ψ^*. $n(r)$ is the local electron density, namely the probability of finding an electron in r within the ion i. $E[n(r)]$ is the energy of the total system. The $\hbar$ is the Planck constant, m is the particle's mass. $\varepsilon_{\mathrm{xc}}[n(\boldsymbol{r})]$ is the exchange–correlation energy of a homogeneous electron gas with the local electron density $n(r)$. $E_{\mathrm{xc}}[n(\boldsymbol{r})]$ refers to the exchange and correlation energies, for example, the local electron density approximation, which is one of the exchange–correlation functions, only takes the uniform electron gas density as a variable, while the generalized gradient approximation method considers the electron density and the gradient of the density as the variables. $v(r)$ is the potential energy of ion i in the position of r. Hence the first item $T[n(r)]$ in Eq. (1) refers to the kinetic energy, the second item $\int v(r)n(r)\mathrm{d}^3r$ is the external potential. The last item in Eq. (1) refers to the Hartree energy (electron–electron repulsion), where r' is the coordinate perturbation relative to r, and $\boldsymbol{r}$ represents the vector of r. ∇ is the vector differential operator, and ∇^2 is the laplacian for coordinate derivation.

A self-consistent iterative procedure is described as follows.

Given an initial electron density $n(r)$ obtained from all occupied orbitals by an arbitrary wave function $\Psi_0(r)$:

$$n(r) = \sum_{i=1}^{\mathrm{occ.}} \psi_0^*(r)\psi_0(r) \tag{5}$$

where occ. refers to the number of occupied orbitals, then

$$H[n_0(r)]\psi_1(r) = \varepsilon\psi_1(r) \tag{6}$$

where H refers to the Hamiltonian for the wave function ψ with its energy represented by ε, then a new electron density can be obtained by

$$n_1(r) = \sum_{i=1}^{\mathrm{occ.}} \psi_1^*(r)\psi_1(r) \tag{7}$$

Then

$$H[n_1(r)]\psi_2(r) = \varepsilon\psi_2(r) \tag{8}$$

...

$$H[n_n(r)]\psi_{n+1}(r) = \varepsilon\psi_{n+1}(r) \tag{9}$$

The iterative procedure will exit prematurely when $\psi_{n+1}(r) - \psi_n(r)$ reached the minimum convergence standard required, and the E_{ad} can be calculated by the energy gap between the $E_{\mathrm{adsorbate\_slab}}$ and $E_{\mathrm{bare\_slab}} + E_{\mathrm{adsorbates}}$.

2.3.2. The theoretical model for machine learning

The convolution-fed Gaussian process (CFGP) [37] is a method that the pooled outputs of the convolutional layers of the network are used to supply features to a GP regressor [38], which then makes training to produce both mean and predictions on the adsorption energies. The CNN is applied by Chen et al. [39] and Xie and Grossman [40] on top of a graph representation of bulk crystals to predict various properties, and further modified by Back et al. [31], to collect neighbor information using Voronoi polyhedral [32] for the application in predicting binding energies (for example, the adsorption energy) on heterogeneous catalyst surfaces. In the CFGP method, a complete CNN is first trained to create the final fixed network's weights. Then all the pooled outputs of the convolutional layers are used as features in a new GP. The GP would then be trained to use these features to produce both mean and uncertainty predictions on the adsorption energies.

In the CFGP method, the crystal structure is represented by a crystal graph G, where the atoms and edges representing connections between atoms in a crystal are encoded by the nodes with the information of atomic features and neighbor features, and then a CNN is constructed on the top of the undirected multigraph [40]. Due to the characteristics of periodicity for the crystal graphs, multiple edges are allowed between the same pair of end nodes, the number of each node is marked by i, and each node i can be represented by a feature vector $\boldsymbol{v}_i$. Similarly, each edge $(i,j)_k$ can be represented by the feature vector $\boldsymbol{u}_{(i,j)_k}$, which corresponds to the kth bond connecting atom i and atom j. Considering the differences of interaction between each atom feature and the neighbors, the first convolutional layers iteratively update the atom feature by

$$\boldsymbol{z}_{(i,j)_k} = \boldsymbol{v}_i \oplus \boldsymbol{v}_j \oplus \boldsymbol{u}_{(i,j)_k} \ (i,j)_k \in G \tag{10}$$

where $\boldsymbol{z}_{(i,j)_k}$ is the updated atom feature of atom i and atom j connected by kth bond in crystal graph G. $\oplus$ denotes the concatenation of atom and bond feature. Then a nonlinear graph convolution function is defined as follows:

$$\boldsymbol{v}_i^t = \boldsymbol{v}_i^{t-1} + \sum_{j,k} \sigma\left(\boldsymbol{z}_{(i,j)_k}^{t-1} W_{\mathrm{f}}^{t-1} + b_{\mathrm{f}}^{t-1}\right) \odot g\left(\boldsymbol{z}_{(i,j)_k}^{t-1} W_{\mathrm{s}}^{t-1} + b_{\mathrm{s}}^{t-1}\right) \tag{11}$$

where $\odot$ denotes an element-wise multiplication, σ is a sigmoid function, and g is a nonlinear activation function (for example, the "Leaky ReLu" or "Softplus"); W and b denote weights and biases of the neural networks, respectively. The $\sigma(\bullet)$ function is a learned weight matrix to different interactions between neighbors; f and s represent the abbreviation of first and self, respectively. After R convolutional layers, resulting vectors are then fully connected via K hidden layers, followed by a linear transformation to scalar values. Then, the distance filters collected by the connection distances are applied to exclude contributions of atoms that are too far from the adsorbates. A mean pooling layer is then used for producing an overall feature vector $\boldsymbol{v}_{\mathrm{c}}$, which can be represented by a pooling function,

$$\boldsymbol{v}_{\mathrm{c}} = \mathrm{Pool}\left(\boldsymbol{v}_0^{(0)}, \boldsymbol{v}_1^{(0)}, ..., \boldsymbol{v}_N^{(0)}, ..., \boldsymbol{v}_N^{(R)}\right) \tag{12}$$

