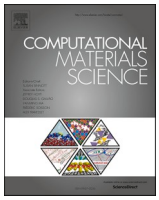
Computational Materials Science 252 (2025) 113775

Contents lists available at ScienceDirect

Computational Materials Science

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



Full Length Article

Fully automated high-throughput computer-based catalytic material screening framework and its application on the new-generation Tianhe supercomputer

Can Leng [a,b,c,1], Xuguang Chen [b,1], Jie Liu [b], Chunye Gong [b], Bo Yang [b], Zhuo Tang [c,d], Wangdong Yang [d], Wei-Qing Huang [e], Yi-Ge Zhou [e], Mengxia Mo [b,d], Kenli Li [c,d,*], Keqin Li [f]

[a] School of Intelligent Manufacturing, Hunan First Normal University, Changsha, China
[b] Collerge of Computer Science, National University of Defense Technology, Changsha, China
[c] National Supercomputer Center in Changsha, Hunan University, Changsha, China
[d] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[e] Department of Applied Physics, School of Physics and Electronics, Hunan University, Changsha China
[f] Department of Computer Science, State University of New York, NY, USA

ARTICLE INFO

Keywords:
High-throughput computing framework
Large-scale supercomputer
Density functional theory calculations
Machine learning
HER catalyst materials

ABSTRACT

The integration of high-performance computing with machine learning (ML) has established a transformative scientific paradigm that significantly enhances the efficiency of material discovery, particularly in the search for catalysts in alternative energy research. However, significant challenges remain in the utilization of available computational resources to accelerate the screening of catalyst materials. In this study, we implement a high-throughput framework on the new-generation Tianhe supercomputer, featuring the development of a Ping-Fault Recovery algorithm, single-task optimization for Density Functional Theory (DFT) to maximize efficiency, and enhanced task scheduling using a two-level scheduling strategy to ensure efficient utilization of the abundant computational resources of the supercomputer. This framework facilitates the identification of 2,028 candidate surfaces across 868 intermetallics from 2,713,897 unique adsorption sites, achieving a screening speed 193 times faster than traditional methods. Alloys composed of Mo, Nb, and V are used as case studies to provide a detailed elucidation of the process of identifying the most effective catalytic surfaces. The framework achieved the best single-day candidate hit performance on 18,106 nodes, completing in one day what previously took a year. This supercomputer-based framework optimizes the use of computational resources, driving innovation in catalyst material discovery.

1. Introduction

Earthshaking changes in human society are inextricably linked to the exploration of nature [1–6]. With the development of various tools and cutting-edge methods, natural exploration—particularlyin materials science—has become a popular research topic, directly influencing societal development [7–12]. Recently, the growth of high-performance computing (HPC), combined with the rapid rise of artificial intelligence (AI) and machine learning (ML), has driven significant advancements in materials science. These developments are characterized by data-driven approaches to processing increasingly large-scale material datasets, either computationally or experimentally through innovative technologies [13–15]. As the scale of data-driven processing in computational and experimental materials research continues to expand, the design and optimization of innovative technologies have become critical challenges, essential for further advancing materials science.

The ongoing transformation towards high-throughput research in materials science is predominantly driven by advancements in informatics tools and ML techniques [16–21], which build upon the foundational work of traditional computational methods such as Density Functional Theory (DFT). Although DFT remains essential for accurate quantum–mechanical atomistic simulations, it is time-consuming and computationally expensive [22]. Python-based informatics tools such as Luigi, FireWorks (FWS), and Python Materials Genomics (Pymatgen) are essential for managing and processing compute-intensive tasks [22–25],

* Corresponding author.
[1] These authors contributed equally to this work and should be considered co-first authors.

https://doi.org/10.1016/j.commatsci.2025.113775
Received 22 September 2024; Received in revised form 9 January 2025; Accepted 9 February 2025

a necessity for high-throughput screening of materials such as catalysts—substances that lower energy barriers in chemical reactions to guide the formation of desired products. The Open Quantum Materials Database (OQMD), developed by Wolverton et al., exemplifies how high-throughput DFT-calculated crystal structures can be integrated with advanced informatics tools, thereby overcoming the limitations of traditional methods. This enhances research capabilities and facilitates the exploration of complex applications such as battery combination catalysts [26]. Furthermore, general AI frameworks including Scikit-learn, TensorFlow, and PyTorch have been extensively utilized in data classification, regression, clustering, and the development of complex neural networks to predict material properties, which are critical for optimizing experimental outcomes [26–35]. Similarly, AFLOW-ML offers a RESTful API for ML predictions of material properties, further advancing the integration of ML techniques with high-throughput data processing to accelerate material discovery and optimization [36]. This integration not only addresses the limitations of traditional DFT methods but also fosters innovative research paradigms, enhancing the scope of materials science investigations.

Despite these advancements, challenges persist, particularly the limited availability of material data, which stems from the high costs and time demands associated with empirical data collection. Additionally, ML is hindered by the scarcity of robust datasets. There is substantial opportunity for optimizing AI utilization with high-throughput computing (HTC) using large-scale resources in HPC systems. For example, the combination of HTC and ML on platforms such as the Materials Artificial Technology Cloud (Matcloud) has been applied to the development of next-generation batteries. However, there remains considerable opportunity for improving the speed of this process [37]. Similarly, Ulissi et al. developed a Generalized Adsorption Simulation (GASpy) platform, which integrates ML with HTC and DFT calculations to design advanced electrocatalytic materials. Although the platform is capable of screening hundreds of candidates annually [38,39], it has the potential to screen more candidates with enhanced processing capabilities and further optimization. Addressing the speed limitations in candidate material screening and accelerating the generation of effective datasets are crucial for overcoming the challenges in traditional experimental methods and mitigating the constraints of ML due to insufficient data [26,38–40]. Therefore, optimizing AI-based HPC systems such as the new-generation Tianhe supercomputer presents an appealing solution. As illustrated in Fig. 1, HTC processes that rely on extensive computational resources encounter several challenges, including dataset management, fault tolerance, scheduling, and task optimization. These issues affect the efficiency of utilizing large-scale computing resources in supercomputing systems, thereby impacting both the accumulation of effective datasets and the progress of candidate material screening. In summary, the efficient integration of HTC with AI to accelerate catalyst material screening using abundant computational resources remains a major challenge.

In this study, we constructed a fully automated high-throughput computational supercompter-based framework for catalytic material screening. By developing several advanced methods, we optimized the utilization of the abundant computational resources available on the new-generation Tianhe supercomputer to validate the efficiency of the material-screening process. This framework enables the efficient screening process of millions of catalytic materials and is demonstrated through specific alloy examples to elucidate the identification of high-performance catalytic materials. Our research revealed an exceptional screening efficiency and produced significant results, highlighting the compelling potential of large-scale computing resources.

The remainder of this study is organized as follows. First, we provide an overview of the framework techniques and computational details. Next, the workflow and specific screening cases are elaborated. Finally, we conclude with a summary of the findings and their implications.

2. Framework techniques and computational details overview

2.1. Informatics tools of the framework and automated material construction approaches

The functionality of the framework is supported by a suite of informatics tools and Python packages, such as Pymatgen, Atomic Simulation Environment (ASE), FWS, Luigi, and MongoDB. These form the core infrastructure of the framework [41,42], as detailed in our previous study [2]. Pymatgen facilitates high-throughput material calculations, ASE manages atomistic simulations, FWS manages HTC jobs, and Luigi orchestrates complex streaming batch tasks. MongoDB, which supports the JavaScript Object Notation (JSON) format, serves as a flexible data storage solution. Supercomputers often employ the Lustre file system for storage [43].

