

机器学习势在正极材料电化学性质预测中的评测

蒋冰鉴^{1,2}, 杨 涛^{1,2}, 路 宽³, 张艺馨⁴

(1. 北京信息科技大学计算机学院, 北京 100101; 2. 北京信息科技大学, 北京材料基因工程高精尖创新中心, 北京 100101; 3. 中国科学院山西煤炭化学研究所, 煤炭高效低碳利用全国重点实验室, 太原 030001; 4. 宁夏大学化学化工学院, 银川 750021)

摘 要: 系统评估了 8 种具有代表性的预训练机器学习势(MACE、M3GNet、CHGNet、ALIGNN、SevenNet、AlphaNet、GRACE 和 Orb)在典型 Na/Li 正极材料(层状氧化物、磷酸盐及其他快离子导体)上的性能, 涵盖电压预测、离子迁移与热稳定性 3 大关键任务, 旨在为不同电化学性质预测任务中机器学习势的合理选择提供定量依据。结果显示, M3GNet 在电压预测中整体误差最小, 其次为 AlphaNet 与 CHGNet; 离子迁移的 MSD 分析表明, MACE 与 SevenNet 给出的扩散系数最接近参考值; 热稳定性对比表明, M3GNet 与 MACE 对释氧温度及阶梯分解特征的预测更为准确。上述体系覆盖 Li/Na 化学体系、层状与聚阴离子结构、多种过渡金属氧化还原对, 为定量基准测试提供了可靠的 DFT/实验数据支撑。

关键词: 机器学习势; 钠离子电池; 锂离子电池; 电压曲线; 离子扩散; 热稳定性

中图分类号: TB30 文献标志码: A 文章编号: 0454-5648(2026)07-2338-14

网络出版时间: 2026-04-10



1 Introduction

Lithium- and sodium-ion batteries underpin today's portable electronics, electric vehicles, and grid-scale energy-storage technologies^[1-2]. The ever-growing need for higher energy density, longer cycle life, and safer operation continues to drive the discovery of new electrode materials. First-principles methods, most notably density functional theory (DFT), have been indispensable for evaluating redox potentials, ion-migration barriers, and thermal stability. However, the steep $\mathcal{O}(N^3)$ scaling of conventional DFT places practical limits on the size of supercells and on the length/time scales accessible to ab initio molecular dynamics (AIMD) calculations^[3-4], especially when hundreds to thousands of candidate structures must be screened^[5-6]. Beyond the issue of scaling, the intrinsic complexity of electrochemical processes such as multiphase equilibria during

intercalation, strongly anisotropic diffusion pathways, and temperature-dependent oxygen release requires models that can simultaneously capture thermodynamics and kinetics across diverse chemistries. In practice, screening cathode candidates for commercial relevance demands the evaluation of hundreds of phase diagrams, voltage curves, and diffusivities, which is infeasible using DFT alone^[7]. The development of surrogate models that retain near-first-principles fidelity while enabling high-throughput workflows is thus critical for accelerating electrode discovery^[8].

Recent progress in machine-learning interatomic potentials (MLIPs) offers an attractive route to circumvent these bottlenecks. By learning the potential-energy surface from a finite DFT training set, MLIPs such as M3GNet, CHGNet, ALIGNN, MACE, and their successors can deliver near-DFT accuracy at several orders of magnitude lower cost, enabling nanosecond-scale molecular dynamics and high-throughput energy evaluations on

收稿日期: 2025-11-28。 修订日期: 2025-12-24。

基金项目: 国家自然科学基金项目(22272009)。

第一作者: 蒋冰鉴(2000—), 男, 硕士研究生。

通信作者: 杨 涛(1980—), 女, 博士, 助理研究员。

Received date: 2025-11-28. Revised date: 2025-12-24.

First author: JIANG Bingjian (2000—), male, Master candidate.

E-mail: 2023020608@bistu.edu.cn

Correspondent author: YANG Tao (1980—), female, Ph.D., Assistant Professor.

E-mail: ytyangmei@bistu.edu.cn

commodity hardware. More broadly, advances in materials representations and feature engineering remain crucial for machine-learning models of crystalline materials, especially for coupling local environments with global chemistry–structure information^[9].

Despite this promise, a systematic benchmark of the latest graph-neural-network potentials across multiple, functionally distinct electrochemical tasks has been missing. Crucially, the fidelity required to predict, for example, an average voltage profile is not necessarily the same as that for migration kinetics or decomposition temperatures, and model performance may vary accordingly. In this work, we benchmark eight state-of-the-art pretrained MLIPs (MACE, M3GNet, CHGNet, ALIGNN, SevenNet, AlphaNet, GRACE, and Orb) on representative Na- and Li-based cathodes. The voltage curves for NaMnO₂, LiCoO₂, NaFePO₄ and Na₃V₂(PO₄)₃ show that M3GNet best reproduces the reference trends and phase-transition points on average, followed by AlphaNet and CHGNet, whereas Orb and ALIGNN exhibit larger bias and larger curve-shape errors. For ion migration in Li₃Sc₂(PO₄)₃, the MSD-derived diffusivity is predicted most accurately by MACE and SevenNet (with Orb and GRACE also close to the reference), while CHGNet and M3GNet tend to overestimate *D* and AlphaNet underestimates it; moreover, ALIGNN-FF could not complete the long-timescale MD runs due to memory limitations. These comparisons highlight the task-dependent strengths and constraints of each pretrained model in realistic materials-screening workflows.

The present framework provides a consistent and extensible route for evaluating machine-learning interatomic potentials across multiple, functionally distinct electrochemical and thermochemical metrics. In this work, three representative tasks: voltage prediction, ion migration, and thermal stability are examined under matched conditions. By jointly assessing thermodynamic, kinetic, and high-temperature behaviors, this approach offers a holistic understanding of model transferability and reliability in realistic battery environments, establishing a foundation for integrated evaluation of energy, transport, and stability

performance.

We deliberately couple each evaluation task to material families that exercise distinct bonding, structural motifs, and redox chemistry. For voltage profiles, we use two layered oxides (NaMnO₂, LiCoO₂) and two phosphate frameworks [NaFePO₄, Na₃V₂(PO₄)₃]. The layered oxides probe strong TM–O covalency and ordering effects, whereas the polyanion hosts test inductive effects and corner/edge-sharing PO₄ networks. Together they cover both Na and Li chemistries and produce multi-plateau *V* – *x* curves with established DFT/experimental references.

For ion migration, we select Li₃Sc₂(PO₄)₃ as a representative fast-ion conductor featuring well-defined diffusion channels and stable molecular-dynamics behavior at elevated temperatures, minimizing confounding effects from transition-metal magnetism while stressing long-timescale sampling. For thermal stability, we benchmark MnPO₄ and FePO₄ (olivine and trigonal), where oxygen release proceeds in discrete steps toward pyrophosphate products; these systems offer clear experimental trends and are widely used to study high-temperature decomposition. This mapping of task to material family ensures that architectural and data-coverage differences among MLIPs are tested across thermodynamic, kinetic, and high-*T* regimes under realistic, well-documented baselines.

Because voltage prediction, ion migration, and thermal stability benchmarks rely on different reference datasets and experimental observables, the three tasks are necessarily evaluated on different (though representative) material families in this work. Therefore, any cross-task ranking reflects a combined effect of the target property/workflow and the chemistry–structure distribution of the benchmark set. To reduce this confounding, we benchmark voltage on four cathodes spanning layered and polyanion structures, and thermal stability on three phosphate polymorphs/reaction pathways. Accordingly, conclusions are stated as conditional on the selected benchmark families and protocols. A fully decoupled benchmark where all tasks are evaluated on the same overlapping set of materials is an important direction for future work.

2 Computational Methods

2.1 Dataset Preparation

A concise overview of the benchmark material systems and the rationale for their selection is provided in the Introduction (“Benchmark Material Systems and Selection Rationale”). Here we detail the supercell choices, enumeration procedures, and simulation settings for each task.