The training is performed by the cost function $J\left(E_{\mathrm{ad}}, \hat{E}_{\mathrm{ad}}\right)$, then the whole process produces the function f parametrized by weights W that maps a crystal C to the target property $\hat{E}_{\mathrm{ad}}$. Using backpropagation and stochastic gradient descent (SGD), the following optimization problem can be solved by iteratively updating the weights with DFT calculated data:

$$\min_W J(E_{\mathrm{ad}}, f(C;W)) \tag{13}$$

Here, the penultimate layer of the pooling outputs and the corresponding learning weights W rather than the target property E_{ad} is further extracted out as features in the GP. Hence the descriptor for nodes is $V = \left[\boldsymbol{v}_0^0, \boldsymbol{v}_1^0, ..., \boldsymbol{v}_N^R\right]$ and is trained with their corresponding energies (E_{ad}). The prediction function is

$$f(\boldsymbol{v}) \sim \mathrm{GP}[P(\boldsymbol{v}), k(\boldsymbol{v}, \boldsymbol{v}')] \tag{14}$$

where $P(\boldsymbol{v})$ is the constant mean of prior function and $k(\boldsymbol{v}, \boldsymbol{v}')$ is the Matern kernel with the length scale trained by the maximum likelihood estimation method, $\boldsymbol{v}$ and $\boldsymbol{v}'$ refer to different feature vector, respectively. All training and predictions were done with Tesla P100-PCIE GPU acceleration as implemented in GPyTorch [41].

2.4. Iteration between machine learning and numerical calculations

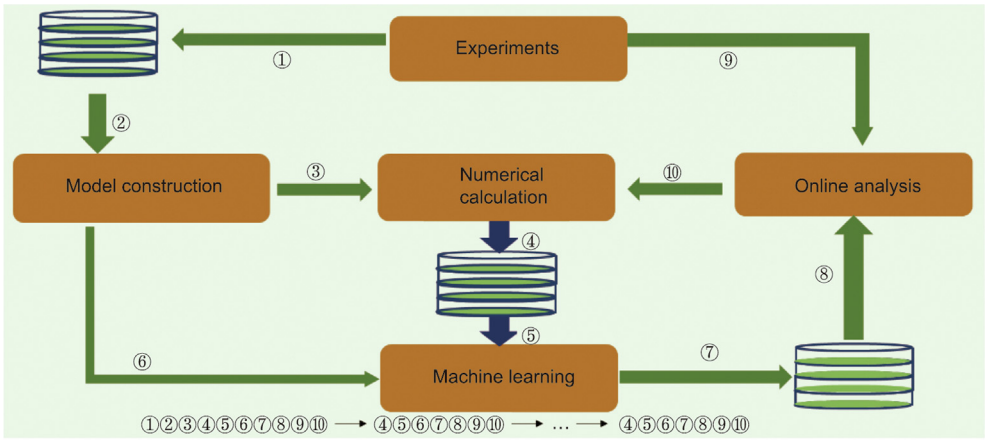
The intelligence-driven, knowledge-centric nature of the fifth paradigm platform can be well depicted by the iterations between machine learning and numerical calculation concatenated by the interdisciplinary knowledge of "volcano plot." This breaks through the new material bottleneck of artificial screening research between machine learning and numerical calculation and realizes the mutual promotion of scientific experiments and AI, as shown in Fig. 4(a). The experiments involve the process of fetching the primitive crystals (or primitive cells) from the Material Project website to be stored in the database, as well as the information about "volcano plot." Then, the model is automatically reconstructed to create a bulk of adsorption energy calculation models. Through numerical calculation (i.e., *ab initio* DFT calculation), the optimized model and adsorption energy data are stored in the database, and fingerprints are extracted from it to train a suitable machine-learning model. Then, the trained model can use the fingerprint extracted from the bulk materials that have not been theoretically calculated to predict their adsorption energy and can be stored in the database again. Adsorption energy prediction results are intelligently analyzed through "volcano plot" to screen models that require further DFT calculations. Then the entire loop is ①②③④⑤⑥⑦⑧⑨⑩, ④⑤⑥⑦⑧⑨⑩, ..., ④⑤⑥⑦⑧⑨⑩.

The cycle stops only when all the materials delivered in the framework are calculated in the machine learning or DFT processes. The characteristics of the fifth paradigm platform are well reflected in these steps. The step ⑤ indicates that the dataset obtained by numerical calculation supplements the problem of no dataset and fewer datasets in the machine-learning process. The step ⑩ indicates that the bulk of numerical calculations can be abandoned with the help of machine-learning prediction and the "volcano plot" to accelerate the entire DFT calculation. Moreover, the results of machine learning can be intelligently analyzed through the "volcano plot" that integrates the knowledge of experimental and theoretical scientists (the synergy of interdisciplinary experts), forming a knowledge-centric fifth paradigm driven by intelligence.

