

PhononBench-MP40: A spectrum-resolved benchmark dataset for phonon stability

Wen-Kao Li(李文考), Ze-Feng Gao(高泽峰), and Zhong-Yi Lu(卢仲毅)
Citation: Chin. Phys. B, 2026, 35 (**8**): 086301. DOI: 10.1088/1674-1056/ae843f
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Finite-size effects on phonon-mediated thermal transport across Si-Ge interfaces: Spectral analysis and parameter optimization for molecular dynamics simulations

Zhicong Wei(魏志聪), Haoqiang Li(李浩强), Jianlian Huang(黄建廉), Weikuang Li(李唯宽), Yijuan Li(李艺娟), Yajuan Cheng(程亚娟), and Shiyun Xiong(熊世云)
Chin. Phys. B, 2025, 34 (**11**): 116301. DOI: 10.1088/1674-1056/ae0432

Thermal transport properties of 2D narrow bandgap semiconductor Ca_3N_2, Ba_3P_2, and Ba_3As_2: Machine learning potential study

Wenlong Li(李文龙), Yu Liu(刘余), Zhendong Li(李振东), Pei Zhang(张培), Xinghua Li(李兴华), and Tao Ouyang(欧阳滔)
Chin. Phys. B, 2025, 34 (**9**): 096302. DOI: 10.1088/1674-1056/ade5a0

Lattice and phonon properties in semiconductors $FeSb_2$ and $RuSb_2$

Meng Zhang(张萌), Shengnan Dai(戴胜男), Ranran Zhang(张冉冉), Mingfang Shu(舒明方), Wei Xu(徐威), Jinfeng Zhu(朱金峰), Xianglin Liu(刘祥麟), Yixuan Luo(罗伊轩), Toru Ishigaki, Bo Duan(段波), Yanfeng Guo(郭艳峰), Zhe Qu(屈哲), Jiong Yang(杨炯), and Jie Ma(马杰)
Chin. Phys. B, 2025, 34 (**8**): 086302. DOI: 10.1088/1674-1056/adda0e

Dynamic response of a thermal transistor to time-varying signals

Qinli Ruan(阮琴丽), Wenjun Liu(刘文君), and Lei Wang(王雷)
Chin. Phys. B, 2024, 33 (**5**): 056301. DOI: 10.1088/1674-1056/ad2dcc

Evaluation of performance of machine learning methods in mining structure—property data of halide perovskite materials

Ruoting Zhao(赵若廷), Bangyu Xing(邢邦昱), Huimin Mu(穆慧敏), Yuhao Fu(付钰豪), and Lijun Zhang(张立军)
Chin. Phys. B, 2022, 31 (**5**): 056302. DOI: 10.1088/1674-1056/ac5d2d

---

DATA PAPER

PhononBench-MP40: A spectrum-resolved benchmark dataset for phonon stability

Wen-Kao Li(李文考)[1,2,†], Ze-Feng Gao(高泽峰)[1,2,†‡], and Zhong-Yi Lu(卢仲毅)[1,2,§]

[1] *School of Physics and Key Laboratory of Quantum State Construction and Manipulation (Ministry of Education), Renmin University of China, Beijing* 100872, *China*

[2] *AI-Driven Quantum Materials Research Center, School of Physics, Renmin University of China, Beijing* 100872, *China*

(Received 4 June 2026; revised manuscript received 26 June 2026; accepted manuscript online 30 June 2026)

Imaginary phonon modes remain a practical bottleneck in computational materials screening because otherwise plausible structures can be locally dynamically unstable under a chosen workflow. Here we present PhononBench-MP40, a spectrum-resolved benchmark dataset of Materials Project-derived crystals for workflow-defined phonon stability. The dataset starts from 47969 MP40 workflow tasks and provides 46899 completed records with paired stability labels and local phonopy YAML spectra, including 16683 stable records and 30216 completed-phonon unstable records. A further 1067 relaxation failures are reported separately rather than merged into the completed phonon denominator. The release centers on the local YAML spectrum: the stability label, the lowest sampled frequency and any threshold-dependent relabeling are derived from that spectrum. The dataset is openly available through Science Data Bank at https://doi.org/10.57760/sciencedb.38735. A companion GitHub repository provides the calculation code and lightweight access utilities. PhononBench-MP40 provides an auditable reference for workflow-defined stability classification, minimum-frequency analysis, threshold studies and failure-aware triage, while keeping the reference workflow, data schema and interpretation boundaries explicit.

Keywords: phonon stability, benchmark dataset, phonon spectra, materials informatics

PACS: 63.20.–e, 63.20.D–, 07.05.Mh, 02.70.–c

DOI: 10.1088/1674-1056/ae843f **CSTR:** 32038.14.CPB.ae843f

1. Introduction

Large-scale crystal generation and screening have changed the rate at which candidate materials are proposed.[1–4] The bottleneck is no longer only how to generate structures, but how to decide which of them deserve expensive follow-up. Formation energy, charge balance, symmetry and structural validity are useful filters, but they do not by themselves establish local dynamical stability. Within a specified interatomic-potential and phonon workflow, the relevant question is sharper: does the relaxed structure have a sampled phonon mode with negative curvature, i.e., an imaginary frequency? This makes phonon stability a natural target for reusable computed data and for model evaluation.

The field already has extensive computed-data infrastructure. Materials Project, AFLOW, OQMD, JARVIS, NOMAD, and Materials Cloud organize large computational materials spaces.[5–11] Workflow and analysis tools such as pymatgen, ASE, FireWorks, atomate, and AiiDA support reproducible workflows.[12–16] High-throughput phonon resources and software have established practical routes to vibrational and thermal properties, including DFPT phonons, AFLOW-AAPL, ShengBTE, ALAMODE, TDEP, SCAILD, and recent anharmonic-phonon databases.[17–25] Most existing open phonon datasets contain spectra or vibrational properties for thousands to tens of thousands of materials. Recent data-paper and platform efforts have also emphasized the importance of curated, task-specific materials datasets, including HTSC-2025 for ambient-pressure high-temperature-superconductor critical-temperature prediction, an *ab initio* dataset of size-dependent effective thermal conductivity for advanced technology transistors, and MaterialsGalaxy for fusing experimental and theoretical condensed-matter data.[26–28] PhononBench-MP40 extends this data-resource logic to 46899 completed MP-derived phonon records, while preserving the spectral object behind each stability label.

This design matters because a binary stability flag is convenient but incomplete. A shallow soft mode, a deep imaginary branch, and a failed calculation are different records for both physics and workflow auditing. Collapsing them into one undifferentiated "unstable" class hides the evidence behind the decision and obscures the denominator used in later statistics. PhononBench-MP40 keeps these cases separate. It pairs labels with local phonopy YAML spectra, fixes the completed label+YAML cohort, and reports missing-spectrum relaxation

[†]These authors contributed equally to this work.

[‡]Corresponding author. E-mail: zfgao@ruc.edu.cn

[§]Corresponding author. E-mail: zlu@ruc.edu.cn

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

failures outside the completed phonon denominator. The result is both a data release and a benchmark target: users can work with a fixed PhononBench-defined label, but can also inspect the spectrum, extract a scalar frequency diagnostic, or move the stability threshold for their own application.

PhononBench-MP40 should be interpreted as a workflow-defined harmonic phonon dataset. An imaginary mode in this release indicates a sampled negative-curvature mode under the adopted 0 K MatterSim/uMLIP–phonopy finite-displacement workflow. It is not an absolute statement of thermodynamic instability or finite-temperature instability. Strongly anharmonic materials, superionic conductors, and high-symmetry phases near structural phase transitions may show imaginary harmonic modes at 0 K but become dynamically stabilized at finite temperature. This boundary is central to the intended use of the dataset: the released spectra provide auditable workflow evidence, while borderline soft-mode cases remain candidates for follow-up calculations when stricter physical conclusions are required.

2. Dataset construction and workflow

Figure 1 summarizes the construction. The starting point is an audited MP40 task cohort containing 47969 recorded tasks within the local $N_{\mathrm{atoms}} \leq 40$ workflow scope. Atom-count windows drawn from Materials Project are common in crystal-generation benchmarks, including MP-20, MP-21–40 and MP-40 settings.[1,29,30] Here the cutoff also has a direct phonon-workflow meaning. Under the $2 \times 2 \times 2$ finite-displacement supercell used in the workflow, a 40-atom recorded cell corresponds to an up-to-320-atom force-evaluation cell before symmetry reduction. The MP40 cutoff therefore defines a practical release domain: it includes many inorganic prototypes, ordered derivatives and multicomponent cells, while keeping the large-scale finite-displacement calculation computationally controlled. It is a benchmark-domain definition rather than a physical stability boundary.

Each task follows the PhononBench route. Structures are relaxed, finite-displacement supercells are constructed, forces are evaluated with MatterSim/uMLIP, and phonopy is used to build the dynamical matrix and phonon outputs.[31–34] High-symmetry paths are generated through the seekpath/spglib ecosystem.[35,36] The top row of Fig. 1 is therefore shown as an explicit workflow rather than as a stand-alone label source: the reference target is a computed phonon result under a specified route. The force evaluations are therefore MLIP-dependent. MatterSim/uMLIP defines the force model of the reference workflow and enables the present scale of finite-displacement calculations, but the resulting spectra inherit the accuracy and generalization limits of the adopted potential. Chemical spaces, structural prototypes, oxidation states, magnetic or strongly correlated systems, and distorted local environments that are underrepresented in the training distribution may be more susceptible to systematic or statistical force errors. These errors can propagate into force constants, ω_{min}, and especially near-threshold stability labels. PhononBench-MP40 should therefore be regarded as a documented MLIP-based reference workflow rather than a replacement for material-specific first-principles or anharmonic validation.

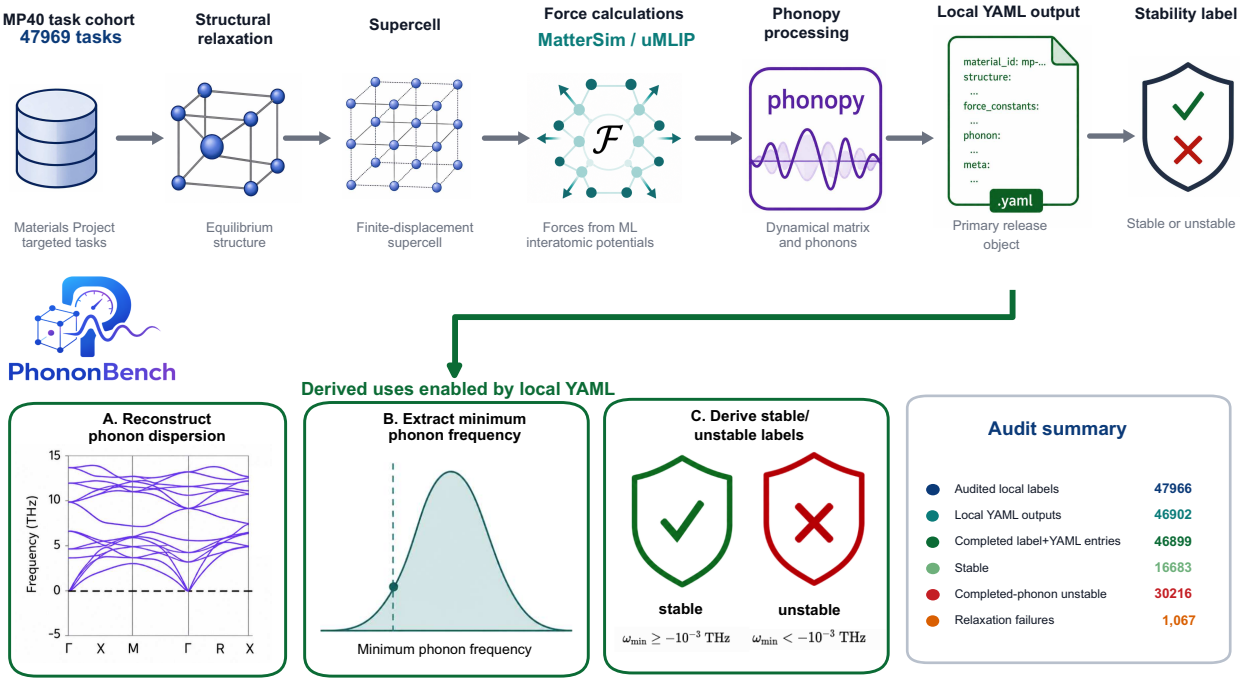


Fig. 1. Workflow and release object of PhononBench-MP40. The dataset starts from 47969 MP40 tasks and follows the PhononBench route through structural relaxation, finite-displacement supercell construction, MatterSim/uMLIP force evaluation and phonopy processing. The local YAML output is the primary release object because it enables dispersion reconstruction, minimum-frequency extraction and stability-label derivation. The audit summary reports 47966 audited local labels, 46902 local YAML outputs, 46899 completed label + YAML entries, 16683 stable entries, 30216 completed-phonon unstable entries and 1067 relaxation failures.

Table 1. Cohort accounting. Counts for the MP40 audit and completed spectral cohort.

Quantity	Count
Audited MP40 tasks	47969
Recovered workflow labels	47966
Recovered phonopy YAML outputs	46902
Completed label+YAML records	46899
stable completed records	16683
Completed-phonon unstable records	30216
Relaxation failures without YAML	1067

The primary release object is the local phonopy YAML file. It contains the spectral information needed to reconstruct the dispersion, extract the lowest sampled phonon frequency, and derive a stable/unstable decision. The label is thus a derived view of the spectrum. A completed record in PhononBench-MP40 is defined by the intersection of two local objects: a recovered workflow stability label and a matching local YAML output. This rule yields 46899 completed label + YAML records from 47966 audited local labels and 46902 local YAML outputs. Within the completed cohort, 16683 records are stable and 30216 records are completed-phonon unstable. The audit also reports 1067 relaxation failures separately, because no matching local YAML spectrum is available for those records. This separation is essential for denominator-aware reuse: a completed imaginary-mode spectrum and a missing-spectrum relaxation failure answer different scientific and computational questions. Release-file definitions, common fields, workflow settings, and the label rule are provided in the supplementary information, Sections S1–S3.

3. Relation to PhononBench

The present release builds on the PhononBench workflow introduced in Ref. [34]. PhononBench was designed to evaluate the dynamical stability of AI-generated crystal structures under a MatterSim/phonopy route. In contrast, PhononBench-MP40 applies the same documented phonon route to an MP-derived $N_{\mathrm{atoms}} \leq 40$ cohort of known database structures and releases the completed spectral records as a reusable data resource.

The distinction is therefore in the source cohort, target task, and release object. PhononBench focuses on generation-model evaluation and reports stability statistics for generated crystals. PhononBench-MP40 focuses on MP-derived reference spectra: it fixes a completed label + YAML denominator, releases local phonopy YAML spectra for 46899 completed records, provides ω_{min} values and a formula-group split, and reports 1067 relaxation failures separately. The shared workflow is intentional because it makes generated-crystal evaluation and MP-derived reference spectra comparable under a common phonon route. The new contribution of PhononBench-MP40 is the spectrum-resolved MP40 data release, threshold-relabeling target, failure-cohort accounting, and robustness analyses, rather than the original definition of the PhononBench workflow.

4. Data records

The released data are organized to make the stable/unstable decision traceable. The archival Science Data Bank record contains the completed label table, the failure-accounting table, the reference split table, release metadata, and the phonopy YAML spectra for completed records. Table 2 summarizes the data objects that define the public release. The completed table is the main machine-learning table: it links each completed record to a structure identifier, formula, space group, unit-cell atom count, workflow label, completed label, YAML path, ω_{min} value, and split assignment. The failure table is kept separate because a relaxation failure without a YAML spectrum is a workflow outcome, not a completed phonon-instability label.

Table 2. Released data records. The public release is defined by tabular metadata plus phonopy YAML spectra. Counts refer to the release described in this work.

Data object	Records	Main use
Audited MP40 task table	47969	Defines the recorded MP-derived task scope and denominator for audit accounting.
Completed label+YAML table	46899	Main spectral cohort linking structure identifiers, workflow labels, completed labels, YAML paths, ω_{min} values and split assignments.
Phonopy YAML spectra	46899	Primary scientific object for reconstructing dispersions, extracting frequency diagnostics and re-deriving threshold-dependent labels.
Relaxation-failure table	1067	Separate workflow-failure cohort for failure-aware triage and denominator-aware reporting.
Reference split table	46899	Formula-group train/validation/test assignment for benchmark comparisons without exact-formula leakage.
Metadata and integrity files	–	Version, repository, file-integrity and reuse information for reproducible access.

This separation is important for reuse. Users who need a fixed binary target can train on the completed label+YAML cohort. Users who need more physically resolved targets can reconstruct the band frequencies from the YAML files, calculate ω_{min}, or impose an application-specific threshold. Users

interested in high-throughput workflow robustness can include the 1067 relaxation failures in a separate failure-aware triage task. The file schema, field definitions, split construction, and example parsing workflow are given in the supplementary information, Sections S1, S2, S5, and S9.

5. Spectra behind the labels

The completed stability label is derived from high-symmetry-path band frequencies. A completed entry is labeled unstable if any sampled band frequency is below -10^{-3} THz; otherwise it is labeled stable. This threshold turns a phonon dispersion into a machine-readable target, but the spectrum is the more informative object. We use $\omega_{\min}$ as the lowest sampled frequency along the released path. For negative values, $\omega_{\min}$ measures the severity of an imaginary branch under the reference workflow; near-zero values identify threshold-sensitive cases. It is a compact diagnostic of the YAML spectrum, not a replacement for the full dispersion and not a model-independent stability law. For severity analysis, we further classify the completed spectra using the YAML-derived minimum frequency $\omega_{\min}$ along the released high-symmetry path. The primary stable/unstable label remains unchanged and follows the workflow threshold: spectra with $\omega_{\min} < -10^{-3}$ THz are labeled unstable, whereas the remaining completed spectra are labeled stable within tolerance. For interpretation and robustness-test stratification, we use three unstable severity bins: near-threshold unstable (-0.05 THz $\leq \omega_{\min} < -10^{-3}$ THz), shallow soft-mode unstable (-1.0 THz $< \omega_{\min} < -0.05$ THz), and deep-imaginary unstable ($\omega_{\min} \leq -1.0$ THz). These bins are technical severity categories derived from the released spectra, not universal physical phase-stability boundaries.

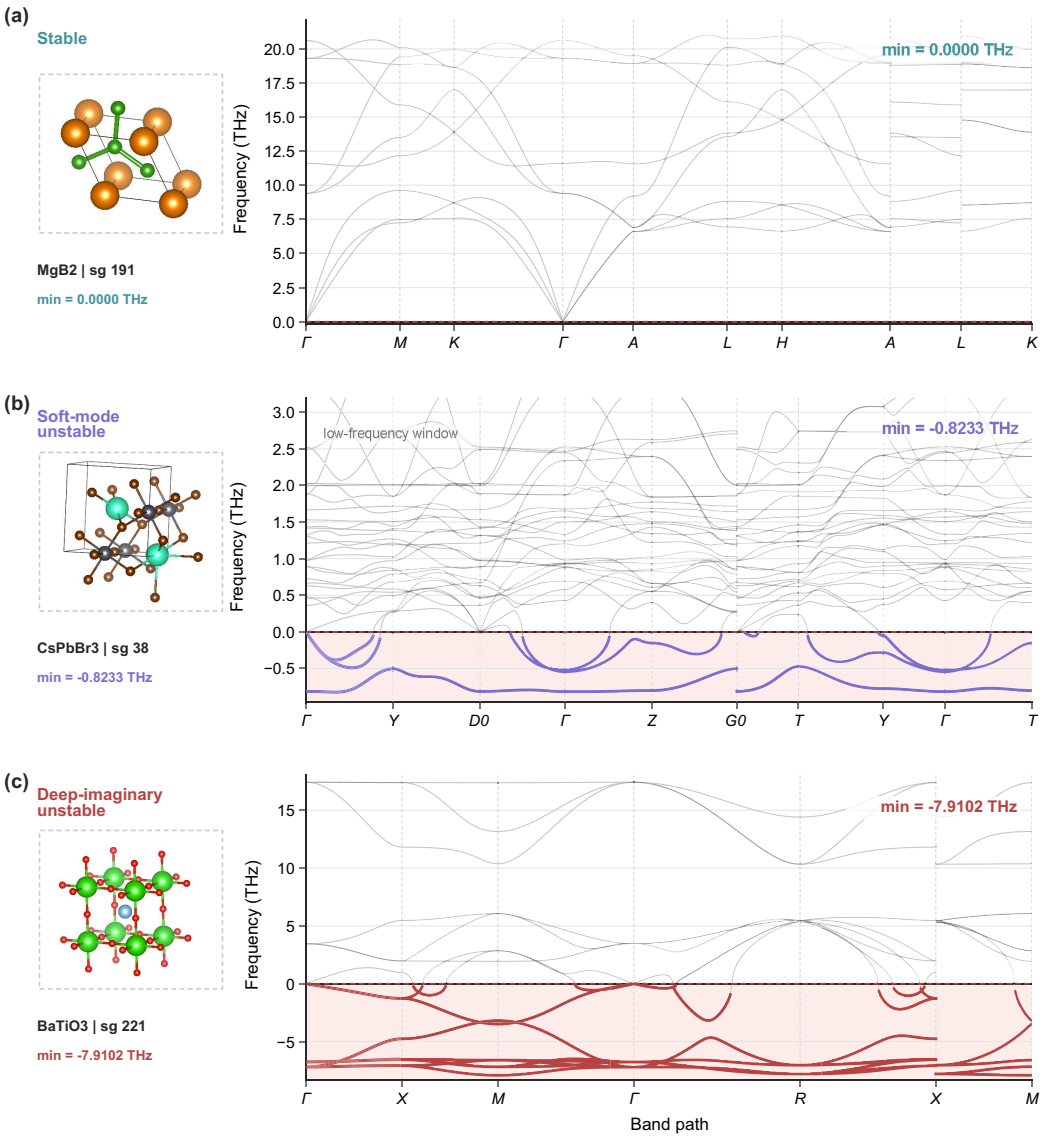


Fig. 2. Representative phonon spectra behind the stability labels. Three YAML-derived dispersions illustrate the label rule and the information retained beyond the binary decision. (a) MgB_2 is labeled stable, with $\omega_{\min} = 0.0000$ THz within the adopted tolerance. (b) $CsPbBr_3$ is a shallow soft-mode completed-phonon unstable example with $\omega_{\min} = -0.8233$ THz. (c) $BaTiO_3$ is a deep-imaginary completed-phonon unstable example with $\omega_{\min} = -7.9102$ THz. Grey branches denote non-negative modes, colored branches denote imaginary modes and the dashed line marks zero frequency.

Figure 2 illustrates why the spectral object matters. MgB_2 is labeled stable within the adopted tolerance, with $\omega_{\min} = 0.0000$ THz. $CsPbBr_3$ is a shallow soft-mode completed-phonon unstable example with $\omega_{\min} = -0.8233$ THz, falling in the -1.0 THz $< \omega_{\min} < -0.05$ THz severity bin. Such shallow imaginary modes are often the cases most sensitive to relaxation details, threshold choice, supercell construction, and low-symmetry distortions. $BaTiO_3$ is also unstable, but with a much deeper imaginary branch, $\omega_{\min} = -7.9102$ THz, indicating a stronger negative-curvature signal under the same reference route. These records share the same binary decision, yet the underlying spectra carry different physical and computational information. A label table alone would hide this distinction; a YAML-backed release keeps it available for inspection, threshold analysis, and learning tasks. The example metadata are listed in the supplementary information, Section S4.

The spectrum-resolved design also defines how the data can be reused. A classifier may learn the PhononBench-defined stable/unstable target. A regression model may predict $\omega_{\min}$ to distinguish near-threshold modes from deep imaginary branches. A calibration study may move the -10^{-3} THz threshold and quantify label changes. These tasks all refer back to the same released YAML objects rather than to disconnected post-processing tables.

6. Chemical and structural coverage

Figure 3 summarizes the completed cohort rather than the planned task list. All panels use the same denominator: the 46899 completed label+YAML records. The completed cohort spans 87 observed elements and 182 observed space groups. Oxygen appears in 37.2% of completed records, making it the dominant element in the occurrence count. This O-rich character is chemically natural for MP-derived inorganic crystals: O^{2-} forms robust coordination polyhedra with many metal cations, stabilizes a wide range of oxide and oxyanion frameworks, and supports charge-balanced multication chemistries. The extended element, space-group and chemical-complexity statistics are reported in the supplementary information, Section S6.

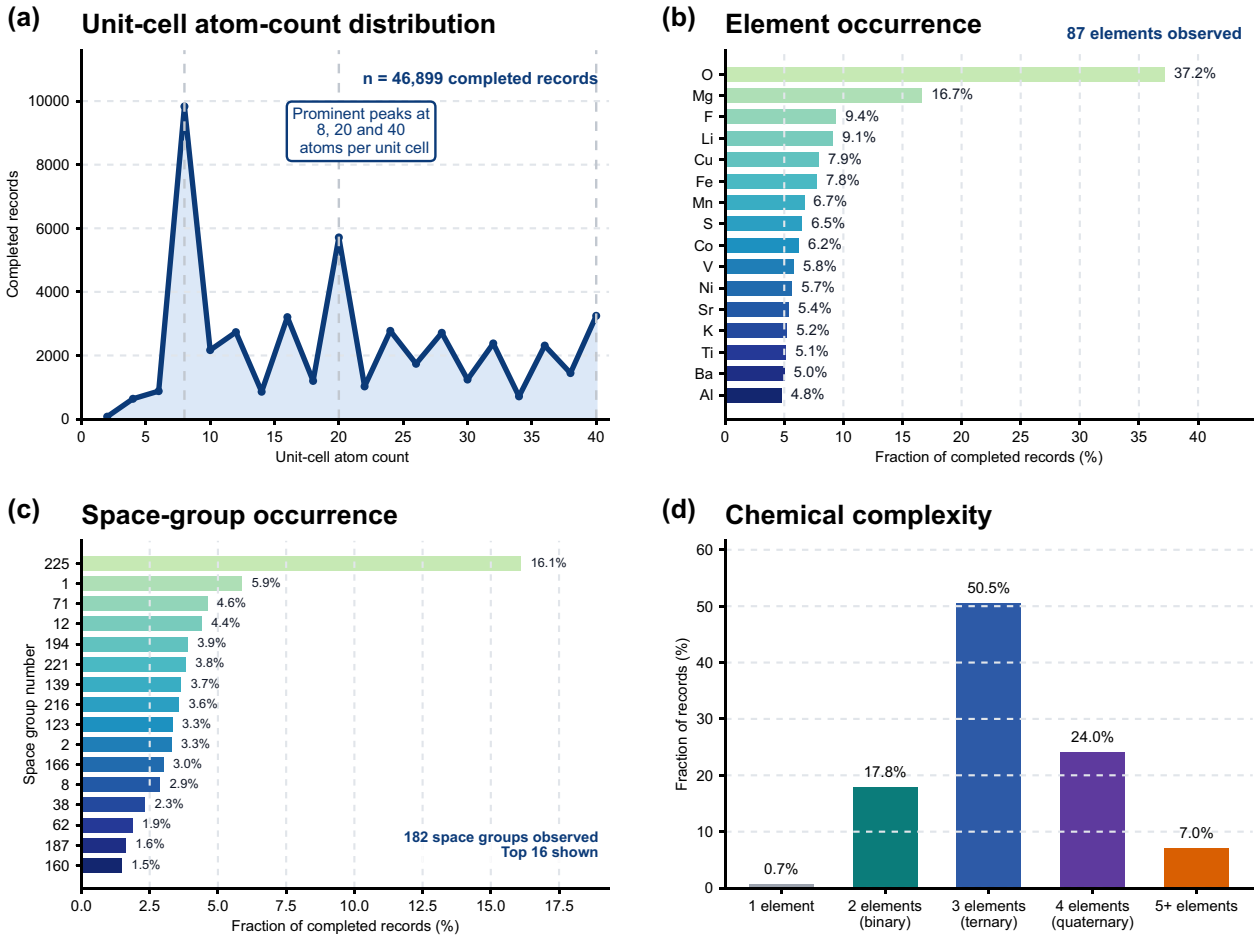


Fig. 3. Chemical and structural coverage of the completed cohort. All panels are computed over the 46899 completed label+YAML records. (a) Distribution of recorded unit-cell atom counts within the MP40 scope, with prominent peaks at 8, 20 and 40 atoms. (b) Element occurrence, counted once per completed record when an element appears in the parsed formula. (c) Ranked space-group occurrence, showing broad but uneven crystallographic coverage. (d) Chemical complexity measured by the number of distinct elements in the parsed formula, with ternary and quaternary compounds forming the largest groups.

The chemical-complexity distribution is dominated by ternary and quaternary compounds: 50.5% of records contain three elements and 24.0% contain four. This distribution is relevant for reuse and model evaluation. Binary compounds provide important prototypes, but additional cation or anion species expand charge-compensation routes, site-substitution degrees of freedom, and coordination environments. In oxides and oxyanion frameworks, for example, robust oxygen coordination can coexist with multiple metal sublattices, substitutions, or charge-balancing species. For machine learning, the dataset therefore extends beyond simple binary compounds and contains many chemically richer compositions for which formula-level cues alone are less likely to capture the relevant lattice-dynamical behavior.

The unit-cell atom-count and space-group panels should be read as coverage statistics of the completed MP40 cohort, not as stability trends. Peaks at 8, 20, and 40 atoms reflect the MP40 source space, conventional-cell choices, and crystallographic representation. The 40-atom edge is especially important: it admits ordered superstructures and multicomponent inorganic cells that are more demanding than very small-cell benchmarks, while avoiding a release dominated by very large phonon supercells. The ranked space-group plot shows broad but uneven crystallographic coverage, with high-symmetry prototypes and lower-symmetry distorted or ordered structures both present. Reporting these distributions is part of the audit: users can see the chemical and structural support over which the stability labels and spectra are defined.

7. Technical validation

PhononBench-MP40 is validated as an auditable workflow dataset rather than as a claim of model-independent absolute dynamical stability. The first validation layer is cohort accounting: task records, recovered local labels and recovered YAML outputs are intersected explicitly, giving 46899 completed label + YAML records and identifying 1067 relaxation failures outside the completed phonon denominator. This accounting prevents failed relaxations from being silently merged with completed imaginary-mode spectra.

The second validation layer is label traceability. For each completed record, the released label is derived from the sampled high-symmetry-path frequencies in the corresponding phonopy YAML file. The label rule is deterministic: any sampled band frequency below -10^{-3} THz gives the completed-phonon unstable label, while all other completed spectra are labeled stable. Because the YAML file is released, users can re-extract $\omega_{\min}$, reproduce the binary label, and test alternative thresholds without relying on a disconnected label table.

The third validation layer is distributional and physical sanity checking. The release reports the completed-cohort element distribution, space-group distribution, unit-cell atom-count distribution, and chemical-complexity distribution using the same 46899-record denominator. Representative spectra in Fig. 2 further check that the label rule distinguishes stable spectra, shallow soft-mode spectra, and deep-imaginary spectra in a physically interpretable way. Additional validation and recommended integrity checks are described in the supplementary information, Sections S6, S10, and S11.

To assess the sensitivity of the workflow label to the finite-displacement supercell, we performed an additional $2 \times 2 \times 2$ versus $3 \times 3 \times 3$ robustness test on a stratified small-cell subset of 60 completed entries with unit-cell sizes no larger than 6 atoms. The subset contains 20 workflow-stable structures, 20 near-threshold or shallow soft-mode unstable structures, and 20 deep-imaginary unstable structures. For the $3 \times 3 \times 3$ reruns, the MatterSim relaxation protocol, displacement amplitude of 0.01 Å, seekpath high-symmetry path generation, and stability threshold of -10^{-3} THz were kept unchanged. All 60 $3 \times 3 \times 3$ calculations completed successfully. The binary labels agree for 57/60 cases between the original $2 \times 2 \times 2$ workflow and the $3 \times 3 \times 3$ rerun. All 20 deep-imaginary cases remain unstable, whereas the three label changes occur only in the stable or shallow-soft-mode boundary regimes. This subset test is not intended to prove full supercell convergence for the entire MP40 dataset, but it supports the interpretation that deep imaginary branches are more robust under the larger-supercell rerun, while borderline soft modes are more sensitive to workflow choices and should be interpreted with caution.

The 1067 relaxation failures were further analyzed as a separate workflow-failure cohort. The direct technical cause of these records is relaxation non-convergence under the adopted MatterSim/ASE FIRE protocol: all 1067 cases failed the final force criterion before a force-constants YAML file was generated. Therefore, they are not completed phonon calculations and are kept outside the completed stable/unstable denominator. Compared with the 46899 completed records, the failure cases are biased toward larger and chemically more complex structures. The median unit-cell size of the failure cohort is 28 atoms, compared with 20 atoms for the completed cohort. In addition, 78.4% of the failure records contain more than 20 atoms, whereas this fraction is 41.7% for the completed cohort. The failure cohort is also enriched in quaternary compositions: 46.3% of failures are quaternary, compared with 24.0% in the completed cohort. Element-wise occurrence analysis shows higher relative occurrence of several oxide/sulfide and transition-metal-containing chemistries, including records containing S, Cr, Mg, Co, V, W, Mn, O, Ti, Fe, Ni and Mo. We interpret these trends as indicators of more challenging relaxation landscapes for the fixed MLIP-based workflow, rather than as intrinsic physical instability of these chemistries.

We also tested the sensitivity of the finite-displacement amplitude on an unstable-only subset. The subset contains 60 completed-phonon unstable records selected from the original

0.01 Å workflow spectra, including 15 near-threshold unstable, 25 shallow soft-mode unstable, and 20 deep-imaginary unstable structures. To isolate the effect of displacement amplitude, the relaxed structures were fixed, and only the phonopy displacement amplitude was changed from the reference value of 0.01 Å to 0.02 Å and 0.03 Å. The MatterSim calculator, $2\times2\times2$ supercell, force-constant symmetrization, seekpath high-symmetry path, and -10^{-3} THz stability threshold were kept unchanged. All 120 recalculations completed successfully. The unstable label was preserved for 59/60 cases at 0.02 Å and 57/60 cases at 0.03 Å. All 20 deep-imaginary cases remained unstable at both larger amplitudes, whereas the few label flips occurred only in near-threshold or shallow soft-mode cases. This result supports the use of 0.01 Å as the reference displacement amplitude for the released workflow, while reinforcing the need to treat borderline soft modes as workflow-sensitive records.

8. Benchmark tasks and reuse

PhononBench-MP40 is useful as a data release because spectra, labels and accounting are distributed together. It can also be used as a benchmark target because the completed cohort and label rule are fixed. For users who wish to compare models under a shared setting, we provide a deterministic formula-group reference split, while recognizing that it is not the only valid way to use the release. The split assigns 37633 training records, 4724 validation records and 4542 test records so that every one of the 35348 formula groups appears in exactly one partition. This prevents exact-formula leakage and gives future studies a common starting point. If external materials databases are used for pretraining or feature generation, overlap with test structures, formulas or Materials Project identifiers should be reported because it changes how model scores should be interpreted. Split construction and reporting suggestions are detailed in the supplementary information, Sections S5 and S7.

The most direct benchmark task is PhononBench-defined stability classification on the completed label + YAML cohort. Since the completed cohort is imbalanced, with 30216 completed-phonon unstable records and 16683 stable records, accuracy alone is not a sufficient metric. Balanced accuracy, macro-F1, AUROC, AUPRC, and class-specific recall are more informative for future comparisons. The released spectra also support ω_{min} prediction and threshold-dependent relabeling. Separately, the 1067 relaxation failures define a failure-aware triage problem: predicting whether a workflow produces an auditable spectrum is useful, but it should not be confused with predicting a completed phonon instability.

For reproducible baseline reporting, we recommend that future model reports include at least one simple composition-only baseline, one structure-aware baseline and, when computationally feasible, a graph-neural-network baseline such as CGCNN, MEGNet or ALIGNN.[37–39] The purpose of these baselines is not to rank all model families in the present data paper, but to make the benchmark entry point clear: the split, labels, frequency target and failure-aware task are fixed, while methods can be compared under transparent reporting rules. The supplementary information, Section S7, lists the recommended baseline protocol and metrics.

This benchmark should be interpreted with the right boundary. PhononBench-MP40 does not claim model-independent absolute dynamical stability. It fixes a documented PhononBench/MatterSim–phonopy reference workflow. Other potentials, denser sampling, larger supercells, anharmonic treatments, or first-principles phonon calculations may change borderline spectra. That limitation is also why the YAML object is released: users can inspect imaginary branches, measure threshold sensitivity, and validate high-confidence candidates with the workflow or higher-fidelity calculations when needed. Additional interpretation boundaries are summarized in the supplementary information, Section S8.

9. Relation to machine-learning targets

Materials machine learning increasingly relies on fixed tasks, transparent splits, and reproducible metrics, as illustrated by Matbench, Matbench Discovery, and OC20.[40–42] PhononBench-MP40 brings the same logic to workflow-defined phonon stability while retaining the spectral evidence behind each target. Descriptor-based materials learning, matminer workflows, crystal graph neural networks, SchNet, MEGNet, and ALIGNN are natural model families for future structure-to-label and structure-to-frequency studies.[37–39,43–45] Universal interatomic potentials such as M3GNet, CHGNet, NequIP, MACE, and MatterSim have made large atomistic workflows increasingly practical.[33,46–50] Recent work also shows that phonons remain a demanding and informative test for such models.[51] PhononBench-MP40 gives this ecosystem a concrete reference target for phonon-stability screening: not an informal pass/fail label, but a spectrum-backed workflow result. The present release does not rank model families; it supplies the shared target, split files, and spectral evidence needed for future comparisons.

10. Usage notes and limitations

The release is intended for three common entry points. First, users can read the completed table, follow the `yaml_relative_path` field to a phonopy YAML file, reconstruct the released dispersion and reproduce the label rule. Second, benchmark users can join the completed table with the reference split table and evaluate models on the fixed formula-group split. Third, workflow users can combine the completed table and the failure table to study whether a structure is likely

to produce an auditable phonon spectrum under the reference route.

Several limitations should be retained in downstream use. PhononBench-MP40 is not a universal statement about thermodynamic stability, finite-temperature stability or DFT-level phonon stability. It records the result of a specified MP-derived PhononBench/MatterSim–phonopy workflow, a specified supercell choice, a specified path sampling procedure and a specified numerical threshold. Borderline structures, especially those with shallow soft modes, should be treated as candidates for follow-up calculation rather than final physical classifications. The released spectra and failure accounting are meant to make these limitations measurable instead of hidden. The additional supercell- and displacement-amplitude robustness tests further support this interpretation: clear deep-imaginary cases are more robust under the tested workflow changes, whereas near-threshold and shallow soft-mode cases remain the sensitivity-dominated regime.

11. Conclusion

PhononBench-MP40 converts a large MP-derived PhononBench calculation into a reusable spectrum-resolved benchmark dataset for workflow-defined phonon stability. The main release contains 46899 completed label + YAML records, including 16683 stable records and 30216 completed-phonon unstable records, with 1067 relaxation failures reported separately. Its central contribution is the release object: labels, spectra, splits and failure accounting are kept together, so each completed stability decision can be traced back to phonon evidence. This makes PhononBench-MP40 useful for stability classification, minimum-frequency analysis, threshold-dependent relabeling and failure-aware screening, while making the reference workflow and its boundaries explicit.

Data availability statement

The data presented in this paper are openly available in Science Data Bank at the following link https://doi.org/10.57760/sciencedb.38735.

Acknowledgments

Project supported by the National Natural Science Foundation of China (Grant Nos. 12434009 and 62476278) and Beijing Natural Science Foundation (Grant Nos. Z260002 and Z250005). Computational resources were provided by the Physical Laboratory of High Performance Computing at Renmin University of China.

References

[1] Xie T, Fu X, Ganea O E, Barzilay R and Jaakkola T 2022 arXiv:2110.06197
[2] Jiao R, Huang W, Lin P, Han J, Chen P, Lu Y and Liu Y 2023 arXiv:2309.04475
[3] Merchant A, Batzner S, Schoenholz S S, Aykol M, Cheon G and Cubuk E D 2023 *Nature* **624** 80
[4] Zeni C, Pinsler R, Zügner D, Fowler A, Horton M, Fu X, Wang Z, Shysheya A, Crabbé J, Ueda S, Sordillo R, Sun L, Smith J, Nguyen B, Schulz H, Lewis S, Huang C W, Lu Z, Zhou Y, Yang H, Hao H, Li J, Yang C, *et al.* 2025 *Nature* **639** 624
[5] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K A 2013 *APL Mater.* **1** 011002
[6] Curtarolo S, Setyawan W, Hart G L W, Jahnátek M, Chepulskii R V, Taylor R H, Wang S, Xue J, Yang K, Levy O, Mehl M J, Stokes H T, Demchenko D O and Morgan D 2012 *Comput. Mater. Sci.* **58** 218
[7] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *JOM* **65** 1501
[8] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 *npj Comput. Mater.* **1** 15010
[9] Choudhary K, Garrity K F, Reid A C E, DeCost B, Biacchi A J, Hight Walker A R, Trautt Z, Hattrick-Simpers J, Kusne A G, Centrone A, Davydov A, Jiang J, Pachter R, Cheon G, Reed E, Agrawal A, Qian X, Sharma V, Zhuang H L, Kalish I, Vogel D, Carrete J, Yan Q, Cao C, Yu R, Doak J, Neaton J B, Hautier G and Tavazza F 2020 *npj Comput. Mater.* **6** 173
[10] Draxl C and Scheffler M 2018 *MRS Bull.* **43** 676
[11] Talirz L, Kumbhar S, Passaro E, Yakutovich A V, Granata V, Gargiulo F, Borelli M, Uhrin M, Huber S P, Zoupanos S, Adorf C S, Andersen C W, Schütt O, Pignedoli C A, Passerone D, VandeVondele J, Schulthess T C, Smit B, Pizzi G and Marzari N 2020 *Sci. Data* **7** 299
[12] Ong S P, Richards W D, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier V L, Persson K A and Ceder G 2013 *Comput. Mater. Sci.* **68** 314
[13] Larsen A H, Mortensen J J, Blomqvist J, Castelli I E, Christensen R, Dulak M, Friis J, Groves M N, Hammer B, Hargus C, Hermes E D, Jennings P C, Jensen P B, Kermode J, Kitchin J R, Kolsbjerg E L, Kubal J, Kaasbjerg K, Lysgaard S, Maronsson J B, Maxson T, Olsen T, Pastewka L, Peterson A, Rostgaard C, Schiotz J, Schütt O, Strange M, Thygesen K S, Vegge T, Vilhelmsen L, Walter M, Zeng Z and Jacobsen K W 2017 *J. Phys.: Condens. Matter* **29** 273002
[14] Jain A, Ong S P, Chen W, Medasani B, Qu X, Kocher M, Brafman M, Petretto G, Rignanese G M, Hautier G, Gunter D and Persson K A 2015 *Concurr. Comput.: Pract. Exp.* **27** 5037
[15] Mathew K, Montoya J H, Faghaninia A, Dwaraknath S, Aykol M, Tang H, Chu I H, Smidt T, Bocklund B, Horton M, Dagdelen J, Wood B, Liu Z K, Neaton J B, Ong S P, Persson K A and Jain A 2017 *Comput. Mater. Sci.* **139** 140
[16] Pizzi G, Cepellotti A, Sabatini R, Marzari N and Kozinsky B 2016 *Comput. Mater. Sci.* **111** 218
[17] Petretto G, Dwaraknath S, Miranda H P C, Winston D, Giantomassi M, van Setten M J, Gonze X, Persson K A, Hautier G and Rignanese G M 2018 *Sci. Data* **5** 180065
[18] Plata J J, Nath P, Usanmaz D, Carrete J, Toher C, de Jong M, Asta M, Fornari M, Nardelli M B and Curtarolo S 2017 *npj Comput. Mater.* **3** 45
[19] Carrete J, Li W, Mingo N, Wang S and Curtarolo S 2014 *Phys. Rev. X* **4** 011019
[20] Mounet N, Gibertini M, Schwaller P, Merkys A, Castelli I E, Cepellotti A, Pizzi G and Marzari N 2018 *Nat. Nanotechnol.* **13** 246
[21] Li W, Carrete J, Katcho N A and Mingo N 2014 *Comput. Phys. Commun.* **185** 1747
[22] Tadano T, Gohda Y and Tsuneyuki S 2014 *J. Phys.: Condens. Matter* **26** 225402
[23] Hellman O, Abrikosov I A and Simak S I 2011 *Phys. Rev. B* **84** 180301
[24] Souvatzis P, Eriksson O, Katsnelson M I and Rudin S P 2009 *Comput. Mater. Sci.* **44** 888
[25] Ohnishi M, Deng T, Torres P, Xu Z, Tadano T, Zhang H, Nong W, Hanai M, Wang Z, Morita M, Tian Z, Hu M, Ruan X, Yoshida R, Suzumura T, Lindsay L, McGaughey A J H, Luo T, Hippalgaonkar K and Shiomi J 2026 *npj Comput. Mater.* **12** 150
[26] Han X Q, Gao Z F, Wang X D, Ouyang Z, Guo P J and Lu Z Y 2025 *Chin. Phys. B* **34** 100301
[27] Xie H, Jia R, Xia Y, Li L, Hu Y, Xu J, Sheng Y, Wang Y and Bao H 2025 *Chin. Phys. B* **34** 046501
[28] Zhu T, Fang Z, Wu Q and Weng H 2025 *Chin. Phys. B* **34** 120702

[29] Xiao H, Li R, Shi X, Chen Y, Zhu L, Chen X and Wang L 2023 *Nat. Commun.* **14** 7027
[30] Park H, Onwuli A and Walsh A 2025 *Nat. Commun.* **16** 4379
[31] Togo A and Tanaka I 2015 *Scr. Mater.* **108** 1
[32] Togo A 2023 *J. Phys. