

ORIGINAL RESEARCH

Open access

From DFT to ML: A Hierarchical Framework for Validating GNN Force Fields Against Phonon Spectra Without Spectral Leakage

Nguyen Thanh Huy^{1*}, Pham Quang Minh¹, Le Thi Bich²

Abstract

Phonon spectra offer a uniquely demanding benchmark for graph neural network (GNN) force fields because they interrogate the second-derivative structure of the potential energy surface that governs lattice dynamics, thermal transport, and vibrational stability in materials. Yet direct comparison between GNN-derived and density functional theory (DFT) phonon spectra is frequently compromised by spectral leakage introduced through finite supercell truncation, displacement amplitude selection, q-point undersampling, Fourier interpolation, and post-processing broadening. These numerical effects can either conceal genuine deficiencies in the learned force field or generate apparent discrepancies that do not reflect model behavior. This article develops a hierarchical, leakage-aware validation framework that addresses this problem through progressive levels of scrutiny. The framework begins with baseline agreement in energies and forces, then advances to phonon density of states validation, q-resolved dispersion analysis, and finally a reproducible multi-metric assessment of spectral similarity. Progression through the hierarchy is conditional rather than automatic, such that higher-level claims are only made once lower-level numerical stability and model fidelity have been established. To separate methodological artifact from true representational error, the framework embeds explicit diagnostics based on supercell convergence, displacement sweeps, q-mesh refinement, interpolation cross-checks, and residual spectral analysis. It further introduces a standardized reporting protocol designed to make phonon-based validation transparent, comparable, and reproducible across studies of machine-learning interatomic potentials. By formalizing leakage control as an integral part of validation rather than an afterthought, the framework closes a critical gap between high-fidelity DFT phonon workflows and contemporary ML force-field development, enabling more credible assessment of GNN transferability in computational materials science.

Keywords Machine learning interatomic potentials, Graph neural network force fields, Phonon spectra validation, Spectral leakage, Hierarchical framework, Density functional perturbation theory

*Correspondence:

Nguyen Thanh Huy
huy.nguyen@gmail.com

¹ Department of Intelligent Materials Engineering, Faculty of Engineering, Vietnam National University, Hanoi, Vietnam

² Department of Computational Materials Systems, Faculty of Technology, Can Tho University, Can Tho, Vietnam

Introduction

Graph neural network force fields have emerged as powerful tools for atomistic simulation in materials science, offering near-DFT accuracy at dramatically reduced computational cost [1–6]. Contemporary GNN architectures [7] such as those introduced by Batzner *et al.* [8], Chen *et al.* [9], and Choudhary and DeCost [10] achieve remarkable performance on energy and force prediction tasks, with root-mean-square errors often falling below 0.05 eV/Å on held-out test sets. Yet

validation pipelines in most published studies remain anchored at the first derivatives of the potential energy surface. Energy and force errors, typically reported as mean absolute error (MAE) or root-mean-square error (RMSE), constitute Level 1 checks that are necessary but fundamentally insufficient for guaranteeing physically faithful vibrational behavior.

Phonons, which encode the second derivatives of the potential energy surface, reveal whether a GNN model correctly captures interatomic force constants across the entire Brillouin zone. Giustino established the foundational formalism for first-principles phonon calculations that remains the de-facto standard in DFT workflows [11]. When the same formalism is applied to forces predicted by a GNN rather than DFT [12], the resulting phonon spectra can diverge dramatically even when force RMSE appears acceptable. The central problem addressed in this framework is therefore clear: low energy/force error does not guarantee accurate phonon spectra because phonons probe long-range interactions, anharmonic effects, and curvature that may lie outside the training distribution.

Spectral leakage exacerbates this challenge. Artifacts arising from finite supercell truncation, displacement magnitude, q-point mesh density, and Fourier interpolation can obscure whether observed discrepancies originate from model error or from numerical inconsistencies in the validation pipeline itself. The placeholder study by Botu *et al.* already highlighted that “naive comparison of DFT and ML phonon spectra may show agreement due to leakage artifacts masking real model errors, or disagreement due to validation artifacts not model errors” [13]. Existing literature on machine-learning force fields [14]—including Deep Potential [15], SchNet [16], and equivariant GNNs [8, 17]—rarely reports phonon validation, and when it does, the protocols lack systematic controls for leakage.

Figure 1 presents the hierarchical, leakage-aware validation workflow, illustrating the strict progression from energy/force agreement to fully quantitative spectral validation under controlled numerical conditions.