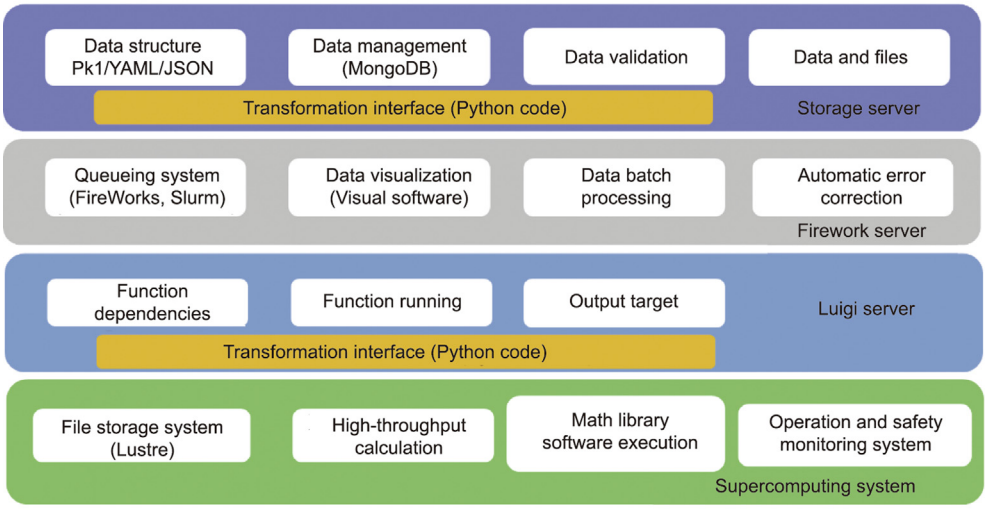
2.5. Information science tools

The framework of the fifth paradigm is built by using various Python packages, for example, Python Materials Genomics (pymatgen), the automic simulation environment (ASE), FireWorks, Luigi, and MongoDB [42–45]. Pymatgen is one of the powerful program packages supported by Python for high-throughput material calculations. It standardizes the initialization settings required before running high-throughput calculations and provides process analysis of the data generated by the calculations. The ASE aims to set up, steer, and analyze atomistic simulations. The function of FireWorks is to perform job management in high-throughput computing workflows running on high-performance computing clusters. Luigi can be used to build complex batch job pipelines, handle dependency resolution, and conduct workflow management. MongoDB is written in the C++ language and is used for real-time data storage and can jointly meet the JavaScript Object Notation data-exchange format.

As shown in Fig. 4(b), the data-intensive DFT calculations can be done on the Tianhe-1 supercomputer using Lustre as the file-storage system [46]. High-throughput computing jobs can be realized by running the security-monitoring system deployed on the cluster. Luigi is used to building the various physical models through dependency resolution (function dependencies, running, and output target), which are then configured and calculated by the task management through FireWorks and batched processing performance through the resource management Slurm in the supercomputer [47]. These two task-management systems can automatically correct errors, re-run a single job, and simultaneously visualize the data through the installed visualization tools.



(a)



(b)

Fig. 4. The architecture of this fifth paradigm case. (a) The repeated iterations framework includes machine learning and numerical computing in the fifth paradigm platform. Steps ① and ② and steps ⑦ and ⑧ are pulling the experimental results and machine learning results into and out of the database, respectively. Step ③ shows the constructed model prepared for *ab initio* calculation. Step ④ refers to the storage process of the calculation results, and steps ⑤ and ⑥ are the fingerprints extracted from calculated results and experimental results, respectively. Step ⑨ refers to an online analysis of machine learning results through "volcano plot." Step ⑩ shows the remaining models after online analysis (screening) which require further numerical calculations. (b) The realization of services and functions are based on the fifth paradigm platform of the Tianhe-1 supercomputer. Typic components dedicated to services in GASpy contain Storage server, Firework server, and Luigi server. The basic environment in the supercomputing system is at the software level.

3. Performance evaluation

To illustrate the performance of the fifth paradigm platform in catalytic materials screening, we conducted a comparison test to explain how the machine-learning process accelerates numerical calculations and how the process of numerical calculations provides trainable samples for machine-learning iterations. In this article, we do not use the updated dataset containing the online DFT calculation process in the learning cycle of each model, but instead, we use the DFT calculated dataset to extract the corresponding fingerprints for research. Because target prediction is not directly related to the structure of DFT calculation, it is related to the fingerprint extracted from the initial structure without any simulation process. We believe that it will not affect the evaluation of the platform.

The dataset we prepare to test the cross-validation process comes from Github[‡]. Five adsorbates of H, CO, OH, O, and N are consisted, of which the main dataset comes from the first two adsorbates (21 269 and 18 437). The method of CFGP is used to create a model to compare the impact of different machine learning models and the total number of the dataset on the accuracy of the catalyst screening through the performance metrics of the R^2, and the mean absolute error (MAE), as well as the root-mean-square error (RMSE). Hyperparameters for the dataset have been tuned by Back et al. [31] and Tran et al. [37], while the research in this paper focuses on the performance of different models under the same method, thus these hyperparameters are still applicable. In our work, the statement of the learning problem is determined by the famous "volcano plot" to evaluate the size and activity level of its adsorption energy. Taking the H adsorbate as an example, the hydrogen evolution reaction (HER) is a method that uses adsorption energies to predict catalytic performance. The optimal adsorption energy ΔE_H is −0.27 eV [48], and the near-optimal range of the "volcano plot" is defined as [−0.37 eV, −0.17 eV]. Therefore, if the result of each cycle reaches a range close to the optimal range (it can also be defined as a hit in the near-optimal range), then it is selected as a candidate to continue the DFT calculation before the start of the next cycle.

One realization of the mutual feedback between machine learning and numerical calculation is that the trainable sample provided

[‡] https://github.com/ulissigroup/uncertainty_benchmarking

by DFT calculations can supplement machine-learning iterations. In this platform, once an iteration occurs, the dataset containing the target features is determined, which means the machine learning model for the corresponding iteration is determined. In addition, as a typical case of the fifth paradigm platform, the performance comparison of each iterative process is derived from the model comparison under the same data generation conditions. As shown in Table 1, the entire dataset is first randomly shuffled and split into ten models, and 10% of the total dataset is taken as the first model dataset, and then added in increments until 100% of the total dataset is taken as the tenth model to form the datasets corresponding to ten models. The dataset of the previous model is encompassed in the dataset of the next model. For the cross-validation process, the train/validate/test ratio of each model is 64/16/20, and all the monometallic slabs are added to the training set, as described by Tran et al. [37]. The cross-validation and its results are listed in Table 1 and Fig. 5. The violin in Fig. 5(a) refers to the correlation coefficient (R^2) of the training and testing samples. The greater the difference between the two values, the slenderer it becomes. Otherwise, it turns out to be stubby. If the two are the same, it can be a line. Therefore, the slender violins of models 1, 2, 5, 6, and 9 are indicators of overfitting or underfitting, followed by models 3, 4, 7, and 10, and model 8 performs best. As the dataset increases, the MAE and RMSE in Table 1 gradually decrease, while the R^2 trend of the validation and testing process in Fig. 5(a) gradually increases, which indicates that the training model is more accurate than the previous models. In addition, the hit numbers of H adsorbates verified by the DFT calculation (N_{DFT}) and machine learning prediction (N_{ML}) are also listed, and their trend also increases with the expansion of the dataset, as shown in Fig. S1 (in Appendix A). The dataset of model 1 to be hit is set as the baseline. To find the performance of increasing trainable samples provided by the numerical calculation of machine-learning iteration, a formula is defined as follows:

$$\eta = \frac{D_n - D_1}{M_n - M_1} \quad n \in \mathbf{N},\ 1 < n \leq 10 \tag{15}$$

where η represents the increment of N_{DFT} compared with the N_{ML}. D_n and M_n refer to N_{DFT} and N_{ML} of model n in the near-optimal range (namely the hit number). With the expansion of the dataset, the trend of η becomes larger and approaches 1, indicating that the hit number N_{ML} is slowly approaching the hit number N_{DFT}, which shows that the larger the training sample of numerical calculation, the higher the accuracy of the machine-learning model. Furthermore, η fits well in Fig. 5(b), even if some points are not in the linear range. For example, η of model 4 is very small compared to other points, which we attribute to the compensation of larger values in models 5 and 6.

