

ORIGINAL RESEARCH

Open access

A Framework for Validating ML Potentials against Experimental Phonon Dispersion and Elastic Constants

Sophie Laurent^{1*}, Pierre Dubois¹, Marc Lefevre², Claire Moreau¹

Abstract

Machine learning interatomic potentials (MLIPs) are widely validated against density functional theory (DFT), yet this practice risks reproducing DFT's systematic biases rather than true experimental behavior. We present a concise four-level hierarchical framework for rigorous experimental validation focused on phonon dispersion and elastic constants. Level 1 evaluates static room-temperature elastic constants; Level 2 assesses zone-center phonon frequencies (Raman/IR); Level 3 examines full phonon dispersion curves from inelastic neutron/X-ray scattering; and Level 4 tests temperature-dependent shifts and softening. For each level we provide key observables, simple comparison metrics, uncertainty-aware criteria, and typical failure modes. A practical validation protocol and minimum reporting standards (Level 1 on ≥ 3 materials, Level 2 on ≥ 1) are proposed. This framework complements DFT benchmarks by anchoring MLIPs in experimental reality, improving reliability for thermal, mechanical, and vibrational applications in materials science.

Keywords Machine learning interatomic potentials, Experimental validation, Phonon dispersion, Elastic constants, Phonon transport, Materials modeling framework

*Correspondence:

Sophie Laurent
sophie.laurent@gmail.com

¹ Department of Data-Driven Materials Modeling, Faculty of Engineering, Sorbonne University, Paris, France

² Department of Computational Materials Engineering, Faculty of Technology, École Polytechnique, Paris, France

Introduction

Machine learning potentials have emerged as powerful tools in materials science, offering the ability to perform molecular dynamics simulations at scales far beyond traditional ab initio methods while retaining much of their accuracy. These potentials, trained on datasets of atomic configurations and associated energies or forces, approximate the potential energy surface in a data-driven manner. Applications range from defect studies and phase transitions to thermal transport and mechanical properties. Recent developments illustrate the breadth, including machine-learning potentials tailored for phonon transport in AlN with defects in multiple charge states [1], machine learning force field based phonon dispersion prediction [2], a machine learning potential for graphene [3], an energy-free machine learning force field for aluminum [4], machine learning for thermal transport and phonon high-pressure behavior in boron arsenide [5], machine-learning-based interatomic potential for phonon transport in perfect crystalline Si and crystalline Si with vacancies [6], training machine learning interatomic potentials for accurate phonon properties [7], accurate machine learning force fields via experimental validation [8], a machine learning framework for elastic constants predictions in multi-principal element alloys [9], machine learning-based prediction of elastic properties using reduced datasets of accurate calculations results [10], physics-informed and interpretable machine learning for elastic constants of RHEAs [11], ab-initio trained machine learning potential for MAX phases [12], benchmarking machine learning interatomic potentials via phonon anharmonicity [13], on-the-fly machine learning of interatomic potentials for elastic property modeling in Al–Mg–Zr solid solutions [14],

composition-transferable machine learning potential for LiCl-KCl molten salts validated by high-energy x-ray diffraction [15], transferability of machine-learning interatomic potential to α -Fe nanocrystalline deformation [16], machine learning interatomic potential with DFT accuracy for general grain boundaries in α -Fe [17], machine learning interatomic potential for simulations of carbon at extreme conditions [18], robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling [19], machine learning-based interatomic potential development and phase transition analysis of ferroelectric hafnium dioxide [20], accurate deep potential model of temperature-dependent elastic constants in phosphorus-doped silicon [21], machine learning interatomic potential [22], prediction model of elastic constants of BCC high-entropy alloys based on first-principles calculations and machine learning techniques [23], machine learning interatomic potential for the structural properties of iron oxides [24], machine learning interatomic potential can infer electrical response [25], a general purposed machine learning interatomic potential for Mg-Al-Si-O system suitable for Earth materials at high pressure and temperature [26], efficient machine learning interatomic potentials robust for liquid and multiple solid polymorphs of NaF and KF [27], how to validate machine-learned interatomic potentials [28], and benchmarking universal machine learning interatomic potentials for rapid analysis of inelastic neutron scattering data [29].

Despite these advances, a critical gap persists in how these models are evaluated. The dominant validation paradigm relies almost exclusively on direct comparison to density functional theory results. Researchers commonly report metrics such as force root-mean-square error or energy error relative to a chosen DFT functional and basis set, interpreting close agreement as evidence that the machine learning potential is accurate [1, 2, 6, 7]. While this DFT-centric approach is convenient and computationally tractable, it is insufficient for establishing true physical fidelity. DFT itself carries systematic errors that vary with the choice of exchange-correlation functional, pseudopotentials, treatment of dispersion interactions, and handling of electronic correlations. A model that perfectly emulates one DFT setup may inherit or even amplify those biases when confronted with experiment [8, 28].

Phonon-related properties and elastic constants are particularly sensitive test cases. Phonons govern thermal transport, vibrational spectroscopy, and many dynamical processes, while elastic constants determine mechanical stiffness and sound velocities. High-quality experimental data exist for both: elastic constants from ultrasonic or resonant ultrasound methods, and phonon dispersion from inelastic neutron scattering. These measurements provide direct access to the second derivatives of the energy with respect to atomic displacements, probing the very essence of the interatomic potential.

This paper proposes a conceptual framework for validating machine learning interatomic potentials against experimental phonon dispersion and elastic constants. The framework emphasizes four hierarchical levels of validation that escalate in both scientific depth and practical demands on data availability. It is deliberately conceptual, focusing on definitions, component relationships, validation protocols, and strategies for integrating experimental information rather than on specific numerical simulations or performance numbers.

By prioritizing experiment as the ultimate reference, the framework addresses a fundamental limitation in current practices. It encourages developers and users to move beyond the comfort of DFT agreement and confront the realities of measurement uncertainty, finite-temperature effects, sample imperfections, and long-range interactions. In doing so, it aims to foster machine learning potentials that are not only consistent with a computational reference but genuinely predictive of observable material behavior [8, 13, 29].

Why Experimental Validation is Essential

Relying solely on DFT for validating machine learning potentials introduces several conceptual problems. Different DFT functionals, such as the widely used PBE or more advanced meta-GGA variants, can produce noticeably different phonon frequencies or elastic tensors even for the same material. Pseudopotential choices and convergence parameters further modulate results. These variations mean that a machine learning model optimized against one DFT reference may not generalize well to others, let alone to experiment [7, 28]. Moreover, standard DFT calculations are performed at zero

temperature for perfect, periodic crystals, whereas real experiments occur at finite temperature on samples that inevitably contain defects, impurities, or surfaces [1, 5, 6].

The consequences are significant. A potential that achieves excellent agreement with DFT may still misrepresent key physical features when applied to properties sensitive to subtle energy landscape details. For instance, phonon dispersion curves encode information about force constants over various length scales; small errors in long-range interactions can distort branch shapes or acoustic slopes even if zone-center values appear reasonable [2, 13]. Similarly, elastic constants reflect the overall curvature of the energy surface and are directly comparable to ultrasonic measurements [9–11, 14, 23].

Experimental validation overcomes these issues by providing a ground-truth reference that is independent of any particular computational approximation. Inelastic neutron scattering yields phonon dispersion relations across the Brillouin zone, directly testing the dynamical matrix derived from the interatomic potential [29]. Ultrasonic, resonant ultrasound, or Brillouin scattering methods supply elastic constants that probe uniform strain responses. Both sets of observables are sensitive to the accuracy of the potential in ways that go beyond simple energy or force matching [8].

Phonon dispersion, in particular, tests long-range interactions and dynamical stability. Acoustic branches near the zone center relate to sound velocities and thus connect to elastic properties, while optical branches and zone-boundary modes reveal local bonding characteristics. Temperature-dependent measurements further expose anharmonic effects that harmonic DFT calculations inherently miss [13, 21]. Elastic constants offer a complementary, lower-cost entry point: they are available for a wide range of materials and can be computed from the potential via finite differences or strain-energy relations [9–11, 14, 23].

When machine learning potentials are validated only against DFT, there is a risk of “overfitting” to the reference method’s biases. The model learns to reproduce DFT errors rather than the underlying physics. Experimental benchmarks break this closed loop. They force the potential to demonstrate transferability to real-world conditions, including finite-temperature effects and, in some cases, defect influences if experimental samples are not perfectly pristine [1, 6, 15–17].

Several studies have already developed machine learning potentials with explicit attention to phonon properties or elastic behavior in systems such as silicon with vacancies [6], multi-principal element alloys [9], refractory high-entropy alloys [11], MAX phases [12], molten salts [15], α -Fe grain boundaries [17], ferroelectric hafnium dioxide [20], phosphorus-doped silicon [21], and iron oxides [24]. Yet even these works often stop at DFT comparison or limited experimental checks [8, 28]. A systematic framework that places experiment at the center can unify these efforts and raise the bar for what constitutes a well-validated potential [8, 29].

In summary, experimental validation is essential because it anchors computational models in measurable reality. It reveals limitations hidden by DFT agreement, tests aspects of the potential critical for applications in thermal and mechanical properties, and provides clear criteria for model improvement. Phonon dispersion and elastic constants are ideal targets because they are both fundamental to the interatomic potential and richly documented in the experimental literature [13, 29].

Experimental Data Sources and Uncertainties

High-quality experimental data on elastic constants and phonon dispersion are available for many materials, though with varying coverage and precision. For elastic constants, common measurement techniques include ultrasonic pulse-echo methods, resonant ultrasound spectroscopy (RUS), and Brillouin light scattering. Ultrasonic techniques propagate acoustic waves through single crystals or polycrystals and extract velocities that relate directly to stiffness tensor components. RUS excites mechanical resonances of a sample and fits frequencies to determine the full elastic tensor, often with high precision for small specimens. Brillouin scattering uses light to probe acoustic phonons and is particularly useful for opaque or microscale samples.

Typical uncertainties for elastic constants range from about one to five percent, depending on the method, sample quality, and crystal symmetry. Simpler cubic materials like metals or basic semiconductors often achieve the lower end of this range, while lower-symmetry or softer materials may show larger scatter. Comprehensive compilations exist in reference works such as the Landolt-Börnstein series, which tabulates evaluated experimental data alongside measurement conditions. Additional sources include specialized handbooks for metals and ceramics, and individual papers on specific compounds.

Phonon dispersion data come primarily from inelastic neutron scattering (INS) on single crystals, with supplementary information from inelastic X-ray scattering (IXS) or optical spectroscopies for zone-center modes. INS maps energy and momentum transfer to extract phonon frequencies and eigenvectors across the Brillouin zone. It is the method of choice for full dispersion curves because neutrons interact directly with nuclei and have wavelengths comparable to interatomic distances. IXS offers higher resolution in some cases and does not require large samples, but access is more limited [29].

Uncertainties in phonon frequencies typically fall between 0.1 and 1 THz, equivalent to roughly 1 to 5 cm^{-1} in spectroscopic units, although values can be larger near zone boundaries or for damped modes. Factors influencing uncertainty include instrumental resolution, sample mosaic spread, temperature control, and the fitting procedure used to extract peak positions from scattering intensities. Many classic dispersion curves were measured decades ago on high-purity single crystals and remain reference standards. Newer measurements benefit from brighter neutron sources and improved detectors, reducing uncertainties and extending coverage to more complex materials.

Challenges in using experimental phonon data for validation include the requirement for sizable single crystals in traditional INS, the time-intensive nature of measurements, and the fact that raw scattering intensities must often be processed through model fitting (for example, Born-von Karman force-constant models) to obtain dispersion relations. Reported data should ideally include temperature, pressure, sample purity or isotope composition, and estimated uncertainties.