DFT is widely used in quantum–mechanical simulations to predict atomic properties and serves as a bridge between experimental data and theoretical models [22]. Automated DFT calculations over large search spaces are streamlined through Luigi for task management and FWS for job execution. Adsorption energy calculations composed of DFT tasks, such as bulk retrieval and relaxation, bulk expansion, surface cutting, bare slab preparation and relaxation, and adsorbate slab preparation and relaxation, are managed by Luigi. Automatic structure construction is achieved using the Luigi outputs, whereas DFT-related workflows are managed by FWS. These workflows generate computing tasks and monitor their progress through the SLURM scheduler [44]. The results are continuously updated in the database and utilized in subsequent tasks to ensure both efficiency and fault tolerance. Upon job completion, the LaunchPad of the MongoDB pings the “heartbeat” thread via FWS to verify task status, extracts results, and updates the materials database. This process ensures a self-consistent closed-loop system, maintaining

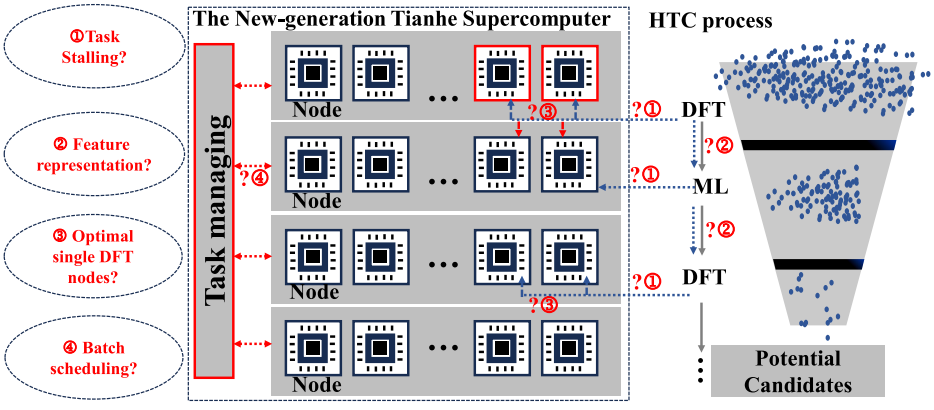


Fig. 1. HTC process composed of DFT + ML iteration in the new-generation Tianhe Supercomputer.

an efficient research workflow.

2.2. ML and DFT computational details

To accelerate computational screening, mutual feedback between ML and DFT calculation are used by compensating for the lack of samples in machine learning and replacing some numerical calculations caused by insufficient computing resources. For the ML process, we combined surrogate-based optimization with active learning. Surrogate-based optimization employs a surrogate model (SSM) to approximate the objective function, reducing the computational cost by optimizing the surrogate instead of a more expensive model. This approach is particularly well-suited to time-consuming objective functions [45,46]. Active learning, a type of ML, efficiently interacts with a small amount of existing labeled data to select the most valuable samples for further labeling [47]. The criterion for evaluating samples is based on the volcano scaling relationships, a classical standard traditionally used to assess the performance of catalytic materials [48–50]. In the active learning process, this standard is incorporated by applying a Gaussian distribution around the optimal values to prioritize the labeling of samples with higher uncertainty. These uncertain materials are selected for DFT calculations, labeled, and subsequently used to retrain the model, enabling iterative exploration of the surrogate model's optimal space.

During the ML process, the fingerprints of the adsorption site environment obtained via DFT, consist of vectors such as: ($[Z_1, X_1, CN_1, \Delta E]$, $[Z_1, X_1, NCN_1, \Delta E]$, …, $[Z_{i+1}, X_{i+1}, CN_{i+1}, \Delta E]$, $[Z_{i+1}, X_{i+1}, NCN_{i+1}, \Delta E]$), as depicted in Table 1. The labeled features are used to create surrogate models, which predict unlabeled fingerprints. These predictions are then used in active learning to select valuable samples for further DFT calculations, significantly accelerating the screening process. By continuously enhancing the surrogate model through iterative feedback from active learning based on DFT calculations, this approach reduces the amount of labeled data required.

In catalyst material research, the physicochemical properties of structures—particularly the adsorption sites where adsorbates attach—are critical for evaluating catalytic efficiency. The key metric, referred to as "candidate hit performance," compares the time required for catalyst screening using the framework against traditional calculation methods. As defined in Equation (1), the percentage P_{all} represents the proportion of materials filtered out using active learning, relative to the total number of raw materials N_r. N_f denotes the number of materials screened using the ML method. The number of candidate materials identified through the iterative feedback between ML and DFT calculations is represented by N_h. Thus, the "candidate hit performance" can be defined using Equation (2).

$$P_{all} = \frac{N_r - N_f}{N_r} \tag{1}$$

$$S = \frac{\frac{N_h}{N_f}}{\frac{N_h}{N_r}} = \frac{N_r}{N_f} \tag{2}$$

Catalysts that adhere to volcano scaling relationships and fall within the optimal adsorption energy range are considered ideal candidates. This study analyzed a wide range of elements, as depicted in Fig. 2, covering 2,771 crystal structures, which led to 205,046 unique surfaces and 2,713,897 unique adsorption sites. These surface-adsorbed gases include H, O, OH, OOH, and CO (Fig. 2), with hydrogen adsorption being the most extensively studied DFT calculation. Each adsorption site on a surface represents a potential catalytic material. For the DFT calculation configuration, we used revised Perdew–Burke–Ernzerhof (PBE) functionals for pseudopotentials [49]. For selected materials, adsorption energy was calculated using DFT to evaluate catalytic efficiency using three relaxation types: bulk/gas relaxation (E_{bulk}/E_{gas}) and slab relaxation with and without adsorbates (E_{adslab} and E_{bare_slab}). The adsorption energy, E_{ads}, was determined as $E_{adslab} - E_{bare\_slab} - E_{gas}$. This study focuses on screening catalytic materials for the adsorption of renewable and environmentally friendly gases, particularly those used in hydrogen evolution reaction (HER) catalysts. Optimal adsorption energies for different chemical element combinations are determined by volcano scaling relationships, with the ideal ΔE_H for HER being −0.27 eV [50–52]. The optimal range for adsorption energy fluctuations is defined as [-0.37, −0.17] eV, and materials within ± 0.1 eV of this optimal value are considered ideal candidates.

Table 1
Details of the fingerprints. Z represents the atomic number, X represents the atomic Pauling electronegativity, CN/NCN represents the number of atoms coordinated/neighbor-coordinated to the adsorbate, and the target is represented by the adsorption energy ΔE. The variable i is related to the maximum number of material components. For atoms without coordination, a dummy value is used ($Z = 0$, and the other features are averages).

No.	feature	Definition (Physical meaning)	Independent
1	Z_1	Atomic number of the first atom in the material composition	Y
	X_1	Pauling electronegativity of the first atom in the material composition	Y
	CN_1	Number of atoms to which the first atom in the material composition is coordinated to the adsorbate	Y
	ΔE	Adsorption energy	N
2	Z_1	Atomic number of the first atom in the material composition	Y
	X_1	Pauling electronegativity of the first atom in the material composition	Y
	NCN_1	Number of the neighbor atoms to which the first atom in a material composition is coordinated to the adsorbate	Y
	ΔE	Adsorption energy	N
i	Z_{i+1}	Atomic number of the $(i+1)^{th}$ atom in the material composition	Y
	X_{i+1}	Pauling electronegativity of the $(i+1)^{th}$ atom in the material composition	Y
	CN_{i+1}	Number of atoms to which the $(i+1)^{th}$ atom in the material composition is coordinated to the adsorbate	Y
	ΔE	Adsorption energy	N
$i+1$	Z_{i+1}	Atomic number of the $(i+1)^{th}$ atom in the material composition	Y
	X_{i+1}	Pauling electronegativity of the $(i+1)^{th}$ atom in the material composition	Y
	NCN_{i+1}	Number of the neighbor atoms to which the $(i+1)^{th}$ atom in the material composition is coordinated to the adsorbate	Y
	ΔE	Adsorption energy	N($[Z_1, X_1, CN_1]$, $[Z_1, X_1, NCN_1]$, …, $[Z_{i+1}, X_{i+1}, CN_{i+1}]$, $[Z_{i+1}, X_{i+1}, NCN_{i+1}]$, …) $i > 0$, $i \in Z$

2.3. HTC system and environment

We used a new-generation Tianhe supercomputer equipped with domestic high-performance heterogeneous multicore processors (MT-SQ) and connected via a proprietary high-speed network (TH-E3). Approximately 160,000 DFT tasks were executed with execution times ranging from a few seconds to two days. These tasks were distributed across 18,016 compute nodes and performed using the VASP software (version 5.4.4) [53]. An AutoML tree-based pipeline optimization tool (TPOT) was used during the ML process [54].

