xxx *et al. J. Mater. Inf.* **2026**;*x*:xx
DOI: xxx/jmi.xxxx.xx

Journal of materials Informatics

Original Article

Open Access

PhyMLP: an automated strategy for machine-learning potential construction via data fusion and adaptive point-sampling

Yiqun Guo[1,†], Shihui Ning[2,†], Guoqi Guo[2], Yangchun Chen[2,*], Bowen Huang[2,*], Shifang Xiao[1,*], Wangyu Hu[2,*]

[1]School of Physics and Electronics, Hunan University, Changsha 410082, Hunan, China.
[2]College of Materials Science and Engineering, Hunan University, Changsha 410082, Hunan, China.

Correspondence to: Dr. Shifang Xiao, School of Physics and Electronics, Hunan University, Changsha 410082, Hunan, China. E-mail: shifangxiao@hnu.edu.cn; Dr. Yangchun Chen, Dr. Bowen Huang, Prof. Wangyu Hu, College of Materials Science and Engineering, Hunan University, Changsha 410082, Hunan, China. E-mail: ychchen@hnu.edu.cn; bowen_huang@hnu.edu.cn; wyuhu@hnu.edu.cn

Received: 8 April 2026 **First Decision:** 7 May 2026 **Revised:** 15 July 2026 **Accepted:** 16 July 2026

Abstract

This paper presents an automated strategy for constructing machine-learning potential (MLP) datasets through a physics-strengthened point-sampling scheme called physics-strengthened machine-learning potential (PhyMLP). By integrating heterogeneous data sources - including the Rose equation of state, experimental pressure-volume (P - V) measurements, traditional empirical potentials, and first-principles calculations - this method effectively circumvents computational bottlenecks and reduces the reliance on the exhaustive density functional theory (DFT) computations inherent in conventional training-set construction. The PhyMLP employs an adaptive, physics-guided sampling strategy that leverages intrinsic material responses and requires only a limited number of critical DFT calculations to efficiently characterize the potential energy surface. Using body-centered cubic tungsten as a benchmark system, the moment tensor potential trained on the PhyMLP-generated dataset exhibited exceptional predictive accuracy across a wide spectrum of material properties, ranging from fundamental physical constants to complex defect energetics and kinetic behavior. Ultimately, this study established a systematic and computationally efficient paradigm for developing accurate transferable MLPs, thereby offering a robust framework for large-scale atomistic simulations and advanced material modeling.

Keywords: Machine-learning potentials, multi-source data fusion, automated dataset construction, tungsten, molecular dynamic

INTRODUCTION

Driven by a paradigm shift in physics and materials science from empirical trial-and-error to data-driven design, the advent of machine-learning potentials (MLPs) has profoundly accelerated the timescale and fidelity of atomistic simulations[1–3]. Traditional empirical models, such as the Finnis-Sinclair[4] and embedded atom method[5] potentials, along with their modified variants[6], are inherently constrained by fixed mathematical forms, which restrict their ability to capture complex material behaviors. In contrast, MLPs bypass these limitations by bridging the quantum-mechanical accuracy of density functional theory (DFT) with the linear-scaling efficiency of empirical models, thereby achieving exceptional predictive power and transferability across diverse structural environments[7–10].

Driven by the demand for customized material design, machine-learning techniques are increasingly being deployed to elucidate the implicit relationships between atomic structures encoded via diverse descriptors and target properties[11–13]. Because the configuration space required to map the potential energy surface (PES) scales with the structural and compositional complexities, the accuracy and transferability of MLPs are fundamentally governed by the fidelity, completeness, and physical representativeness of their training datasets[14,15]. However, conventional dataset construction typically relies on intuitive estimates or exhaustive DFT calculations, which are computationally prohibitive and require extensive manual intervention. This ad hoc approach introduces substantial inefficiency, lacks systematic rigor, and frequently omits critical regions of the PES, thereby creating a data bottleneck that severely impedes the deployment of high-performance potential[16–19].

To alleviate these constraints, several automated frameworks have been developed to streamline MLP training-set generation, including DP-GEN for deep potentials and AutoPLEX for Gaussian approximation potentials (GAPs)[20,21]. For instance, the DeePKS method bridges low-fidelity DFT functionals with neural network corrections to mimic high-fidelity first principles results at a reduced computational cost[22]. Concurrently, the NepTrain tool automates dataset generation for neuroevolution potential models[23], whereas complementary active learning schemes have been designed to capture short-range chemical ordering in complex alloys and liquids[24]. Despite these advances, existing frameworks often suffer from suboptimal uncertainty sampling strategies and incur substantial computational overhead owing to the dense DFT validations required[25,26]. Among the various prominent MLP architectures, Moment tensor potential (MTP)[27] has emerged as an exceptionally efficient descriptor-based framework. While both the GAP and MTP models achieve top-tier accuracy, rendering the lowest mean absolute error (MAE) for energies and forces, GAP models scale poorly in terms of the computational cost, whereas MTP offers an optimal Pareto frontier between quantum-level accuracy and computational efficiency. Nevertheless, the development of robust MTPs still relies heavily on exhaustive DFT sampling to construct the initial training sets. To date, no comprehensive, low-cost, and fully automated framework tailored for MTP training-set construction has been developed[28-30].

MLPs strike a balance between accuracy and computational efficiency in atomistic simulations by fitting reference data derived DFT calculations. MLPs are widely applied in structural relaxation, molecular dynamics, and defect evolution studies because they map the local atomic environment within a specified cutoff radius to physical quantities such as energies and forces. The quality of the training dataset directly dictates the generalization capability of the MLP. Specifically, the data must comprehensively sample critical regions of the PES; otherwise, the model is prone to extrapolation failures.

The MTP[27] atomic energy is expressed as follows:

$$V(\mathbf{n_i}) = \sum_{\alpha} c_{\alpha} B_{\alpha}(\mathbf{n_i}), \tag{1}$$

where the basis functions, B_{α}, are constructed using moment tensors as follows:

$$M_{\mu,\nu}(\mathbf{n_i}) = \sum_{j} f_{\mu,\nu}(r_{ij}), \mathbf{r}_{ij}^{\otimes \nu}, \tag{2}$$

where $f_{\mu,\nu}(r)$ is a radial function that ensures the smoothness and cutoff behavior. These basis functions are invariant to the rotations, reflections, and permutations of atoms, making them physically meaningful descriptors of atomic environments. Linear coefficients, c_{α}, are obtained from the reference data via regularized regression.

These descriptors naturally satisfy the rotational, reflectional, and permutational invariance requirements and are computationally efficient. This study validated the proposed dataset generation method, physics-strengthened machine-learning potential (PhyMLP), using the MTP framework.

To address this fundamental challenge, we developed an automated dataset construction framework that introduces an adaptive sampling strategy achieved through the seamless integration of the Rose equation of state, experimental pressure-volume ($P - V$) data, and PES analysis results [31]. Initially, a high-fidelity equation of state (EOS) curve was constructed using experimentally measured $P - V$ data in conjunction with the equilibrium volume parameters derived from first-principles calculations. This approach effectively synergized the physical authenticity of the experimental measurements with the quantum mechanical accuracy of the first-principles computations [32-35]. Within this framework, the nonlinear and sensitive regimes of the PES - including the bond-breaking configurations and lattice instabilities - were identified to efficiently characterize the PES and ensure accurate potential fitting. This heterogeneous data fusion strategy has three pivotal advantages. First, integrating experimental $P - V$ data substantially enhances the physical fidelity of the training set near equilibrium states, ensuring the reliability of the resulting MLP under standard conditions. Second, by leveraging the extrapolation capability of the EOS, only a sparse set of critical configurations requires high-fidelity DFT calculations to efficiently characterize the PES, thereby substantially reducing the computational overhead. Finally, the framework enables targeted sampling in critical nonlinear regimes, effectively capturing the intrinsic PES features that are frequently overlooked by traditional methods.

A notable aspect of this methodology is that the incorporation of experimental data alleviates the exclusive reliance on computationally expensive DFT evaluations by leveraging empirically validated physical laws [36,37]. This integration not only renders the sampling process highly targeted but also profoundly enhances the physical credibility of the resulting datasets. By seamlessly embedding the experimental data into an automated workflow and extending its physical reach via the Rose EOS, this methodology establishes a robust paradigm for the efficient and accurate training of MLPs. While active learning represents a prevalent contemporary approach - iteratively querying DFT calculations based on model uncertainty to explore the PES - it remains a highly computationally intensive strategy. In contrast, PhyMLP adopts a physics-guided data-fusion paradigm that efficiently characterizes the PES *a priori* by integrating established physical knowledge: the Rose EOS (calibrated with sparse DFT data), experimental $P - V$ measurements, and MD snapshots from empirical potentials. Emphasizing data efficiency through physical priors, this approach is intended to complement existing active learning frameworks rather than replace them. Ultimately, PhyMLP provides a highly efficient and cost-effective solution for generating accurate, transferable MLPs, thereby paving the way for large-scale MD simulations, precise thermodynamic calculations, defect behavior analysis, phase transition studies, and the atomistic design of advanced materials.

MATERIALS AND METHODS

The proposed PhyMLP framework delineates an efficient approach to construct training datasets for machine-learned interatomic potentials via a heterogeneous data fusion strategy. As illustrated in Fig. 1, the workflow integrates three primary data categories. First, by leveraging the Rose EOS, the key parameters derived from a sparse set of first-principles calculations facilitate the large-scale generation of energy–volume datasets, encompassing elemental polymorphs, defective structures, and alloy configurations. Second, experimentally measured $P - V$ relationships are incorporated, calibrated, and refitted against equilibrium volumes derived from first-principles calculations, thereby generating the corresponding stress-strain configurational data. Third, validated empirical interatomic potentials are deployed to rapidly sample representative atomic configurations - encompassing perfect crystals, defective systems, and liquid states - across broad temperature and pressure ranges. Subsequently, these sampled configurations undergo single-point first-principles evaluations to extract the high-fidelity energies, atomic forces, and virial stresses. Finally, the essential configurations, including isolated atoms, dimers, and lattice-perturbed structures, are supplemented via additional DFT calculations. Collectively, these components yield a high-fidelity dataset that offers broad phase-space coverage and strong physical representativeness, drastically reducing the computational cost required to efficiently characterize the PES.

Machine-learned interatomic potentials trained on this dataset require systematic multiscale validation to rigorously assess their reliability and extrapolative capabilities. Consequently, the comprehensive validation framework covers a wide spectrum of material properties, ranging from fundamental equilibrium characteristics to complex responses under extreme conditions. These include lattice and elastic constants, the formation energetics, the kinetic evolution of structural defects (e.g., vacancies, interstitials, dislocations, and grain boundaries), lattice dynamic and