For effective machine learning potential validation, researchers need not only tabulated frequencies or constants but also context: exact measurement conditions, uncertainty estimates, and preferably access to raw or processed datasets when possible. Where multiple independent measurements exist, consensus values or ranges should be used to define a reliable experimental window. For materials with known discrepancies between different studies, the validation protocol should explicitly discuss possible sources of variation, such as impurity levels or thermal history.

Overall, while elastic constant data cover hundreds of inorganic crystals across metals, semiconductors, and ceramics, full phonon dispersion is available for fewer systems, often simpler structures like elemental solids, binary compounds, or well-studied insulators. This scarcity defines one boundary of the proposed framework: Level 1 and 2 validation can apply broadly, while Level 3 is more selective, and Level 4 remains aspirational for many materials. Curating centralized, machine-readable repositories of critically evaluated experimental phonon and elastic data would greatly accelerate adoption of experiment-based validation [29].

Framework: Four Levels of Experimental Validation

The proposed framework structures experimental validation of machine learning potentials into four levels that increase in both the depth of physical insight and the demands placed on data availability and computational effort.

Figure 1 presents the proposed experimental validation architecture as a strictly hierarchical progression from accessible, low-data tests of static elasticity to demanding, high-stringency evaluation of temperature-dependent vibrational behavior.

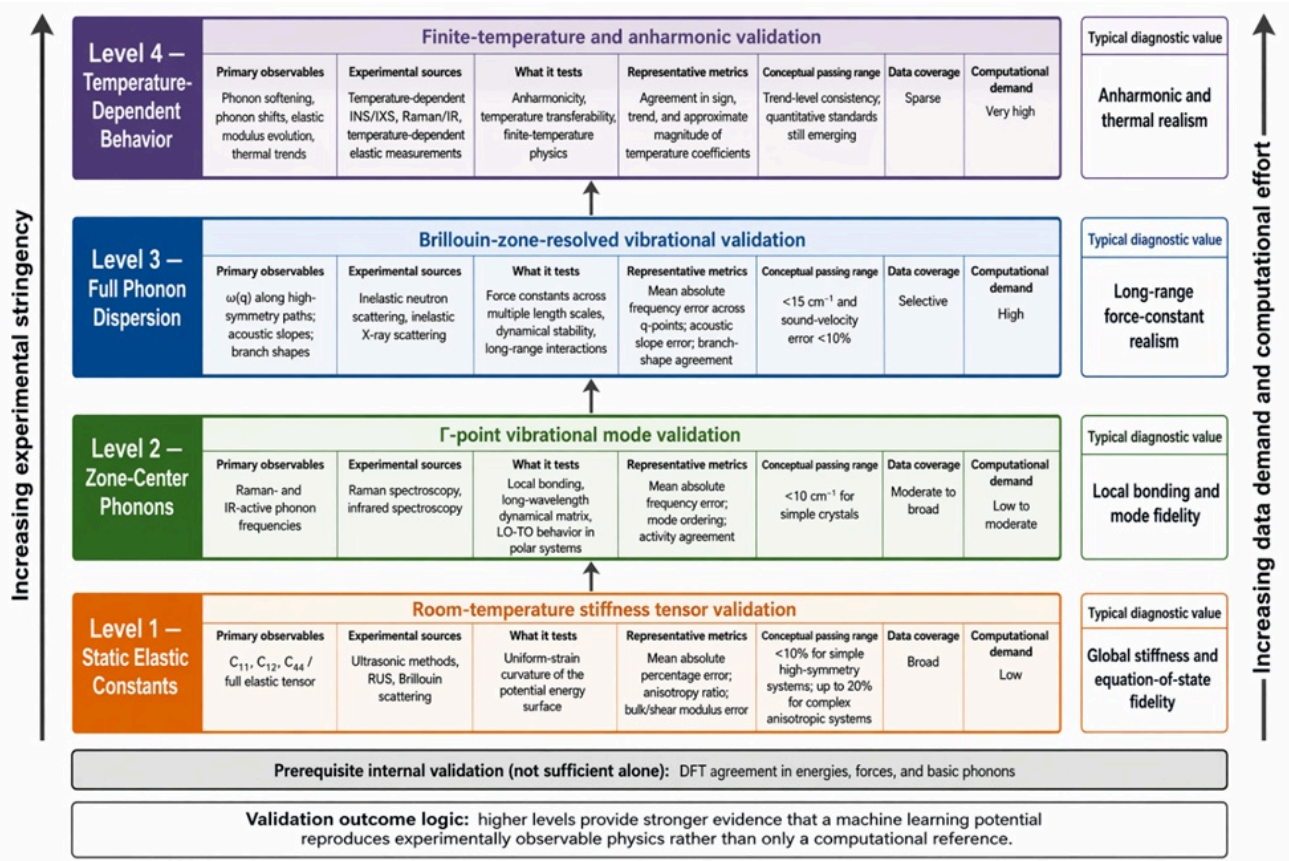


Figure 1. Hierarchical experimental validation architecture for machine learning interatomic potentials: from static elasticity to temperature-dependent phonon behavior

Each level targets specific observables directly linked to the interatomic potential while remaining conceptually straightforward.

Table 1 consolidates the analytical logic of the four validation levels by distinguishing the observable tested, the physical aspect constrained, the evidence strength provided, and the principal failure mode each level is most likely to expose.