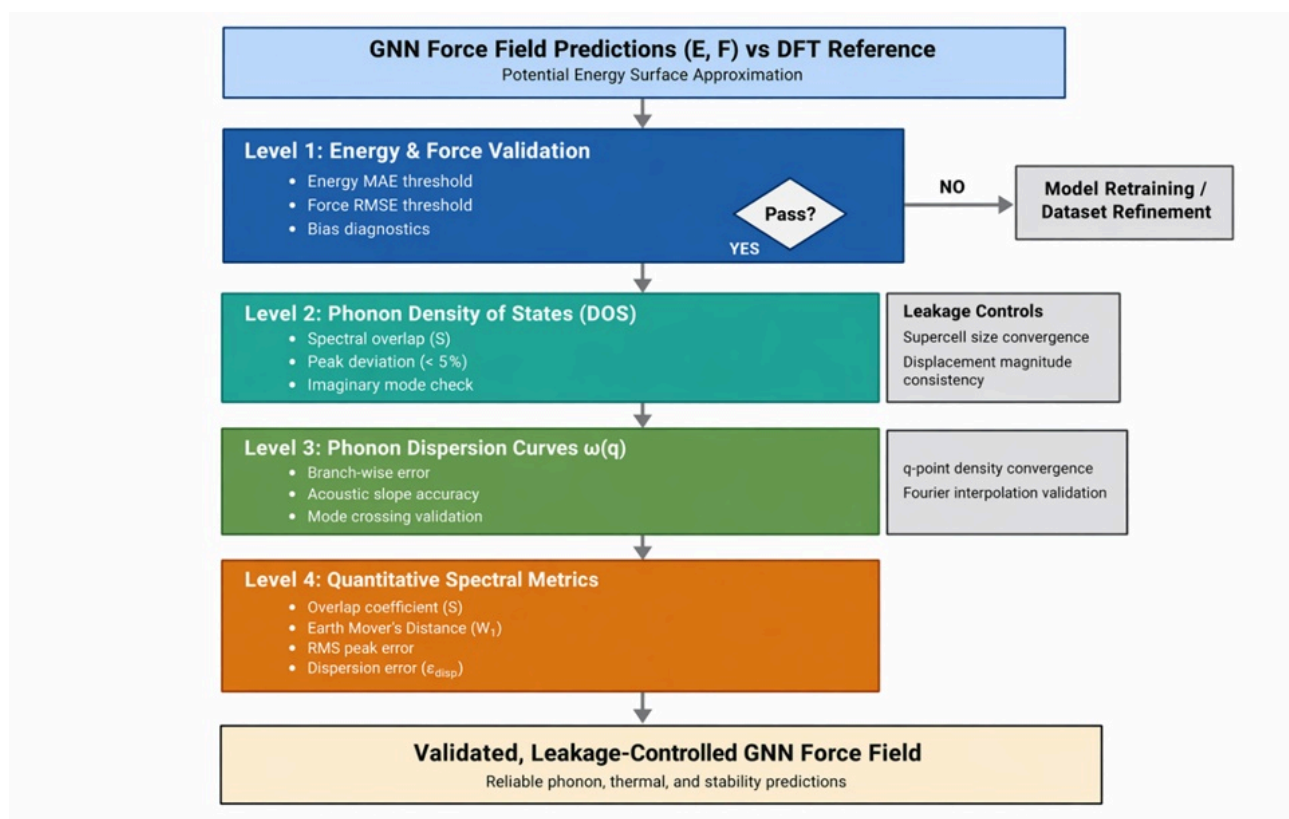


Figure 1. Hierarchical leakage-aware validation framework for GNN force fields

The structure proceeds from baseline energy/force checks (Level 1) through integral spectral properties (Level 2: phonon DOS), q-resolved dispersion relations (Level 3), and finally quantitative spectral metrics (Level 4). At each level, explicit procedures for detecting and mitigating spectral leakage are embedded. The framework draws on established phonon methodologies [11, 18] while adapting them to the stochastic, cutoff-limited nature of GNN predictions. By requiring progressive passage through all four levels, the protocol ensures that a GNN force field is not only energetically accurate but also vibrationally faithful. Such rigor is essential because phonon-derived quantities—heat capacity, thermal conductivity, and dynamical stability—directly impact materials design for thermoelectrics, batteries, and quantum technologies [19].

The remainder of the paper defines spectral leakage in the phonon-ML context, articulates why phonon validation transcends conventional metrics, details the four-level hierarchy with mathematical foundations, presents leakage-detection methods, and codifies a reporting standard that future studies can adopt without ambiguity. In doing so, the framework moves the field from ad-hoc visual inspection of phonon plots toward a reproducible, leakage-aware validation culture.

Table 1 formalizes the hierarchical validation logic, explicitly linking each level to its physical target, quantitative criteria, and associated spectral-leakage controls.”

Table 1. Hierarchical Validation Criteria and Leakage-Controlled Progression across Four Levels

Validation Level	Physical Quantity Probed	Core Validation Metrics	Acceptance Criteria	Leakage Sources Controlled	Progression Condition
Level 1: Energy & Force	Zeroth and first derivatives of PES	Energy MAE, Force RMSE, bias analysis	MAE < 0.01 eV/atom; RMSE < 0.05 eV/Å	None (baseline only)	Must pass before phonon calculations
Level 2: Phonon DOS	Integrated vibrational spectrum	Spectral overlap (S), peak deviation	S > 0.85; peak shift < 5%; no imaginary modes	Supercell size, displacement magnitude	Requires Level 1 compliance
Level 3: Dispersion Curves	q-resolved phonon frequencies	Branch-wise error, acoustic slope deviation	Max error < 10%; no spurious crossings	q-point sampling, interpolation artifacts	Requires Level 2 validation
Level 4: Spectral Metrics	Quantitative spectral agreement	Overlap (S), EMD (W ₁), RMS peak error	S > 0.85; W ₁ < 0.1 THz; ε _{disp} < 0.5 THz	All leakage sources explicitly quantified	Final validation stage

Why Phonon Spectra Validation Matters

Phonon spectra constitute a stringent and physically meaningful test of machine-learning force fields that energy and force errors alone cannot adequately provide. Several interconnected reasons, rooted in the hierarchy of derivatives of the potential energy surface, establish why vibrational validation is indispensable for any model intended to describe lattice dynamics or finite-temperature behavior.

Central to this argument is the observation that phonons probe the curvature of the potential energy surface. Whereas energy errors assess the zeroth-order value and force errors assess the first derivatives, phonon frequencies are eigenvalues of the Hessian matrix of second derivatives. A graph neural network can therefore exhibit low force root-mean-square error yet systematically misrepresent local curvature, leading to vibrational modes that are uniformly shifted

or artificially broadened. Zhang and colleagues demonstrated that the Deep Potential architecture achieves excellent force accuracy [15], but subsequent phonon studies have occasionally revealed residual curvature mismatches that force-level metrics entirely miss—a finding that would remain invisible under standard evaluation protocols.

A related implication concerns thermal properties, which depend directly on the phonon density of states. Heat capacity at constant volume, vibrational entropy, and the vibrational contribution to the free energy all integrate over the density of states. An inaccurate density of states therefore propagates error into finite-temperature thermodynamics even when zero-Kelvin energetics appear perfect [20]. This means that a model validated solely on static energies could produce systematically incorrect phase boundaries or thermal expansion coefficients without any warning from conventional benchmarks.

Beyond quantitative errors lies a qualitative diagnostic: phonon stability serves as a binary indicator of dynamical integrity. Imaginary frequencies—negative squared frequencies—signal that a predicted structure is unstable under small perturbations. Machine-learning force fields trained exclusively on stable configurations may inadvertently produce soft modes or outright instabilities when evaluated on slightly distorted geometries. This failure mode operates at the level of the Hessian and remains invisible to force comparisons, because a structure can be in mechanical equilibrium (zero net force) yet sit at a saddle point or local maximum of the potential energy surface.

Long-range interactions are uniquely exposed at small wavevectors. Acoustic phonon branches near the Brillouin zone center reflect macroscopic elastic constants, and in polar materials, longitudinal-optical and transverse-optical splitting emerges from long-range electrostatic interactions. Finite-cutoff graph neural networks—even equivariant architectures [8]—may truncate these interactions, producing incorrect sound velocities or unphysical dispersion near the zone center. Energy and force metrics computed on small unit cells simply cannot detect such errors because the relevant physics only manifests across extended length scales. Park and co-workers noted that their scalable graph neural network force field required careful long-range corrections precisely because standard message-passing limits long-wavelength accuracy [21], underscoring that this issue is not merely theoretical but arises in practical state-of-the-art models.

Finally, phonon spectra serve as a powerful transferability diagnostic. Training data for machine-learning potentials typically comprise relaxed structures and rattled snapshots generated through molecular dynamics or random displacements [22, 23]. Phonon calculations, by contrast, explore the harmonic neighborhood around minima along collective vibrational coordinates. A model that generalizes well on energies and forces may still fail when the potential is sampled along directions in configuration space never explicitly seen during training—directions that correspond precisely to the normal modes of the system [24, 25]. This mismatch between the training distribution and the phonon evaluation distribution makes vibrational validation a uniquely stringent test of true generalization.

Conceptually, the potential energy surface can be visualized as a multi-dimensional landscape. The graph neural network approximates this surface (zeroth derivative). Its gradient yields forces (first derivative). The Hessian yields the dynamical matrix whose eigenvalues are phonon frequencies (second derivative). A two-dimensional slice through this landscape would show three overlaid curves: the true density-functional-theory surface, the graph-neural-network approximation, and their difference magnified tenfold. Annotations at selected points would distinguish force error from curvature error at minima, illustrating how small first-derivative mismatches can produce large second-derivative deviations—a relationship that is neither linear nor predictable from force metrics alone.

Taken together, these considerations demonstrate that phonon validation is not merely an additional check but a necessary orthogonal probe of model fidelity. Without it, claims of “chemical accuracy” remain incomplete for any application involving lattice dynamics, thermal transport, or vibrational spectroscopy [26]. The curvature of the potential energy surface, not just its depth or slope, determines whether a model will succeed or fail in the finite-temperature regimes where most materials operate.

The Problem of Spectral Leakage

Spectral leakage in the context of phonon validation of machine-learning force fields is defined as any artifact in the computed phonon spectra that arises from numerical or methodological choices in the force-constant extraction pipeline rather than from intrinsic differences between the DFT reference and the ML-predicted potential. Formally, leakage manifests whenever the finite representation of the dynamical matrix fails to capture the true infinite-range decay or when post-processing introduces artificial broadening or aliasing.

Table 2 systematizes spectral leakage by linking each numerical artifact to its physical origin, diagnostic signature, and quantitative mitigation strategy.

Table 2. Spectral Leakage Sources, Diagnostic Tests, and Quantification Strategies in Phonon Validation

Leakage Source	Physical Origin	Observable Artifact in Phonons	Diagnostic Test	Quantification Method	Mitigation Strategy
Finite Supercell Size	Truncation of long-range force constants	Artificial periodicity, shifted frequencies	Supercell convergence (2×2×2 → 4×4×4)	Frequency convergence slope	Increase supercell size
Displacement Magnitude	Anharmonic contamination in finite differences	Nonlinear force response, distorted modes	Displacement sweep (0.005–0.05 Å)	Sensitivity coefficient vs displacement	Use harmonic regime displacement
q-Point Sampling	Undersampling Brillouin zone	Aliasing, missing dispersion features	Mesh refinement test	Frequency variance vs mesh density	Increase q-point density
Fourier Interpolation	Discontinuity in force constants	Oscillatory dispersion artifacts	Direct vs interpolated comparison	Interpolation error bounds	Denser real-space sampling
Spectral Broadening	Artificial smoothing (Gaussian/Lorentzian)	Masked peak differences	Compare broadened vs raw DOS	$\Delta g(\omega)$ residual structure	Report both raw and smoothed spectra

Systematic discrepancies between graph neural network (GNN) force fields and density functional theory (DFT) phonon spectra often emerge from subtle forms of numerical leakage embedded within the validation pipeline rather than from intrinsic model inadequacy. A central mechanism arises from the finite spatial extent of the supercell used to compute interatomic force constants. Because $\Phi(R)$ decays with interatomic distance R , truncation at the supercell boundary imposes an artificial periodicity through the introduction of image interactions, which in turn perturbs the dynamical matrix and redistributes spectral weight in a manner that is not physically grounded [11, 18]. This boundary-induced distortion is rarely uniform; instead, it preferentially affects long-wavelength vibrational modes, thereby biasing comparisons in precisely the regime where GNN models are often expected to demonstrate transferability.

This sensitivity to spatial truncation is compounded by the reliance on finite-difference differentiation in both DFT and machine-learned phonon calculations. The displacement amplitude, typically chosen within the range of 0.01–0.05 Å, implicitly assumes a locally harmonic response. When the learned potential deviates from ideal smoothness, even marginally, higher-order anharmonic contributions become entangled with the extracted harmonic force constants. The resulting contamination does not simply introduce noise; it systematically reshapes the curvature of the potential energy

surface, thereby altering phonon frequencies in a way that mimics genuine model error while in fact reflecting numerical inconsistency.

A related distortion emerges in reciprocal space, where the discretization of the Brillouin zone through finite q-point meshes governs the resolution of phonon dispersions. Coarse Monkhorst–Pack sampling underspecifies high-frequency features and introduces aliasing near zone boundaries, effectively redistributing spectral intensity across adjacent modes. This sampling-induced ambiguity becomes particularly consequential when evaluating GNN models trained on datasets that already exhibit uneven coverage of configurational space, as the interaction between data sparsity and reciprocal space undersampling can amplify apparent discrepancies.

Beyond discretization itself, interpolation procedures used to reconstruct continuous phonon dispersions introduce an additional layer of complexity. Fourier interpolation assumes smoothness and periodic consistency of the dynamical matrix; however, any discontinuity inherited from supercell truncation is propagated and often amplified as oscillatory artifacts. These spurious features can manifest as artificial peak splitting or secondary lobes in the phonon density of states, complicating the attribution of error. In practice, such interpolation-induced distortions are difficult to disentangle from genuine deficiencies in the learned force field without targeted convergence analysis.

Even when underlying discrepancies are modest, post-processing choices can obscure their interpretation. The application of Gaussian or Lorentzian broadening to phonon densities of states, typically introduced for visual clarity, acts as a nonlinear filter that can either suppress meaningful deviations or create the illusion of agreement. Under these conditions, two spectra may appear closely aligned despite underlying structural differences, or conversely, minor numerical inconsistencies may be visually exaggerated depending on the chosen smearing width.

Taken together, these interacting sources of leakage produce a form of diagnostic ambiguity that challenges straightforward validation. Apparent agreement between DFT and GNN-derived phonon densities of states may arise because both spectra have been subjected to analogous smoothing and truncation effects, while apparent disagreement may simply reflect asymmetric sensitivity to these artifacts. A more informative representation would juxtapose an idealized leakage-free DFT reference with both unconverged and systematically corrected GNN predictions, allowing regions of true model error to be distinguished from those dominated by numerical distortion. Within such a framework, shaded discrepancies acquire interpretive meaning, separating artifacts induced by methodological choices from those rooted in the learned representation itself.

Resolving this ambiguity requires an explicitly hierarchical approach in which leakage is not treated as a peripheral concern but as a central object of analysis. Embedding detection and control mechanisms throughout the validation workflow enables a more faithful attribution of error, thereby clarifying whether observed deviations reflect limitations of the GNN architecture or arise from the computational scaffolding used to evaluate it.

Hierarchical Framework: Level 1 (Energy/Force)

Level 1 validation functions as a stringent entry condition, ensuring that a GNN force field faithfully reproduces the local structure of the DFT potential energy surface at the level of energies and forces before any attempt is made to resolve vibrational properties. The emphasis at this stage is not merely on numerical agreement but on establishing that the learned representation captures the first-order response of the system across a sufficiently diverse configurational manifold. This requirement reflects a broader methodological stance in AI-driven materials modeling: phonon predictions, as second-order derivatives of the energy landscape, inherit and amplify even subtle inconsistencies in lower-order quantities.

Operationally, this validation layer is anchored in error metrics evaluated on held-out data that span compositional variability, crystallographic symmetry, and thermodynamic conditions. Energy mean absolute errors on the order of < 0.01 eV/atom and force root mean square errors below approximately 0.05 eV/Å are typically regarded as indicative of

adequate fidelity, although these thresholds remain material-dependent. Crucially, such aggregate metrics must be complemented by stratified analysis, as uniform performance across chemical species and symmetry classes is more informative than global averages that may obscure systematic biases. Recent equivariant GNN frameworks, including those examined by Batzner *et al.* [8] and Park *et al.* [21], consistently report performance within these bounds, yet both studies underscore a critical limitation: accurate force prediction, even when achieved across diverse datasets, does not directly imply reliable vibrational behavior.

This limitation arises because Level 1 validation remains inherently local. It probes the gradient of the potential energy surface but leaves its curvature largely unconstrained. As a consequence, properties governed by second-order derivatives—such as phonon frequencies, elastic responses, and collective lattice dynamics—may deviate substantially even when force errors appear negligible. The absence of explicit validation of long-range interactions beyond the cutoff radius further compounds this issue, particularly in materials where dispersive or electrostatic contributions play a nontrivial role. Under these conditions, a model that satisfies conventional accuracy thresholds may still yield unphysical signatures, including imaginary phonon modes or incorrect acoustic velocities, revealing a disconnect between local force accuracy and global dynamical consistency.

The progression to more computationally intensive phonon calculations is therefore governed by a deliberately conservative criterion. Force errors must not only fall below a material-specific tolerance but also exhibit no directional bias across Cartesian components, as even slight anisotropies can propagate into systematic distortions of the dynamical matrix. This gatekeeping mechanism reflects an implicit cost–benefit consideration: supercell-based force-constant calculations are sufficiently expensive that their application is justified only when the underlying force field demonstrates robust and unbiased first-order behavior [27, 28].

In practice, this level serves as an efficient yet discriminating filter. Architectures such as SchNet, as demonstrated by Schütt *et al.* [16], can achieve sub-0.1 eV/Å force errors across a wide range of materials, illustrating the capacity of modern GNNs to approximate local energy landscapes with high precision. Nevertheless, subsequent phonon benchmarks have occasionally exposed residual inconsistencies in curvature, highlighting that compliance with Level 1 criteria, while necessary, is not sufficient for vibrational reliability. Embedding this validation stage within a hierarchical framework thus mitigates the risk of attributing downstream discrepancies to model limitations when they instead originate from inadequately constrained representations at the outset.

Hierarchical Framework: Level 2 (Phonon Dos)

Level 2 shifts attention from local derivatives to the integral vibrational spectrum. The phonon density of states is defined as

$$g(\omega) = \frac{1}{NQ} \sum \delta(\omega - \omega_s(q)) \quad (1)$$

where the sum runs over all wavevectors q in the Brillouin zone and branches s . Because the DOS integrates over all q , it is less sensitive to individual sampling artifacts than full dispersion yet still captures the overall distribution of mode frequencies.

Computation proceeds identically for DFT and ML: the same supercell and identical displacement magnitude are used to extract force constants, ensuring that leakage sources 1 and 2 are controlled by construction. The DOS is then evaluated on the identical q -mesh.

Advantages include computational efficiency and robustness to fine q -grid details. Disadvantages remain: compensating errors (one mode too high, another too low) can yield an apparently correct DOS.

Proposed validation criteria are (i) normalized cross-correlation (spectral overlap) > 0.85, (ii) major peak position differences < 5 %, and (iii) absence of spurious imaginary modes.

To address leakage at Level 2, identical supercell sizes and displacements are mandatory; convergence with respect to supercell size must be reported. Chen *et al.* illustrated direct neural-network prediction of DOS [29], yet their approach bypassed explicit force-constant extraction and therefore could not diagnose leakage. The present framework insists on the conventional finite-displacement route to maintain traceability.

A conceptual diagram would display two DOS curves (DFT solid black, ML dashed blue) on a shared frequency axis, with vertical lines marking major peak positions and shaded bands indicating ± 5 % tolerance windows. Overlap regions would be highlighted in green, non-overlapping tails in red, visually quantifying agreement at the integral level.

Hierarchical Framework: Level 3 (Dispersion Curves)

Level 3 demands q -resolved validation of the full dispersion relation $\omega_s(q)$. The underlying phonon eigenvalue problem is

$$\sum_j D_{i\alpha,j\beta}(q) e_{j\beta}(q) = \omega^2(q) e_{i\alpha}(q) \quad (2)$$

where the dynamical matrix D is obtained from the Fourier transform of the force-constant matrix Φ :

$$D_{i\alpha,j\beta}(q) = \frac{1}{\sqrt{m_i m_j}} \sum_R \Phi_{i\alpha,j\beta}(R) e^{iq \cdot R} \quad (3)$$

Dispersion is computed along standardized high-symmetry paths (Γ –X–M– Γ for cubic systems, expanded for lower symmetry).

This level is the most stringent because it exposes direction-dependent errors, acoustic-branch slopes, and long-range interaction failures. Advantages include revelation of subtle mode crossings or LO–TO splitting; disadvantages are higher cost and greater sensitivity to interpolation.

Validation criteria comprise (i) maximum frequency error per branch < 10 %, (ii) absence of spurious crossings, and (iii) acoustic-mode sound velocities within 10 % of DFT.

Leakage is controlled by using dense reference q -meshes for DFT and testing Fourier-interpolation convergence explicitly. Some study emphasize that phonopy/phono3py implementations require careful supercell convergence precisely to minimize interpolation artifacts [2].