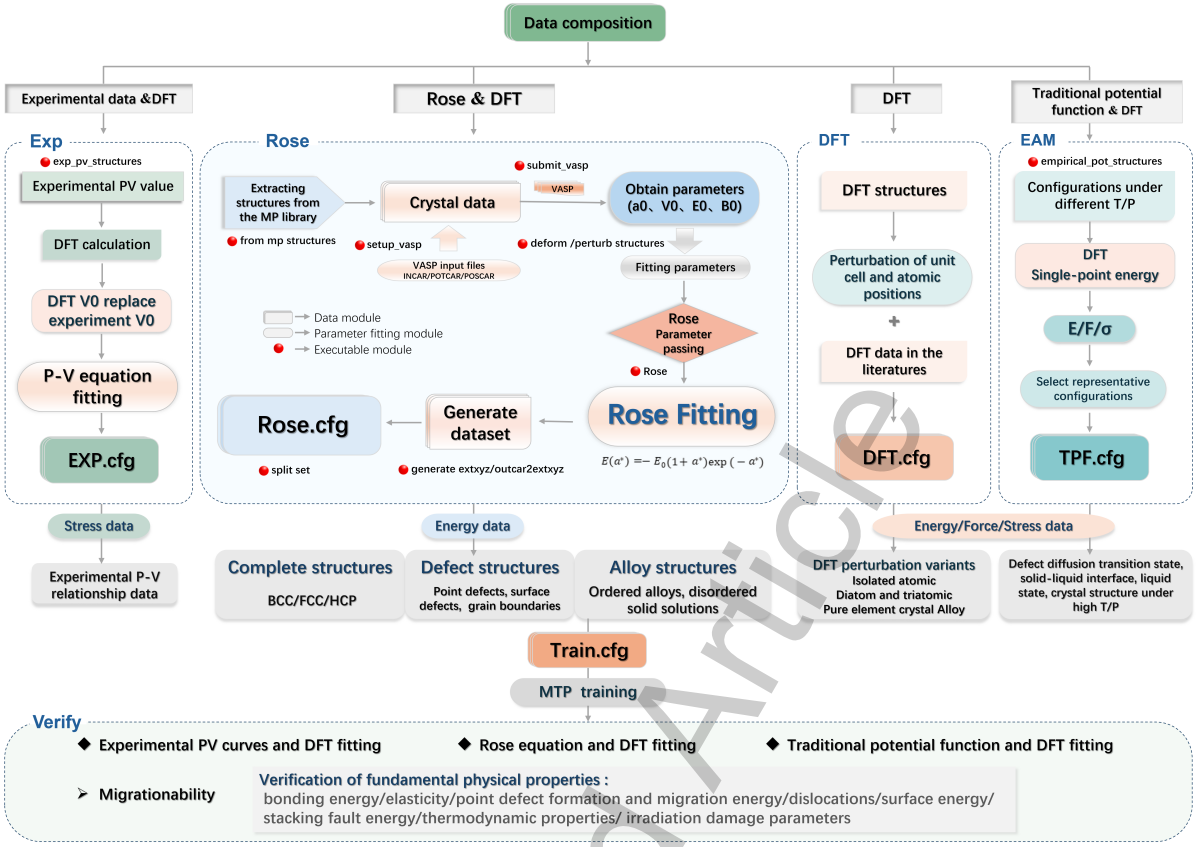

Figure 1. The workflow of PhyMLP automated sampling starts with file preparation and is divided into four sections: experiment, Rose, DFT, and traditional potential function. In this diagram, E represents energy, F represents force, and σ represents stress. PhyMLP: Physics-strengthened machine-learning potential; DFT: density functional theory; P-V: pressure-volume; VASP: Vienna Ab initio Simulation Package; BCC: body-centered cubic; FCC: face-centered cubic; HCP: hexagonal close-packed; EAM: Embedded-Atom Method.

thermodynamic properties (e.g., phonon dispersion spectra and thermal expansion coefficients), and high-energy or phase-transition phenomena, including melting behaviors and displacement threshold energies. Comprehensive benchmarking against first-principles results and experimental measurements demonstrated that potentials trained via the PhyMLP strategy not only reproduced near-equilibrium properties with high fidelity, but also exhibited exceptional predictive accuracy across extensive configurational spaces and extreme physical regimes. This rigorous validation established a robust foundation for deploying this potential in subsequent large-scale high-fidelity material simulations.

Rose-generated training set

Rose et al. conducted an in-depth investigation of the essence of metallic bonding [31]. Their work predicted that the binding energy as a function of interatomic distance can be quantitatively described using a two-parameter scaling formulation of a universal function in conjunction with established values for the equilibrium specific volume in atomic units. They found that although the binding-energy curves of different metals have various shapes, they can collapse onto a universal curve after appropriate scaling transformations. This suggests that a universal physical essence exists for the binding energy of metals. The following basic assumptions are made.

1. The binding energy of metals mainly depends on the electron density distribution.
2. The shape of the binding energy curve is determined by a single characteristic length scale.
3. The energy variation near the equilibrium point has a universal form.

The scaling variable is defined as follows:

$$a^* = \frac{a/a_0 - 1}{\alpha}, \tag{3}$$

where α denotes the determined scaling factor.

Characteristic length l is determined by the bulk modulus as follows:

$$l = \left(\frac{\Delta E}{12\pi B r_{\mathrm{WSF}}} \right)^{1/2}, \tag{4}$$

where ΔE is the cohesive energy (positive value), B is the bulk modulus, and r_{WSF} is the equilibrium Wigner-Seitz radius.

Combining this with the previous result, we obtain the following:

$$a^* = \frac{a/a_0 - 1}{\left(\frac{\Delta E}{9BV_0} \right)^{1/2}}. \tag{5}$$

Based on physical considerations and numerical fitting, the following exponential decay form can be selected:

$$E^*(a^*) = -f(a^*)e^{-a^*}, \tag{6}$$

where $f(a^*)$ is a polynomial function satisfying the boundary conditions.

Fitting the thermal expansion data yields $c_3 = 0.05$.

$$f(a^*) = 1 + a^* + 0.05a^{*3} \tag{7}$$

In a simplified version, neglecting the cubic term results in the following:[31]

$$E(a^*) = -E_0(1 + a^*)e^{-a^*}. \tag{8}$$

This approach was seamlessly integrated into the overarching data generation framework as follows. First, the target crystal structures - encompassing pure elemental polymorphs, defective configurations, and alloy systems - were subjected to structural optimization and elastic constant evaluations via first-principles methods, which yielded four key fundamental parameters with high accuracy: equilibrium lattice constant a_0, atomic volume V_0, cohesive energy per atom E_0, and bulk modulus B_0. These parameters were subsequently substituted into the universal $E - V$ relation governed by the Rose equation, enabling the rapid numerical evaluation and high-throughput generation of system energies across a broad spectrum of volumes (or lattice constants) for each configuration.

A primary advantage of this strategy lies in its ”single-calibration, high-throughput generation” paradigm. By requiring only a minimal set of first-principles calculations to ascertain the Rose equation parameters for each structural prototype, the resulting parameterization facilitates the efficient, low-cost generation of large-scale energy datasets spanning extensive volumetric regimes. This enables the construction of foundational training datasets that capture diverse crystalline environments, thereby efficiently characterizing the PES and substantially accelerating the development of robust MLPs.

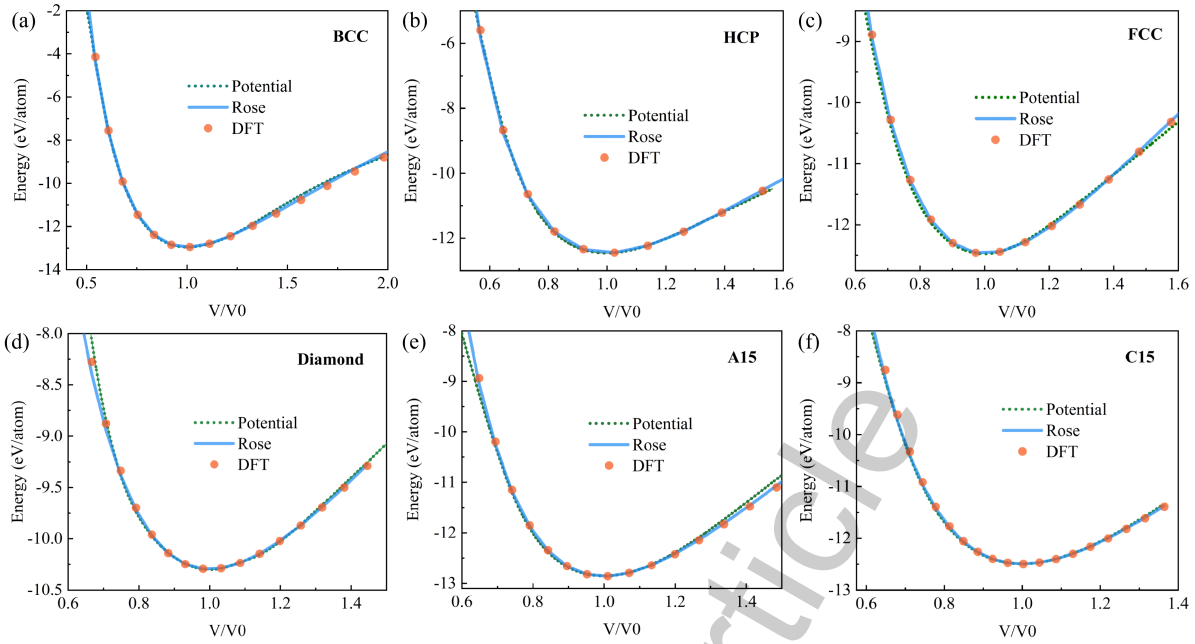

Figure 2. Comparison of the energy-volume ($E - V$) relationships for various pure tungsten polymorphs calculated using the Rose EOS, first-principles DFT calculations, and trained MLP (called "Potential" in the figure). The horizontal axes represent the relative volume (V/V0), and the vertical axes represent the energy per atom (eV/atom), illustrating the structural stability under isotropic deformation. (a) Body-centered cubic (BCC) ground-state structure. (b) Hexagonal close-packed (HCP) structure. (c) Face-centered cubic (FCC) structure. (d) Diamond structure. (e) A15 structure. (f) C15 structure.

To validate the feasibility of employing the Rose equation for the initial training data generation, systematic benchmarking was performed. Fig. 2 presents the results of a comparative analysis of the energies for various tungsten configurations evaluated using the Rose EOS, first-principles methods, and conventional empirical potentials. Furthermore, Fig. 3 illustrates the results of a detailed comparison of pure tungsten ground-state body-centered cubic (BCC) structures - encompassing mono-vacancies, di-vacancies, interstitials, free surfaces, and grain boundaries - which were calculated using the Rose EOS and DFT. These results demonstrated that the Rose EOS predictions were in excellent agreement with the first-principles calculations. While retaining near-DFT accuracy, its analytical nature ensures that the computational efficiency drastically outpaces explicit first-principles methods, thereby providing a highly robust and instantaneous alternative. This confirms that the Rose EOS can reliably and efficiently characterize the PES across diverse structural configurations, establishing it as a highly feasible and superior baseline for constructing high-fidelity MLP training sets.

Extension of the rose equation to complex multi-component systems

To demonstrate the universality and robustness of the PhyMLP data fusion strategy beyond simple refractory metals, the Rose equation methodology was systematically extended to complex multi-principal element alloys. Our tests covered the following systems.

W-Re Alloy Systems: The W-Re binary system includes W-Re disordered BCC solid solutions, B2 virtual ordered phases, and precipitated σ- and χ-phase structures.

Unary Metals (Ti, Al, V, Cr, and Zr): These are pure elemental configurations categorized by their ground-state crystal structures, including Al (FCC), Cr (BCC), Ti (HCP), V (BCC), and Zr (HCP). For each unary metal, the validation encompassed pristine bulk phases, various surface facets, and grain boundaries.

Binary Alloys: These include a diverse range of binary intermetallics and solid solutions across multiple stoichiometric ratios, specifically covering Al-Cr (Al_3Cr, Al_8Cr_5, $Al_{45}Cr_7$, $AlCr_2$), Al-V (Al_3V, Al_8V_5, $Al_{10}V$, $Al_{23}V_4$), Ti-Al (Ti_2Al, Ti_3Al, Ti_5Al_{11}, TiAl), Ti-Cr (Ti_4Cr, $TiCr_2$), Ti-V (TiV), V-Cr (V_3Cr, VCr), Zr-Al (Zr_2Al, Zr_2Al_3, Zr_4Al_3, Zr_5Al_3), Zr-Cr ($ZrCr_2$), Zr-Ti (ZrTi, $ZrTi_2$), and Zr-V (Zr_3V, ZrV_2).

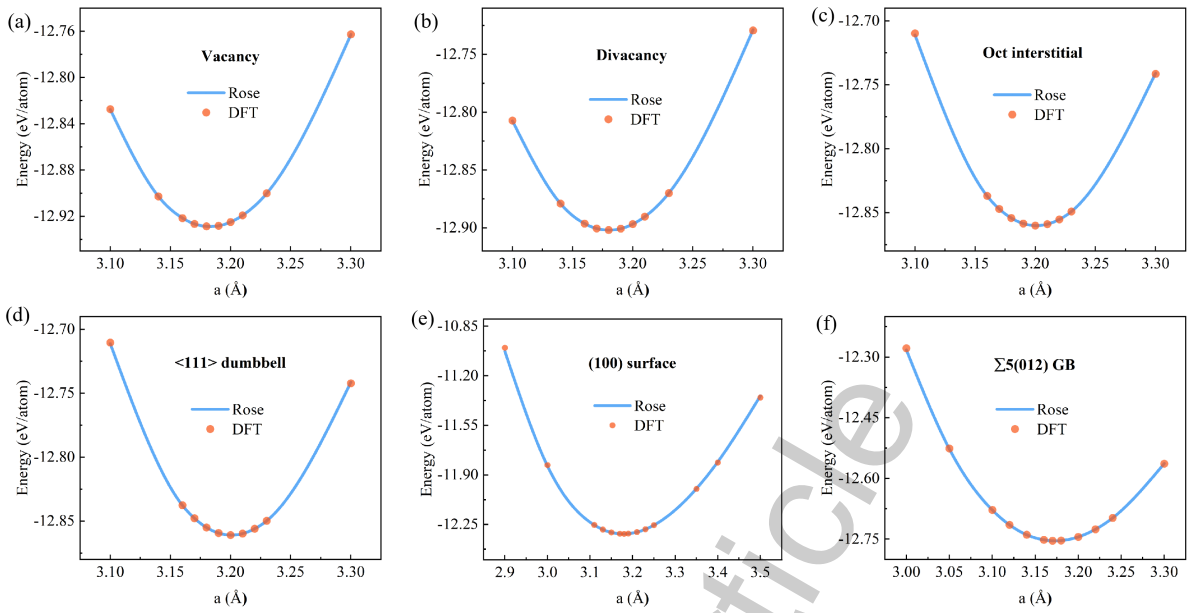

Figure 3. Validation of the Rose equation predictions against first-principles DFT calculations for various defect structures in the BCC tungsten lattice. The horizontal axes denote lattice parameter a (Å), and the vertical axes denote the structural energy (eV/atom), demonstrating the accuracy of the method in capturing local energetic perturbations caused by defects. (a) Single vacancy configuration. (b) Divacancy configuration. (c) Octahedral interstitial atom. (d) $< 111 >$ dumbbell interstitial structure. (e) (100) surface structure. (f) Σ5(012) grain boundary (GB) structure.

Ternary Alloys: These are complex three-component atomic arrangements and structures, including Ti-Al-Cr (multiple Ti_2AlCr geometric configurations), Ti-Al-V (Ti_2AlV, $TiAl_2V$), and Zr-Ti-Al (Zr_2TiAl, $ZrTi_2Al$).

$Ni_{55-x}Cr_{20}Fe_{20}Al_5P_x$ Quinary Alloy: The quinary alloy is a random solid solution, and the occupational randomness is accounted for by generating five quasi-random structures using the sqs-generator (SQS) software[38]. The E–V data for these five structures are then fitted.

The efficacy of this framework was first demonstrated in binary alloy systems, specifically in the W–Re system. As illustrated in Fig. 4, the $E - V$ profiles of W–Re disordered solid solutions, B2 virtual ordered phases, and intermetallic precipitated σ- and χ-phases evaluated via the Rose EOS were benchmarked against explicit DFT calculations. The comparison revealed a remarkable consensus between the analytical Rose EOS predictions and first-principles data, thereby substantiating the fidelity of this approach in describing multi-component alloy thermodynamics.

The Rose EOS methodology was systematically extended to complex multi-principal element alloys (MPEAs) to establish the universality and robustness of the PhyMLP data fusion paradigm beyond elemental systems. Such refractory MPEAs represent the vanguard of material design for next-generation aerospace and nuclear engineering applications. Given their exceptional high-temperature specific strength and resilience in extreme environments, efficient characterization of their PESs is paramount for advancing structural components in space propulsion systems and advanced nuclear reactors. Accordingly, the $E - V$ scaling relationships for a comprehensive suite of configurations within the Ti–Al–V–Zr–Cr system were evaluated. As detailed in the Supplementary Materials (Figs. S1–S19), the energies analytically predicted by the Rose EOS exhibited excellent agreement with the reference DFT calculations across all multi-component configurations. Furthermore, the E–V characteristics of the $Ni_{55-x}Cr_{20}Fe_{20}Al_5P_x$ quinary alloy were investigated (Fig. S20). This alloy possesses highly intricate chemical environments—comprising multiple transition metals alongside the nonmetal P - and adopts a face-centered cubic (FCC) lattice that induces intense many-body interactions spanning the first and second coordination shells. Thus, it posed a stringent test for the extrapolative robustness of the Rose method.

Across all evaluated multi-component configurations, the synergy between the Rose EOS and sparse DFT sampling

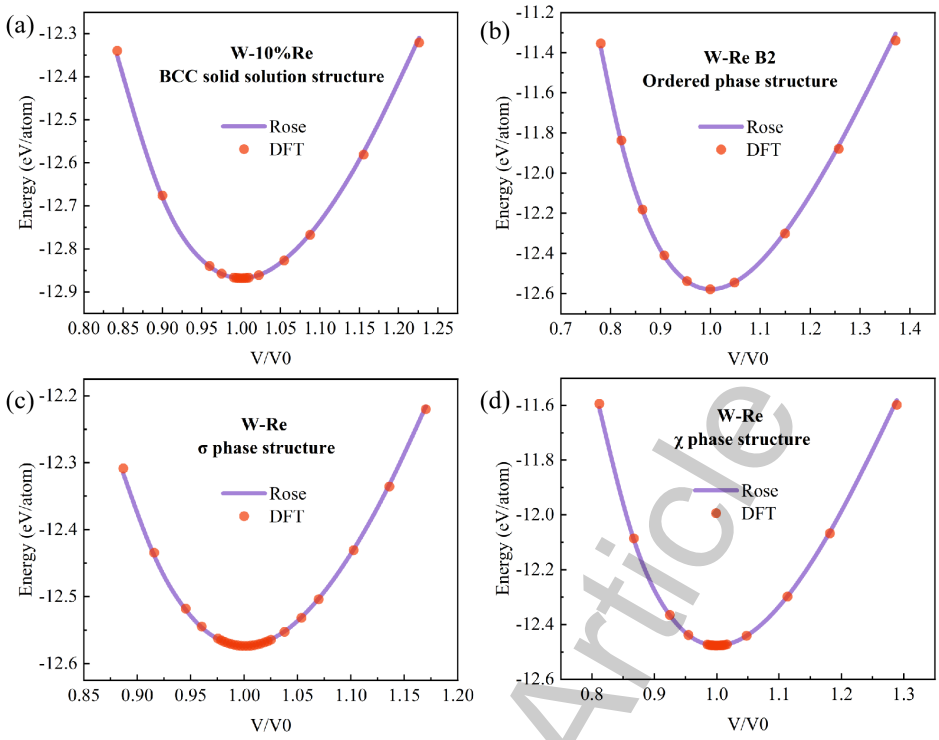

Figure 4. Comparison of the $E - V$ scaling relationships for binary W-Re alloy configurations calculated using the Rose equation and reference DFT methods. The horizontal axes show the relative volume (V/V_0), and the vertical axes represent the energy (eV/atom), confirming the generalizability of the Rose equation to multi-component systems. (a) W-10 %Re disordered BCC solid solution structure. (b) W-Re B2 virtual ordered phase structure. (c) W-Re precipitated σ-phase structure. (d) W-Re precipitated χ-phase structure.

established a highly generalized paradigm, successfully overcoming the computational bottlenecks inherent in traditional dataset construction for complex modern alloys.

Experimentally derived training set

To further fortify the training dataset, this multi-source data fusion strategy seamlessly integrated the experimentally measured $P - V$ relationships with the data generated using the empirical Rose equation. Mechanistically, while preserving the experimental pressure and relative volume ratio (V/V_0), the experimentally determined equilibrium volume was replaced by a high-fidelity V_0 derived from first-principles calculations. The modified $P - V$ relationship was subsequently fitted to a polynomial equation, facilitating the rapid, high-throughput generation of configurational data mapping pressure variations across diverse volumetric states (as delineated in Fig. S21). This approach drastically amplified the data generation efficiency, enabling the rapid exploration of expansive configurational spaces while substantially curtailing the computational overhead associated with exhaustive DFT sampling. By anchoring the dataset to empirical experimental data, the physical credibility and numerical accuracy of the training set were significantly increased, thereby enhancing the predictive reliability of the resulting MLP under extreme thermodynamic conditions.

Crucially, the experimental and DFT data sampled distinct non-overlapping regimes of the configurational space, ensuring complementary integration and precluding the risk of internal inconsistencies. The DFT sampling was primarily concentrated near the equilibrium volume, providing a high-fidelity quantum mechanical baseline to efficiently characterize the PES. Conversely, the experimental $P - V$ data predominantly populated the high-pressure regime - a domain in which conventional DFT calculations become computationally prohibitive and highly sensitive to functional approximations. Bridged by the continuous $E - V$ trajectories generated via the Rose equation, the experimental data seamlessly supplemented the high-pressure regimes, where the DFT coverage remained sparse. Even in the hypothetical scenario of phase-space overlap, both sources fundamentally described the same physical system; any minor discrepancies merely reflected the inherent methodological variances, rather than contradictions.

Within the comprehensive training set, the experimental configurations accounted for only 106 of the 8,542 structures. Their primary function was to extend the extrapolative validity of the model to extreme-pressure domains, rather than compete with the DFT data near equilibrium. This synergistic sampling ensured that the empirical information enriched, rather than obfuscated, the DFT-derived training signals.

It is imperative to clarify that incorporating experimental data was not intended to ”replace” DFT calculations, nor did it introduce additional generation costs. The ”cost” discussed herein refers strictly to the computational overhead - specifically, the thousands of CPU hours typically demanded by exhaustive DFT configuration sampling. In contrast, the literature-derived experimental data incurred zero computational costs; their acquisition required only a negligible, one-time effort for digitization and curation. Therefore, our assertion regarding ”bypassing expensive DFT calculations” pertains exclusively to circumventing the massive number of computational hours normally required by utilizing the Rose equation (anchored by sparse DFT calibrations) for high-throughput data generation, not replacing fundamental quantum mechanics with empirical measurements. While the experimental fusion module served as a powerful, optional augmentation leveraging the available literature, the core PhyMLP methodology - the synergy of Rose-equation-based generation and targeted DFT sampling - remains fully generalized and broadly applicable to any material system.

Considering pure tungsten as a representative sample, Fig. S22 presents a comparison between the $P-V$ relationship derived using this methodology (solid red curve) and direct experimental measurements (scatter points). This striking concordance validates the efficacy of the hybrid data generation strategy. Furthermore, for cubic crystal systems, because the normal stress components along the three principal axes are numerically equivalent to the hydrostatic pressure, this framework natively provides exact virial stress tensors for the corresponding configurations as high-fidelity training labels. This naturally enriches the diversity and physical completeness of the training dataset.

Empirical potential derived training set

To construct a machine-learned interatomic potential framework, a multiscale sampling strategy was deployed to systematically curate a highly representative training dataset. First, foundational atomic configurations—including pristine crystal structures spanning broad ranges of temperatures and pressures—were rapidly generated using validated empirical potentials. Subsequently, single-point first-principles calculations were performed on the sampled configurations to extract high-fidelity energies, atomic forces, and virial stress tensors. To efficiently characterize the PES, classical molecular dynamics (MD) simulations were executed across a wide range of thermodynamic conditions to sample the defective configurations and liquid states. Representative configuration snapshots extracted from these trajectories captured the intrinsic dynamic evolution of the system and served as inputs for subsequent first-principles evaluations. Compared with sampling approaches that rely exclusively on computationally expensive *ab initio* molecular dynamics (AIMD), this hybrid high-throughput paradigm drastically mitigates the computational overhead while simultaneously maximizing the phase-space diversity and physical completeness of the dataset. Consequently, the finalized training set comprehensively encompassed pivotal configurations encountered across extreme regimes—including high-temperature and high-pressure crystalline states, defect migration pathways, solid–liquid coexisting interfaces, and bulk liquid phases—thereby establishing a robust foundation that guaranteed the extrapolative reliability and transferability of the resulting MLP within the target material system.

Dataset visualization and training results

In this study, an integrated dataset comprising 8,542 tungsten structures was constructed using the PhyMLP automated dataset generation framework, and the subsequent validation constituted a rigorous multi-physics assessment. The dataset combined structures extracted from the Materials Project (MP) database, configurations generated using empirical potentials, experimentally fitted volume–pressure data, and $E-V$ curves derived from the Rose EOS. To assess the structural diversity, we employed t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction based on smooth overlap of atomic positions (SOAP) descriptors, where proximity in this space indicates higher feature similarity (Fig. 5). The visualization revealed a broad, elliptical distribution of points in a 2D projection, demonstrating that the training set encompassed diverse and complementary structural features, which were keys to enhancing the generalizability of the MLP for this elemental system.

The different data sources exhibited distinct distribution patterns. The Materials Project structures clustered near the center, suggesting that perturbing these could expand the feature space. Both the volume–pressure fits and Rose-equation data formed continuous trajectories, representing continuous deformation paths in stress and energy that extended the boundaries of the dataset without overlapping. Notably, the grain boundary and surface structures

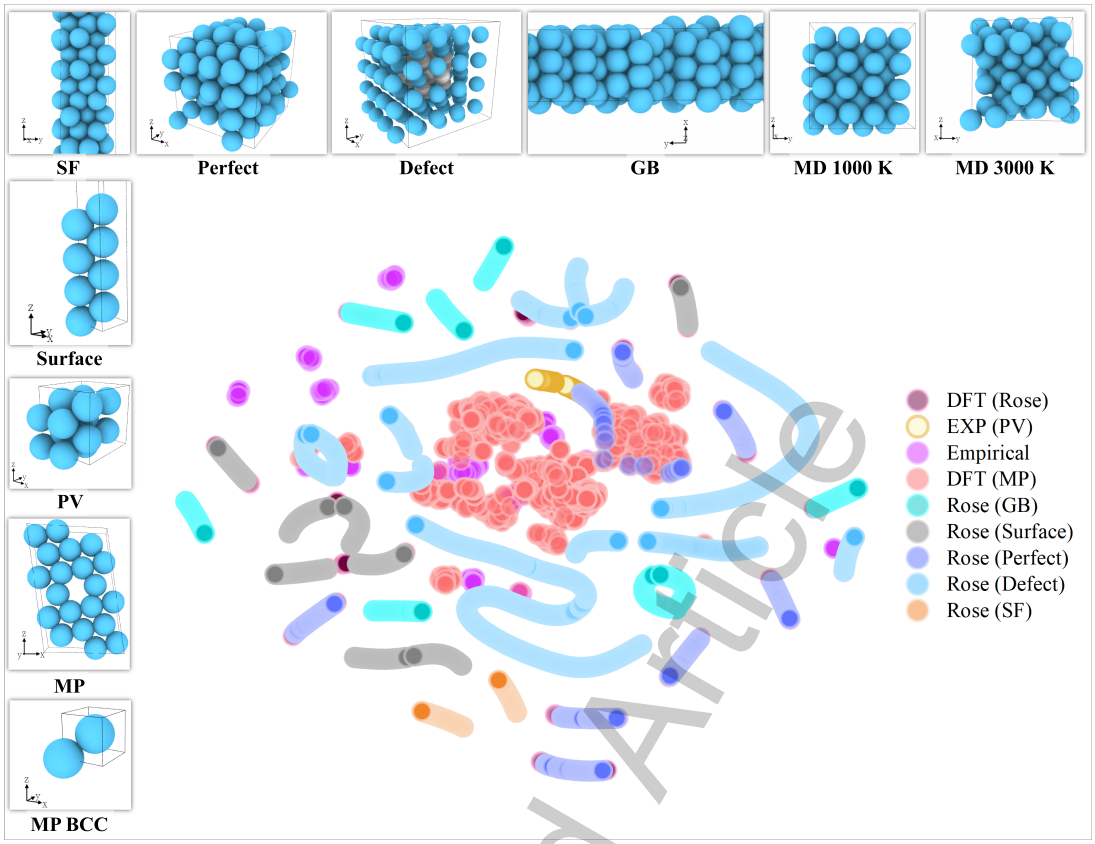

Figure 5. T-SNE projections of the W dataset based on the SOAP descriptor. t-SNE: t-Distributed stochastic neighbor embedding; SOAP: smooth overlap of atomic positions; SF: surface; GB: grain boundary; MD: molecular dynamics; PV: pressure-volume; MP: Materials Project; BCC: body-centered cubic; DFT: density functional theory.

obtained via the Rose method were clearly separated, confirming the nonredundant feature representations. Empirical potential data spanned the broadest region in the feature space owing to the time-resolved sampling acquired during extended MD relaxation, which efficiently captured temperature-dependent features while avoiding significant costly first-principles MD sampling. In summary, the constructed training set systematically encompassed equilibrium, perturbed, phase-transformed, and temperature-informed structures, providing a comprehensive and feature-rich foundation for developing highly transferable MLPs.

Table 1. Detailed breakdown of energy RMSE values by physical properties.

	Data Source	RMSE_E (meV/atom)
Physical Properties	Perfect and large-scale lattice scaling	20.1
	Surface and large-scale lattice scaling	39.0
	Grain_boundary	13.4
	Interstitial	1.2
	Stacking_fault	14.5
	Vacancy	2.8
	High-T-P MD	3.4
	Melts and liquids	7.6
	Elastic_phonon	3.0
	Dimer	50.8

In this study, we trained an automated dataset of 8542 tungsten structures using an MTP framework. The SOAP descriptor was constructed using the following parameters: the number of radial basis functions n_{max}=8, maximum degree of spherical harmonics l_{max}=4, cutoff radius r_{cut}=4.4 Å, and Gaussian smearing width σ=0.55 Å. Gaussian-type orbitals (GTO) were used as the radial basis functions, and periodic boundary conditions were considered. The

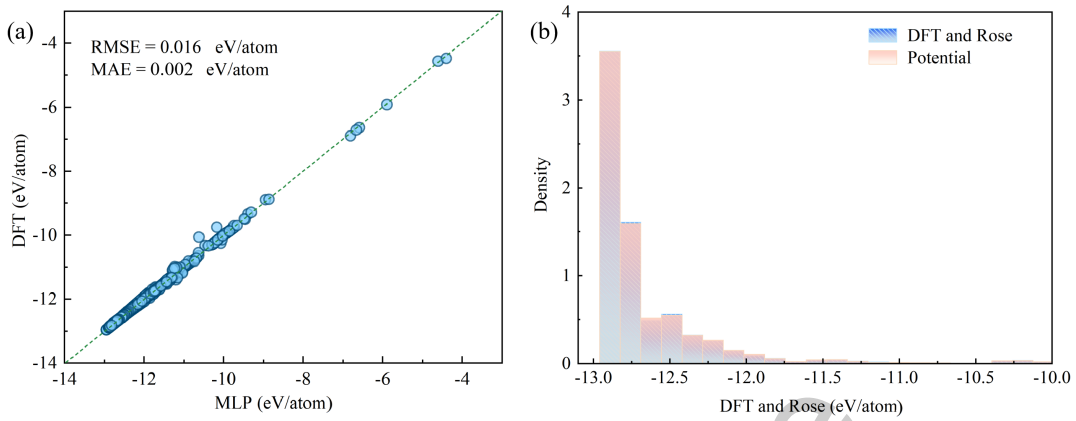

Figure 6. Quantitative validation of the energy predictions made by the trained MLP against the reference dataset (comprising DFT and Rose equation data). (a) Scatter plot comparing the MLP-predicted energies (horizontal axis) with the reference DFT calculations (vertical axis). The dashed diagonal line represents perfect agreement. The inset provides the quantitative prediction accuracy metrics, with a root-mean-square error (RMSE) of 0.016 eV/atom and MAE of 0.002 eV/atom. (b) Probability density distribution histogram of the structural energies in the dataset. The horizontal axis represents the energy range (eV/atom), and the vertical axis represents the density of the configurations, demonstrating that the energetic predictions of the MLP closely map the structural distribution of the original DFT and Rose reference data.

resulting feature vector dimensionality was $n_{\max}^2 * (l_{\max} + 1) = 320$.

The global energy RMSE of 16.3 meV/atom (Fig. 6) listed in Table 1 does not indicate a general deficiency in the model accuracy; rather, it represents an inherent and inevitable mathematical compromise arising from the multi-source data fusion strategy. Our model achieved an exceptionally high precision for near-equilibrium configurations, including perfect crystals, standard surfaces, and static point defects. To ensure comprehensive phase space coverage and transferability, our perfect-crystal dataset was not strictly limited to the stable BCC ground state; it explicitly incorporated multiple hypothetical crystalline phases—both metastable and high-energy—including FCC, HCP, simple cubic (SC), diamond, and A15 structures. Similarly, our surface dataset encompassed a diverse array of crystallographic orientations, ranging from low-index fundamental facets—such as (100), (110), and (111)—to complex, high-index stepped surfaces with inherently higher surface energies—including (112), (210), (211), (221), (310), (311), and (321). Critically, these configurations were not restricted to the conventional relaxed states. We deliberately incorporated large-scale lattice-scaling structures, subjecting diverse perfect crystals and various surface facets to extreme volumetric compression and tension. This deliberate scaling drove these nominally fundamental configurations far from equilibrium and into highly strained regimes, thereby capturing the steep short-range Pauli repulsion and near-dissociation limits. The primary reason for the apparent elevation of the overall error was the deliberate inclusion of extreme, high-energy, and non-equilibrium subsets in the training dataset—specifically dimers, isolated atoms, high-temperature liquids, and extremely high-pressure shock MD trajectories. These extreme configurations exhibited severe thermal fluctuations, chaotic bond-breaking dynamics, and strong short-range repulsions, resulting in highly rugged PESs and substantially elevated absolute energy baselines. Forcing a neural network to map both smooth, low-temperature quantum ground states and highly chaotic, high-temperature states (on the order of thousands of degrees Kelvin) within the same high-dimensional feature space inevitably yields larger absolute fitting deviations in high-variance regions. A prime example is the dimer subset, where interatomic distances are extensively sampled down to the 1.2–2.2 Å range. These extreme close-contact configurations are physically essential for capturing severe short-range Pauli repulsions; however, they inherently exhibit enormous energy magnitudes that disproportionately amplify the absolute fitting error. Therefore, the overall error of 16.3 meV/atom was mathematically dominated and skewed by these extreme high-energy components. This is not a flaw in the model architecture but rather a necessary physical trade-off—a dynamic global balance—required to ensure that the potential remains highly robust and free from catastrophic failure when simulating extreme service environments, such as fusion collision cascades and strong shock loading.

The absence of force and stress labels from certain subsets in the training dataset does not imply that the predicted forces or stresses are zero or unphysical. The loss function employed in MTP training explicitly incorporated contributions from energies, atomic forces, and stresses, with the availability of each label type depending on the

Table 2. Summary of dataset composition and labeling.

Dataset Component	Sampling Strategy & Structural Description	Number	Labels	Label Contribution
Rose Equation	**Adaptive point-sampling along E-V curves**	**5,700**	E	**Total energy only**
1. Pristine Crystals	Bulk structures under different volumes (BCC, FCC, HCP, etc.)	700		
2. Crystal Defects	Vacancies, interstitials, and defect clusters	3,200		
3. Grain Boundaries	Symmetric tilt grain boundaries ($\Sigma3, \Sigma5, \Sigma7, \Sigma9$)	800		
4. Surfaces & Stacking Faults	Low-index surfaces and generalized stacking faults	1,000		
Experimental P-V	**Macroscopic high-pressure EOS data**	**106**	S	**Box virial stress only**
Traditional Potentials	**Displacement/diSIA/diVAC/liquid/surf/VAC/SIA/temperature**	**384**	E, F, S	**Configurations from MD; labels (energy, forces, stress) from DFT**
DFT Calculations	**Direct first-principles and active learning datasets**	**2,352**	E, F, S	**Full labels (energy, force, stress)**
1. DFT_Rose	DFT single-point validation on Rose-derived configurations	492		
2. DFT Sampling	AIMD sampling for DFT	1,351		
3. Box Distorted	Lattice and box perturbed configurations	500		
4. Isolated Atom	Reference state for isolated atom energy	1		
5. Dimers	Short-range repulsion dimer configurations	8		
Total Dataset	**PhyMLP Data Fusion Total**	**8,542**	—	**E: 8,436; F: 2736; S: 2842; Total dataset: 8,542**

E-V: Energy-volume; BCC: body-centered cubic; FCC: face-centered cubic; HCP: hexagonal close-packed; P-V: pressure-volume; EOS: equation of state; VAC: vacancy; SIA: self-interstitial atom; DFT: density functional theory; AIMD: ab initio molecular dynamics; PhyMLP: physics-strengthened machine-learning potential.

specific data source. Structures generated via the Rose equation (5,700 configurations) provided only total energy labels, while those derived from experimental P–V fitting (106 configurations) provided only stress labels via the fitted EOS. Structures sampled from empirical-potential MD trajectories (384 configurations) were subsequently evaluated via single-point DFT calculations, yielding complete sets of energy, force, and stress labels; and directly perturbed DFT structures (2,352 configurations) provided full energy, force, and stress labels. The training set contains a total of 8,542 labels, of which 8,436 are energy labels, as listed in Table 2. For evaluation metrics beyond the energy labels, the force RMSE on the training set was 13.18 meV/Å, and the stress RMSE was 0.598 GPa. These quantitative indicators, together with the strong performance of the force-sensitive downstream properties—including the phonon spectra, defect-cluster energetics, thermal expansion, and displacement threshold energies—collectively confirmed that the MTP trained on the PhyMLP-generated dataset reliably captured both the energetic and mechanical responses of the target system.

RESULTS AND DISCUSSION

Using tungsten as a case study, this section details the fully automated PhyMLP workflow for generating moment-tensor potential training sets. We present the implementation steps and provide a comprehensive validation of the resulting potential. Tungsten is widely used in various extreme environments because of its exceptional physical properties[39]. As a refractory metal with a BCC crystal structure, tungsten exhibits strong metallic bonding, which results in an extraordinarily high melting point (3,422 °C) and remarkable high-temperature strength[40,41]. Additionally, its high Debye temperature and the efficient electron transport pathways in its pure crystalline lattice endow it with superior thermal conductivity at elevated temperatures[42-44]. In terms of its mechanical behavior, tungsten maintains a stable crystal structure even at elevated temperatures, exhibiting very low diffusion coefficients and self-diffusion rates, which explain its outstanding creep resistance and low vapor pressure[45,46]. From a nuclear engineering perspective, the high atomic number of tungsten (Z = 74) makes it effective at shielding against X-rays and gamma-rays. Its favorable neutronic characteristics and low mobility of irradiation-induced defects further ensure reliable performance in high-radiation environments[47-49].

Construction of tungsten training set

Data Generation via the Rose Equation: Static energy data covering multiple crystal structures and defect configurations were generated based on the Rose EOS. Specifically, this dataset comprises a total of 5,700 configurations, which provide comprehensive foundational $E - V$ mapping.

The dataset included the energies of pristine crystal structures across different volumes, encompassing seven typical phases: BCC, FCC, HCP, simple cubic (SC), diamond, A15, and C15. This section describes 700 configurations. The unit cells of these crystal structures were constructed and structurally optimized using the first-principles software VASP to obtain equilibrium lattice constant a_0, atomic volume V_0, and ground-state energy E_0. Following the structural optimization, the elastic constants of each system were calculated using VASP, from which bulk modulus B_0 was derived. These four parameters were then incorporated into the Rose EOS, and a self-developed Fortran program was executed to systematically compute the energy data for various volumes. The calculation process is illustrated in Fig. S23.

Crystal defect configurations were extensively sampled to describe localized distorted environments, contributing 3,200 configurations. This subset encompassed BCC structures containing various point defects and clusters, including mono-vacancies, di-vacancies, $\langle 111 \rangle$ dumbbells, $\langle 110 \rangle$ dumbbells, $\langle 100 \rangle$ dumbbells, $\langle 111 \rangle$ crowdions, octahedral interstitials, tetrahedral interstitials, and di-interstitial configurations.

Planar and interfacial defects were incorporated to handle the boundary and surface behaviors. These included symmetric tilt grain boundaries across multiple misorientation axes, such as the $\Sigma 3(110)$, $\Sigma 3(112)$, $\Sigma 5(012)$, $\Sigma 5(013)$, $\Sigma 7(132)$, and $\Sigma 9(221)$ configurations, yielding 800 structures. Furthermore, low-index surfaces—such as (100), (110), (111), (210), (211), (311), (320), and (321)—along with generalized stacking faults (e.g., on the (110) and (112) planes) were sampled to contribute another 1,000 configurations. The calculation process for the defect-containing surface and grain-boundary configurations is shown in Fig. S24.

The following implementation strategy was used to generate the Rose equation training data for defect-containing and interface structures. First, based on the constructed atomic configurations containing defects or interfaces, structural optimization was performed using VASP to obtain the equilibrium parameters, a_0, V_0, and E_0. Subsequently, box-length scaling deformation was applied along a specific crystallographic direction while preserving both the cell shape ratio and relative atomic positions. Subsequently, single-point energy calculations were performed on the deformed configurations using VASP, and bulk modulus B_0 was derived by fitting the corresponding E–V curves. These four parameters were substituted into the Rose EOS, and a custom-developed Fortran program was employed for batch generation of the energy data over a range of volumes. This approach required only a limited number of first-principles calculations to determine the key Rose equation parameters, thereby enabling efficient large-scale data generation and substantially enhancing both the phase-space coverage and overall construction efficiency of the training dataset. Furthermore, for alloy structures—including disordered solid solutions and ordered phases composed of different elements within the target potential framework—the $E - V$ relationship can be similarly generated via the Rose equation, as illustrated in Fig. S25.

Data Generation from Experiments and DFT Calculations: To enhance the ability of the potential function to describe continuous deformation, active learning evolutions, and extreme electronic structure conditions, high-pressure experimental EOS data were incorporated, as well as an extensive set of fundamental quantum mechanical calculations.

The experiment incorporated macroscopic high-pressure P–V relationship data, contributing exactly 106 configurations. These configurations represented macroscopic compressed or expanded box states, generating specialized box virial stress (S) labels derived directly from experimental EOS parameters to constrain the long-range behavior under shock conditions.

The quantum-mechanical dataset evaluated directly via first-principles methods comprised 2,352 configurations in total, which were systematically classified into five sub-categories according to their structural roles. (1) DFT-Rose configurations: A subset of 492 configurations selected from the Rose-equation scaling datasets were evaluated via single-point DFT calculations to serve as a direct cross-validation baseline between the analytical Rose equation and first-principles energy surfaces. (2) DFT-sampled configurations: These were configurations obtained from active learning iterations and data-driven sampling, and included 1,351 structures. These snapshots were iteratively

selected during active learning cycles, and generalized automated searches were performed to effectively map the low-energy valleys and high-energy barriers of the PES. (3) Box-distorted structures: A set of 500 configurations featuring randomly perturbed lattice constants and severely distorted box vectors were designed to map a broad range of elastic and shear deformation tensors. (4) Isolated atom: One configuration represented a single isolated atom in a large vacuum box, providing an absolute reference baseline for the atomic energy. (5) Dimers: Eight configurations captured short-range dimer interactions, in which the interatomic distance between two tungsten atoms ranged from 1.2 to 2.2 Å, enabling the MLP to accurately handle hard-core repulsions encountered during high-energy irradiation cascades[32].

Data Generation Using Traditional Interatomic Potentials: Atomic-scale data at finite temperatures and in complex thermodynamic phases were obtained by combining traditional empirical potentials with first-principles calculations. MD simulations were performed using the reported empirical tungsten potential[50].

Using this traditional interatomic potential sampling scheme, 384 representative dynamic snapshots were extracted from various thermodynamic ensembles. This trajectory-derived dataset included MD snapshots of perfect bulk BCC crystals operating under wide ranges of finite temperatures and pressures, nonequilibrium configuration pathways containing point defects such as vacancies and self-interstitial atoms (SIAs) under extreme thermal fluctuations, and highly disordered snapshots representing solid-liquid coexistence interfaces and pure liquid melts.

First-principles DFT calculations were subsequently applied to these 384 representative configurations via single-point evaluations to extract accurate quantum mechanical energy and atomic force information. This strategic integration effectively extended the coverage of the training set to complex dynamic pathways and non-equilibrium fluid states, ensuring high fidelity during phase transition simulations.

All of the first-principles calculations employed consistent parameter settings, including a plane-wave cutoff energy of 500 eV, a KSPACING of 0.15, and an energy convergence criterion of 10^{-5} eV. The interaction between ionic cores and valence electrons was described using the projector augmented wave (PAW) method, and the exchange-correlation functional was treated within the Perdew-Burke-Ernzerhof (PBE) formulation of the generalized gradient approximation (GGA). The W_sv pseudopotential (version 04Sep2015), which includes semi-core (5p) states, was employed to accurately describe the electronic structure of W atoms. Finally, using the multi-component training set constructed above, an MTP was trained with a selected model level of lev18, successfully yielding a high-accuracy MLP for the tungsten system.

Verification of MLP for tungsten

The trained tungsten potential was used for comprehensive validations of a wide range of fundamental and defect-related properties. These included the lattice constant, bulk modulus, elastic constants, formation and migration energies of point defects, binding energies of vacancies and interstitial clusters, relative stability of interstitial dislocation loops, surface energy, stacking fault energy, grain boundary energy, core structure of screw dislocations, phonon spectrum, thermal expansion behavior, melting point, and displacement threshold energy. The validations covered multiple critical scales ranging from lattice thermodynamics to defect evolution, demonstrating the accuracy and generalizability of the developed potential within this elemental system.

Table 3 compares the fundamental bulk and elastic properties predicted by the proposed potential with those obtained using existing mainstream MLPs (DP-HYB-22, GAP-19, and SNAP-23), DFT benchmarks, and experimental data. Despite being trained using a low-cost automated database construction scheme, the developed potential yielded predictions that closely aligned with first principles and experimental references. Notably, for shear elastic constant C_{44} - which is challenging to reproduce - the proposed potential achieved a remarkably close agreement with the experimental value (161 GPa), outperforming several computationally intensive potentials. Similarly, Table S1 presents a comparison of the calculated vacancy formation, vacancy migration, divacancy binding, and self-interstitial atom formation energies in various configurations with the reference DFT results. The strong consistency further confirms the reliability of the present potential for describing point-defect properties.

Vacancies and interstitial clusters, which are fundamental microstructural defects in metallic materials, have formation and binding energies that serve as critical physical parameters for assessing the defect stability and evaluating the performance of interatomic potentials in complex defect environments[58-60]. As shown in Fig. 7, this study systematically compared the formation and binding energies of vacancy clusters, $< 111 >$ interstitial clusters, and

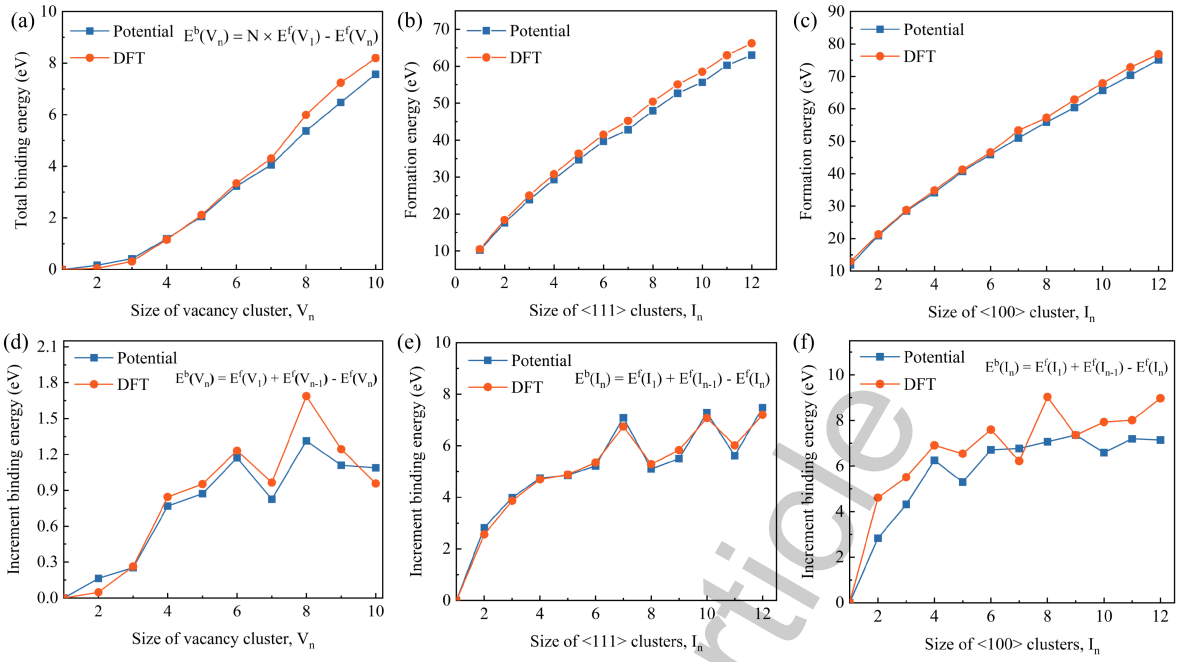

Figure 7. Comparison of formation and binding energies for point-defect clusters between the machine-learned potential and DFT calculations. (a) Total and (d) per-vacancy binding energies of vacancy clusters. (b) Formation and (e) binding energies of $< 111 >$ interstitial clusters. (c) Formation and (f) binding energies of $< 100 >$ interstitial clusters.

$< 100 >$ interstitial clusters predicted by the machine-learned tungsten potential against the first-principles reference data. Figs. 7a and d present the total and per-vacancy binding energies of the vacancy clusters, respectively, as a function of the cluster size. The results demonstrated that the machine-learned potential was in good agreement with DFT data across all cluster sizes considered, accurately reproducing the energetic evolution trend of vacancy clusters. Figs. 7b and e show the predicted formation and binding energies of $< 111 >$ interstitial clusters, respectively. For all the cluster sizes investigated, the machine-learned potential exhibited close agreement with the first-principles data, demonstrating its capacity to accurately describe the energetics of $< 111 >$ interstitial configurations. Figs. 7c and f display the formation and binding energies of $< 100 >$ interstitial clusters, respectively. The computational results confirmed that the potential successfully reproduced the size-dependent energy variation of $< 100 >$ interstitial clusters, further validating its broad applicability across various self-interstitial atomic configurations. In summary, the machine-learned potential demonstrated good predictive performances for all types of point-defect clusters, providing a reliable computational foundation for the multi-scale modeling of vacancy- and interstitial-related defects.

An accurate description of the dislocations, surface energies, and stacking faults is essential for reliably predicting the mechanical response and thermodynamic stability of materials[61–63]. Validations of the machine-learned tungsten potential for extended defects are shown in Fig. 8. Fig. 8a shows the relationship between the formation energy and size of the $1/2< 111 >$ and $< 100 >$ interstitial dislocation loops. The formation energy of the $1/2< 111 >$ loops is systematically lower than that of the $< 100 >$ loops, indicating that the $1/2< 111 >$ configuration is more stable at

Table 3. Comparison of lattice constants, cohesive energies, bulk moduli, and elastic constants calculated using the tungsten potential function (bold) and other MLPs, DFT, and experimental results.

Property	Potential (This work)	DP-HYB-22[51]	GAP-19[52]	SNAP-23[53–55]	DFT (This work)	DFT[53]	Exp.[56,57]
a (Å)	**3.184**	3.172	3.185	3.182	3.185	3.185	3.165
E_c (eV/atom)	**8.34**	8.47	8.39	8.91	8.39	8.39	8.9
B (GPa)	**301**	312	304	283	304	304	310
C_{11} (GPa)	**506**	543	526	405	507	522	522
C_{12} (GPa)	**198**	203	200	231	202	195	204
C_{44} (GPa)	**161**	141	149	126	155	148	161

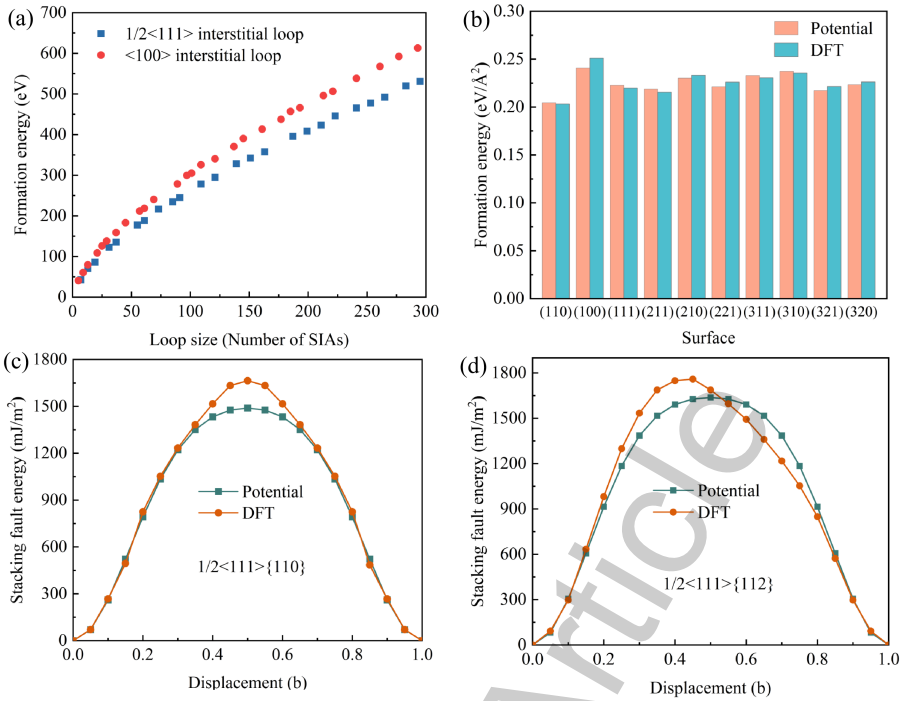

Figure 8. Validation of the machine-learned tungsten potential for extended defects. (a) Formation energies of 1/2< 111 > and < 100 > interstitial dislocation loops as a function of size. (b) Surface energies for ten different crystallographic planes compared with DFT references. Generalized stacking fault energy curves for (c)1/2< 111 >110 and (d)1/2< 111 >112 slip systems.

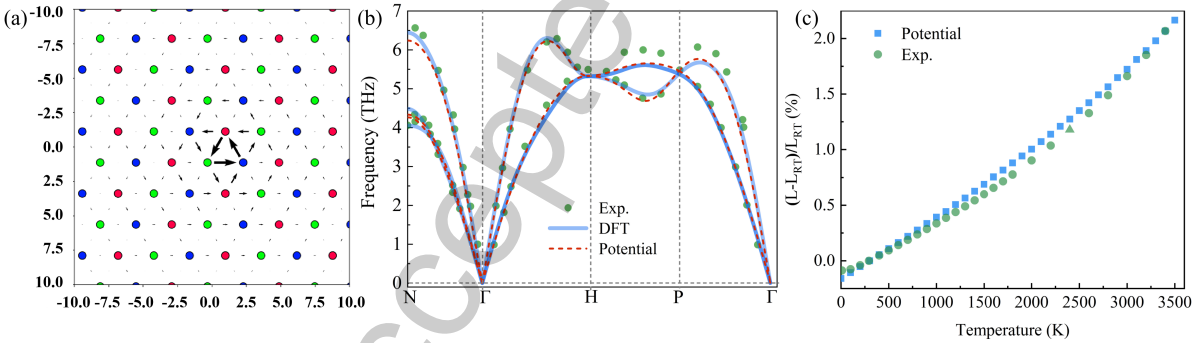

Figure 9. Validation of the machine-learned tungsten potential for dislocation core properties, lattice dynamics, and thermal behavior. (a) The core structure of a 1/2< 111 > screw dislocation predicted by the potential shows a non-degenerate compact configuration. The different colors of the atoms (red, green, and blue) denote three consecutive (111) atomic planes along the dislocation line. (b) Comparison of phonon dispersion curves with experimental measurements and DFT calculations. (c) Comparison of the change in the thermal expansion coefficient with temperature and experimental data.

low temperatures, which is consistent with the experimental observations under irradiation. Fig. 8b lists the surface energies of ten typical low-index crystal planes, including (110), (100), (111), (211), (210), (221), (311), (310), (321) and (320). The potential is in close agreement with DFT reference values across all of the surfaces considered. Figs. 8c and d present the generalized stacking fault energy curves for the 110< 111 > and 112< 111 > slip systems in BCC tungsten, respectively. The potential accurately reproduced the DFT-predicted γ surfaces for both primary slip systems, confirming its ability to model the plastic deformation behavior. Fig. S26 shows a comparison of the calculated grain boundary energies with DFT results for four different grain boundaries: $\Sigma3(112)$, $\Sigma3(111)$, $\Sigma5(013)$, and $\Sigma5(012)$.

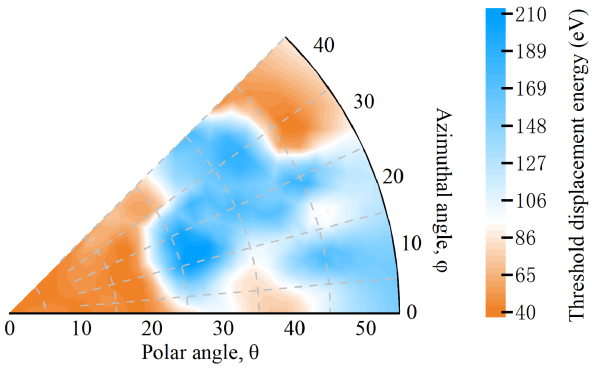

Figure 10. Displacement threshold energies along various crystallographic directions predicted using the machine-learned tungsten potential.

To assess the generalizability of the potential across diverse physical properties within the same elemental system, we validated its performance in describing the dislocation core structures, lattice dynamics, and finite-temperature behavior. The validation results when using the machine-learned tungsten potential to predict key material properties are presented in Fig. 9. Fig. 9a shows the predicted core structure of a 1/2< 111 > screw dislocation, revealing a nondegenerate compact configuration that aligned with the DFT predictions. Fig. 9b presents the phonon dispersion curves, which are in close agreement with the experimental measurements and first-principles calculations across all branches. Fig. 9c shows the thermal expansion behavior as a function of the temperature, for which the potential accurately reproduced the experimental data across the entire temperature range. Collectively, these results verify the potential accuracy of modeling complex material behaviors spanning the mechanical, dynamic, and thermal properties. Fig. 10 shows the displacement threshold energies along the different crystallographic directions, yielding a minimum value of 44 eV (consistent with the DFT result of 42 eV) and an average value of 93 eV, which is in close agreement with the ASTM-recommended value of 90 eV. These findings confirm the reliability of the model for simulating high-temperature and radiation damage phenomena[64,65].

The PhyMLP package

PhyMLP is an automated tool based on multi-physics sampling that generates essential datasets for constructing MLPs. The framework integrates tensor descriptors to build potentials from the sampling process, while the generated datasets are compatible with other MLP architectures. This passive learning strategy offers significant advantages for modeling complex material systems or when high generalization and extrapolation capabilities are required for MLPs.

The tool was implemented primarily in Python, utilizing widely adopted computational materials science libraries, including numpy, scipy, ase, and pymatgen. For detailed installation instructions and a comprehensive user guide, please refer to https://phymlp.readthedocs.io/en/latest/index.html. The framework interfaces with established computational tools, including the Vienna Ab initio Simulation Package (VASP) for first-principles calculations, the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) to validate MLP properties, optimized MTPs for MLPs fitting and dataset reliability assessment, and the Materials Project database (MP) for crystallographic structures[66-68]. PhyMLP can be installed via the command ***pip install . /phymlp-package***, although the current distribution requires manual uploading prior to installation. All the available functions and their detailed usage instructions can be accessed through ***phymlp-kit –help***.

Fig. 1 illustrates the complete workflow of the software. All the input parameters are specified in the input.yaml file, beginning with the user-defined chemical elements for interatomic interaction description. The framework automatically retrieves all relevant structures (including unary and binary compounds) from the MP database, while also accepting user-provided structures. These structures undergo systematic expansion through perturbations, surface cleavages, and tensile and compressive deformations, followed by first-principles static calculations. The Rose module then fits the $E - V$ relationships from the deformation data, generating additional data points through the fitted curves. Finally, all of the first-principles datasets and rose-fitted datasets are consolidated, where the energy and atomic force thresholds can be applied to filter unphysical configurations.

This section details all the functional modules within PhyMLP, including their specific capabilities, user-configurable input parameters, and output specifications. The independence and interconnectivity of the modular architecture are systematically described, with practical implementation cases discussed in Section 4.

Structure acquisition from MP module

This module serves as the initial component of the software and is executed via the command ***phymlp-kit from_mp_structures −−input_file xx/input.yaml***. The ***from_mp_structures*** directive invokes the module, whereas ***−−input_file*** specifies the input file path. It provides fundamental dataset acquisition through two channels: user-supplied structures and automated retrieval from an MP database. The input parameters are configured in a YAML file and include ***api:*** a user-defined authentication key for MP access (obtained from https://next-gen.materialsproject.org/api); ***elements:*** chemical elements for structure queries (e.g., elements: W); ***paths:*** a custom directory for storing the retrieved structures (default: ***./materials_project***); and ***user_poscar:*** a directory containing user-provided structures in the ***.vasp*** format (e.g., ***user_poscar: ./input-poscar***). The retrieved structures are systematically organized by material ID (e.g., ***xx/W/W-mp-91.vasp***). A summary.txt file generated in the output directory provides a statistical overview of the acquired structures, which is particularly valuable for multi-element materials.

Perturbation module

The perturbation module applies structural distortions to the database-retrieved or user-uploaded configurations. This is executed using ***phymlp-kit perturb_structures −−input_file xx/input.yaml***, where ***perturb_structures*** activates the module and ***−−input_file*** specifies the parameter file. To prevent excessive atomic or box distortions, the unit cells are first expanded into supercells controlled by the ***max_atoms*** and ***max_supercell*** parameters (default: 150 atoms and 4× expansion for VASP efficiency). The number of perturbed configurations per structure is set by ***n_structures***, while ***cell_pert_fraction*** (default: 0.05) and ***cell_pert_distance*** (default: 0.05 Å) define the maximum box deformation and atomic displacement magnitudes, respectively. All output structures are organized in the ***./perturb directory***, with filenames following the pattern ***xx/perturb/W-mp-91_supercell_perturb_0001.vasp***. For a standalone operation independent of the automated workflow, users can enable perturb_independent_process:true and specify the input directory via ***perturb_independent_dir: xx/xx***.

Deformation module

The deformation module generates structurally deformed configurations through tensile and compressive strains, serving two primary purposes: introducing strained structures with non-zero pressure states into the dataset and producing a continuous deformation series for subsequent Rose EOS fitting. This module is executed using ***phymlp-kit deform_structures −−input_file xx/input.yaml***. Similar to the perturbation module, supercell expansion is performed prior to deformation using comparable ***max_atoms*** and ***max_supercell*** parameters. Critical deformation control is achieved using the ***scaling_factors*** parameter (default: [0.8, 0.9, 1.0, 1.1, 1.2]), which represents the lattice scaling coefficients for compression (0.8, 0.9) and tension (1.1, 1.2). The energy and atomic force thresholds applied during dataset compilation automatically filter out structures that exhibit excessive stress. All the deformed structures are systematically organized in the ***./deform directory***, using the filename convention ***xx/deform/W-mp-91/W-mp-91_scaled_0.8*** **.vasp**. For a standalone operation independent of the automated workflow, users can enable ***deform_independent_process:true*** and specify the input directory via ***deform_independent_dir: xx/xx***.

VASP job submit module

The structures generated using the aforementioned modules require static calculations to be performed using VASP. Given the substantial number and diversity of configurations, systematic job management is implemented to prevent computational errors caused by resource limitations or process blocking.

The workflow comprises two distinct commands: ***phymlp -kit setup_vasp −−input_file xx/input*** .yaml creates calculation directories using the INCAR parameters specified in ***incar_parameters*** in the input file. This generates a ***vasp_calculations directory*** containing all the pending jobs, each populated with INCAR, POSCAR, POTCAR, and submit.sh files. ***phymlp-kit submit_vasp −−input_file xx/input.yaml*** monitors the calculation progress in real time and is controlled by ***max_batch_size*** (maximum concurrent jobs) and ***check_interval*** (workload verification frequency). Successful completion yields standard VASP output files, including OUTCAR and CONTCAR.

Rose module

The Rose fitting module processes deformation structures with completed first-principles calculations by utilizing the $E - V$ data obtained from VASP outputs. The module is executed using ***phymlp-kit Rose −−input_file input.yaml***.

Key parameters include ***initial_guess*** in the fitting section, which provides initial values for the Equation (8) parameters; ***optimization*** controls, which includes the maximum iteration count and convergence tolerance (default: 1000 and 10^{-6}); and ***output*** configuration ,which generates a graphical representation of fitting results (.png format) with annotated fitted parameters, as shown in Fig.S27. Upon completion, the module generates a ***Rose_fit directory*** containing the detailed fitting results for each processed curve.

Set merge module

The dataset consolidation module constitutes the final stage of the automated dataset generation workflow. It integrates datasets from all the preceding modules using the command ***phymlp-kit generate_extxyz --input_file input.yaml***. The procedure comprises two parallel processes. During first-principles data collection, the ***outcar_search_dir*** parameter (e.g., "***xx/materials_project/vasp_calculations***") specifies the VASP output directories for extracting structural data in the ***.extxyz*** format from OUTCAR files.

To obtain Rose augmented data, fitted curves (Fig.S28) are systematically sampled to generate additional structures, deriving the atomic coordinates and energies from the volume-energy relationships. This approach significantly expands a dataset while circumventing the computational costs of first-principles calculations. The sampling density is controlled by ***n_structures_per_fit*** (default: 50) with optional pressure filtering via ***pressure_range*** (e.g., [-100, 100] GPa). All of the consolidated datasets are organized in the ***./Extxyz_set directory***, resulting in a unified ***merged_datasets.extxyz*** file ready for MLP construction, and thereby completing the automated dataset generation pipeline.

MLP Training Module

During the automated workflow, the generated datasets are stored in the ***.extxyz*** format, which accommodates richer structural information and facilitates easier migration to other MLP software platforms. The automated workflow was integrated with the MTP software interface to validate the quality of the automatically sampled datasets. Optimization was performed within the native MTP framework to enable direct training using the ***.extxyz*** format datasets.

In the potential construction process, training data are required from first-principles calculations, Rose EOS fitting, and P-V experimental data extraction. These datasets include the atomic types, coordinates, energies, and other relevant information. The local energies of the system are predicted using tensor-based descriptors derived from the inputs. The MTP is the local potential, and energy E^{MTP} is the sum of the contributions, $V^{MTP}(n_i)$, of the atomic neighborhoods (ni) for N atoms[27].

$$E^{\mathrm{MTP}} = \sum_{i=1}^{N} V^{\mathrm{MTP}}(n_i) \tag{9}$$

Each neighborhood is a tuple:

$$n_i = \left\{(r_{ij}, z_i, z_j)\right\}_{j=1}^{N_{\mathrm{nbh}}}, \tag{10}$$

where r_{ij} are the relative atomic positions; z_i and z_j are the types of central and neighboring atoms, respectively; and N_{nbh} is the number of atoms in the neighborhood. Each V^{MTP} can be represented by a basis function:

$$V^{\mathrm{MTP}}(n_i) = \sum_{\alpha=1}^{N_{lin}} \xi_\alpha B_\alpha(n_i), \tag{11}$$

where B_α are the MTP basis functions,ξ_α are the linear parameters to be determined, and N_{lin} is the number of parameters.

Empirical potential extraction module

This module primarily processes ***.dump*** files generated by empirical potentials, and extracts structurally representative configurations from a large ensemble using the farthest-point sampling algorithm. The selected structures are then converted into POSCAR files, which are compatible with VASP, and used for subsequent first-principles calculations.

The module is executed using the command ***phymlp-kit empirical_pot_structures −−input_file input.yaml***, where ***emp irical_pot_structures*** calls the module and ***−−input_file*** specifies the path to the input YAML configuration file. Typically, users generate ***. dump*** files using MD software such as LAMMPS and define the file path in the input YAML, e.g., ***input_path: ". /dump_files***." In addition, the atomic-type mapping in the ***. dump*** file must be specified; for example, ***atom_type_mapping: 1: W*** indicates that atom type 1 corresponds to tungsten. The number of structures to be extracted can be configured using ***n_dump_structures***. It is important to ensure that the specified number does not exceed the total number of structures available in the provided path because the FPS operates on the available pool of configurations. Finally, the extracted structures are saved in a user-defined output directory, e.g., ***output_dir: "./dump_to_poscar_output***". All output files are formatted as POSCAR files and used for first-principles static calculations to obtain accurate energy, force, and stress data.

P-V experimental data extraction module

This module utilizes experimentally available P–V data, which are readily accessible from the literature and commonly reported in published studies. The module is executed using the command ***phymlp-kit exp_pv_structures −−input_file input.yaml***.

The relationship between the pressure and volume compression ratio (V/V_0) can be derived using the experimental P-V data. Using the equilibrium lattice constant (a_0) obtained from first-principles calculations, this relationship can be extended to correlate the pressure with the lattice compression ratio (a/a_0). A polynomial fitting function is then applied to establish a continuous curve describing the lattice parameter as a function of the pressure, thereby generating a large set of strained configurations for training. The polynomial expression used for fitting is as follows:

$$P = \sum_{k=0}^{4} B_k a^k, \tag{12}$$

where P denotes the pressure in gigapascals, a represents the lattice parameter in Å, k indicates the polynomial order, and B is the fitting coefficient. In the current implementation, the default maximum polynomial degree is set to four (i.e., ***polynomial_degree: 4***). The essential input parameters include the path to the experimental data file specified by the ***EXP_data_file*** parameter, which contains two columns: the pressure values and volume compression ratios (V/V_0); and equilibrium lattice constant a_0, which is obtained from first-principles calculations using the ***DFT_a0*** parameter in Å. Users can optionally define the pressure range for fitting, e.g., ***pressure_range: [0, 100]***. After configuration, an output directory must be designated, for instance, ***output_directory: ". /exp_pv_fitting".***

Upon completion of the fitting procedure, the resulting lattice parameter versus pressure data points are stored in the file ***fit_a_P.data***. The detailed fitting parameters are recorded in ***fitting_coefficients.txt***, and a graphical representation of the fitted curve along with the discrete data points is saved as ***pressure_fitting.png***, as seen in Fig.S27.

This automated tool has been integrated as a core module of the machine learning-accelerated toolkit for material simulation (MATS). MATS is a full-stack computational software package for materials science that enables flexible multi-level material modeling and simulation within a single framework, while achieving both high accuracy and high efficiency. The software integrates three core modules: first-principles calculation (MATS-DFT), AI-driven potential training (MATS-MLP; i.e., the PhyMLP work presented here), and MD simulation (MATS-MD). MATS-DFT provides high-precision electronic structure calculations based on the DFT, including hybrid functionals and ABMD. MATS-MD supports large-scale MD simulations that can handle millions of atoms with variable time-stepping and cascade collision simulation capabilities. Together, these three modules form a closed-loop workflow ranging from quantum mechanical accuracy to meso/micro-scale dynamics, offering a powerful and efficient cross-scale simulation solution for materials science research and industrial applications.

CONCLUSION

This paper introduced PhyMLP, an automated and physics-guided framework for constructing training datasets for MLPs. By integrating the universal scaling relation of the Rose EOS, empirical experimental measurements, and DFT calculations, PhyMLP enables targeted sampling of critical configurational regimes to efficiently characterize a PES. Compared to conventional empirical and computationally intensive configuration generation methods, this data fusion approach provides a targeted and data-efficient paradigm. This substantially reduces the required number of explicit DFT evaluations while preserving the structural and physical representativeness of the dataset. Systematic validation using the BCC tungsten benchmark demonstrated that the MTP trained via PhyMLP yielded predictions consistent with experimental measurements and DFT reference data across diverse material properties. These included the lattice and elastic constants, point defect formation and migration energetics, dislocation core structures, free surface and stacking fault energies, thermodynamic properties, and primary irradiation damage responses. These results confirmed the predictive accuracy and transferability of the potential, thereby validating the structural representativeness of the training set generated using the PhyMLP methodology.

PhyMLP provides a systematic solution to the data bottleneck inherent in the development of MLPs. Consequently, it facilitates high-throughput large-scale MD simulations and atomic-scale material design, offering a scalable framework for applications in computational materials science.

DECLARATIONS

Acknowledgments

The authors acknowledge support from the National Supercomputer Center in Changsha (NSCC).

Authors' contributions

Software, Conceptualization, Writing–review & editing, Writing–original draft, Methodology, and Investigation: Yiqun Guo, Shihui Ning, Guoqi Guo.

Software, Conceptualization, Methodology, Writing–review & editing: Yangchun Chen, Bowen Huang, Shifang Xiao.

Review & editing, Supervision, Resources, Conceptualization: Wangyu Hu.