Table 1. Analytical structure of the four-level experimental validation framework for machine learning interatomic potentials

Validation level	Primary experimental observable	Core physical quantity constrained	Main scientific question answered	Typical experimental access	Dominant comparison logic	Evidence strength for real-world fidelity	Principal failure mode
Level 1: Static Elastic Constants	Room-temperature elastic tensor components; derived bulk/shear moduli; anisotropy ratios	Second derivative of energy with respect to uniform strain	Does the potential reproduce macroscopic stiffness and symmetry-consistent	Ultrasonic methods; resonant ultrasound spectroscopy; Brillouin scattering	Component-wise percentage error and consistency of derived moduli	Foundational: establishes whether the potential has correct global curvature under small	Inconsistent equation of state; stress response; symmetry; inconsistency

			strain response?			homogeneous deformation	uns cur unde
Level 2: Zone-Center Phonons	Raman- and IR- active Γ -point phonon frequencies; mode ordering; activity assignment	Long- wavelength dynamical matrix and local bonding environment	Does the potential reproduce local vibrational physics at the zone center?	Raman spectroscopy; infrared spectroscopy	Absolute frequency error, mode ordering, and qualitative activity agreement	Intermediate: moves beyond static strain response to dynamical accuracy at q $= 0$	Loca consta pol rel defic incorr TO be m inve
Level 3: Full Phonon Dispersion	Phonon branches along high- symmetry paths; acoustic slopes; branch crossings/curvature	q -dependent dynamical matrix and force constants across multiple length scales	Does the potential capture spatially resolved vibrational physics throughout the Brillouin zone?	Inelastic neutron scattering; inelastic X-ray scattering	Branch-wise frequency agreement, slope agreement, and shape agreement across q	Strong: provides direct evidence that the model reproduces experimentally resolved lattice dynamics	Missi ra intera inc br top unst dis disp poor ve cons
Level 4: Temperature- Dependent Behavior	Thermal shifts in phonon frequencies; phonon softening; temperature- dependent elastic moduli	Anharmonic free-energy landscape and finite- temperature transferability	Does the potential remain physically valid when thermal effects alter the energy landscape?	Temperature- dependent INS/IXS; Raman/IR; temperature- dependent elastic measurements	Trend agreement in sign, slope, and approximate magnitude of thermal response	Highest: shows whether the potential captures experimentally relevant finite- temperature physics rather than only harmonic structure	Inad anhar inc the exp coupli temp transf

Elastic constants (Static validation)

This foundational level compares the predicted elastic stiffness tensor components (such as C11, C12, C44 for cubic crystals, or the full set for lower symmetry) to room-temperature experimental values. Elastic constants probe the response of the energy surface to uniform strain and serve as a sensitive test of overall potential accuracy. Data are relatively abundant for many materials, making this level broadly applicable. Recommended metrics include the mean absolute percentage error averaged over independent components and errors in derived quantities such as anisotropy

ratios or bulk and shear moduli. Passing criteria can be set conceptually as mean absolute percentage error below 10 percent for high-symmetry simple crystals and up to 20 percent for more complex or anisotropic systems, allowing for typical experimental and methodological variations. Suitable test materials include aluminum, copper, silicon, magnesium oxide, titanium dioxide, and various perovskites, all of which have well-documented experimental elastic tensors [9-11, 14, 21, 23].

Zone-center phonons (Raman/IR accessible modes)

At this level, validation focuses on phonon frequencies at the Brillouin zone center (Γ point), which are measurable by Raman and infrared spectroscopy for optically active modes. These frequencies test the dynamical matrix at long wavelength and reveal local bonding and polarity effects, including LO-TO splitting in polar materials. Metrics center on mean absolute frequency error in wavenumbers, together with qualitative checks on mode ordering and activity. A conceptual threshold of mean absolute error below 10 cm^{-1} for simple crystals provides a reasonable target. The advantage of Level 2 is the wide availability of Raman and IR databases, enabling validation for many compounds without requiring neutron facilities [7, 13].

Full phonon dispersion

This more stringent level evaluates the entire phonon dispersion relation $\omega(\mathbf{q})$ along high-symmetry directions in the Brillouin zone, typically obtained from inelastic neutron scattering. It tests force constants over multiple length scales, acoustic branch slopes (which must match sound velocities from elastic constants), and features such as Kohn anomalies or LO-TO splitting. Metrics include mean absolute frequency error across sampled $\mathbf{q}$ -points, relative errors in acoustic slopes, and agreement in branch shapes or avoided crossings. Suggested criteria involve mean absolute error below 15 cm^{-1} and sound velocity error below 10 percent. Although data exist for only roughly one to two hundred well-studied crystals, these provide critical tests of long-range behavior and dynamical stability that shorter-range or DFT-only checks may miss [2, 5, 6, 29].

Temperature-dependent properties

The highest level incorporates finite-temperature effects by comparing predicted temperature evolution of elastic constants, phonon frequencies (softening or shifts), or related quantities such as thermal expansion to experimental trends. This level probes anharmonicity, which harmonic or quasi-harmonic approximations often fail to capture fully. Metrics are more qualitative at present—agreement in the sign and approximate magnitude of temperature coefficients—because quantitative standards are still maturing. Challenges include the need for temperature-aware potentials or explicit anharmonic calculations. Few materials have extensive temperature-dependent phonon or elastic datasets, limiting this level to well-characterized cases [13, 21].

A conceptual diagram would depict the four levels as a stepped pyramid. The base (Level 1) is wide, representing broad data availability and lower computational demand, with arrows indicating increasing validation stringency and data requirements upward to the apex (Level 4). Each step is labeled with the primary observable, key metrics, and example material classes. Side annotations highlight data sources (ultrasonic/Raman/INS) and typical uncertainties. The pyramid visually conveys that robust validation requires progressing through levels as far as data permit, with higher levels providing stronger evidence of physical accuracy.

This hierarchical structure allows flexible application: basic potentials can target Level 1 and 2, while advanced claims of transferability or thermal transport accuracy should demonstrate success at Level 3 or 4. By defining clear conceptual components and progression, the framework guides consistent evaluation across different machine learning architectures and material classes [8, 28].