The datasets with multiple adsorbates including H adsorbate are used for train/validation/test, and MAE and RMSE are used to evaluate the performance of the machine learning model. The number of surfaces for which low-coverage H adsorption energies in near-optimal activity in the "volcano plot" are verified by the DFT calculation and machine learning prediction, represented by N_{DFT} and N_{ML}, respectively. η is used to evaluate the trend of model performance changes.

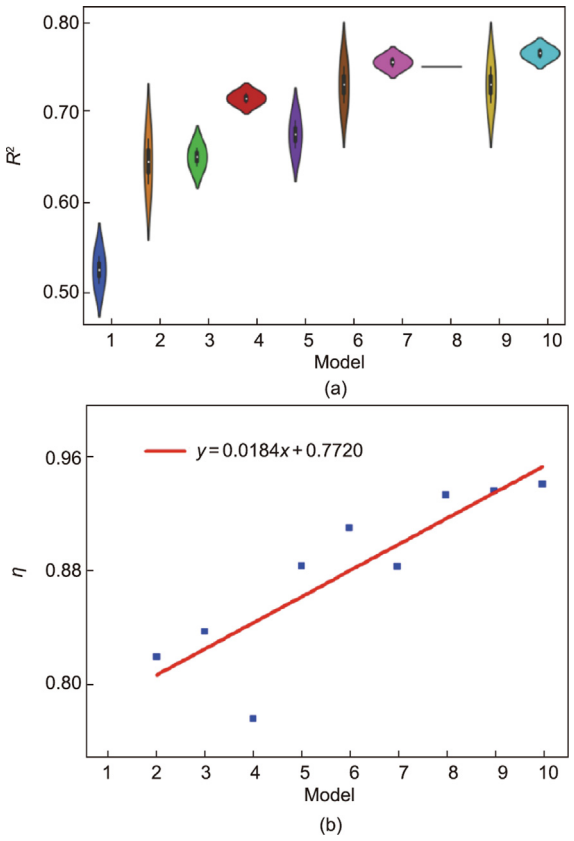


Fig. 5. Performance metrics evaluation of the learning model in the fifth paradigm platform. (a) The R^2 correlation coefficient of the validation and testing process in the ten models; (b) the linear fit of η for all the models.

To illustrate the realization of mutual feedback between machine learning and numerical calculation (e.g., machine learning solves the time-consuming problem of massive models caused by insufficient computing resources in numerical calculations, and the numerical calculation process provides machine learning training samples), we prepared three types of prediction cases to understand the performance of the model trained and validated as described above. The dataset that we used in the prediction process is from the work of Tran and Ulissi [28], which encompassed 22 675 H adsorbates DFT results. To be honest, it has covered most of the 21 269 H-dataset mentioned above. However, we believe

Table 1
Ten models constructed from the entire dataset to evaluate the performance of the fifth paradigm platform.

Model	Dataset	MAE (eV)	RMSE (eV)	H-dataset	Near-optimal activity		η
					N_{DFT}	N_{ML}	
1	4 728	0.30	0.52	2 090	72	60	–
2	9 456	0.26	0.49	4 155	141	144	0.82
3	14 184	0.24	0.43	6 293	208	222	0.84
4	18 912	0.22	0.40	8 384	275	321	0.78
5	23 640	0.23	0.44	10 530	351	375	0.89
6	28 368	0.22	0.41	12 678	417	438	0.91
7	33 096	0.21	0.37	14 756	512	557	0.89
8	37 824	0.21	0.38	16 926	598	622	0.94
9	42 552	0.22	0.41	19 086	658	684	0.94
10	47 290	0.19	0.36	21 269	709	735	0.94

that it doesn't matter of the repeated dataset, because our goal is to compare the performance of machine learning models generated on samples of different sizes and find out the acceleration behavior of machine learning under prediction samples of different sizes. Moreover, the material structure corresponding to the dataset to be predicted does not depend on whether simulation calculations have been performed. Therefore, the decision that this machine learning prediction dataset is taken from the DFT calculation dataset will not affect the overall evaluation of the intelligent driving process.

In terms of the characteristics of the platform, the DFT calculations performed in each cycle (except the first cycle) are obtained from the machine-learning results. Three types of methods are designed for prediction in Table 2: Hit_no_split, No_hit_with_split, and No_hit_no_split. The No_hit_with_split method refers to the incremental dataset from 10% to 100% of the total prediction dataset corresponding to the machine learning model from model 1 to model 10 formed above. In addition, the entire prediction dataset can also be kept the same in each cycle, as defined by the No_hit_no_split method. As for the Hit_no_split method, it means that the model predicted by machine learning in the optimal range is discarded in the next model prediction. The process is as follows: Starting from model 1, 4960 models predicted by machine learning are found to be hits in the entire 22 675 models predicted. When using model 2 to make predictions, 4960 models predicted by machine learning will be removed from 22 675 models predicted, leaving only 17 715 (22 675 – 4960 = 17 715) models. Then, model 2 finds that 860 models predicted by machine learning were hits, and provides another simplified sample 16 855 (17 715 – 860 = 16 855) for model 3 prediction. The hit and drop will not end until the predictions of the ten models are completed. Note that N_{Hits} should be equal to N_{ML}, but certain materials in the samples must be excluded from the near-optimal activity process.