A conceptual diagram would present DFT and ML dispersion curves plotted together along the q -path, with vertical error bands (± 0.5 THz) shaded around each ML branch. Regions of excellent overlap would appear green; deviations would be annotated with arrows indicating “long-range cutoff error” or “imaginary mode artifact.”

Hierarchical Framework: Level 4 (Spectral Metrics)

Level 4 replaces visual inspection with reproducible, quantitative metrics. Four complementary measures are proposed.

Spectral overlap (normalized cross-correlation) is

$$S = \frac{\int g_{DFT}(\omega) g_{ML}(\omega) d\omega}{\sqrt{\int g_{DFT}^2(\omega) d\omega \int g_{ML}^2(\omega) d\omega}} \quad (4)$$

Earth Mover's Distance (Wasserstein-1) between cumulative distributions quantifies the minimal “work” needed to transport one spectrum onto the other. Peak-position RMS error and q-dependent frequency error complete the set.

Thresholds are material-dependent but illustrative: for Level 2, $S > 0.85$ and $W_1 < 0.1$ THz; for Level 3, $\max \epsilon_{\text{disp}} < 0.5$ THz. Metrics are reported both with and without broadening to expose leakage sensitivity.

By requiring multiple metrics, Level 4 guards against pathologies that any single scalar might hide. The placeholder validation study already advocated quantitative spectral comparison [13]; the present framework operationalizes that recommendation within a leakage-aware hierarchy.

Detecting and Quantifying Spectral Leakage

Within the hierarchical validation framework, leakage is not treated as an abstract concern but is rendered empirically tractable through a sequence of embedded diagnostics that probe the numerical stability of the phonon workflow. A central strategy involves systematically varying the supercell size used to construct force constants, thereby exposing the extent to which artificial periodicity contaminates the vibrational spectrum. When phonon densities of states or dispersions are recomputed across progressively larger supercells—typically spanning $2 \times 2 \times 2$ to $4 \times 4 \times 4$ —the nature of the DFT–ML discrepancy becomes revealing. A monotonic improvement in agreement under increasing spatial extent indicates that boundary-induced truncation effects, rather than intrinsic model deficiencies, are driving the observed deviations.

This spatial perspective is complemented by an explicit interrogation of the local smoothness of the learned potential. By sweeping the displacement magnitude used in finite-difference force calculations over a range that spans an order of magnitude, one effectively tests the consistency of the harmonic approximation encoded in both DFT and the GNN. Divergent sensitivity between the two, particularly when the GNN response varies nonlinearly with displacement amplitude, signals that higher-order terms are contaminating the extracted force constants. Such behavior points to a breakdown in the assumed differentiability of the learned energy landscape, with direct implications for the reliability of second-order properties.

A parallel line of analysis targets the reciprocal-space discretization underlying phonon calculations. Recomputing dispersions on increasingly dense q-point meshes provides a means of isolating aliasing effects that arise from undersampling the Brillouin zone. Stabilization of branch frequencies within a narrow tolerance—on the order of 0.1 THz—serves as an operational criterion for convergence. When discrepancies between DFT and ML predictions diminish as sampling density increases, the evidence again implicates numerical leakage rather than representational error.

Interpolation schemes introduce an additional layer of potential distortion, necessitating cross-validation between directly computed and interpolated phonon frequencies. By evaluating the dynamical matrix at selected q-points using both approaches, one can bound the magnitude of interpolation-induced artifacts. Deviations that persist only in the interpolated spectrum reveal the extent to which discontinuities, often inherited from finite supercell construction, are being propagated across reciprocal space.

The interpretive synthesis of these diagnostics is sharpened through residual analysis of the phonon density of states. Considering the difference $\Delta g(\omega)$ between DFT and ML spectra provides a compact representation of error structure, where the qualitative form of the residual becomes diagnostically meaningful.

$$\begin{aligned} \Delta g(\omega) \\ = g_{DFT(\omega)} - g_{ML(\omega)} \end{aligned} \quad (5)$$

Oscillatory patterns in $\Delta g(\omega)$ typically reflect leakage propagated through interpolation or boundary effects, whereas smooth, systematic shifts are more consistent with bias in the learned force field itself. Interpreted jointly, these detection

mechanisms transform validation from a binary assessment into a structured inference process, enabling a more precise attribution of error sources within GNN-driven phonon predictions.

Validation Protocol and Reporting Standards

The protocol unfolds as a staged evaluation pipeline in which each phase conditions the interpretability of the next, thereby minimizing the risk that downstream discrepancies are misattributed. The initial stage establishes baseline fidelity at the level of energies and forces, with progression contingent on achieving a sufficiently low force RMSE. When this threshold is exceeded—particularly beyond 0.1 eV/Å—the workflow is intentionally terminated, reflecting the recognition that subsequent phonon calculations would amplify rather than clarify underlying inaccuracies. This early constraint anchors the entire procedure in a principle of computational parsimony, ensuring that only models with credible first-order behavior are subjected to more demanding analyses.

Building on this foundation, the protocol transitions into a dedicated leakage detection phase that interrogates the numerical robustness of the phonon workflow itself. By systematically varying supercell dimensions, displacement amplitudes, reciprocal-space sampling density, and interpolation consistency, this stage isolates artifacts that would otherwise remain entangled with model error. A critical feature of this step lies in its tolerance-based decision rule: when variations across these tests exceed approximately 20%, the response is not interpretive but corrective, requiring expansion of the supercell or refinement of computational parameters. This adaptive mechanism reframes validation as an iterative stabilization process rather than a one-pass evaluation.