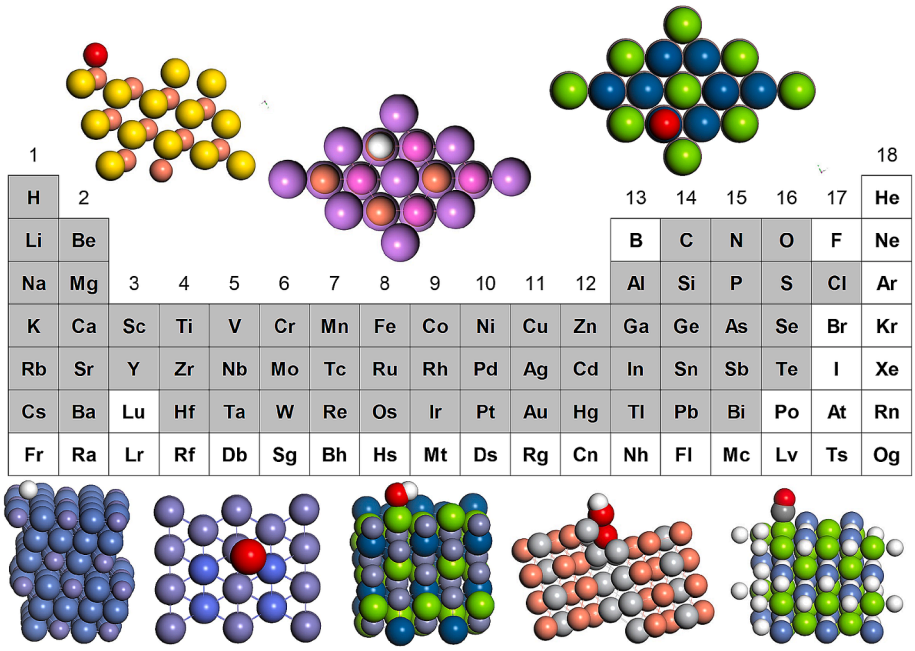


Fig. 2. Morphology and elemental distribution of catalyst materials considered in this screening. The gray elements in the periodic table represent those covered by the catalytic material screening process.

3. Results and Discussion

3.1. Workflow

The integration of informatics tools and supercomputing system architecture enabled high-throughput job generation in the supercomputing environment, data processing from first-principles DFT calculation results, ML, and interactive feedback to screen electrocatalytic material efficiently, as shown in Fig. 3. The DFT calculations provide an effective dataset and machine learning predictions of adsorption capabilities accelerate the computational process. The process includes the extraction of raw data and task scheduling management functions to support comprehensive screening of candidate catalysts on the supercomputer. The screening process began with

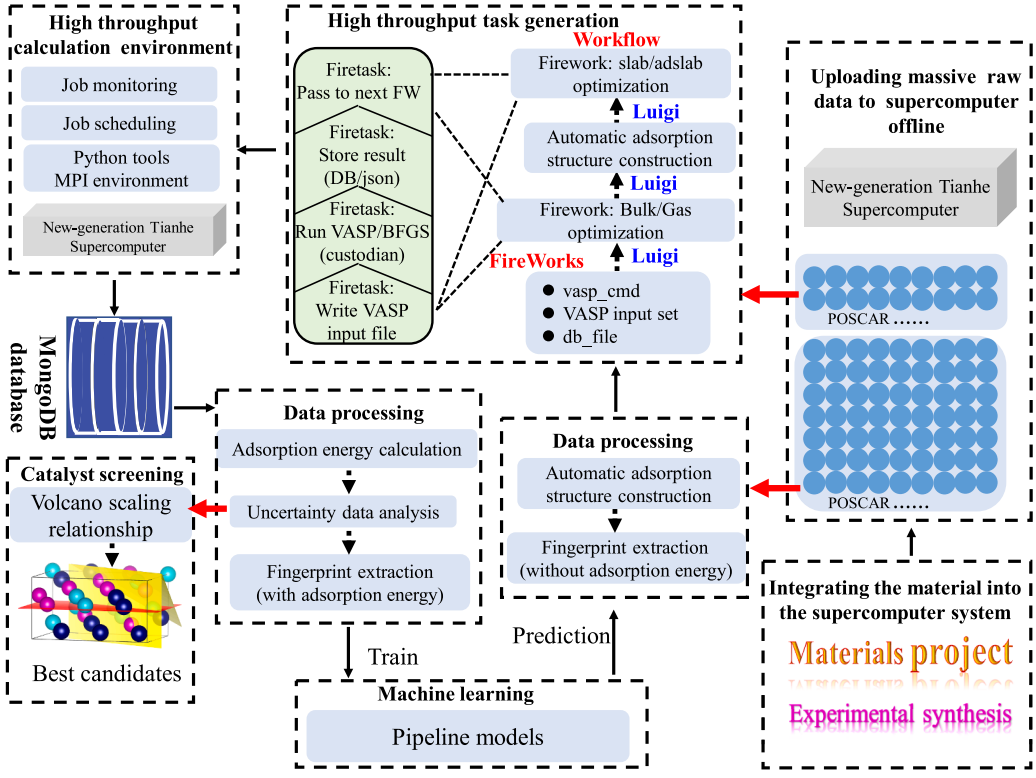


Fig. 3. Overview of the framework. It includes components for data extraction (POSCAR files), high-throughput task generation, monitoring and scheduling, data processing, machine learning, and catalyst screening.

extracting numerous material structures (POSCAR files) from the Materials Project or other experimental results provided by researchers [55,56]. These structures were then transferred to the supercomputer and divided into two parts: a smaller subset for task generation and a larger subset for data processing. Luigi managed the output dependencies of the physical structure, creating a pipelined task for structure generation. These structures were first optimized using VASP calculations managed by FWS.

Automatic adsorption structure generation followed, producing various adsorption structures with different Miller indices for FireTasks. FSW handled task management, setting up VASP input files, checking task requirements, and determining whether VASP should run. The results were stored in MongoDB in JSON format, allowing efficient management of large-scale workloads. For ML, fingerprints extracted from the raw data were used to train a pipeline model using TPOT, which predicted the properties of materials after pre-processing. Combined with volcano scaling relationships, the best candidates were determined using robust relaxation schemes to assess the stability of the structures. By leveraging machine learning, this approach significantly accelerated the material screening process. Instead of performing exhaustive DFT calculations on all materials, the model was able to predict key material properties with high accuracy, allowing for the efficient selection of promising candidates for further analysis. This reduced the computational cost and time required for the material screening, facilitating a more rapid identification of optimal materials. HTC jobs were monitored by a security system deployed on the cluster to ensure only the most promising materials, based on their lowest energy configurations, progressed.

The supercomputer system provided VASP and Python environments and monitored and scheduled jobs created by FWS through the SLURM resource manager. SLURM managed the specific VASP jobs, and FWS recorded their runtime status as 'RUN,' 'COMPETED,' 'FIZZLED,' or other states. All results were queried and stored in the MongoDB database. The DFT results stored in MongoDB were also used to connect VASP calculations with the ML process.

3.2. Fault tolerance recovery

Task failures are common during execution due to various faults. Typically, a failure occurs when the "heartbeat" signal from LaunchPad stops, causing the task to be marked as "FIZZLED" after being in the "RUNNING" state in FWS. System faults, such as supercomputer maintenance leading to ping failures, or situations where the "heartbeat" thread continues to send pings while the actual task itself is in a "dead" state in FWS, can exacerbate the issue. In both cases, the tasks submitted to supercomputer computing nodes continued to be executed normally; however, upon completion, the results could not be updated in the database. This prevents the initiation of subsequent tasks within the high-throughput material computation framework. To address these issues and minimize computational losses from ping failures or tasks in a "dead" state despite normal pings, we have developed a Ping-Fault Recovery algorithm, outlined in Algorithm 1. The algorithm operates as follows.

Algorithm 1: Ping-Fault Recovery

Input: SLURM id (S_IDs), and the FWS id (F_IDs)
Output: Completed calculation task (Comp_task)
FW.json: Documentation that collects the information of tasks in FWS

```
for each SLURM_ID in S_IDs do
    Retrieve the directory associated with SLURM_ID
    Extract the fwid from FW.json in the directory
    match_found = False
    for each FWS_ID in F_IDs do
        if fwid == FWS_ID then
            if status of FWS_ID is FIZZLED then
                Kill the SLURM_ID
                Rerestart the task
```

(continued on next column)

(continued)

Algorithm 1: Ping-Fault Recovery

```
            end if
            if status of FWS_ID is RUNNING then
                match_found = True
            end if
            break
        end if
    end for
    if not match_found and status of FWS_ID is RUNNING then
        Change the status of FWS_ID to FIZZLED
        Restart the task
    end if
end for
return Comp_task
```

When SLURM assigns high-throughput tasks to compute nodes, the algorithm traverses all submitted SLURM IDs, locates their respective directories, and retrieves the associated forward IDs. It then initializes match_found to false and compares the SLURM IDs with all fwid IDs, which are marked either as "FIZZLED" or "RUNNING." For tasks marked as FIZZLED, which are presumed interrupted due to system maintenance, the algorithm promptly terminates the matched SLURM ID and restarts it to resume the task. For the tasks marked as "RUNNING," match_found is set to true. If match_found remains false even though the task is "RUNNING," the task is considered "dead" despite normal pings. These tasks are marked as "FIZZLED" and immediately resubmitted. This approach effectively identifies and addresses tasks that cannot be updated, thereby reducing computational resource waste and ensuring smooth progression of dependent tasks.

3.3. Time to solution for DFT calculations and high-throughput task execution

To optimize the efficiency of the DFT calculations, several performance tests were conducted on the most frequent slab computation tasks to identify the optimal number of processes and threads. As shown in Fig. 4, speedup was achieved by comparing computation times across multiple nodes relative to a single compute node. Initially, as the number of processes increased, speedup improved, reaching a factor of 60 with 32 computing nodes. This performance level was maintained as the number of nodes increased to 128, with speedup peaking at 78. However, beyond this point, further increases in node count led to diminished efficiency, with speedup dropping to 2.5 when scaling up to 4096 compute nodes due to increased communication costs. Therefore,

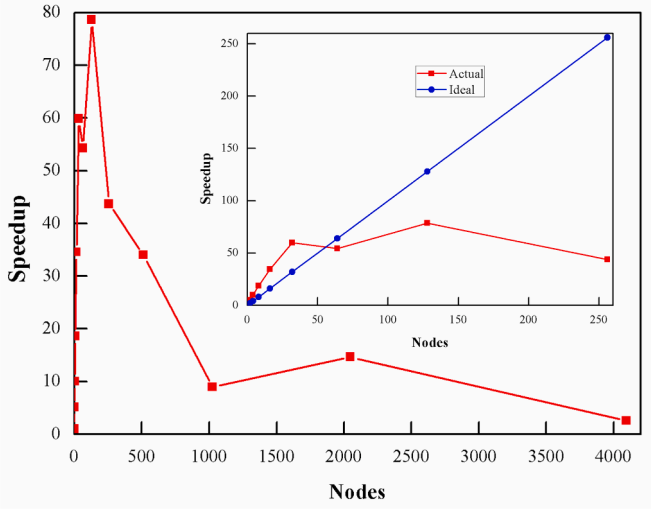


Fig. 4. Speedup on multiple nodes compared to a single node.

while the speedup ratio was satisfactory within 128 computing nodes, a four-fold increase in node count resulted in only a 1.3-fold increase in speedup, which was considered inefficient. As a result, 32 compute nodes were identified as the most effective choice for slab calculation tasks.

The task scheduling process for high-throughput material computing on the new-generation Tianhe supercomputer is illustrated in Fig. 5. In this system, tasks managed by FWS are submitted to SLURM for scheduling purposes. Once SLURM completes scheduling and calculations, the results are sent back to FWS, which stores the completed task results in a material database. Subsequently, additional tasks generated from the ML processes are submitted to SLURM for further scheduling. This system ensures consistency between the statuses of tasks managed by FWS and those submitted to SLURM, thereby embodying the two-level scheduling strategy. A total of 18,016 compute nodes were utilized for high-throughput task execution, whereas the remaining nodes were dedicated to ML tasks.

3.4. Accelerating catalysts screening

To efficiently screen a vast number of catalytic materials for the HER, we analyzed the relaxation numbers on a new-generation Tianhe supercomputer, covering 5367 bulk/gas relaxations from crystal structure expansion and 155,467 bare and adsorbate slab relaxations, as shown in Table 2. For an adsorption site not previously studied for any adsorbate, two DFT calculations (bare and adsorbate slabs) were required. When different adsorbates are adsorbed at the same site, only one additional DFT calculation is required, as the bare slab had already been calculated. Although bulk relaxation requires more time, the number of slab calculations is three times higher, making slab relaxation the predominant DFT computational task.

In the context of HER, this study involved 2,446 distinct bulk structures out of 2,771 crystal structures, resulting in 104,251

Table 2
Relaxation types of high-throughput tasks, DFT calculation numbers, and average time per task.

relaxation	*Bulk/Gas*	*Slab*
DFT calculation numbers	5367	155467
Average time per task (s)	6470	3894