Table 2 lists the results of the three methods in the near-optimal range. In the Hit_no_split method, because the N_{ML} predicted by the previous model is deducted from the prediction samples of the next model (except for model 1), N_{DFT}, N_{ML}, and N_{Hits} from model 1 to model 10 are also reduced accordingly. In the No_hit_with_split method, as the prediction samples increase, N_{DFT} and N_{ML} expand gradually. In the No_hit_no_split method, N_{ML} fluctuates between 4177 and 4556, while N_{DFT} remains unchanged. We infer that this is caused by the different accuracies of the machine-learning model. Meanwhile, the more datasets there are in the model, the more N_{ML} hits there are. From an acceleration point of view, the Hit_no_split method can ensure that the predicted reasonable samples will not be predicted again (of course, provided that it is reasonable), while the other two methods involve repeated predictions of the predicted samples. Therefore, ideally, the Hit_no_split method should be able to optimize the use of all samples that must be predicted to accelerate predictions and provide a faster machine-learning process for accelerating numerical calculations.

To evaluate the difference of these methods in accelerating DFT calculations, we compared the number of N_{DFT} replaced by N_{ML}, as well as the value of N_{ML}/N_{DFT} in Fig. 6. The replacement of machine learning to replace DFT calculations is defined as follows:

$$\mathrm{RE} = \begin{cases} T_n - M_n & n = 1 \\ T_n - M_n - M_{n-1} & n \in \mathbf{N},\ 1 < n \le 10 \end{cases} \tag{16}$$

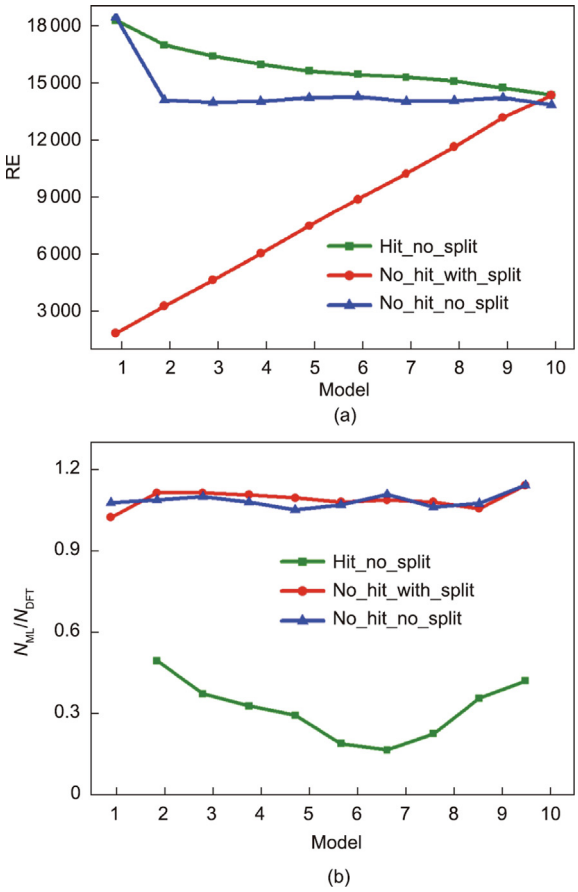


Fig. 6. The predictive performance of all models constructed in the fifth paradigm platform. (a) The number of DFT calculations (N_{DFT}) replaced by the number of machine learning predictions (N_{ML}); (b) the change of N_{ML}/N_{DFT} in the near-optimal range for different models within the prediction process. In the Hit_no_split method, model 1 is abandoned because of its baseline function to the other models.

Table 2
Three types of prediction methods in the near-optimal range and their performance of all models constructed in the fifth paradigm platform.

Model	Hit_no_split				No_hit_with_split			No_hits_no_split		
	Dataset	N_{DFT}	N_{ML}	N_{Hits}	Dataset	N_{DFT}	N_{ML}	Dataset	N_{DFT}	N_{ML}
1	22 675	4 027	4 446	4 960	2 268	389	392	22 675	4 027	4 282
2	17 715	1 707	775	860	4 536	774	853	22 675	4 027	4 331
3	16 855	1 484	485	539	6 804	1 178	1 299	22 675	4 027	4 380
4	16 316	1 336	374	419	9 072	1 573	1 722	22 675	4 027	4 293
5	15 897	1 264	309	324	11 340	1 970	2 133	22 675	4 027	4 177
6	15 573	1 194	161	174	13 608	2 417	2 578	22 675	4 027	4 249
7	15 399	1 157	127	141	15 876	2 846	3 058	22 675	4 027	4 413
8	15 258	1 118	193	210	18 144	3 231	3 448	22 675	4 027	4 215
9	15 048	1 058	328	352	20 412	3 650	3 799	22 675	4 027	4 273
10	14 696	968	365	383	22 675	4 027	4 556	22 675	4 027	4 556

The dataset refers to the total number of data sets for each model. The N_{Hits} is the number of machine learning predictions that do not exclude certain materials.

where RE and T_n are the number of DFT calculations replaced by machine learning and all prediction datasets in each model. As shown in Fig. 6(a), the replacement amount of all models of Hit_no_split is more than 15 000, and the replacement amount from model 1 to model 10 is slightly reduced, but compared with other methods, it has the largest N_{DFT} replacement. For the No_hit_with_split method, the number of replacements increases linearly from 1800 to the same as other methods in model 10. For the No_hit_no_split method, except for model 1, the number of replacements for all the models is approximately 14 000, and there is a slight downward trend. For a large number of replacements of model 1 in the No_hit_no_split method and the subsequent sudden decrease, we believe that it is caused by underfitting because model 1 uses a small amount of the dataset to train the model to predict an ever-larger dataset. In these methods, the Hit_no_split can replace the maximum N_{DFT}, as we expected.

The reason that we compared the value N_{ML}/N_{DFT} in Fig. 6 is that it can reflect the performance of each model in another view. The ideal N_{ML}/N_{DFT} values should all be equal to 1. In the No_hit_with_split and No_hit_no_split methods, the N_{ML}/N_{DFT} is slightly increased to close to 1, which indicates that the prediction behavior of the two methods is similar and is suitable for accelerating DFT calculation. In the Hit_no_split method, except for model 1 set as the baseline, the N_{ML}/N_{DFT} value is gradually reduced from model 2 to model 7 and then gradually increased in the remaining models, all below 0.5. On one hand, we infer that these smaller values are caused by changes in the accuracy of the machine-learning model since smaller datasets lead to underfitting. On the other hand, as the number of hits of the prediction samples decreases, the number N_{ML} that can hit in the next model gradually decreases. In addition, for the No_hit_with_split and No_hit_no_split methods, the number of hits in the previous model will be removed in each model, and the N_{ML} that can be hit in the next model will gradually decrease. Since these methods does not involve hit material to be hit again in other iterations, the advantage in terms of speed then are more obvious.