Availability of data and materials

Results supporting this study are presented in the Supplementary Material. The other raw data supporting the findings of this study are available from the corresponding author upon request.

AI and AI-assisted tools statement

During the preparation of this manuscript, the AI tool Gemini(version Gemini 3.1 Pro, released Google) was used solely for language editing. The tool did not influence the study design, data collection, analysis, interpretation, or the scientific content of the work. All authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

This study was financially supported by the National Key Research and Development Program of China (2023YFB3003000).

Conflicts of interest

The authors declare no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable

REFERENCES

1. Ge W, De Silva R, Fan Y, Sisson SA, Stenzel MH. Machine learning in polymer research. *Adv Mater* 2025;37:2413695. DOI
2. Bartók AP, Kermode J, Bernstein N, Csányi G. Machine learning a general-purpose interatomic potential for silicon. *Phys Rev X* 2018;8:041048. DOI
3. Rohskopf A, Seyf HR, Gordiz K, Tadano T, Henry A. Empirical interatomic potentials optimized for phonon properties. *Npj Comput Mater* 2017;3:27. DOI
4. Finnis M, Sinclair J. A simple empirical N-body potential for transition metals. *Philos Mag A* 1984;50:45. DOI
5. Daw MS, Baskes MI. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Phys Rev B* 1984;29:6443. DOI
6. Baskes MI. Modified embedded-atom potentials for cubic materials and impurities. *Phys Rev B* 1992;46:2727. DOI
7. Fedik N, Zubatyuk R, Kulichenko M, Lubbers N, Smith JS, et al. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat Rev Chem* 2022;6:653–72. DOI
8. Mortazavi B, Zhuang X, Rabczuk T, Shapeev AV. Atomistic modeling of the mechanical properties: the rise of machine learning interatomic potentials. *Mater Horiz* 2023;10:1956–68. DOI
9. Kulichenko M, Nebgen B, Lubbers N, Smith JS, Barros K, et al. Data generation for machine learning interatomic potentials and beyond. *Chem Rev* 2024;124:13681–714. DOI
10. Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. *Nat Comput Sci* 2022;2:718–28. DOI
11. Ito K, Yokoi T, Hyodo K, Mori H. Machine learning interatomic potential with DFT accuracy for general grain boundaries in α-Fe. *Npj Comput Mater* 2024;10:255. DOI
12. Ji Y, Liang J, Xu Z. Machine-learning interatomic potentials for long-range systems. *Phys Rev Lett* 2025;135:178001. DOI
13. Wan K, He J, Shi X. Construction of high accuracy machine learning interatomic potential for surface/interface of nanomaterials—a review. *Adv Mater* 2024;36:2305758. DOI
14. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* 2015;349:255–60. DOI
15. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *science* 2006;313:504–7. DOI
16. Sharifani K, Amini M. Machine learning and deep learning: A review of methods and applications. *World Inf Technol Eng J* 2023;10:3897–904. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4458723.
17. Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, et al. Machine learning and the physical sciences. *Rev Mod Phys* 2019;91:045002. DOI
18. Wang Q, Yao Y. Harnessing machine learning for high-entropy alloy catalysis: a focus on adsorption energy prediction. *Npj Comput Mater* 2025;11:91. DOI
19. Liu X, Zhang J, Pei Z. Machine learning for high-entropy alloys: Progress, challenges and opportunities. *Prog Mater Sci* 2023;131:101018. DOI
20. Zhang Y, Wang H, Chen W, Zeng J, Zhang L, et al. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput Phys Commun* 2020;253:107206. DOI
21. Galib M, Isiet M, Ponga M. AtomProNet: Data flow to and from machine learning interatomic potentials in materials science. *arXiv preprint arXiv:250114039* 2025. Available from: https://doi.org/10.48550/arXiv.2501.14039.
22. Li W, Ou Q, Chen Y, Cao Y, Liu R, et al. DeePKS+ ABACUS as a bridge between expensive quantum mechanical models and machine learning potentials. *J Phys Chem A* 2022;126:9154. DOI
23. Chen C, Li Y, Zhao R, Liu Z, Fan Z, et al. NepTrain and NepTrainKit: Automated active learning and visualization toolkit for neuroevolution potentials. *Comput Phys Commun* 2025:109859. DOI
24. Cao Y, Sheriff K, Freitas R. Capturing short-range order in high-entropy alloys with machine learning potentials. *npj Comput Mater* 2025;11:268. DOI
25. Röcken S, Zavadlav J. Accurate machine learning force fields via experimental and simulation data fusion. *npj Comput Mater* 2024;10:69. DOI
26. Sauceda HE, Gálvez-González LE, Chmiela S, Paz-Borbón LO, Müller KR, et al. BIGDML—Towards accurate quantum machine learning force fields for materials. *Nat Commun* 2022;13:3733. DOI
27. Shapeev AV. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model Simul* 2016;14:1153–73. DOI
28. Liu J, Byggmästar J, Fan Z, Qian P, Su Y. Large-scale machine-learning molecular dynamics simulation of primary radiation damage in tungsten. *Phys Rev B* 2023;108:054312. DOI
29. Kang S, Kim J, Park T, Won J, Baik C, et al. Toward fast and accurate machine learning interatomic potentials for atomic layer deposition precursors. *Mater Today Adv* 2024;21:100474. DOI
30. Mortazavi B, Novikov IS, Podryabinkin EV, Roche S, Rabczuk T, et al. Exploring phononic properties of two-dimensional materials using machine learning interatomic potentials. *Appl Mater Today* 2020;20:100685. DOI
31. Rose JH, Smith JR, Guinea F, Ferrante J. Universal features of the equation of state of metals. *Phys Rev B* 1984;29:2963. DOI
32. Hixson R, Fritz J. Shock compression of tungsten and molybdenum. *J Appl Phys* 1992;71:1721–28. DOI
33. Ma PW, Dudarev S. Universality of point defect structure in body-centered cubic metals. *Phys Rev Mater* 2019;3:013605. DOI
34. Ma PW, Dudarev S. Effect of stress on vacancy formation and migration in body-centered-cubic metals. *Phys Rev Mater* 2019;3:063601. DOI
35. Mason DR, Nguyen-Manh D, Becquart CS. An empirical potential for simulating vacancy clusters in tungsten. *J Phys: Condens Matter* 2017;29:505501. DOI
36. Alexander R, Marinica MC, Proville L, Willaime F, Arakawa K, et al. Ab initio scaling laws for the formation energy of nanosized interstitial

defect clusters in iron, tungsten, and vanadium. *Phys Rev B* 2016;94:024103. DOI
37. Bonny G, Terentyev D, Bakaev A, Grigorev P, Van Neck D. Many-body central force potentials for tungsten. *Model Simul Mater Sci Eng* 2014;22:053001. DOI
38. Gehringer D, Friák M, Holec D. Models of configurationally-complex alloys made simple. *Comput Phys Commun* 2023;286:108664. DOI
39. Gilbert M, Sublet JC. Neutron-induced transmutation effects in W and W-alloys in a fusion environment. *Nucl Fusion* 2011;51:043005. DOI
40. Romaner L, Ambrosch-Draxl C, Pippan R. Effect of rhenium on the dislocation core structure in tungsten. *Phys Rev Lett* 2010;104:195503. DOI
41. Tanno T, Fukuda M, Nogami S, Hasegawa A. Microstructure development in neutron irradiated tungsten alloys. *Mater Trans* 2011;52:1447–51. DOI
42. Li YH, Zhou HB, Liang L, Gao N, Deng H, et al. Transition from ductilizing to hardening in tungsten: The dependence on rhenium distribution. *Acta Mater* 2019;181:110–23. DOI
43. Szlachta WJ, Bartók AP, Csányi G. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys Rev B* 2014;90:104108. DOI
44. Muzyk M, Nguyen-Manh D, Kurzyd lowski K, Baluc N, Dudarev S. Phase stability, point defects, and elastic properties of WV and W-Ta alloys. *Phys Rev B* 2011;84:104115. DOI
45. Bonny G, Bakaev A, Terentyev D, Mastrikov YA. Interatomic potential to study plastic deformation in tungsten-rhenium alloys. *J Appl Phys* 2017;121. DOI
46. Mrovec M, Gröger R, Bailey AG, Nguyen-Manh D, Elsässer C, et al. Bond-order potential for simulations of extended defects in tungsten. *Phys Rev B* 2007;75:104119. DOI
47. Featherston FH, Neighbours J. Elastic constants of tantalum, tungsten, and molybdenum. *Phys Rev* 1963;130:1324. DOI
48. Wang Y, Chen D, Zhang X. Calculated equation of state of Al, Cu, Ta, Mo, and W to 1000 GPa. *Phys Rev Lett* 2000;84:3220. DOI
49. Ferroni F, Yi X, Arakawa K, Fitzgerald SP, Edmondson PD, et al. High temperature annealing of ion irradiated tungsten. *Acta Mater* 2015;90:380–93. DOI
50. Chen Y, Li YH, Gao N, Zhou HB, Hu W, et al. New interatomic potentials of W, Re and W-Re alloy for radiation defects. *J Nucl Mater* 2018;502:141–53. DOI
51. Wang X, Wang Y, Zhang L, Dai F, Wang H. A tungsten deep neural-network potential for simulating mechanical property degradation under fusion service environment. *Nucl Fusion* 2022;62:126013. DOI
52. Byggmästar J, Hamedani A, Nordlund K, Djurabekova F. Machine-learning interatomic potential for radiation damage and defects in tungsten. *Phys Rev B* 2019;100:144105. DOI
53. Lide DR. CRC handbook of chemistry and physics: a ready-reference book of chemical and physical data. CRC press; 1995.
54. Qi X, Cai N, Wang S, Li B. Thermoelastic properties of tungsten at simultaneous high pressure and temperature. *Journal of Applied Physics* 2020;128. DOI
55. Renault PO, Badawi K, Bimbault L, Goudeau P, Elkaim E, et al. Poisson's ratio measurement in tungsten thin films combining an x-ray diffractometer with in situ tensile tester. *Applied physics letters* 1998;73:1952–54. DOI
56. Kittel P. Enthalpy, entropy, and exergy flow losses in pulse tube cryocoolers. In: Cryocoolers 13. Springer; 2005. pp. 343–52. DOI
57. Simmons G, Wang H. Single Crystal Elastic Constants and Calculated Aggregate Properties. Cambridge, MA: MIT Press; 1971.
58. You Y, Zhang D, Wu F, Cao X, Sun Y, et al. Principal component analysis enables the design of deep learning potential precisely capturing LLZO phase transitions. *npj Comput Mater* 2024;10:57. DOI
59. He S, Mang EH, El Atwani O, Nathanie J, Leff AC, et al. Complex dislocation loop networks as natural extensions of the sink efficiency of saturated grain boundaries in irradiated metals. *Sci Adv* 2024;10:eadj8395. DOI
60. Borges PP, Ritchie RO, Asta M. Electronic descriptors for dislocation deformation behavior and intrinsic ductility in bcc high-entropy alloys. *Sci Adv* 2024;10:eadp7670. DOI
61. Hou J, Peng D, Kong XS, Deng H, Hu W, et al. Hydrogen modulated dislocation reaction and defect accumulation in bcc metals. *Acta Mater* 2025:121524. DOI
62. Parakh A, Lee S, Harkins KA, Kiani MT, Doan D, et al. Nucleation of dislocations in 3.9 nm nanocrystals at high pressure. *Phys Rev Lett* 2020;124:106104. DOI
63. Li G, Wen H, Zhang Y, Jiang H, Huang J, et al. Electromechanical coupled modulation of dislocation nucleation and annihilation in ferroelectric oxide films. *Phys Rev B* 2025;111:224101. DOI
64. Kanhaiya K, Kim S, Im W, Heinz H. Accurate simulation of surfaces and interfaces of ten FCC metals and steel using Lennard–Jones potentials. *npj Comput Mater* 2021;7:17. DOI
65. Zhu LF, Srinivasan P, Gong Y, Hickel T, Grabowski B, et al. Melting properties of the refractory metals V and W and the binary VW alloy fully from first principles. *Phys Rev B* 2024;109:094110. DOI
66. Kresse G, Hafner J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys Rev B* 1994;49:14251. DOI
67. Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys Rev B* 1996;54:11169. DOI
68. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. *J Comput Phys* 1995;117:1–19. DOI