2.1.1 Initial sources All crystallographic prototypes used in this study were retrieved from the Materials Project (MP) and, when applicable, cross-checked against the experimental literature to ensure phase realism and consistent stoichiometry^[10]. For each chemistry, the as-reported conventional cell was taken as the parent structure for subsequent configuration generation.

2.1.2 Voltage-Profile set For the four benchmark cathodes: NaMnO₂, LiCoO₂, NaFePO₄ and Na₃V₂(PO₄)₃, each MP structure was expanded to the smallest supercell containing approximately 8–12 alkali ions (Na or Li) on the A sublattice. This choice provides a practical composition grid ($\Delta x \approx 1/12 - 1/8$ per vacancy) that resolves piecewise voltage segments and a manageable combinatorial space for site-vacancy enumeration. Within each supercell, partially (de)intercalated configurations were generated by removing subsets of alkali ions; distinct site combinations were treated as unique configurations. The resulting set of configurations serves as input for MLIP energy evaluation and subsequent convex-hull construction in composition–formation–energy space, from which average and piecewise voltages are derived. To generate all possible alkali-vacancy configurations, we performed exhaustive combinatorial enumeration of ion-removal patterns within each expanded supercell.

Given n alkali sites, all distinct subsets C_k of size k ($k=1\dots n$) were constructed using the standard `itertools.combinations` routine, corresponding to all possible compositions A_x with $x=(n-k)/n$. The

resulting set $\mathcal{S} = \bigcup_{k=0}^n C_k$ defines all partially (de)intercalated configurations prior to symmetry reduction. Here, $\mathcal{S}$ denotes the complete set of all

possible deintercalation configurations, C_k is the set of all distinct site combinations after removing k alkali ions, and n is the total number of alkali sites in the supercell.

To avoid redundant configurations arising from crystallographic equivalence, each generated structure was screened using the `spglib` symmetry library. For each trial configuration, the equivalent-atom mapping returned by “`spglib.get_symmetry_dataset()`” was used to form a unique hash key $h = \{(Z_i, e_i)\}_{i=1}^N$, where Z_i and e_i denote the atomic number and equivalent-atom label, respectively. Two configurations sharing identical h were treated as symmetry-equivalent and only one was retained. A numerical tolerance of 10^{-3} Å was applied to fractional coordinates when identifying equivalence. This procedure ensures nonredundant enumeration over the configurational space without the need for manual curation.

Typical supercells (8–12 alkali sites) yield 20–60 nonredundant configurations per composition step ($\Delta x \approx 0.1$), providing sufficient resolution for voltage-curve construction while keeping the total number of MLIP energy evaluations tractable. The same enumeration–deduplication routine was applied consistently across all cathode chemistries.

2.1.3 Ion-migration set To assess diffusional kinetics via molecular dynamics, we selected Li₃Sc₂(PO₄)₃ (Materials Project ID: mp-6565) as a representative fast-ion conductor. The simulation structure was taken directly from MP and prepared for MD runs; supercells were chosen to mitigate finite-size artifacts and image correlations, while preserving the underlying framework topology and ion-conduction channels.

2.1.4 Thermal-stability set To benchmark oxygen-release thermodynamics, the initial structures of MnPO₄ and FePO₄ (olivine and trigonal polymorphs) were taken from the thermodynamically stable entries in the MP. The intermediate and final solids appearing in the stepwise decomposition scheme were also selected as the corresponding MP stable phases. Each plateau in Fig. 3 corresponds to a balanced reaction between MP stable reactants and products with explicit O₂ release. The MP identifiers of all

phases are available upon request.

The reactions used in this benchmark are: $\text{MnPO}_4 \rightarrow 0.5 \text{ Mn}_2\text{P}_2\text{O}_7 + 0.25 \text{ O}_2$; olivine $\text{FePO}_4 \rightarrow 0.1 \text{ Fe}_7(\text{PO}_4)_6 + 0.1 \text{ Fe}_3(\text{P}_2\text{O}_7)_2 + 0.1 \text{ O}_2$; olivine $\text{FePO}_4 \rightarrow (1/6) \text{ Fe}_3(\text{PO}_4)_2 + (1/6) \text{ Fe}_3(\text{P}_2\text{O}_7)_2 + (1/6) \text{ O}_2$; $\text{FePO}_4 \rightarrow 0.5 \text{ Fe}_2\text{P}_2\text{O}_7 + 0.25 \text{ O}_2$. For trigonal FePO_4 , the latter two reactions apply.

2.2 Machine-Learning Potential Models

All eight MLIPs benchmarked in this work are publicly released pretrained models. Each model was trained on extensive DFT datasets that include total energies, atomic forces, and, where applicable, stresses or magnetic moments. The training corpora span equilibrium and perturbed crystal structures across diverse chemistries, enabling chemical transferability without any fine-tuning in this study. Table 1 summarizes the main pretraining datasets and objectives for the evaluated models.

Table 1 Pretraining datasets and training objectives for the eight pretrained MLIPs benchmarked in this work

Model	Pretraining dataset	Targets/loss terms
M3GNet	MPF	Energies, forces, stresses
CHGNet	MPtrj	Energies, forces, stresses, magnetic moments
ALIGNN	MP 2022	Energies, forces
MACE	MPtrj+sAlex	Energies, forces
GRACE	OMat24+sAlex+MPtrj	Energies, forces
SevenNet	OMat24+sAlex+MPtrj	Energies, forces
AlphaNet	OMat24+sAlex+MPtrj	Energies, forces
Orb	OMat24+sAlex+MPtrj	Energies, forces

2.2.1 MACE The MACE model^[11] is an equivariant message-passing neural network constructed on top of the Atomic Cluster Expansion (ACE) framework. Its design extends conventional message passing from pairwise interactions to incorporate explicit higher-order terms, enabling the direct representation of three-body, four-body, and more complex interactions within atomic environments. In this architecture, atomic environments are described using equivariant features that transform consistently under translations, rotations, and reflections, ensuring strict preservation of physical symmetries. The model parameterizes geometric dependencies through tensor products of spherical harmonics and radial basis functions, allowing natural inclusion of angular correlations among atoms. By combining the systematic expansion principles of ACE with the expressive capacity of deep equivariant

networks, MACE is applicable to the construction of potential energy surfaces for both molecular and crystalline systems.

2.2.2 ALIGNN ALIGNN (Atomistic line graph neural network)^[12] augments standard atomistic graphs with their line graphs so that message passing occurs on both bonds and *bond angles*. By explicitly encoding angle information, ALIGNN improves sensitivity to directional bonding and local geometry compared to distance-only GNNs. In its force-field form (ALIGNN-FF), the architecture is trained jointly on energies and forces from large DFT corpora, yielding a chemically broad model (dozens of elements) suitable for fast relaxations and molecular dynamics, with ASE/LAMMPS interfaces for practical workflows^[13–14].

2.2.3 M3GNet M3GNet^[15–16] is a materials graph network that incorporates explicit three-body interactions and treats both atomic coordinates and the crystal lattice so that energies, forces, and stresses are obtained via automatic differentiation. Trained at scale on Materials Project relaxation trajectories, the released “universal potential” supports structure relaxation, MD, and screening workflows; current implementations are provided in the MatGL library with pretrained weights and an ASE calculator for drop-in use in simulation pipelines.

2.2.4 CHGNet CHGNet (Crystal Hamiltonian Graph Neural Network)^[17] is a pretrained, charge-informed MLIP designed for inorganic solids. It is trained on the Materials Project Trajectory dataset including energies, forces, stresses, and magnetic moments; the latter are used to regularize learned atom features and enable inference about local electronic environments. The open-source release provides general-purpose models for broad chemistries together with utilities for charge-aware MD, structure relaxation, and property evaluation.

2.2.5 SevenNet SevenNet^[18] (Scalable equivariance-enabled neural network) is a NequIP-based equivariant interatomic potential with a parallel domain-decomposition scheme for multi-GPU (Graphics processing unit) molecular dynamics, achieving near-ideal strong/weak scaling on large systems

while maintaining rotational equivariance. The accompanying multi-fidelity training framework demonstrates that high-fidelity (*e.g.*, meta-GGA/SCAN) accuracy can be reached with limited high-level data by learning jointly from lower-fidelity datasets; SevenNet provides LAMMPS (Large-scale atomic/molecular massively parallel simulator) coupling and pretrained models to facilitate large-scale MD.

2.2.6 AlphaNet AlphaNet^[19] is a local-frame-based equivariant architecture that constructs rotationally consistent features in learned local coordinate frames. By reducing reliance on high-order spherical harmonics and adopting efficient frame transitions, it targets improved accuracy-throughput trade-offs for large-scale MD while retaining strict symmetry handling. The public implementation includes training/inference code and examples; current results are reported as a preprint with open GitHub resources.

2.2.7 Orb Orb^[20] is a family of universal neural network potentials from Orbital Materials focused on speed, stability, and scalability for large atomistic simulations. The models employ attention-based message passing with physically motivated edge embeddings (*e.g.*, Bessel-spherical harmonic outer-product features) and provide both “direct” and “conservative” (energy gradient) variants. Recent releases introduce compiled inference and uncertainty heads, and expand training to broader datasets (*e.g.*, OMat24), enabling fast relaxations and long-time MD with ASE integration.

2.2.8 GRACE GRACE^[21] (Graph atomic cluster expansion) generalizes the atomic cluster expansion by introducing graph basis functions and tensor decompositions, yielding a formulation that unifies and extends message-passing MLIPs while preserving systematic improvability. The resulting models exhibit linear scaling with neighbor count and layer depth in the graph basis and have been demonstrated on molecules, clusters, and general-purpose elemental systems. An open implementation and tooling support training/deployment of GRACE-based potentials.

3 Electrochemical Property Workflows

We adopt a unified, reproducible pipeline to

evaluate three functionally distinct electrochemical tasks: voltage profiles, ion migration, and thermal stability, so that all machine-learning interatomic potentials (MLIPs) are assessed under matched conditions. Below we restate the governing relations and provide the exact procedures used in this work.

3.1 Voltage Profiles

For four representative cathodes, NaMnO₂^[22], LiCoO₂^[23], NaFePO₄^[24], and Na₃V₂(PO₄)₃^[25], we first retrieve all reported structures with the target composition from the Materials Project. For each parent structure, we generate partially (de)intercalated configurations by removing subsets of alkali ions; distinct site combinations are treated as unique configurations. Each MLIP evaluates the total energy of every configuration (optionally preceded by brief local relaxations to remove residual forces). Formation energies are then computed relative to the alkali-rich and alkali-poor endmembers, and the lower convex hull is constructed in composition–formation–energy space. Only hull phases are retained as thermodynamically accessible at 0 K. Average, piecewise voltages are obtained from adjacent hull vertices, yielding a staircase $V-x$ curve that approximates the experimental voltage profile.

At equilibrium, the cell voltage is given by the Nernst relation,

$$V = -\frac{\Delta G_r}{nF} \quad (1)$$

where ΔG_r is the reaction Gibbs free energy, n is the number of transferred electrons, and F is Faraday’s constant^[26]. Between two ground-state compositions x_1 and x_2 of A_xMX_y ($A = \text{Li, Na}$), the 0 K approximation $\Delta G_r \approx \Delta E_r$ leads to the average voltage.

$$V(x_1, x_2) = -\frac{E(A_{x_1}MX_y) - E(A_{x_2}MX_y) - (x_1 - x_2)E(A)}{(x_1 - x_2)F} \quad (2)$$

Formation energies relative to the endmembers are,

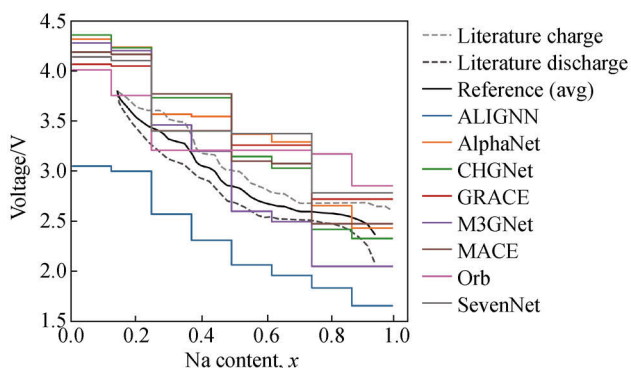
$$E_f(x) = E(A_xMX_y) - xE(A_1MX_y) - (1-x)E(A_0MX_y) \quad (3)$$

and the lower convex hull of E_f vs. x defines thermodynamically accessible two-phase segments whose slopes yield the piecewise voltages *via*

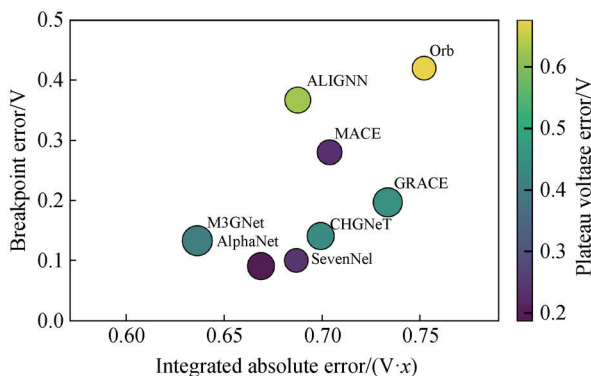
Eqs.(1)–(2) [27–28].

To quantify both pointwise accuracy and curve-shape fidelity, we compute voltage errors on a common composition grid x by linearly interpolating the staircase V - x curves from each MLIP and the reference. In addition to MAE, RMSE, and mean bias, two metrics used in Fig. 1 are defined as follows:

Integrated absolute error (IAE): $\text{IAE} =$



(a)

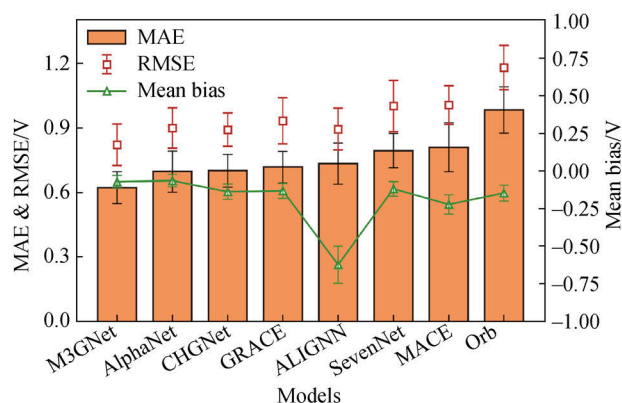


(c)

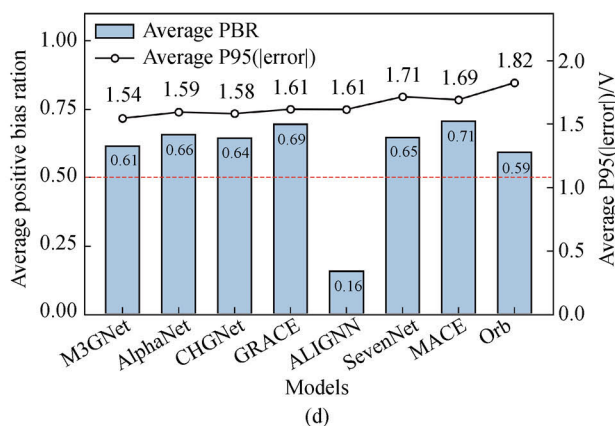
$\int_{x_{\min}}^{x_{\max}} |V_{\text{MLIP}(x)} - V_{\text{ref}(x)}| dx$, approximated discretely as $\sum_{i=1}^N |\Delta V_i| \Delta x_i$. The unit is V·x.

Positive Bias Ratio (PBR): $\text{PBR} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\Delta V_i > 0]$,

the fraction of points where the model overestimates the reference voltage. PBR = 0.5 indicates balanced over-/under-estimation.



(b)



(d)

Fig. 1 Comparison of ML models for voltage prediction. (a) NaMnO₂ model plateaus vs. literature curves^[36]; (b) Overall error levels (MAE/RMSE) and systematic bias; (c) Shape fidelity of V - x curves (IAE and phase-break accuracy); (d) Error sign balance (PBR) and extreme-error risk (P95). Models shown include M3GNet, AlphaNet, CHGNet, GRACE, ALIGNN, SevenNet, MACE, and Orb

3.2 Ion Migration

To assess diffusional kinetics without invoking minimum-energy pathways, we select Li₃Sc₂(PO₄)₃ (mp-6565) as a representative fast-ion host and perform classical molecular dynamics with each MLIP at 600 K for a total trajectory length of 2 ns. Multiple independent initializations are run and ensemble-averaged to reduce initialization variance. Along each trajectory, we accumulate the mean-square displacement (MSD) of Li ions and extract the self-diffusion

coefficient from the diffusive window by linear regression.

The chemical diffusivity is related to microscopic displacements by

$$D = \Theta \frac{1}{2d} \lim_{t \rightarrow \infty} \frac{1}{t} \left\langle \frac{1}{N} \sum_{i=1}^N \|\Delta \mathbf{r}_i(t)\|^2 \right\rangle \quad (4)$$

with dimensionality d , thermodynamic factor $\Theta = \partial \mu / (k_B \partial \ln c)$, and particle displacements $\Delta \mathbf{r}_i(t)$. In the dilute or weakly correlated limit ($\Theta \approx 1$), the tracer (self) diffusivity follows the Einstein

relation^[29–30] For $\text{Li}_3\text{Sc}_2(\text{PO}_4)_3$, Li transport is three-dimensional; therefore we take $d = 3$ when computing the self-diffusion coefficient.

$$D_{\text{self}} = \frac{1}{2d} \frac{d}{dt} \left\langle \frac{1}{N} \sum_{i=1}^N \|\Delta \mathbf{r}_i(t)\|^2 \right\rangle \quad (5)$$

For diagnostics, we partition the MSD into self and collective parts,

$$\text{MSD}_{\text{self}}(t) = \frac{1}{N} \sum_{i=1}^N \|\Delta \mathbf{r}_i(t)\|^2 \quad (6)$$

$$\text{MSD}_{\text{total}}(t) = \frac{1}{N} \left\langle \left\| \sum_{i=1}^N \Delta \mathbf{r}_i(t) \right\|^2 \right\rangle \quad (7)$$

$$\text{MSD}_{\text{distinct}}(t) = \text{MSD}_{\text{total}}(t) - \text{MSD}_{\text{self}}(t) \quad (8)$$

where the Haven ratio $\text{HR} = D_{\text{self}} / D_{\text{total}}$ (with D_{total} inferred from $\text{MSD}_{\text{total}}$) quantifies ion–ion correlations; $\text{MSD}_{\text{distinct}} < 0$ (thus $\text{HR} > 1$) indicates anti-correlated motion that reduces charge transport relative to tracer diffusion^[31].

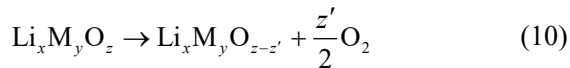
For activated hopping between adjacent sites separated by a , transition-state theory gives

$$k(T) \approx \nu^* \exp\left(-\frac{\Delta G^\ddagger}{k_B T}\right), \quad D \approx g a^2 k(T) \quad (9)$$

with attempt frequency ν^* and geometric factor $g \sim 1$ ^[32]. In the 0 K limit, the activation free energy $\Delta G^\ddagger$ is often approximated by an energy barrier $\Delta E^\ddagger$, leading to Arrhenius behavior $D = D_0 \exp(-E_a / k_B T)$.

3.3 Thermal Stability

Oxygen release in highly charged oxide cathodes is represented by the explicit reaction



for which the reaction free energy (neglecting solid-phase entropy differences at first order) is approximated by

$$\Delta G_r(T) \approx \Delta E - \frac{z'}{2} T S_{\text{O}_2}^\circ(T) \quad (11)$$

with ΔE obtained from MLIP internal energies and $S_{\text{O}_2}^\circ(T)$ the standard molar entropy of O_2 . The onset temperature T^* is defined by $\Delta G_r(T^*) = 0$. To account for atmosphere, we use the oxygen grand potential^[33–34]

$$\phi = G - \mu_{\text{O}_2} N_{\text{O}_2}, \quad \mu_{\text{O}_2}(T, p) = \mu_{\text{O}_2}(T, p^\circ) + k_B T \ln\left(\frac{p}{p^\circ}\right) \quad (12)$$

Here, $\mu_{\text{O}_2}(T, p_{\text{O}_2})$ denotes the intensive chemical

potential of the oxygen reservoir (gas phase) at a given temperature and oxygen partial pressure, and it represents the free-energy cost (per O_2 molecule) of exchanging oxygen with the environment. In contrast, the oxygen grand potential $\phi = G - \mu_{\text{O}_2} N_{\text{O}_2}$ is a Legendre-transformed thermodynamic potential for a system that is open to oxygen uptake/release; it enables a consistent comparison among solid phases with different oxygen contents. At fixed (T, p_{O_2}) , the stable phase minimizes ϕ .

Practically, each oxygen-release step in Fig. 3 corresponds to a crossing of grand potentials (or equivalently, solving $\Delta G_r(T^*) = 0$ for the associated reaction), and the dependence on p_{O_2} can be

incorporated through the ideal-gas term $k_B T \ln\left(\frac{p}{p^\circ}\right)$.

In this benchmark we use ideal-gas O_2 thermochemistry (while neglecting solid-phase entropy differences at first order) and keep a common O_2 reference across MLIPs to avoid confounding errors from gas-phase molecular energetics, enabling stability maps $\text{vs. } (T, p_{\text{O}_2})$ and consistent comparison across phases. In practice, we compute ΔE for reactants and products with each MLIP, determine $\Delta G_r(T)$ from Eq. (11) [optionally using $\mu_{\text{O}_2}(T, p)$ from Eq. (12)], solve $\Delta G_r(T^*) = 0$, and report T vs. moles O_2 released curves for benchmarking against reference trends^[35].

4 Results and discussion

4.1 Voltage Profiles

The model-derived staircase voltage curves broadly reproduce the expected multi-plateau behavior during Na extraction. Around shallow desodiation, several models start slightly above the averaged reference, while in the mid-to-deep regime some curves fall below it, indicating a composition-dependent bias. The number and positions of breakpoints vary across models, reflecting differences in how each potential ranks close-lying cation-vacancy orders.

Beyond NaMnO_2 , we observe qualitatively similar model-to-model voltage behavior for NaFePO_4 and $\text{Na}_3\text{V}_2(\text{PO}_4)_3$: the relative ranking and the dominant bias patterns are largely preserved across these Na cathodes, suggesting that the voltage-

performance disparity is not unique to a single chemistry. By contrast, for LiCoO_2 nearly all MLIPs tend to predict systematically lower voltages than the literature/DFT reference curve, indicating that this layered Co–O system is a more challenging regime for current pretrained potentials. This systematic underestimation contributes to the cross-material bias metrics and should be considered when interpreting averaged performance.

Across four materials, M3GNet attains the lowest MAE and a modest RMSE with a small mean bias. Alphanet and CHGNet follow closely with mild negative bias. GRACE is mid-tier. ALIGNN shows a sizable negative bias despite a moderate MAE, suggesting a near-constant offset across compositions. SevenNet and MACE exhibit higher error, and Orb performs worst.

The difference between MAE and RMSE reflects the shape of the error distribution: RMSE penalizes occasional large deviations more strongly than MAE. For M3GNet, the low average MAE together with a comparatively higher RMSE indicates that most voltage points are predicted with small-to-moderate error, but a small number of compositions [typically near phase-boundary segments or deep (de) intercalation states] show larger deviations that disproportionately increase RMSE. This behavior is consistent with the convex-hull nature of the workflow: a single mis-ranked low-energy configuration can change the selected hull vertices and shift plateau boundaries, creating localized but relatively large voltage errors while leaving the rest of the curve accurate. In contrast, models such as ALIGNN show a very low PBR and large negative mean bias, suggesting a more systematic underestimation of voltages across compositions, which can arise from consistent energy offsets between alkali-rich/poor endmembers or imperfect transferability of Na/Li reference energetics. Orb shows larger MAE/RMSE and larger breakpoint/plateau errors, implying that it more frequently misorders near-degenerate alkali/vacancy configurations; since voltages are computed from small energy differences, such ranking errors are amplified into noticeable plateau-level and topology mismatches in the predicted V - x curves.

IAE captures integrated deviations along composition, while the breakpoint error probes phase-boundary localization. M3GNet, AlphaNet and SevenNet sit in the low-IAE/low-breakpoint-error corner (IAE $0.63 \text{ V}\cdot x$ – $0.69 \text{ V}\cdot x$; breakpoint 0.09 – 0.10 V), indicating good tracking of both plateau levels and transitions. CHGNet is comparable in IAE but with slightly larger breakpoint error (0.14 V). MACE has moderate IAE yet noticeably larger breakpoint error (0.28 V), consistent with occasional misplacement or merging of plateaus. ALIGNN and Orb show the largest breakpoint and plateau errors and larger step-count differences, implying topology mismatches in the predicted V - x curves.

Most models have $\text{PBR} > 0.5$ (more overestimates than underestimates), whereas ALIGNN shows a very low PBR (0.16), consistent with its strong negative mean bias in panel (b). The average $\text{P95}(|\text{error}|)$ ranges from 1.54 V to 1.82 V ; Orb and SevenNet show the largest tail risk, whereas M3GNet and AlphaNet are on the lower end. These statistics emphasize that beyond average accuracy, robustness to outliers is model-dependent.

4.2 Ion Migration

Fig. 2 shows the Li self-MSD traces at 600 K for $\text{Li}_3\text{Sc}_2(\text{PO}_4)_3$ (mp-6565) across the evaluated MLIPs. To minimize early-time transients and late-time noise, the diffusive window used for linear regression is restricted to 200 – 1500 ps ; the slope in this interval yields D_{self} via Eqs. (4)–(5). The reference diffusivity is taken as $3.0 \times 10^{-6} \text{ cm}^2/\text{s}$ for comparison. Overall, traces that remain near-linear and monotonic in this window indicate more stable sampling and typically produce more reliable D estimates.

Table 2 reports the fitted diffusivities and their multiplicative deviation from the reference. The closest-to-reference estimates are obtained by MACE and SevenNet, with Orb and GRACE forming the next cluster. CHGNet and M3GNet overestimate D by a fewfold, while AlphaNet underestimates by an order of magnitude. All models remain within two orders of magnitude of the reference, but the tighter grouping around unity suggests comparatively stronger kinetic fidelity for the former set under these conditions. The

ALIGNN model was excluded from the diffusion benchmark due to repeated out-of-memory and convergence failures encountered during long-timescale molecular dynamics runs. Despite multiple attempts under reduced supercell sizes and shorter trajectories, the ALIGNN calculator consistently exhausted GPU memory resources. As a result, diffusivity data for ALIGNN are not reported. This limitation likely stems from the model's relatively high memory footprint during force evaluations, a known issue when applying ALIGNN to large-scale MD simulations.

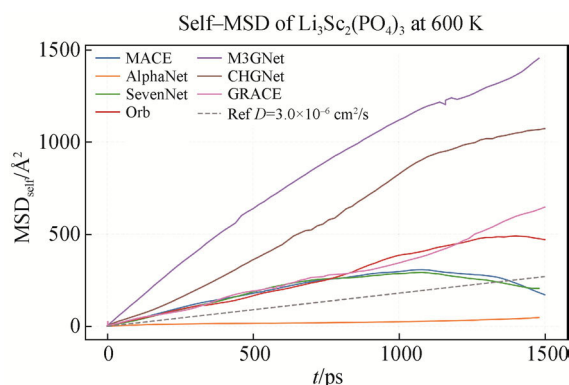


Fig. 2 Self-MSD of $\text{Li}_3\text{Sc}_2(\text{PO}_4)_3$ at 600 K for each MLIP; the regression window (200–1500 ps) is used to extract $D_{\text{self}}^{[37]}$

Table 2 Diffusivity at 600 K from MSD-based regression for $\text{Li}_3\text{Sc}_2(\text{PO}_4)_3$

Model	Fitted $D/(\text{cm}^2\cdot\text{s}^{-1})$	Reference/ $(\text{cm}^2\cdot\text{s}^{-1})$	Factor/Ref.
AlphaNet	4.1×10^{-7}	3.0×10^{-6}	0.14
CHGNet	1.3×10^{-5}	3.0×10^{-6}	4.33
GRACE	7.0×10^{-6}	3.0×10^{-6}	2.33
MACE	1.9×10^{-6}	3.0×10^{-6}	0.63
M3GNet	1.5×10^{-5}	3.0×10^{-6}	5.00
Orb	6.0×10^{-6}	3.0×10^{-6}	2.00
SevenNet	1.8×10^{-6}	3.0×10^{-6}	0.60

Taken together, the MSD stability and the quantitative D errors point to a consistent picture: models whose trajectories maintain a steady diffusive slope within the selected window also yield D values closest to the reference, whereas pronounced curvature or oscillations tend to accompany over-/under-estimation. These trends are complementary to the voltage-profile benchmarking and highlight task-dependent strengths

among MLIPs under identical MD protocols.

The apparent inconsistency, MACE and SevenNet ranking lower for voltage yet yielding the closest-to-reference diffusivity, highlights that thermodynamic and kinetic fidelity probe different aspects of an MLIP. Voltage profiles depend on small energy differences between competing configurations across compositions (hull vertices and their relative ordering); a modest systematic energy bias or occasional misranking of near-degenerate Na/Li–vacancy orderings can shift hull slopes and manifest as plateau-level errors. In contrast, MSD-based diffusivity is extracted from force-driven trajectories at a fixed composition and temperature and is primarily controlled by the local energy landscape and its gradients along Li–hopping pathways. Architectures such as MACE and SevenNet, which are strongly symmetry-equivariant and typically trained with force matching on diverse local environments, tend to produce smooth and accurate short-range forces, enabling stable long-time MD and reliable MSD slopes even when absolute energies are less accurate. This decoupling suggests that selecting an MLIP should be task-specific: models optimized for thermodynamics (voltage/hull) are not necessarily those best suited for migration kinetics. Since diffusivity depends exponentially on migration barriers (*via* Arrhenius behavior), accurate local barrier shapes and forces near transition states can outweigh modest offsets in absolute energies for predicting MSD slopes at a given temperature.

4.3 Thermal Stability

Thermal stability analysis was performed to evaluate the ability of each machine-learning potential to reproduce oxygen-release behavior and decomposition temperatures in representative transition-metal phosphates and oxides. Fig. 3 summarizes the predicted stepwise oxygen evolution for three prototypical systems: MnPO_4 , olivine FePO_4 , and trigonal FePO_4 , together with experimental reference data. Each staircase curve corresponds to the cumulative amount of O_2 evolved per formula unit as a function of temperature, where distinct plateaus represent discrete reduction steps of the transition-metal cations and the concurrent formation of pyrophosphate phases.

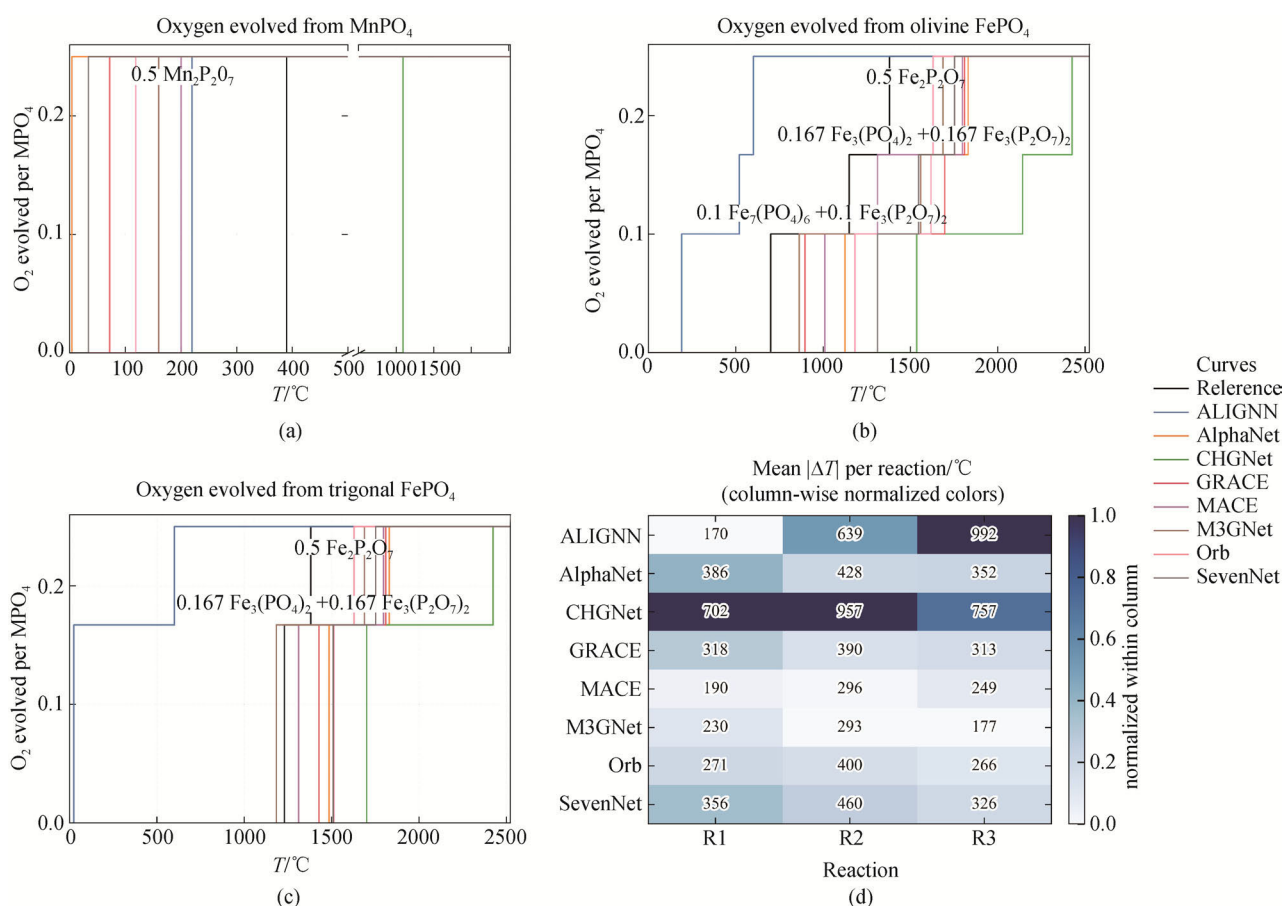


Fig. 3 Comparison between experimental and predicted oxygen-evolution profiles for phosphate systems. Panels (a–c) show the temperature-dependent O₂ release from (a) MnPO₄, (b) olivine FePO₄, and (c) trigonal FePO₄. The black solid line represents the experimental reference^[34]; colored thin lines correspond to eight machine-learning interatomic potentials (ALIGNN, AlphaNet, CHGNet, GRACE, MACE, M3GNet, Orb, and SevenNet). Step annotations indicate the intermediate and final decomposition products. Panel (d) shows a column-normalized heatmap of mean absolute temperature deviations ($|\Delta T|$) for each model and reaction

Across all systems, the ML potentials qualitatively reproduce the experimental multi-step oxygen-release patterns but differ in quantitative accuracy. For MnPO₄, most models capture the single-step oxygen evolution leading to 0.5 Mn₂P₂O₇, with onset temperatures ranging from 4 °C to 1100 °C compared to the experimental 390 °C. The ALIGNN and M3GNet potentials show the closest agreement, yielding deviations below 200 °C. In contrast, AlphaNet and CHGNet significantly overpredict T^* , reflecting their tendency to over-stabilize oxidized Mn³⁺/Mn⁴⁺ states.

The olivine FePO₄ system exhibits three characteristic plateaus experimentally observed at 700, 1150, and 1380 °C. GRACE and MACE successfully reproduce the intermediate step near 1300 °C, while other models either merge or misplace the first transition, resulting in a compressed thermal window.

This outcome suggests that the representation of Fe–O bond energetics and phosphate polymerization pathways remains challenging for purely data-driven force fields, particularly under oxygen-deficient conditions.

For trigonal FePO₄, two major decomposition steps are observed at 1230 °C and 1380 °C. Here, M3GNet and ALIGNN again achieve the smallest errors, accurately reproducing both plateau heights and step positions. The consistent trends between olivine and trigonal FePO₄ highlight that network architectures incorporating explicit three-body or angular terms can better capture the cooperative Fe–O bond rearrangements during oxygen loss.

Quantitatively, the mean absolute temperature deviations in Fig. 3d show that M3GNet and MACE achieve the best overall fidelity, with $|\Delta T| < 300$ °C across all three reactions. GRACE performs

comparably well in reproducing intermediate-temperature transitions, while CHGNet and AlphaNet systematically overestimate decomposition temperatures by several hundred degrees. SevenNet and Orb exhibit moderate-to-large scatter, likely due to their limited high-temperature training coverage. These trends parallel those observed in voltage and diffusion analyses, reinforcing the notion that generalization to extreme thermodynamic conditions requires balanced datasets spanning both equilibrium and metastable configurations.

In summary, M3GNet, ALIGNN, and MACE consistently rank as the most reliable models for predicting high-temperature oxygen evolution, providing both qualitative and quantitative agreement with experimental references. GRACE follows closely, capturing intermediate features with slightly larger deviations, whereas CHGNet and AlphaNet systematically overbind and delay decomposition. The overall results demonstrate that current MLIPs can successfully reproduce the sequence and mechanism of oxygen evolution, though achieving quantitative accuracy in absolute decomposition temperatures remains challenging. Incorporating explicit entropic contributions and oxygen chemical potentials into future MLIP training frameworks will further enhance predictive capability for cathode stability and safety assessments.

4.4 Cross-Model Comparison and Outlook

The observed variations among the eight machine-learning interatomic potentials (MLIPs) can be rationalized by their architectural design, feature representation, and the breadth of their training corpora. In the voltage benchmark, M3GNet achieves the lowest average error, suggesting strong thermodynamic fidelity for ranking near-degenerate alkali/vacancy configurations, while AlphaNet and CHGNet are close followers. In the ion-migration benchmark on $\text{Li}_3\text{Sc}_2(\text{PO}_4)_3$, MACE and SevenNet yield diffusivities closest to the reference value, indicating strong kinetic fidelity under long-timescale MD, with Orb and GRACE forming the next cluster. For thermal stability, M3GNet and MACE best reproduce oxygen-release temperatures and stepwise

decomposition patterns, whereas CHGNet and AlphaNet systematically delay decomposition. These performance patterns highlight that no single model is uniformly optimal across all regimes: the predictive fidelity depends jointly on architectural expressiveness and the diversity of training data encompassing equilibrium, metastable, and high-temperature configurations. Future improvements should combine the physical completeness of equivariant frameworks with data diversification, particularly inclusion of charged, disordered, and finite-temperature states, and explore hybrid or multitask training that couples energy, force, and electronic descriptors. Such developments would enhance both transferability and robustness of universal MLIPs for electrochemical applications.