In addition, since the machine-learning model itself exhibits the characteristics of gradual reduction of poor fitting during the expansion process from small cross-validation samples, there will be a certain degree of accuracy loss in the prediction process from model 2 to model 10. For example, the predicted machine-learning dataset should have been hit but not hit, or the dataset should not be hit but hit, leading to hit data missing or non-hit data increasing in the dataset of the next model. Moreover, it is also possible that the sample size is not large enough, resulting in the underfitting or overfitting of the machine-learning model. Therefore, the Hit_no_split method has the advantage of replacing more DFT calculations, although the evaluation of its accuracy is not suitable for the indicators of N_{ML}/N_{DFT}. However, this by no means indicates that the Hit_no_split method is not applicable to the fifth paradigm platform. We infer that when the prediction model is good enough and the dataset is large enough, it can reduce the repeated prediction process of data while maintaining the reliability of the results to accelerate the advantages of machine learning to, in turn, accelerate numerical calculations.

Based on the results of the three types of methods, the accuracy loss of machine learning prediction relative to DFT calculation is used to evaluate the performance in the fifth paradigm platform. The accuracy loss can be defined as follows:

$$L = \frac{M_n - D_n}{T_n} \quad n \in \mathbf{N},\ 1 \le n \le 10 \tag{17}$$

where L is the accuracy loss. Given that the No_hit_with_split and No_hit_no_split methods have relatively suitable predictive performance, we only consider the accuracy loss of these two methods. As shown in Fig. 7, for No_hit_with_split, although model 1 has the lowest accuracy loss, the dataset is small, and we exclude it and consider that model 9 has the lowest accuracy loss. For the No_hit_no_split method, model 5 has the lowest accuracy loss. Therefore, we believe that, as the dataset expands, machine learning will continue to replace DFT calculations, and there will be varying degrees of accuracy loss. The smallest accuracy-loss point is most conducive to this type of machine learning to accelerate the DFT calculation process.

We believe that the accuracy loss of this fifth paradigm case is related to the size of the sample involving machine learning, theoretical calculations, and experiments fed back from the "volcano plot," which is exactly the knowledge-centric characteristic for the fifth paradigm in terms of precision. As shown in Fig. 7, the accurate fifth paradigm should make machine learning, theoretical

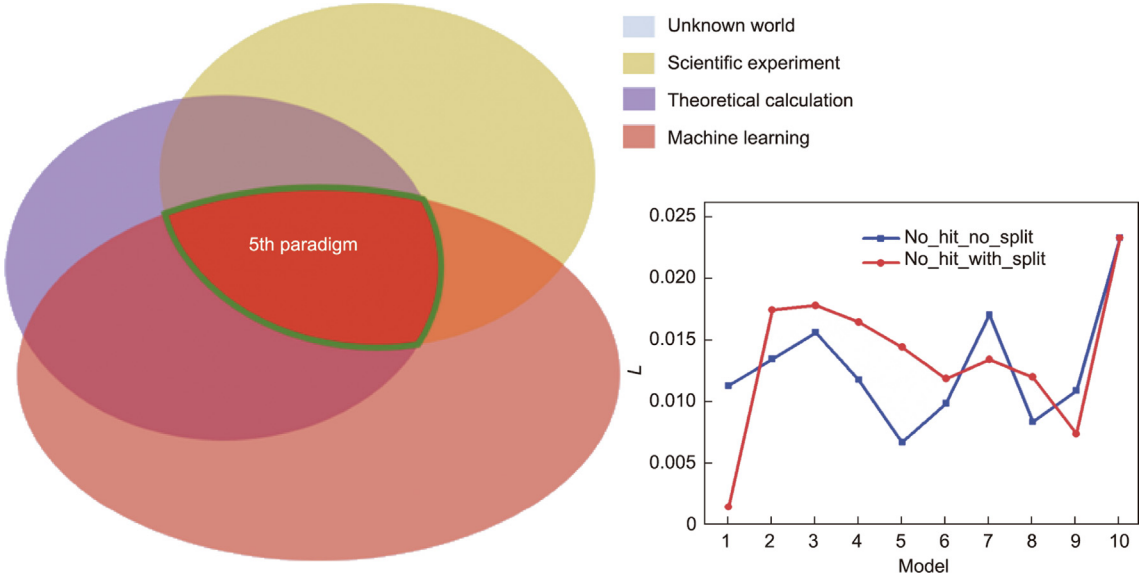


Fig. 7. The accuracy of the fifth paradigm. The mutual verification process of scientific experiment, theoretical calculation, and machine learning in the process of exploring the unknown world represents the accuracy of the fifth paradigm. The accuracy loss (L) of No_hit_no_split and No_hit_with_split methods between machine learning and DFT calculation of all models is constructed in the fifth paradigm platform.

calculation, and scientific experiment unique to the result of the unknown world exploration. Although this standard is very demanding, it is always the ultimate goal for exploring the unknown world.

4. Discussion of the fifth paradigm platform

Automated model construction, automated fingerprint extraction, as well as intelligent coupling of intensive data with DFT calculation and machine learning by the "volcano plot" compose the architecture of the fifth paradigm platform. In the intelligence-driven framework, the workload of traditional modeling construction and calculation is reduced effectively by making full use of the current development of various information tools and methods, greatly simplifying and improving the extremely cumbersome and challenging work in materials research.

One of the challenges this framework faces is the limited application areas implemented in the fifth paradigm. This is because the most typical feature of the fifth paradigm is intelligence-driven, which entails the synergy of interdisciplinary experts to carry out in-depth research. For example, in the materials science introduced in this work, it is necessary to intelligently drive the efficient synergy of experimental experts and theoretical experts, which can be achieved by filtering the machine-learning results through the "volcano plot." For some high-throughput interdisciplinary work, before designing a similar fifth paradigm framework, it is best to first consider appropriate methods of quantifying the collaborative work between these experts in different application fields.