Validation Protocol

A reproducible validation protocol anchors experimental comparisons in consistency and transparency. Selection of validation materials begins with compounds possessing high-quality elastic and phonon reference data, deliberately spanning crystal structures, bonding characters, and known DFT sensitivities such as strong correlation or shallow potentials, thereby probing both transferability and robustness while requiring documented experimental uncertainties and conditions.

From these foundations, machine learning predictions are generated by computing elastic constants through controlled small-strain deformations and phonon properties via finite-displacement force constants in supercells, with explicit reporting of convergence parameters, supercell sizes, and cutoffs to ensure direct comparability across systems [7, 19].

Alignment with experiment then demands temperature matching—through nearest available data or justified quasi-harmonic adjustments—explicit tracking of whether predictions fall within experimental error bars, and deployment of complementary metrics ranging from percentage deviations in elastic constants to frequency errors and dispersion-curve shape similarities, yielding a nuanced rather than scalar assessment [8, 28].

This comparison directly informs discrepancy diagnosis, where Table 2 maps characteristic mismatch patterns onto underlying physical deficiencies, the computational scale at which they emerge, and targeted responses in model development.

Table 2. Discrepancy-to-diagnosis matrix for interpreting experimental validation failures in machine learning interatomic potentials

Observed mismatch pattern	Most likely physical or modeling deficiency	First validation level where it becomes clearly visible	Why DFT-only validation may fail to detect it	Recommended diagnostic follow-up	Most appropriate model-improvement response	Consequence if left unresolved
Elastic constants systematically too stiff or too soft across several tensor components	Misrepresented global curvature of the potential energy surface; poor equation-of-state balance	Level 1	Small force RMSE can remain low even when integrated strain response is biased	Recheck strain amplitude sensitivity, stress consistency, equilibrium volume, and symmetry constraints	Retrain with better deformation coverage; improve stress labeling or equation-of-state sampling	Mechanical predictions, sound velocities, and downstream phonons become unreliable
Good elastic constants but incorrect Raman/IR Γ -point frequencies	Local force constants are inaccurate despite acceptable bulk stiffness	Level 2	DFT-matched energies may not enforce correct local vibrational curvature for optically active modes	Inspect finite-displacement settings, mode eigenvectors, polarity treatment, and LO–TO splitting assumptions	Enrich training around local distortions; include polar or long-range electrostatic physics if needed	Spectroscopic interpretation and mode-resolved lattice dynamics are compromised
Γ -point modes appear reasonable but dispersion	Long-range interactions or nonlocal force-constant	Level 3	Agreement at $q = 0$ does not constrain branch	Compare branch slopes, zone-boundary behavior, and	Extend interaction range; use architectures	Thermal transport and dynamical-stability claims

branches distort away from Γ	transferability are missing		evolution across the Brillouin zone	acoustic-optical separation; test supercell convergence	that better capture nonlocality; expand training on displaced supercells	become weak or misleading
Acoustic branch slopes inconsistent with elastic constants	Internal inconsistency between strain response and vibrational response	Level 3	Separate DFT benchmarks may look acceptable without cross-property coherence checks	Cross-check sound velocities derived from both phonons and elastic tensor	Reconcile stress, force, and displacement training coverage; enforce consistency across observables	The potential lacks physical coherence even if individual benchmarks appear acceptable
Correct branch positions at one temperature but wrong thermal shifts or softening trends	Missing anharmonicity or inadequate finite-temperature representation	Level 4	Standard DFT and harmonic benchmarks are usually zero-temperature and cannot test thermal evolution	Compare temperature coefficients, thermal expansion coupling, and mode-specific softening	Introduce temperature-aware or anharmonic training data; include finite-temperature configurations	Finite-temperature applications such as transport, phase behavior, and thermal response become unreliable
Agreement varies strongly across materials of different bonding types or symmetries	Limited transferability of descriptor or training domain	Levels 1–3, depending on property	Low average DFT error can hide uneven out-of-domain performance	Partition validation by chemistry, symmetry, and bonding regime rather than reporting one pooled error	Diversify training set; use stratified sampling; test chemically distinct materials explicitly	Broad transferability claims are not defensible
Predictions fall just outside experimental uncertainty but within literature scatter	Ambiguous attribution: model error versus experimental variability or condition mismatch	Any level	DFT comparison does not force explicit engagement with measurement uncertainty and sample conditions	Audit temperature, pressure, purity, isotope composition, and provenance of reference data	Report uncertainty-aware interpretation rather than binary pass/fail; use consensus experimental windows	Validation claims become overstated or artificially strict

Excellent DFT agreement yet repeated experimental disagreement	Model is reproducing biases of the computational reference rather than measurable physics	Any experimental level	DFT-only validation is a closed-loop benchmark against the same approximation family	Compare across multiple functionals and prioritize experiment-grounded discrepancies	Treat experiment as the decisive arbiter; introduce multi-fidelity correction strategy	The model remains internally consistent but externally untrustworthy
--	---	-------------------------------	--	--	--	--

When differences arise, systematic diagnosis of their origins becomes essential. Discrepancies in elastic constants often trace to inaccuracies in the equation of state or truncated cutoff radii, while zone-center phonon mismatches typically expose local force-constant errors and full dispersion deviations reveal absent long-range interactions [2, 17]. Temperature-dependent deviations, in turn, frequently highlight insufficient anharmonic treatment. This diagnostic process reframes validation as a mechanism for directed model refinement rather than binary assessment [13, 21].

Subsequent reporting should present direct comparisons in tables that juxtapose experimental and predicted values together with uncertainties and multi-metric assessments, classifying outcomes against level-specific benchmarks while openly discussing limitations such as temperature or sample mismatches and outlining refinement pathways. Publications advancing new potentials must therefore demonstrate successful Level 1 validation on at least three diverse materials and Level 2 on at least one, with claims of broad transferability requiring Level 3 support on a representative system [8, 28].

Conceived as conceptual and adaptable, the protocol deliberately avoids mandating particular software yet insists on methodological transparency and candid disclosure of limitations. When implemented, it transforms validation into an integral driver of potential development, grounding the evaluation of machine learning models in fidelity to physical measurements instead of internal computational consistency alone [28].

Relation to Existing Validation Methods

Existing approaches to validating machine learning interatomic potentials fall into several categories, each serving a distinct but limited role within the broader landscape of materials modeling. The most widespread practice is direct comparison against density functional theory reference data [1, 2, 6, 7]. In this paradigm, agreement in forces, energies, or phonon frequencies computed from the same DFT setup is taken as the primary indicator of model quality. While DFT validation remains necessary for ensuring internal consistency during training and hyperparameter tuning, it is conceptually insufficient as a standalone standard. Because DFT itself introduces systematic approximations related to functional choice and zero-temperature conditions, a machine learning potential that matches DFT closely may still diverge from observable physical behavior [8, 28].

A second category focuses on phonon-specific validation, typically by comparing machine learning predictions to DFT-derived phonon properties such as dispersion curves or zone-center frequencies [7, 13]. These studies emphasize the importance of dynamical matrix accuracy and anharmonicity benchmarks, yet they inherit the same reference limitations as general DFT validation. The present framework extends this line of work by shifting the reference from computed phonon properties to those measured directly in experiment, thereby testing whether the potential captures real-world vibrational physics rather than merely replicating a computational proxy [13, 29].

A third related area involves elastic constant benchmarks, often conducted within high-throughput computational screening of alloys or complex phases [9-11, 14, 23]. Here, machine learning models are assessed against DFT elastic tensors for multi-principal element systems or solid solutions. Again, the reference remains computational, which means the benchmark evaluates fidelity to a particular theoretical level rather than to laboratory measurements of stiffness or sound velocity [8].

Uncertainty quantification methods represent a fourth conceptual strand, where researchers attempt to estimate prediction confidence through ensemble techniques or stratified sampling [19]. These approaches highlight the need to propagate uncertainties into validation conclusions, yet they rarely incorporate experimental uncertainty explicitly when judging final accuracy [28]. The four-level experimental framework proposed here builds directly on uncertainty-aware thinking by requiring that predicted elastic constants and phonon frequencies be compared against reported experimental error bars, turning validation into a statistically grounded process rather than a point-wise match.

Validation protocols for machine-learned potentials more generally stress the importance of diverse test sets and diagnostic tools [28]. The current framework aligns with and strengthens these protocols by embedding experimental phonon dispersion and elastic constants as core, non-negotiable components. It does not supplant DFT-based or phonon-DFT comparisons; instead, it positions them as preparatory steps that precede and inform the decisive experimental checks. By doing so, the framework creates a layered validation hierarchy in which DFT agreement serves as a necessary gatekeeper, while experimental agreement at Levels 1–4 establishes sufficiency for real-world deployment [8, 29].

In conceptual terms, the relation can be viewed as moving from internal consistency (DFT validation) to external grounding (experimental validation). This shift mirrors broader trends in machine learning for materials science, where models trained on high-fidelity computations are increasingly required to demonstrate transferability to measurable quantities such as thermal transport indicators or mechanical response [1, 5, 12, 15, 17, 20, 24, 27]. The present framework therefore unifies and elevates these existing strands into a coherent, experiment-centered standard that can be applied uniformly across the wide variety of machine learning potentials now appearing in the literature.

Challenges and Limitations

Several conceptual and practical challenges must be acknowledged when implementing experimental validation of machine learning interatomic potentials. The first is data scarcity. While elastic constant measurements exist for hundreds of crystals, full phonon dispersion relations from inelastic neutron scattering are available for far fewer systems, typically simpler structures [29]. Temperature-dependent phonon or elastic data are rarer still, restricting Level 4 validation to a handful of well-studied materials. This limitation implies that the framework cannot be applied with equal stringency to every new potential or every material class.

A second challenge arises from temperature mismatch. Most machine learning potentials are constructed within the harmonic or quasi-harmonic approximation at zero kelvin, whereas experimental elastic constants and phonon frequencies are reported at room temperature or above [13, 21]. Direct numerical comparison therefore requires either explicit temperature-aware architectures or additional theoretical corrections, both of which introduce further conceptual layers that must be justified transparently.

Defects and anharmonicity present a third difficulty. Experimental samples inevitably contain impurities, vacancies, or grain boundaries, while the potentials under test usually assume perfect periodicity [1, 6, 16, 17]. Moreover, many current architectures do not inherently capture anharmonic effects that become prominent at finite temperature [13]. Discrepancies at higher validation levels may therefore reflect missing physics in the model, real sample imperfections, or both, complicating unambiguous diagnosis.

Computational cost constitutes a fourth limitation. Obtaining reliable phonon dispersion requires force evaluations on large supercells, and repeating these calculations for multiple materials and levels increases the expense of validation itself [7, 19]. Although this cost is lower than full experimental campaigns, it still constrains how many systems can be examined routinely.

Finally, interpretation of discrepancies demands careful conceptual handling. When predicted values fall outside experimental uncertainty, the source could be model error, reference data variation, or incomplete reporting of measurement conditions [28, 29]. Without standardized uncertainty propagation on both sides, it is conceptually difficult to assign responsibility or to decide whether a potential has “passed” a given level.

These challenges do not invalidate the framework; rather, they define its boundaries and highlight areas for community-level solutions such as shared experimental databases and hybrid temperature-aware potentials. By surfacing these limitations explicitly, the framework encourages honest reporting and incremental progress toward more robust validation practices.

Implications for ML Potential Development

The adoption of the proposed four-level experimental validation framework carries direct implications for three stakeholder groups: model developers, end users, and benchmark designers.

For developers of machine learning interatomic potentials, the framework implies a shift in workflow. Experimental validation at Levels 1 and 2 should become routine alongside DFT benchmarks [8, 28]. Discrepancies uncovered at any level can be treated as diagnostic signals that point to missing physical ingredients—long-range interactions at Level 3 or anharmonicity at Level 4—thereby guiding targeted model improvements [2, 13, 17]. Developers are encouraged to report performance against the conceptual criteria for each level, creating a standardized language that facilitates comparison across different architectures and material systems [7, 19].

Practitioners who apply these potentials in materials design or simulation campaigns gain a clearer decision tool. Rather than accepting a model solely on the basis of low DFT error, users can consult validation summaries to assess suitability for specific applications such as thermal transport or mechanical response [1, 5, 9, 11]. The minimum reporting standard—elastic constants for at least three materials and zone-center phonons for one—provides a practical threshold below which a potential should be treated with caution [28].

Benchmark designers and community initiatives benefit by using the framework as a blueprint for curating reference datasets. Centralized collections of critically evaluated experimental elastic constants and phonon dispersion curves, complete with uncertainties and measurement metadata, would accelerate adoption [29]. Such benchmarks would move the field from fragmented, author-specific validation to a shared, experiment-grounded standard, much as existing DFT benchmarks have done for computational consistency.

Overall, the implications reinforce a cultural change: machine learning potentials must be judged not only by how well they reproduce a chosen computational reference but by how faithfully they reflect measurable reality. This evolution strengthens the trustworthiness of downstream applications in phonon transport, defect dynamics, and phase behavior across the diverse materials covered in recent literature [6, 12, 15, 20, 24, 26, 27].

Toward Experimentally Grounded ML Potentials

The long-term vision enabled by this framework is the routine development of machine learning interatomic potentials that are trained on first-principles data yet validated—and, where appropriate, refined—against experimental phonon dispersion and elastic constants. Discrepancies between model predictions and measurements become drivers for iterative improvement rather than endpoints of evaluation.

Several conceptual research directions support this vision. One is the incorporation of multi-fidelity learning strategies that blend DFT training data with sparse experimental anchors, allowing models to correct systematic reference biases [1, 5]. Another is the design of temperature-dependent or anharmonicity-aware architectures capable of reproducing finite-

temperature shifts observed in experiment [13, 21]. A third direction involves architectures that explicitly learn long-range interactions to improve Level 3 performance across a broader range of crystal symmetries [2, 17].

Success can be conceptualized through clear community criteria: a potential that consistently achieves conceptual agreement with experimental elastic constants within typical measurement scatter and with phonon dispersion within accepted spectroscopic tolerances for a diverse set of test materials. Such models would represent a qualitative leap, offering predictions that practitioners can trust without constant DFT cross-checks.

Progress toward this goal will require collective action—shared experimental databases, standardized reporting templates, and open validation challenges built around the four-level structure [8, 28, 29]. When these elements are in place, machine learning potentials will transition from computationally convenient approximations to experimentally grounded, physically reliable tools for materials discovery and design. The framework presented here supplies the conceptual scaffolding necessary to realize that transition.

Conclusion

Machine learning interatomic potentials are typically validated against density functional theory, yet DFT carries systematic errors that may not align with physical reality. Experimental phonon dispersion obtained from inelastic neutron scattering and elastic constants measured by ultrasonic or spectroscopic methods provide the gold-standard reference for assessing true accuracy. The conceptual framework introduced in this paper organizes validation into four progressive levels: Level 1 compares static elastic constants, Level 2 examines zone-center phonon frequencies, Level 3 evaluates full phonon dispersion along high-symmetry paths, and Level 4 assesses temperature-dependent behaviors. Each level is accompanied by plain-language definitions, comparison strategies, and criteria that reflect increasing rigor and data demand.

The accompanying validation protocol guides researchers through material selection, prediction computation, uncertainty-aware comparison, discrepancy diagnosis, and standardized reporting. A minimum standard for new potentials requires Level 1 results on at least three diverse materials and Level 2 results on at least one material, with broader transferability claims supported by Level 3 evidence. This structure complements rather than replaces DFT validation, anchoring computational models in measurable reality while preserving the efficiency of data-driven training.

By placing experiment at the center of the evaluation process, the framework addresses a critical gap in current practice and raises the bar for what constitutes a well-validated machine learning potential. Its adoption will help the community identify models that genuinely capture the physics of phonon transport and mechanical response, ultimately enabling more reliable simulations across metals, semiconductors, ceramics, and complex alloys. The time has come to move beyond DFT-only benchmarks and embrace experimental validation as an essential component of machine learning for materials science.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Jun 2025

Revised: 21 Sep 2025

Accepted: 07 Dec 2025

Published online: 18 January 2026

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

Dou Y, Shimizu K, Carrete J, Fujioka H, Watanabe S. Machine-learning potential for phonon transport in AlN with defects in multiple charge states. *Phys Rev Mater.* 2025;9(3):034601.

<https://doi.org/10.1103/PhysRevMaterials.9.034601>.

Hwang J, Jin Y, Lee J. Machine learning force field based phonon dispersion prediction. *Curr Appl Phys.* 2024;66:76-80.

<https://doi.org/10.1016/j.cap.2024.07.001>.

Rowe P, Csányi G, Alfè D, Michaelides A. Development of a machine learning potential for graphene. *Phys Rev B.*

2018;97(5):054303.

<https://doi.org/10.1103/PhysRevB.97.054303>.

Kruglov I, Sergeev O, Yanilkin AV, Oganov AR. Energy-free machine learning force field for aluminum. *Sci Rep.* 2017;7(1):8512.

<https://doi.org/10.1038/s41598-017-08455-3>.

Dai L, Li M, Hu Y. Machine learning for thermal transport and phonon high-order anharmonicity in high thermal conductivity materials: A case study in boron arsenide. *Phys Rev Mater.* 2025;9(4):045403.

<https://doi.org/10.1103/PhysRevMaterials.9.045403>.

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>.

Loew A, Wang HC, Cerqueira TFT, Marques MAL. Training machine learning interatomic potentials for accurate phonon properties.