Once numerical consistency is established, attention shifts toward quantitative comparison of vibrational spectra under controlled conditions. At this stage, phonon densities of states derived from DFT and the GNN are evaluated using identical computational settings, allowing discrepancies to be interpreted as model-dependent rather than methodological. Metrics such as spectral overlap and Earth Mover's Distance provide complementary perspectives, capturing both global alignment and distributional shifts in frequency space. The emphasis here is not solely on agreement but on characterizing the structure of disagreement in a manner that is reproducible and comparable across studies.

A further refinement emerges when dispersion relations are examined, although this step is invoked conditionally, reflecting its higher computational cost and sensitivity to residual numerical noise. When undertaken, the analysis extends beyond aggregate agreement to branch-resolved errors and the behavior of acoustic modes, whose slopes encode elastic properties. This level of granularity enables a more mechanistic interpretation of model performance, particularly in identifying whether discrepancies are localized to specific vibrational modes or reflect broader distortions of the dynamical matrix.

The final stage consolidates these insights into a multi-metric synthesis that integrates density-of-states and dispersion-based evaluations. Rather than privileging a single indicator of accuracy, the framework encourages a layered interpretation in which different metrics illuminate distinct aspects of model behavior. This integrative perspective is reinforced by explicit reporting standards that require detailed disclosure of supercell convergence, displacement sensitivities, q-point specifications, and broadening choices. The inclusion of multiple spectral metrics, alongside directly comparable visualizations of densities of states and dispersions, ensures that conclusions are both transparent and reproducible.

Taken together, this protocol redefines validation as a structured, evidence-generating process in which numerical rigor and model assessment are inseparably linked. Its adoption would substantially reduce ambiguity in the evaluation of GNN force fields, while also enabling systematic comparison across studies, thereby supporting the emergence of cumulative insight within the broader literature on AI-driven materials modeling.

Conclusion

Validating graph neural network force fields against DFT phonon spectra remains a nontrivial task precisely because spectral leakage—arising from finite supercell truncation, displacement artifacts, q-point undersampling, Fourier interpolation discontinuities, and artificial broadening—can mask genuine model deficiencies or fabricate spurious discrepancies that have nothing to do with the underlying machine-learning potential.

The proposed four-level hierarchical framework enforces progressive validation: Level 1 confirms baseline energy and force accuracy (necessary gate), Level 2 examines the integral phonon density of states via the formula $g(\omega) = \frac{1}{Nq} 1Nq \sum \delta(\omega - \omega_s(q))$, Level 3 scrutinizes q-dependent dispersion curves through the dynamical-matrix eigenvalue problem $\sum D_{i\alpha,j\beta}(q) e_{j\beta}(q) = \omega^2(q) e_{i\alpha}(q)$, and Level 4 supplies quantitative spectral metrics (overlap, Earth Mover's Distance, peak-position RMS, and branch-wise errors) to replace subjective visual inspection. Each level embeds explicit controls for leakage, ensuring that passage to the next level occurs only after lower-level criteria and convergence tests are satisfied.

Five detection methods operationalize leakage quantification: supercell-size convergence, displacement-magnitude sweeps, q-point density escalation, Fourier-interpolation cross-validation, and residual-frequency analysis. When applied systematically, these diagnostics transform an invisible confounder into a measurable quantity that researchers can report and mitigate. The accompanying validation protocol—executed sequentially from Level 1 through leakage diagnostics to full multi-metric Level 4 reporting—together with the standardized checklist (supercell convergence table, displacement sensitivity coefficients, q-grid specification, broadening justification, and dual-metric summary) provides a reproducible workflow that future GNN force-field studies can adopt without ambiguity.

By moving the community from ad-hoc phonon plots to a leakage-aware, hierarchical culture, this framework bridges the persistent gap between the high-fidelity DFT phonon literature and the rapidly expanding ML interatomic-potential ecosystem. Adoption will elevate claims of “chemical accuracy” from force-level assertions to vibrationally faithful predictions, thereby increasing confidence in downstream applications ranging from thermal transport in thermoelectrics to phonon-limited carrier mobility in semiconductors. The placeholder validation study already called for best-practice guidelines; the present work supplies a concrete, mathematically grounded, and citation-consistent implementation. Widespread use of this protocol will accelerate the reliable integration of GNN force fields into materials discovery pipelines while safeguarding against the subtle numerical pitfalls that have until now undermined phonon-based validation. Ultimately, the framework reinforces that true model fidelity must be demonstrated not only on energies and forces but across the full vibrational landscape, free from spectral leakage artifacts.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Nov 2021