In addition, due to the lack of an ever-larger dataset, there must be an insufficient number of samples during the expansion process of the dataset, resulting in poor generalization ability of the training model. Therefore, more datasets must be accumulated to achieve a high-precision machine-learning process. Fortunately, for this fifth paradigm platform, the Open Catalyst project, jointly researched and developed by Facebook AI Research and the Department of Chemical Engineering of Carnegie Mellon University, has realized the Open Catalyst 2020 [49] dataset containing a dramatic rise in DFT calculation results, and it is still constantly updated online. Finally, the accuracy of the fifth paradigm utilized to realize the exploration of the unknown world is affected by machine learning, theoretical calculation, and scientific experiment. The high-precision fifth paradigm tends to explore the same objective thing from the unknown world through the three kinds of cooperation within the scope of its reasonable discovery, derivation, and judgment. We believe that the dissection of this fifth paradigm case can greatly promote the development of the fifth paradigm of materials science in the future.

5. Conclusions

In this work, we discuss the scientific explanation of the newest paradigm emerging due to the prosperity engendered by AI. Then, a detailed discussion is carried out using a fifth paradigm platform as a typical case, which conforms to a specific and well-defined framework capable of promoting the development of materials science. The interdisciplinary knowledge and intelligence-driven characteristics are the keys to the fifth paradigm, which can be addressed in the work encompassing automatic model construction and verification, automated fingerprint construction, as well as the theoretical model and repeated iteration between machine learning and theoretical calculations. These informatics tools needed for architecting the framework are also discussed in detail. Finally, tests and comparisons are conducted to show how the interaction between AI and numerical calculation in the framework of this fifth paradigm case meaningfully promotes each other to reduce numerical calculation and create more trainable samples in the mutual feedback process. The curation of the numerical calculation and machine-learning models, as well as the techniques, makes the fifth paradigm platform more interpretable.

With the expansion of the dataset, on one hand, the more machine learning replaces the DFT calculation, the faster the screening of materials will be. On the other hand, the more consistent the number of candidate materials predicted by the final machine learning is with the number of candidate materials calculated by DFT, the more accurate the prediction by machine learning is. Under the conditions of satisfying these two judgments, machine learning will continue to replace DFT calculation with different degrees of accuracy loss, and the smallest accuracy loss model is most conducive to machine learning to accelerate the DFT calculation process. This minimum accuracy loss discrimination represents the precise exploration premise of materials research under the scientific fifth paradigm, which requires consistent results when machine learning, theoretical calculation, and scientific experiment are jointly exploring the unknown world.

Although this article provides a scientific explanation for the fifth paradigm platform represented in the fields of catalytic materials, it also acknowledges that much more needs to be discussed. The overall development of the fifth paradigm across various fields still faces challenges in terms of the synergy between interdisciplinary experts and the dramatic rise in demand for data in data-driven disciplines. Despite these challenges, an ongoing endeavor in tandem with all the relevant parties can be envisioned to deepen the combination of AI technology and traditional disciplines, so that each simulation and calculation link has higher intelligence and automation characteristics, and finally runs as a platform to improve the efficiency of traditional scientific computing and promote the development of materials research in a more intelligent and high-precision direction. We believe that a glimpse of the fifth paradigm platform can pave the way for the application of the fifth paradigm in other fields.

Acknowledgments

We thank Prof. Zachary W. Ulissi and Prof. Pari Palizahti at Carnegie Mellon University for providing advice on the platform. This study was supported by the National Key Research and Development Program of China (2021ZD40303), the National Natural Science Foundation of China (62225205 and 92055213), Natural Science Foundation of Hunan Province of China (2021JJ10023) and Shenzhen Basic Research Project (Natural Science Foundation) (JCYJ20210324140002006).

Compliance with ethics guidelines

Can Leng, Zhuo Tang, Yi-Ge Zhou, Zean Tian, Wei-Qing Huang, Jie Liu, Keqin Li, and Kenli Li declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2022.06.027.

References

[1] Barber B. Resistance by scientists to scientific discovery. Science 1961;134(3479):596–602.
[2] Dampier WCD. A history of science, technology and philosophy in the eighteenth century. Nature 1939;143(3613):134–5.
[3] Crombie AC. Scientific change: historical studies in the intellectual, social and technical conditions for scientific discovery and technical invention, from antiquity to the present. London: Heinemann; 1963.