5 Conclusions

In this study, we systematically evaluated eight machine-learning interatomic potentials including MACE, M3GNet, CHGNet, ALIGNN, SevenNet, AlphaNet, GRACE, and Orb on representative Na- and Li-based cathode materials. A unified assessment was conducted across three key regimes: voltage profiles, ion migration, and thermal stability.

The voltage-profile benchmark indicates that M3GNet achieves the lowest average voltage MAE across the four cathodes considered here, with AlphaNet and CHGNet close behind. However, the best model can vary with host chemistry, and LiCoO_2 presents a pronounced negative-bias challenge for all tested MLIPs. For ion migration in $\text{Li}_3\text{Sc}_2(\text{PO}_4)_3$, MACE and SevenNet yield the closest-to-reference diffusivity, whereas some models over-/under-estimate by several fold. For the thermal-stability benchmark, MACE and M3GNet provide the most accurate decomposition temperatures and stepwise oxygen-release patterns. Overall, voltage fidelity and diffusivity fidelity are not strictly correlated, underscoring the need for task- and system-aware MLIP selection.

Taken together, these benchmarks highlight that transferable accuracy in MLIPs depends jointly on architectural design and on the match between a model's training distribution and the target

chemistry/structure/temperature regime. Because each task is benchmarked on a representative set of material families dictated by available reference data, the cross-task comparisons should be interpreted as task-oriented guidance rather than a fully decoupled, property-only ranking. The unified workflow established here provides a practical template for multi-objective evaluation, and it can be extended to larger and more overlapping benchmark sets to further separate property dependence from system dependence in future studies.

6 Acknowledgement

The authors thank Prof. Gian-Marco Rignanese from Université catholique de Louvain, and Prof. Xiaotong Liu from Beijing Information Science and Technology University for helpful discussions.

References:

- [1] TARASCON J M, ARMAND M. Issues and challenges facing rechargeable lithium batteries[J]. *Nature*, 2001, 414(6861): 359–367.
- [2] SLATER M D, KIM D, LEE E, et al. Sodium-ion batteries[J]. *Adv Funct Mater*, 2013, 23(8): 947–958.
- [3] GOEDECKER S. Linear scaling electronic structure methods[J]. *Rev Mod Phys*, 1999, 71(4): 1085–1123.
- [4] BOWLER D R, MIYAZAKI T. $\mathcal{O}(N)$ methods in electronic structure calculations[J]. *Rep Prog Phys*, 2012, 75(3): 036503.
- [5] MARX D, HUTTER J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*[M]. Cambridge, UK: Cambridge University Press, 2009.
- [6] CAR R, PARRINELLO M. Unified approach for molecular dynamics and density-functional theory[J]. *Phys Rev Lett*, 1985, 55(22): 2471–2474.
- [7] JAIN A, ONG S P, HAUTIER G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation[J]. *APL Mater*, 2013, 1: 011002.
- [8] CHOUDHARY K, GARRITY K F, REID A C E, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design[J]. *npj Comput Mater*, 2020, 6: 173.
- [9] XIAO B, TANG Y C, LIU Y. Integrating materials representations into feature engineering in machine learning for crystalline materials: From local to global chemistry-structure information coupling[J]. *WIREs Comput Mol Sci*, 2025, 15(4): e70044.
- [10] ONG S P, RICHARDS W D, JAIN A, et al. Python Materials Genomics (pymatgen): A robust, open-source Python library for materials analysis[J]. *Comput Mater Sci*, 2013, 68: 314–319.
- [11] BATATIA I, KOVACS D P, SIMM G, et al. MACE: higher order equivariant message passing neural networks for fast and accurate force fields [C]//Advances in Neural Information Processing Systems, 2022, 35: 11423–11436.
- [12] CHOUDHARY K, DECOST B. Atomistic Line Graph Neural Network for improved materials property predictions[J]. *npj Comput Mater*, 2021, 7: 185.
- [13] HJORTH LARSEN A, JØRGEN MORTENSEN J, BLOMQVIST J, et al. The atomic simulation environment: A Python library for working with atoms[J]. *J Phys: Condens Matter*, 2017, 29(27): 273002.
- [14] PLIMPTON S. Fast parallel algorithms for short-range molecular dynamics[J]. *J Comput Phys*, 1995, 117(1): 1–19.
- [15] CHEN C, ONG S P. A universal graph deep learning interatomic potential for the periodic table[J]. *Nat Comput Sci*, 2022, 2(11): 718–728.
- [16] KO T W, DENG B W, NASSAR M, et al. Materials Graph Library (MatGL), an open-source graph deep learning library for materials science and chemistry[J]. *npj Comput Mater*, 2025, 11: 253.
- [17] DENG B W, ZHONG P C, JUN K, et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling[J]. *Nat Mach Intell*, 2023, 5(9): 1031–1041.
- [18] PARK Y, KIM J, HWANG S, et al. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations[J]. *J Chem Theory Comput*, 2024, 20(11): 4857–4868.
- [19] YIN B C, WANG J A, DU W T, et al. AlphaNet: Scaling up local-frame-based atomistic interatomic potential[EB/OL]. 2025: arXiv: 2501.07155. <https://arxiv.org/abs/2501.07155>
- [20] RHODES B, VANDENHAUTE S, ŠIMKUS V, et al. Orb-v3: Atomistic simulation at scale[EB/OL]. 2025: arXiv: 2504.06231. <https://arxiv.org/abs/2504.06231>
- [21] BOCHKAREV A, LYSOGORSKIY Y, DRAUTZ R. Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing[J]. *Phys Rev X*, 2024, 14(2): 021036.
- [22] BILLAUD J, CLÉMENT R J, ARMSTRONG A R, et al. β -NaMnO₂: A high-performance cathode for sodium-ion batteries[J]. *J Am Chem Soc*, 2014, 136(49): 17243–17248.
- [23] MIZUSHIMA K, JONES P C, WISEMAN P J, et al. Li_xCoO₂ (0<x<1): A new cathode material for batteries of high energy density[J]. *Mater Res Bull*, 1980, 15(6): 783–789.
- [24] FANG Y J, ZHANG J X, XIAO L F, et al. Phosphate framework electrode materials for sodium ion batteries[J]. *Adv Sci*, 2017, 4(5): 1600392.
- [25] LIM S Y, KIM H, SHAKOOR R A, et al. Electrochemical and thermal properties of NASICON structured Na₃V₂(PO₄)₃ as a sodium rechargeable battery cathode: A combined experimental and theoretical study[J]. *J Electrochem Soc*, 2012, 159(9): A1393–A1397.
- [26] BARD A J, FAULKNER L R, WHITE H S. *Electrochemical Methods*[M]. Third edition. Hoboken: John Wiley & Sons, 2022.
- [27] AYDINOL M K, KOHAN A F, CEDER G, et al. *Ab initio* study of lithium intercalation in metal oxides and metal dichalcogenides[J]. *Phys Rev B*, 1997, 56(3): 1354–1365.
- [28] SUN W H, DACEK S T, ONG S P, et al. The thermodynamic scale of inorganic crystalline metastability[J]. *Sci Adv*, 2016, 2(11): e1600225.
- [29] FRENKEL D, SMIT B, RATNER M A. Understanding molecular simulation: From algorithms to applications[J]. *Phys Today*, 1997, 50(7): 66.
- [30] MAGINN E J, MESSERLY R A, CARLSON D J, et al. Best practices for computing transport properties 1. self-diffusivity and viscosity from equilibrium molecular dynamics [article v1.0][J]. *Living J Comp Mol Sci*, 2019, 1(1): 6324.
- [31] VARGAS-BARBOSA N M, ROLING B. Dynamic ion correlations in solid and liquid electrolytes: How do they affect charge and mass transport?[J]. *ChemElectroChem*, 2020, 7(2): 367–385.
- [32] VINEYARD G H. Frequency factors and isotope effects in solid state

- rate processes[J]. *J Phys Chem Solids*, 1957, 3(1–2): 121–127.
- [33] REUTER K, SCHEFFLER M. First-principles atomistic thermodynamics for oxidation catalysis: Surface phase diagrams and catalytically interesting regions[J]. *Phys Rev Lett*, 2003, 90(4): 046103.
- [34] URBAN A, SEO D H, CEDER G. Computational understanding of Li-ion batteries[J]. *npj Comput Mater*, 2016, 2: 16002.
- [35] WANG L, MAXISCH T, CEDER G. Oxidation energies of transition metal oxides within the GGA+U framework[J]. *Phys Rev B*, 2006, 73(19): 195107.
- [36] MA X H, CHEN H L, CEDER G. Electrochemical properties of monoclinic NaMnO₂[J]. *J Electrochem Soc*, 2011, 158(12): A1307.