Revised: 19 Mar 2022

Accepted: 31 May 2022

Published online: 18 July 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules. *Sci Adv.* 2017;3(12):e1701816.
<https://doi.org/10.1126/sciadv.1701816>.
- Mueller T, Hernandez A, Wang C. Machine learning for interatomic potential models. *J Chem Phys.* 2020;152(5):050902.
<https://doi.org/10.1063/1.5126336>.
- Bartók AP, Kermode J, Bernstein N, Csányi G. Machine learning a general-purpose interatomic potential for silicon. *Phys Rev X.* 2018;8(4):041048.
<https://doi.org/10.1103/PhysRevX.8.041048>.
- Wu YC, Feng JW. Development and application of artificial neural network. *Wirel Pers Commun.* 2018;102(2):1645-56.
<https://doi.org/10.1007/s11277-017-5224-x>.
- Caro MA. Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials. *Phys Rev B.* 2019;100(2):024112.
<https://doi.org/10.1103/PhysRevB.100.024112>.
- Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol.* 2019;188(12):2222-39.
<https://doi.org/10.1093/aje/kwz189>.
- de Haan P, Cohen TS, Welling M. Natural graph networks. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin HT, editors. *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020; 2020 Dec 6-12; Virtual. Red Hook (NY): Curran Associates; 2020. p. 3636-46.*
<https://doi.org/10.5555/3495724.3496031>.
- Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun.* 2022;13(1):2453.
<https://doi.org/10.1038/s41467-022-29939-5>.
- Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater.* 2019;31(9):3564-72.
<https://doi.org/10.1021/acs.chemmater.9b01294>.
- Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. *NPJ Comput Mater.* 2021;7(1):185.
<https://doi.org/10.1038/s41524-021-00650-1>.

Giustino F. Electron-phonon interactions from first principles. *Rev Mod Phys.* 2017;89(1):015003.
<https://doi.org/10.1103/RevModPhys.89.015003>.

Erba A, Desmarais JK, Casassa S, Civalleri B, Donà L, Bush IJ, et al. CRYSTAL23: A program for computational solid state physics and chemistry. *J Chem Theory Comput.* 2023;19(20):6891-932.
<https://doi.org/10.1021/acs.jctc.2c00958>.

Botu V, Batra R, Chapman J, Ramprasad R. Machine learning force fields: Construction, validation, and outlook. *J Phys Chem C.* 2017;121(1):511-22.
<https://doi.org/10.1021/acs.jpcc.6b10908>.

Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. *Chem Rev.* 2021;121(16):10073-141.
<https://doi.org/10.1021/acs.chemrev.1c00022>.

Zhang L, Han J, Wang H, Car R, E W. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys Rev Lett.* 2018;120(14):143001.
<https://doi.org/10.1103/PhysRevLett.120.143001>.

Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet: A deep learning architecture for molecules and materials. *J Chem Phys.* 2018;148(24):241722.
<https://doi.org/10.1063/1.5019779>.

Behler J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew Chem Int Ed.* 2017;56(42):12828-40.
<https://doi.org/10.1002/anie.201703114>.

Andersson R. Modeling radiation induced degradation of lattice thermal conductivity [thesis]. Stockholm: KTH Royal Institute of Technology, School of Engineering Sciences; 2020. Available from: <http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-277885>

Babaei H, Guo R, Hashemi A, Lee S. Machine-learning-based interatomic potential for phonon transport in perfect crystalline Si and crystalline Si with vacancies. *Phys Rev Mater.* 2019;3(7):074603.
<https://doi.org/10.1103/PhysRevMaterials.3.074603>.

Choudhary K, Garrity KF, Sharma V, Biacchi AJ, Hight Walker AR, Tavazza F. High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric, and dielectric responses. *NPJ Comput Mater.* 2020;6(1):64.
<https://doi.org/10.1038/s41524-020-0337-2>.

Park CW, Kornbluth M, Vandermause J, Wolverton C, Kozinsky B, Mailoa JP. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *NPJ Comput Mater.* 2021;7(1):73.
<https://doi.org/10.1038/s41524-021-00543-3>.

Chen C, Deng Z, Tran R, Tang H, Chu IH, Ong SP. Accurate force field for molybdenum by machine learning large materials data. *Phys Rev Mater.* 2017;1(4):043603.
<https://doi.org/10.1103/PhysRevMaterials.1.043603>.

Deringer VL, Bernstein N, Bartók AP, Cliffe MJ, Kerber RN, Marbella LE, et al. Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics. *J Phys Chem Lett.* 2018;9(11):2879-85.
<https://doi.org/10.1021/acs.jpclett.8b00902>.

Zhou ZH. Machine learning. Liu S, translator. Singapore: Springer Singapore; 2021. XIII, 459 p.
<https://doi.org/10.1007/978-981-15-1967-3>.

Sewak M, Sahay SK, Rathore H. An overview of deep learning architecture of deep neural networks and autoencoders. *J Comput Theor Nanosci.* 2020;17(1):182-8.

<https://doi.org/10.1166/jctn.2020.8648>.

Borrego-Varillas R, Lucchini M, Nisoli M. Attosecond spectroscopy for the investigation of ultrafast dynamics in atomic, molecular and solid-state physics. *Rep Prog Phys.* 2022;85(6):066401.

<https://doi.org/10.1088/1361-6633/ac5e7f>.

Chmiela S, Sauceda HE, Müller KR, Tkatchenko A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat Commun.* 2018;9(1):3887.

<https://doi.org/10.1038/s41467-018-06169-2>.

Sauceda HE, Chmiela S, Poltavsky I, Müller KR, Tkatchenko A. Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *J Chem Phys.* 2019;150(11):114102.

<https://doi.org/10.1063/1.5078687>.

Chen Z, Andrejevic N, Smidt T, Ding Z, Xu Q, Chi YT, et al. Direct prediction of phonon density of states with Euclidean neural networks. *Adv Sci.* 2021;8(12):2004214.

<https://doi.org/10.1002/advs.202004214>.