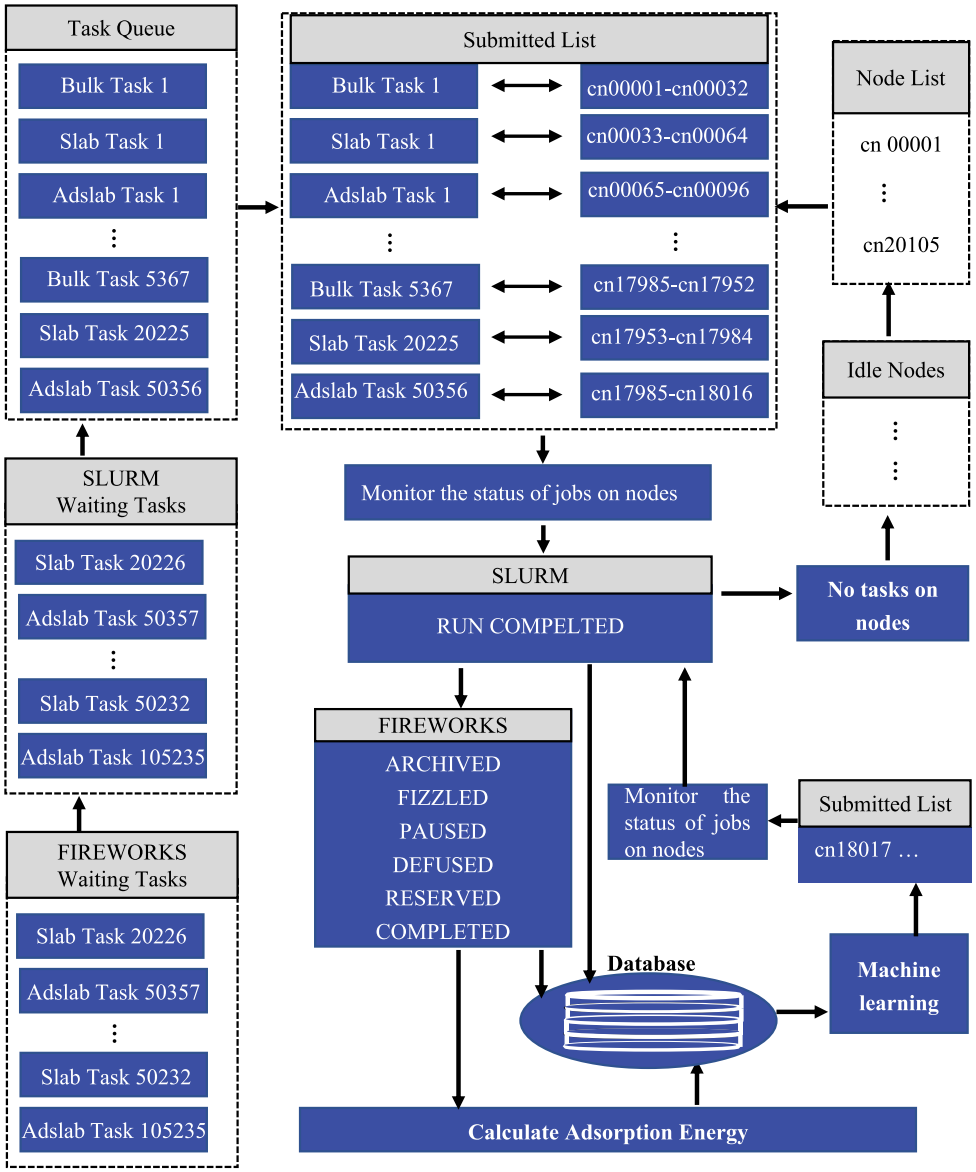


Fig. 5. Hybrid high-throughput computing task management system.

adsorption-energy calculations and encompassing 160,837 DFT calculations. The maximum Miller indices of 2 produced 50 irreducible crystal surfaces, yielding a dataset of 2,098,299 unique adsorption sites. TPOT distributed these adsorption energies, derived from 2,658,820 sites, for ML analysis related to HER, as depicted in Fig. 6(a). The remaining predicted results pertain to adsorption energies beyond the HER. The training dataset included 55,077 different sites composed of 48,704 different surfaces (one surface may have several sites) obtained from our DFT calculations, with the DFT results providing essential data points that form the foundation for machine learning models. Additionally, 22,675H adsorbate DFT results from Tran et al. were incorporated to enhance model training and predictions [38]. Over a period of 20 d, eight surrogate models were generated, each optimizing 60 pipelines. Only 353,441 ML results satisfied the volcano scaling relationship criteria, as shown in Fig. 6(b). Active learning was combined with surrogate-based optimization to enhance the screening process, utilizing active learning to select structures from each surface for DFT calculations. By selecting the most valuable adsorption sites for further labeling

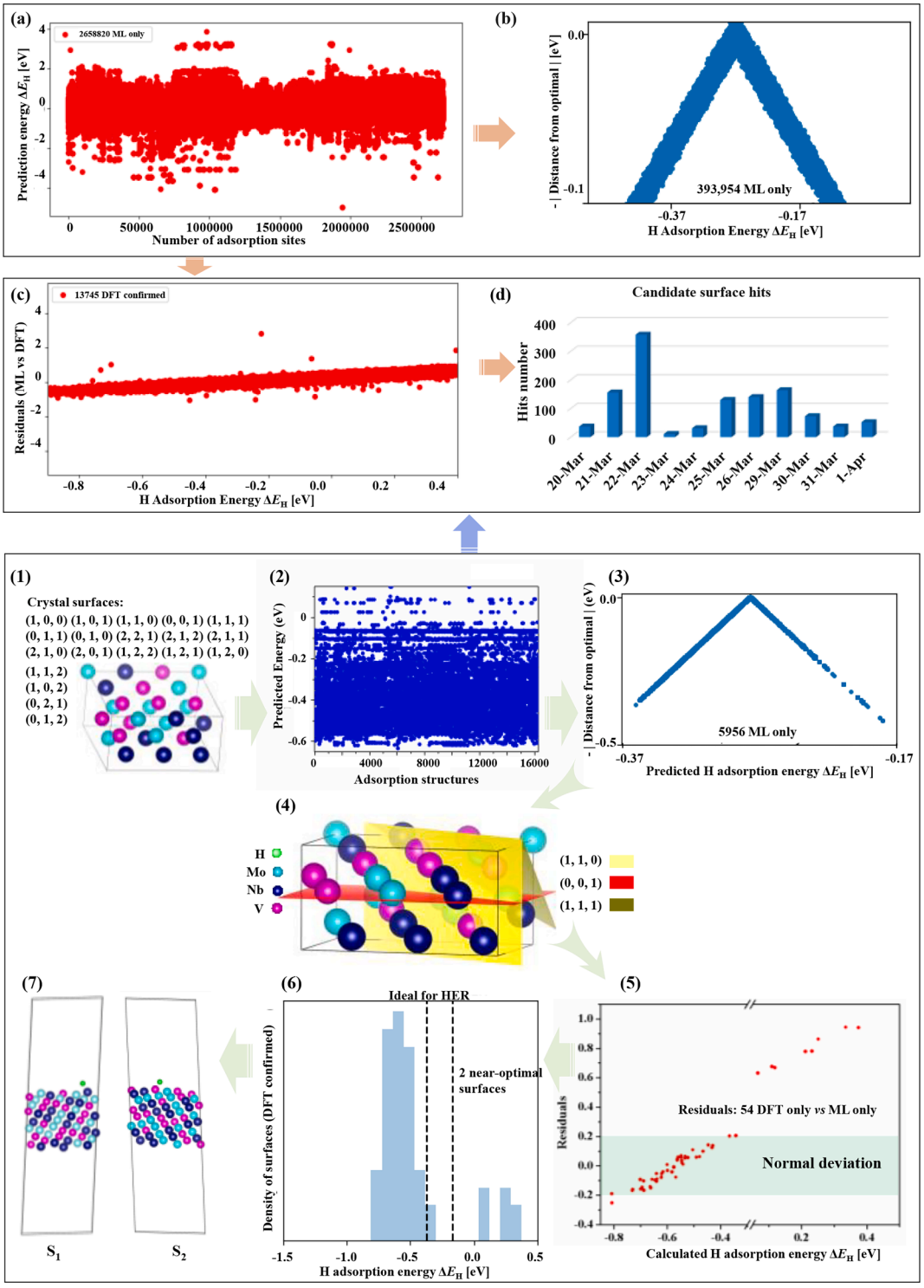


Fig. 6. Catalyst screening process within the framework. The upper figure depicts the overall HER screening process, and the lower figure illustrates a specific case of MoNbV material screening.

and DFT calculations, machine learning predictions guided the efficient selection of surfaces, thus optimizing the number of DFT calculations needed while maintaining high accuracy. This resulted in the generation of eight self-consistent closed loops, each comprising a surrogate model. This process screened 13,745 lowest-energy surfaces, as shown in Fig. 6 (c). The residuals between the DFT results and ML predictions were predominantly close to zero, demonstrating high prediction accuracy and alignment with expected normal deviation. Through robust relaxation strategies, 1,574 candidate surfaces across 868 intermetallics were selected, as shown in Fig. 6(d), with machine learning predictions significantly accelerating the identification of promising candidates, thus minimizing the number of DFT calculations required while maintaining high prediction accuracy. To illustrate the capabilities of this framework, we focused on the HERs of materials composed of Mo, Nb, and V. For the MoNbV material, a maximum Miller index of 2 produced 19 irreducible crystal surfaces, resulting in 16,250 adsorption structures for prediction, as shown in the lower part of Fig. 6(1) and (2). When combined with uncertainty data analysis, only 5956 ML results fell within the optimal range, as shown in Fig. 6(3). These selected surfaces were further screened, and the surface with the lowest adsorption energy was selected for DFT calculations. Fig. 6(4) shows that only three crystal surfaces, namely (1, 1, 1), (1, 1, 0), and (0, 0, 1), were selected, with a total of 54 adsorption structures subjected to DFT calculations. Fig. 6(5) shows the residuals between the DFT results and ML predictions, where most fall within the optimal range, indicating a normal deviation. The two adsorption structures with the best catalytic performance were identified, as shown in Fig. 6(6) and (7).

3.5. Selection results and computational efficiency Evaluation

Based on the above workflow, we analyzed 13,745 DFT calculation results using the average adsorption energy values of pure elements, as depicted in Fig. 7. The average binding energy of hydrogen on Pt is −0.31 eV, closely matching the ideal value of −0.27 eV, indicating that Pt is among the most active elements for the HER. Pt was identified in four distinct regions: weak, weak, strong, weak, and strong. The most active binding occurred in the strong–weak region predicted by ML, followed by the weak-strong region determined by DFT calculations. The weak-weak region generally exhibited minimal element binding and was often inactive due to its weak binding ability.

However, Cu/Pd, Pd/Fe, and Ti/Co exhibited unexpected activity. Cu and Pd, both highly active elements after Pt, may exhibit higher average values due to calculation errors, which could explain their activity in the weak-weak region. Despite their location in this region, these bimetallic combinations followed the volcano scaling relationship and exhibited significant activity. ML predictions identified more active bimetallic compounds than DFT calculations, consistent with Fig. 7, where the surrogate-based model identified more adsorption sites with near-optimal energy. Thus, combinations of the weak-strong and strong–weak regions provide valuable insights into the HER and can guide the synthesis of effective catalysts.

Furthermore, 2,028 surfaces across 868 intermetallics were identified for the HER from 2,713,897 unique adsorption sites, as illustrated in Fig. 8(a). Potential active bimetallic candidates are shown in Fig. 8 (b), based on 21,162 DFT adsorption energy calculations, including several outliers with abnormal values outside the optimal range.

Hundreds of effective catalytic materials are screened daily in Fig. 9. Compared to another study, which required over a year to screen 389

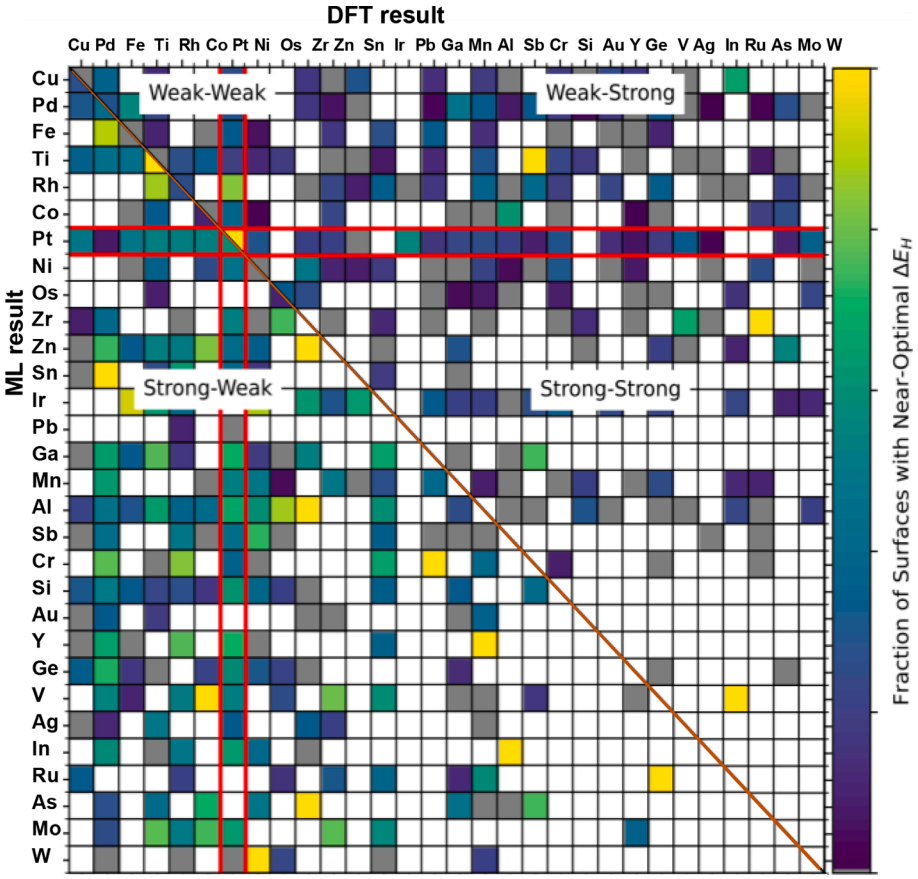


Fig. 7. H2 evolution efficiency map for bimetallics. The colored shading represents possible efficiency for various enumerated surfaces; grey shading indicates bimetallics outside the ± 0.1 eV optimal range, and the white shading denotes surfaces not included in the analyzed dataset. The upper half of this figure shows adsorption energies calculated by DFT, whereas the lower half shows values generated by the surrogate model.

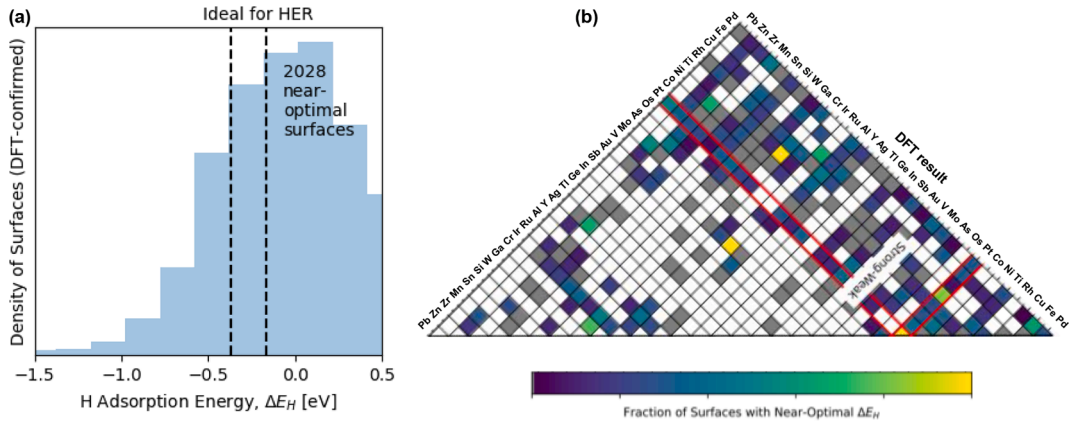


Fig. 8. Number of candidate surface hits and potential active bimetallic candidates in iterative and pure DFT calculations. (a) Total number of candidate surfaces. (b) Possible combinations of active bimetallic candidates in partial screening results. The colored shading indicates the potential efficiency of any enumerated surfaces; grey shading indicates bimetallics outside the ± 0.1 eV optimal range, and the white shading indicates enumerated surfaces not included in the analyzed dataset. All the values created by the adsorption energies were calculated using DFT.

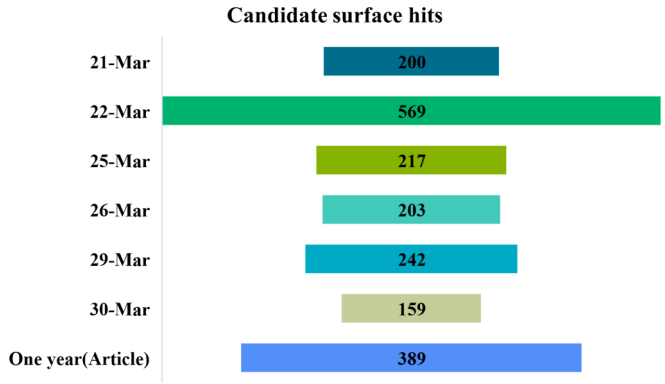


Fig. 9. Number of candidate surface hits per day.

candidate surfaces [38], our method completes this task in a single day. The low hit rate observed during large-scale material screening effectively narrowed the research scope and conserved significant time, demonstrating the high efficiency of this approach.

Efficiency analysis evaluated the acceleration of the material screening process. As shown in Table 3, 2,658,820 adsorption sites (denoted as N_r) were generated. By applying the self-consistent closed loop, only 13,745 adsorption structures (represented by N_f) were deemed suitable for DFT calculations after ML prediction. Ultimately, DFT calculations identified 1,395 theoretically ideal intermetallics (represented by N_h), which are now being further investigated in collaboration with relevant laboratories [57]. This screening process filtered out 99.48 % of the structures, achieving a speedup factor of 193 for identifying the best candidates using this framework.

The framework is designed to fully leverage high-performance distributed computing environments, enabling continuous distribution of multiple computational tasks across various nodes. It utilizes efficient database management systems, such as MongoDB, to meet the demands of processing large datasets, significantly improving the efficiency of catalytic material screening. Additionally, while the framework currently focuses on adsorption energy calculations for catalytic materials, its potential applications can be extended to other areas, such as battery theoretical capacity and charge transfer. The development of these applications can build upon the existing framework by modifying DFT calculations for different microscopic properties and constructing machine learning models based on datasets derived from these properties. This direction represents a key focus for future work, which will further expand upon the capabilities of the current framework.

Table 3
Efficiency of acceleration using ML in the high throughput framework.

	N_r	N_f	N_h	P_{all}	S
Structure numbers	2658820	13745	1395	99.48 %	193

4. Conclusion

In this study, we present a fully automated, high-throughput, computer-based framework deployed on the new-generation Tianhe supercomputer. To effectively utilize its extensive computational resources, we developed a Ping-Fault Recovery algorithm to enhance fault tolerance, reducing computational resource waste and ensuring smooth progression of dependent tasks. We identified 32 compute nodes as the most effective configuration for slab calculations based on time-to-solution for DFT calculations. The task scheduling process for high-throughput material computing was enhanced using FWS and SLURM on a new-generation Tianhe supercomputer. By utilizing this method, the continuous computing capacity of the supercomputer enabled the identification of optimal candidate catalysts from a substantial number of materials by screening their adsorption properties for renewable and environmentally friendly gases, focusing on the HER. We demonstrated the framework using Mo, Nb, and V as case studies to provide a detailed elucidation of the process for identifying the most effective catalytic surfaces. In total, 2,028 candidate surfaces across 868 intermetallics were identified from 2,713,897 unique adsorption sites, achieving a speed-up factor of 193 in identifying the best candidates using this framework. The best single-d ay candidate hit performance using 18,106 nodes allowed us to achieve in one day what previously required a year. These identified candidates hold significant potential for further in-depth research in catalytic materials, contributing to the advancement of more intelligent and high-precision materials research.

The supercomputer-based high-throughput screening framework, emphasizing an automated and uninterrupted computational process that integrates DFT calculations with machine learning feedback loops, serves as an exemplary case for high-throughput computing in complex distributed systems. The inclusion of Fault Tolerance Recovery and High-throughput Task Execution modules ensures the seamless and continuous submission of tasks. This framework, combined with supercomputing capabilities, highlights the crucial role of advanced computational tools in accelerating material screening processes, thereby significantly advancing the development of efficient and sustainable energy solutions to address global energy challenges.

CRediT authorship contribution statement

Can Leng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Data curation, Conceptualization. **Xuguang Chen:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Jie Liu:** Supervision, Software, Resources, Funding acquisition. **Chunye Gong:** Supervision, Software, Resources. **Bo Yang:** Software, Resources, Funding acquisition. **Zhuo Tang:** Supervision, Software, Conceptualization. **Wangdong Yang:** Supervision, Resources, Methodology. **Wei-Qing Huang:** Writing – original draft. **Yi-Ge Zhou:** Software, Resources, Conceptualization. **Mengxia Mo:** Validation. **Kenli Li:** Writing – original draft, Visualization, Validation, Supervision, Software, Resources. **Keqin Li:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research work was supported by the National Key Research and Development Program of China, China (2021YFB0300101). We thank Prof. Zachary W Ulissi and Prof. Pari Palizahti from Carnegie Mellon University, United States, for providing advice on the framework.

Data availability

Data will be made available on request.

References

[1] N.H. Angello, D.M. Friday, C. Hwang, et al., Closed-loop transfer enables artificial intelligence to yield chemical knowledge, Nature 633 (2024) 351–358.
[2] C. Leng, Z. Tang, Y. Zhou, et al., Fifth paradigm in science: a case study of an intelligence-driven material design, Engineering 24 (2023) 126–137.
[3] X. Peng, X. Wang, Next-generation intelligent laboratories for materials design and manufacturing, MRS Bull. 48 (2023) 179–185.
[4] B.A. Koscher, et al., Autonomous, multiproperty-driven molecular discovery: from predictions to measurements and back, Science 382 (2023) eadi1407.
[5] R.K. Saifur, A.R. Dana, P. Anthony, B.W. Michael, Integration of AI and traditional medicine in drug discovery, Drug Discov. Today 26 (4) (2021) 982–992.
[6] J. Glaser, J.V. Vermaas, D.M. Rogers, et al., High-throughput virtual laboratory for drug discovery using massive datasets, The International Journal of High Performance Computing Applications. 35 (5) (2021) 452–468.
[7] S. Biyela, K. Dihal, K.I. Gero, et al., Generative AI and science communication in the physical sciences, Nat. Rev. Phys. 6 (2024) 162–165.
[8] J. Rockstrom, W. Steffen, K. Noone, A. Persson, F.S. Chapin, E.F. Lambin, et al., A safe operating space for humanity, Nature 461 (2009) 472–475.
[9] C. Lv, X. Zhou, Z.L. Zhou, C. Yan, et al., Yan, machine learning: an advanced platform for materials development and state prediction in lithium-ion batteries, Adv. Mater. 34 (2022) 2101474.
[10] S. Singh, J. Ru, Accessibility, affordability, and efficiency of clean energy: a review and research agenda, Environ. Sci. Pollut. Res. 29 (2022) 18333–18347.
[11] M. Abolhasani, E. Kumacheva, The rise of self-driving labs in chemical and materials sciences, Nature Synthesis. 2 (2023) 483–492.
[12] M.T. Warren, C.I. Biggs, A. Bissoyi, et al., Data-driven discovery of potent small molecule ice recrystallisation inhibitors, Nature Communication. 15 (2024) 8082.
[13] T. Huang, Z. Yang, L. Li, H. Wan, C. Leng, G. Huang, W.Y. Hu, W.Q. Huang, Dipole effect on oxygen evolution reaction of 2d janus single-atom catalysts: a case of Rh anchored on the P6m2-Np configurations, The Journal of Physical Chemistry Letters. 15 (2024) 2428–2435.
[14] L. Himanen, A. Geurts, A.S. Foster, P. Rinke, Data-driven materials science: status, challenges, and perspectives, Adv. Sci. 6 (2019) 1900808.
[15] X.Y. Liu, J.P. Xiao, H.J. Peng, et al., Understanding trends in electrochemical carbon dioxide reduction Rates, Nat. Commun. 8 (2017) 15438.
[16] S. Sehrish, J. Kowalkowski, M. Paterno, C. Green, Python and HPC for high energy physics data analyses, in: Proceedings of the 7th Workshop on Python for High-Performance and Scientific Computing, Association for Computing Machinery, Denver, USA, 2017, pp. 1–8.
[17] D. Kossmann, The Global AI Supercomputer, in: Proceedings of the 13th ACM International Conference on Distributed and Event-based Systems, Association for Computing Machinery, New York, NY, USA, 2019, p. 6.
[18] N.K. Nepal, P.C. Canfield, L. Wang, HTESP (High-throughput electronic structure package): a package for high-throughput ab initio calculations, Comput. Mater. Sci 244 (2024) 113247.
[19] E. Heid, K.P. Greenman, Y. Chung, S.-C. Li, et al., Chemprop: a machine learning package for chemical property prediction, ChemRxiv. (2023).
[20] G.J. Wang, L.Y. Peng, K.Q. Li, et al., Alkemie: an intelligent computational platform for accelerating materials discovery and design, Comput. Mater. Sci 186 (2021) 110064.
[21] YG. Wang, K. Li, L.Y. Peng, et al., High-throughput automatic integrated material calculations and data management intelligent platform and the application in novel alloys, Acta Metall. Sin. 58 (1) (2022) 75–88.
[22] W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. 140 (1965) A1133–A1138.
[23] Bernhardsson, E. Freider, E. Rouhani, A. Luigi, a Python Package that Builds Complex Pipelines of Batch Jobs. https://github.com/spotify/luigi, 2012.
[24] F. Yalcin, M. Wolloch, SurfFlow: high-throughput surface energy calculations for arbitrary crystals, Comput. Mater. Sci 234 (2024) 112799.
[25] S.P. Ong, W.D. Richards, A. Jain, et al., Python materials genomics (pymatgen): a robust, open-source python library for materials analysis, Comput. Mater. Sci 68 (2013) 314–319.
[26] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open Quantum Materials Database (OQMD), JOM 65 (11) (2013) 1501–1509.
[27] J.G. Hao, T.K. Ho, Machine learning made easy: a review of scikit-learn package in python programming language, J. Educ. Behav. Stat. 44 (2019) 348–361.
[28] Z. Zeng, L. Xiao, X. Li, Z.Cai, et al., Method for Deploying TensorFlow Model in Mobile Terminal. China Patent CN109933339A, Filed: 2019-02-01, Published: 2019-06-25.
[29] X. Li, A Method for PyTorch Task Training Method, Device, and Quality Setting. China Patent CN110782040A, Filed: 2019-10-12, Published: 2020-02-11.
[30] B.A. Koscher, et al., Autonomous, multiproperty-driven molecular discovery: from predictions to measurements and back, Science 382 (2023) adi1407.
[31] B. Smiti, et al., A high-throughput microfabricated platform for rapid quantification of metastatic potential, Science Advance. 10 (2024) adk0015.
[32] C.A. Patino, et al., Multiplexed high-throughput localized electroporation workflow with deep learning–based analysis for cell engineering, Science Advance. 8 (2022) abn7637.
[33] B.P. MacLeod, F.G.L. Parlane, A.K. Brown, et al., Flexible automation accelerates materials discovery, Nat. Mater. 21 (2022) 722–726.
[34] F. Yalcin, M. Wolloch, SurfFlow: High-throughput surface energy calculations for arbitrary crystals, Comput. Mater. Sci 234 (2024) 112799.
[35] M. Yao, Y. Wang, X. Li, et al., Materials informatics platform with three dimensional structures, workflow and thermoelectric applications, Sci. Data 8 (2021) 236.
[36] G. Eric, T. Cormac, O. Corey, et al., AFLOW-ML: a RESTful API for machine-learning predictions of materials properties, Comput. Mater. Sci 152 (2018) 134–145.
[37] X.Y. Yang, Z.G. Wang, X.S. Zhao, et al., Matcloud: a high-throughput computational infrastructure for integrated management of materials simulation, data and resources, Comput. Mater. Sci 146 (2018) 319–333.
[38] M. Zhong, K. Tran, Y.M. Ming, et al., Accelerated discovery of CO2 electrocatalysts using active machine learning, Nature 581 (2020) 178–183.
[39] K. Tran, Z.W. Ulissi, Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution, Nat. Catal. 1 (2018) 696–703.
[40] R. Feng, C.A. Zhang, M.C. Gao, et al., High-throughput design of high-performance lightweight high-entropy alloys, Nat. Commun. 12 (2021) 4329.
[41] A.H. Larsen, J.J. Mortensen, J. Blomqvist, et al., The Atomic Simulation Environment Python Library for Working with Atoms, Journal of Physics-Condensed Matter. 29 (27) (2017) 273002.
[42] S. Agrawal, J.P. Verma, B. Mahidhariya, N. Patel, A. Patel, Survey on MongoDB: An Open-Source Document Database, Int. J. Adv. Res. Eng. Technol. 6 (2015) 1–11.
[43] Y. Wang, Y. Lu, C. Qiu, P. Gao, J. Wang, Performance evaluation of a infiniband-based lustre parallel file system, Procedia Environ. Sci. 11 (2011) 316–321.
[44] A.B. Yoo, M.A. Jette, M. Grondona, Slurm: Simple linux utility for resource management, in: Workshop on job scheduling strategies for parallel processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 44–60.
[45] M. Pilát, R. Neruda, An evolutionary strategy for surrogate-based multiobjective optimization, in: 2012 IEEE congress on evolutionary computation, IEEE, 2012, pp. 1–7.
[46] H.M. Gutmann, A radial basis function method for global optimization, J. Glob. Optim. 19 (2001) 201–227.
[47] J. Yao, Y. Wu, J. Koo, B. Yan, H. Zhai, Active learning algorithm for computational physics, Phys. Rev. Res. 2 (2020) 013287.
[48] B. Hammer, J.K. Nørskov, Theoretical surface science and catalysis—calculations and concepts, Adv. Catal. 45 (2000) 71–129.
[49] B. Hammer, L.B. Hansen, J.K. Nørskov, Improved adsorption energetics within density-functional theory using revised perdew-burke-ernzerhof functionals, Phys. Rev. B 59 (1999) 7413–7421.
[50] F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, et al., Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces, Phys. Rev. Lett. 99 (1) (2007) 016105.
[51] F. Calle-Vallejo, J.I. Martínez, J.M. García-Lastra, et al., Physical and chemical nature of the scaling relations between adsorption energies of atoms on metal surfaces, Phys. Rev. Lett. 108 (11) (2012) 116103.
[52] J.K. Norskov, T. Bligaard, A. Logadottir, et al., Trends in the exchange current for hydrogen evolution, J. Electrochem. Soc. 152 (2005) J23–J26.

[53] G. Kresse, J. Furthmüller, Efficiency of Ab-Initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci 6 (1996) 15–50.
[54] T.T. Le, W.X. Fu, J.H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector, Bioinformatics 36 (1) (2020) 250–256.
[55] M.C. Sorkun, S. Astruc, J. Koelman, et al., An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery, npj Comput. Mater. 6 (1) (2020) 7.
[56] A. Jain, S.P. Ong, G. Hautier, et al., Commentary: the materials project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002.
[57] J.H. Shi, Z.X. Yang, H. Wan, et al., Rapid construction of double crystalline prussian blue analogue hetero-superstructure, Small 20 (33) (2024) e2311267.