- [37] WANG C H, AOYAGI K, WISESA P, et al. Lithium ion conduction in cathode coating materials from on-the-fly machine learning[J]. *Chem Mater*, 2020, 32(9): 3741–3752.

作者贡献声明:

蒋冰鉴: 实验测试, 设计论文框架, 撰写论文

杨涛: 提出研究方向, 提供实验设备与环境, 完善论文框架, 修订论文

路宽: 提供理论指导帮助, 参与数据分析

张艺馨: 绘制图表, 参与数据分析

Comparative Evaluation of Machine-Learning Potentials for Electrochemical Properties of Cathodes

JIANG Bingjian^{1,2}, YANG Tao^{1,2}, LU Kuan³, ZHANG Yixin⁴

(1. School of Computer, Beijing Information Science and Technology University, Beijing 100101, China; 2. Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, China; 3. Institute of Coal Chemistry, Chinese Academy of Sciences, State Key Laboratory of Coal Conversion, Taiyuan 030001, China; 4. School of Chemistry and Chemical Engineering, Ningxia University, Yinchuan 750021, China)

Extended Abstract

Introduction Lithium-ion and sodium-ion batteries underpin portable electronics, electric vehicles, and grid-scale energy storage. Their performance and safety are strongly constrained by cathode behavior, and practical screening repeatedly requires three types of predictions, *i.e.*, voltage profiles during (de)intercalation, alkali-ion migration kinetics, and thermal stability against oxygen release and decomposition. Density functional theory (DFT) can provide accurate baselines, but its cost limits supercell sizes, configuration sampling, and the time/length scales accessible to *ab initio* molecular dynamics (AIMD). Pretrained machine-learning interatomic potentials (MLIPs) offer near-DFT accuracy at far lower cost, thus enabling high-throughput evaluation and long-time molecular dynamics. However, the fidelity required for voltage thermodynamics, migration kinetics, and high-temperature decomposition is not identical, and a model that performs well on one task may be unreliable on another. In this study, eight representative pretrained MLIPs (*i.e.*, MACE, M3GNet, CHGNet, ALIGNN, SevenNet, AlphaNet, GRACE, and Orb) were systematically benchmarked across these three cathode-relevant tasks under matched protocols.

Methods All crystallographic prototypes were obtained from the Materials Project. All the MLIPs were used as released pretrained models without fine-tuning to reflect a “drop-in” screening scenario. For voltage profiles, four representative cathodes spanning layered oxides and polyanion frameworks were selected, *i.e.*, NaMnO₂, LiCoO₂, NaFePO₄, and Na₃V₂(PO₄)₃. Each parent structure was expanded to a supercell containing 8–12 alkali sites. Partially (de)intercalated configurations were generated by exhaustive enumeration of alkali-vacancy patterns and symmetry-deduplicated with a space-group analysis tool. Each MLIP evaluated energies for all configurations. Formation energies were referenced to alkali-rich/poor endmembers, and lower convex hulls were constructed to obtain piecewise average voltages. Accuracy and curve-shape fidelity were quantified using metrics including mean absolute error (MAE), root mean square error (RMSE), integrated absolute error (IAE), and bias indicators.

For ion migration, Li₃Sc₂(PO₄)₃ was chosen as a representative fast-ion conductor. Molecular dynamics was performed at 600 K for 2 ns with each MLIP, and the Li self-diffusion coefficient was extracted from the mean-square displacement (MSD) *via* the Einstein relation in three dimensions. A fixed regression window (200–1500 ps) was applied consistently to reduce transient and finite-sampling effects.

For thermal stability, oxygen-release/decomposition was benchmarked on phosphate systems with clear stepwise pathways, *i.e.*, MnPO₄ and FePO₄ in olivine and trigonal polymorphs. Each step was represented by balanced solid-state reactions with explicit O₂ release. Reaction energies from MLIPs were combined with an ideal-gas oxygen chemical potential in an oxygen grand-potential framework to estimate onset temperatures and cumulative O₂-evolution curves.

Results and discussion The benchmark reveals a pronounced task dependence among pretrained MLIPs. For voltage profiles, most models reproduce the qualitative multi-plateau behavior, yet quantitative accuracy differs because voltages are sensitive to small energy differences among near-degenerate vacancy orderings that determine convex-hull vertices and plateau boundaries. Across the four cathodes, M3GNet achieves the minimum overall voltage error with a small systematic bias, followed by AlphaNet and CHGNet. GRACE is generally mid-tier, whereas Orb and ALIGNN exhibit larger offsets and shape errors. A notable system dependence occurs for LiCoO₂, where most MLIPs tend to underestimate voltages relative to the reference, indicating that layered Co–O chemistry

remains a challenging regime for the existing universal pretrained models.

For ion migration in $\text{Li}_3\text{Sc}_2(\text{PO}_4)_3$ at 600 K, the ranking differs from the voltage benchmark. MACE and SevenNet yield diffusivities closest to the reference, with Orb and GRACE moderately higher. CHGNet and M3GNet overestimate a diffusivity by severalfold, while AlphaNet underestimates it more strongly. Importantly, ALIGNN can not reliably complete the long-timescale molecular-dynamics runs under our resource constraints, highlighting a practical limitation (memory footprint) that can matter in screening pipelines. These support the interpretation that diffusion is governed primarily by local force accuracy and barrier-region smoothness along hopping pathways, whereas voltage profiles depend on consistent ranking of many competing configurations across composition.

For thermal stability, all the MLIPs qualitatively reproduce the existence of stepwise oxygen evolution and pyrophosphate-forming decomposition, but predicted onset temperatures can deviate substantially. For MnPO_4 , deviations span a wide range. ALIGNN and M3GNet provide the smallest deviations for this system, while AlphaNet and CHGNet delay oxygen release. Across the full phosphate set, M3GNet and MACE deliver the optimum overall agreement with experimental step positions and temperatures, and GRACE often captures intermediate steps with moderate errors. The results indicate that generalization to high-temperature, oxygen-deficient pathways remains sensitive to both training coverage and architectural bias.

Conclusions This work established a unified, extensible benchmark for comparing pretrained MLIPs on cathode-relevant thermodynamics, kinetics, and high-temperature stability. No single MLIP could be universally optimal. M3GNet was most reliable for voltage prediction on the tested cathodes, MACE and SevenNet were strongest for long-time MSD-based diffusion in $\text{Li}_3\text{Sc}_2(\text{PO}_4)_3$, and M3GNet together with MACE best captured oxygen-release/decomposition trends in representative phosphate systems. The proposed workflow could provide a practical guidance for task-aware MLIP selection and a foundation for larger, more decoupled benchmarks in future studies.

Keywords machine-learning interatomic potentials; sodium-ion batteries; lithium-ion batteries; voltage profile; ion diffusion; thermal stability