[4] Bidney M, Piekielek N. Towards a new paradigm in map and spatial information librarianship. J Map Geogr Libr 2018;14(2–3):67–74.
[5] Li J, Huang W. Paradigm shift in science with tackling global challenges. Natl Sci Rev 2019;6(6):1091–3.
[6] Tolle KM, Tansley DSW, Hey AJG. The fourth paradigm: data-intensive scientific discovery. Proc IEEE 2011;99(8):1334–7.
[7] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518 (7540):529–33.
[8] Bainbridge WS. The scientific research potential of virtual worlds. Science 2007;317(5837):472–6.
[9] Zubarev DY, Pitera JW. Cognitive materials discovery and onset of the 5th discovery paradigm. In: Pyzer-Knapp EO, Laino T, editors. Machine learning in chemistry: data-driven algorithms, learning systems, and predictions. Washington, DC: American Chemical Society; 2019. p. 103–20.
[10] Malitsky N, Castain R, Cowan M. Spark–MPI: approaching the fifth paradigm of cognitive applications. 2018. arXiv:1806.01110.
[11] Woinaroschy A. A paradigm-based evolution of chemical engineering. Chin J Chem Eng 2016;24(5):553–7.
[12] Si Y, Wu HY, Yang K, Lian JC, Huang T, Huang WQ, et al. High-throughput computational design for 2D van der Waals functional heterostructures: fragility of Anderson's rule and beyond. Appl Phys Lett 2021;119(4):043102.
[13] Li B, Peng W, Zhang J, Lian JC, Huang T, Cheng N, et al. High-throughput one-photon excitation pathway in 0D/3D heterojunctions for visible-light driven hydrogen evolution. Adv Funct Mater 2021;31(18):2100816.
[14] Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci 2019;6(21):1900808. Corrected in: Adv Sci 2020;7(2):1903667.
[15] Hardian R, Liang ZW, Zhang XL, Szekely G. Artificial intelligence: the silver bullet for sustainable materials development. Green Chem 2020;22 (21):7521–8.
[16] Xu X, Ma WP, Yan B. An electrodeposited nano-porous and neural network-like Ln@HOF film for SO_2 gas quantitative detection via fluorescent sensing and machine learning. J Mater Chem A 2021;9(46):26391–400.
[17] Kumar S, Ignacz G, Szekely G. Synthesis of covalent organic frameworks using sustainable solvents and machine learning. Green Chem 2021;23(22):8932–9.
[18] Ding WL, Lu YM, Peng XL, Dong H, Chi WJ, Yuan X, et al. Accelerating evaluation of the mobility of ionic liquid-modulated PEDOT flexible electronics using machine learning. J Mater Chem A 2021;9(45):25547–57.
[19] Vandenberg P. The fourth industrial revolution. J Asia Pac Econ 2020;25 (1):194–6.
[20] Feng R, Zhang C, Gao MC, Pei Z, Zhang F, Chen Y, et al. High-throughput design of high-performance lightweight high-entropy alloys. Nat Commun 2021;12 (1):4329.
[21] Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering 2021;7(9):1201–11.
[22] Zhou T, Song Z, Sundmacher K. Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 2019;5(6):1017–26.
[23] Chen S, Zhang S, Shang J, Chen B, Zheng N. Brain inspired cognitive model with attention for self-driving cars. 2017. arXiv:1702.05596.
[24] Xu Z. Principle analysis of computer vision and its application research. In: Proceedings of the 2018 7th International Conference on Advanced Materials and Computer Science; 2018 Dec 21–22; Dalian, China. Ottawa: Clausius Scientific Press; 2018. p. 478–82.
[25] Itaya K, Takahashi K, Nakamura M, Koizumi M, Arakawa N, Tomita M, et al. BriCA: a modular software platform for whole brain architecture. In: Hirose A, Ozawa S, Doya K, Ikeda K, Lee M, Liu D, editors. Neural information processing. Cham: Springer International Publishing; 2016. p. 334–41.
[26] US Department of Energy. Synergistic challenges in data-intensive science and exascale computing. Summary report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee. Washington, DC: US Department of Energy, Office of Science; 2013.
[27] Wang C, Yu F, Liu Y, Li X, Chen J, Thiyagalingam J, et al. Deploying the Big Data Science Center at the Shanghai Synchrotron Radiation Facility: the first superfacility platform in China. Mach Learn Sci Technol 2021;2(3):035003.
[28] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nat Catal 2018;1 (9):696–703.
[29] Kresse G, Furthmüller J. Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mater Sci 1996;6(1):15–50.
[30] Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh CT, et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. Nature 2020;581 (7807):178–83.
[31] Back S, Yoon J, Tian N, Zhong W, Tran K, Ulissi ZW. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. J Phys Chem Lett 2019;10(15): 4401–8.
[32] Wigner E, Seitz F. On the constitution of metallic sodium. Phys Rev 1933;43 (10):804–10.
[33] Abild-Pedersen F, Greeley J, Studt F, Rossmeisl J, Munter TR, Moses PG, et al. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. Phys Rev Lett 2007;99(1):016105.
[34] Calle-Vallejo F, Martínez JI, García-Lastra JM, Rossmeisl J, Koper MTM. Physical and chemical nature of the scaling relations between adsorption energies of atoms on metal surfaces. Phys Rev Lett 2012;108(11):116103.
[35] Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev 1964;136(3B): B864–71.
[36] Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. Phys Rev 1965;140(4A):A1133–8.
[37] Tran K, Neiswanger W, Yoon J, Zhang Q, Xing E, Ulissi ZW. Methods for comparing uncertainty quantifications for material property predictions. Mach Learn Sci Technol 2020;1(2):025006.
[38] Garrido Torres JA, Jennings PC, Hansen MH, Boes JR, Bligaard T. Low-scaling algorithm for nudged elastic band calculations using a surrogate machine learning model. Phys Rev Lett 2019;122(15):156001.
[39] Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater 2019;31 (9):3564–72.
[40] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 2018;120(14):145301.
[41] Gardner JR, Pleiss G, Bindel D, Weinberger KQ, Wilson AG. In: GPyTorch: blackbox matrix–matrix Gaussian process inference with GPU acceleration. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, editors. Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3–8. Montréal, QC, Canada. Red Hook: Curran Associates Inc.; 2018. p. 7587–97.
[42] Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci 2013;68:314–9.
[43] Hjorth Larsen A, Jørgen Mortensen J, Blomqvist J, Castelli IE, Christensen R, Dułak M, et al. The atomic simulation environment—a Python library for working with atoms. J Phys Condens Matter 2017;29(27):273002.
[44] Jain A, Ong SP, Chen W, Medasani B, Qu X, Kocher M, et al. FireWorks: a dynamic workflow system designed for high-throughput applications. Concurr Comp Pract E 2015;27(17):5037–59.
[45] Jiao YQ, Li YJ, Li B, Song YG, inventors; Goertek Inc., assignee. [MongoDB-based test data storage query method and system]. Chinese patent CN 105550333A. 2021 May 4. Chinese.
[46] Wang Y, Lu Y, Qiu C, Gao P, Wang J. Performance evaluation of a infiniband-based lustre parallel file system. Proc Environ Sci 2011;11(Pt A):316–21.
[47] Yoo AB, Jette MA, Grondona M. SLURM: simple Linux utility for resource management. In: Feitelson D, Rudolph L, Schwiegelshohn U, editors. Job scheduling strategies for parallel processing. Berlin: Springer; 2003. p. 44–60.
[48] Nørskov JK, Bligaard T, Logadottir A, Kitchin JR, Chen JG, Pandelov S, et al. Trends in the exchange current for hydrogen evolution. J Electrochem Soc 2005;152(3):J23–6.
[49] Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, Riviere M, et al. Open catalyst 2020 (oc20) dataset and community challenges. ACS Catal 2021;11 (10):6059–72.