Mach Learn Sci Technol. 2024;5(4):045019.

<https://doi.org/10.1088/2632-2153/ad86a1>.

Röcken S, Zavadlav J. Accurate machine learning force fields via experimental and simulation data fusion. *npj Comput Mater.*

2024;10:69.

<https://doi.org/10.1038/s41524-024-01251-4>.

Linton N, Aidhy DS. A machine learning framework for elastic constants predictions in multi-principal element alloys. *APL Mach Learn.* 2023;1(1):016109.

<https://doi.org/10.1063/5.0129928>.

Sidnov K, Konov D, Smirnova EA, Ponomareva AV, Belov MP. Machine learning-based prediction of elastic properties using reduced datasets of accurate calculations results. *Metals*. 2024;14(4):438.
<https://doi.org/10.3390/met14040438>.

Bobbili R. Physics-informed and interpretable machine learning for elastic constants of RHEAs. *Materialia*. 2026;45:102682.
<https://doi.org/10.1016/j.mtla.2026.102682>.

Ghosh A, Moitra A, Saha-Dasgupta T. Ab-initio trained machine learning potential for MAX compound Ti₂AlC: Construction, validation, and study of non linear elasticity. *J Phys Mater*. 2025;8(2):025001.
<https://doi.org/10.1088/2515-7639/ada996>.

Bandi S, Jiang C, Marianetti CA. Benchmarking machine learning interatomic potentials via phonon anharmonicity. *Mach Learn Sci Technol*. 2024;5(3):030502.
<https://doi.org/10.1088/2632-2153/ad674a>.

Volkmer L, Sandonas LM, Grimm P, Hufenbach JK, Cuniberti G. On-the-fly machine learning of interatomic potentials for elastic property modeling in Al-Mg-Zr solid solutions. *Acta Mater*. 2025;305:121844.
<https://doi.org/10.1016/j.actamat.2025.121844>.

Guo J, Ward L, Babuji Y, Hoyt N, Williamson M, Foster I, et al. Composition-transferable machine learning potential for LiCl-KCl molten salts validated by high-energy x-ray diffraction. *Phys Rev B*. 2022;106(1):014209.
<https://doi.org/10.1103/PhysRevB.106.014209>.

Ito K, Yokoi T, Hyodo K, Mori H. Transferability of machine-learning interatomic potential to α -Fe nanocrystalline deformation. *Int J Mech Sci*. 2025;291-292:110132.
<https://doi.org/10.1016/j.ijmecsci.2025.110132>.

Ito K, Yokoi T, Hyodo K, Mori H. Machine learning interatomic potential with DFT accuracy for general grain boundaries in α -Fe. *npj Comput Mater*. 2024;10:255.
<https://doi.org/10.1038/s41524-024-01451-y>.

Willman JT, Nguyen-Cong K, Williams AS, Belonoshko AB, Moore SG, Thompson AP, et al. Machine learning interatomic potential for simulations of carbon at extreme conditions. *Phys Rev B*. 2022;106(18).
<https://doi.org/10.1103/PhysRevB.106.L180101>.

Qi J, Ko TW, Wood BC, Pham TA, Ong SP. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. *npj Comput Mater*. 2024;10:43.
<https://doi.org/10.1038/s41524-024-01227-4>.

Hetti Kankanamalage Y, Xi Y, Zhang S, Singh S, Abdolrahim N. Machine learning-based interatomic potential development and phase transition analysis of ferroelectric hafnium dioxide. *Phys Rev Mater*. 2026;10(1):014414.
<https://doi.org/10.1103/gkvf-t1vl>.

Gao M, Bie X, Wang Y, Li Y, Zhai Z, Lyu H, et al. Accurate deep potential model of temperature-dependent elastic constants for phosphorus-doped silicon. *Nanomaterials (Basel)*. 2025;15(10):769.
<https://doi.org/10.3390/nano15100769>.

Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. *Adv Mater*. 2019;31(46):1902765.
<https://doi.org/10.1002/adma.201902765>.

Hayashi G, Suzuki K, Terai T, Fujii H, Ogura M, Sato K. Prediction model of elastic constants of BCC high-entropy alloys based on first-principles calculations and machine learning techniques. *Sci Technol Adv Mater Methods*. 2022;2(1):381-91.
<https://doi.org/10.1080/27660400.2022.2125853>.

Torres A, de Oliveira AB, Barbosa MDS, Lelovsky LV, de Resende VG, Dutra FDC, et al. Machine learning interatomic potential for the structural properties of iron oxides. *Sci Rep.* 2026;16(1):8576.
<https://doi.org/10.1038/s41598-026-38096-4>.

Zhong P, Kim D, King DS, Cheng B. Machine learning interatomic potential can infer electrical response. *npj Comput Mater.* 2025;11:384.
<https://doi.org/10.1038/s41524-025-01911-z>.

Zhong X, Li Y, John T. A general purposed machine learning interatomic potential for Mg-Al-Si-O system suitable for Earth materials at high pressure and temperature conditions. *npj Comput Mater.* 2026;12:141.
<https://doi.org/10.1038/s41524-026-02056-3>.

Fan Z, Whittaker ML, Asta M. Efficient machine learning interatomic potentials robust for liquid and multiple solid polymorphs of NaF and KF. *Phys Rev Mater.* 2025;9(10):103406.
<https://doi.org/10.1103/xbfm-clgd>.

Morrow JD, Gardner JLA, Deringer VL. How to validate machine-learned interatomic potentials. *J Chem Phys.* 2023;158(12):121501.
<https://doi.org/10.1063/5.0139611>.

Han B, Cheng Y. Benchmarking universal machine learning interatomic potentials for rapid analysis of inelastic neutron scattering data. *Mach Learn Sci Technol.* 2025;6(3):030504.
<https://doi.org/10.1088/2632-2153/adfa68>.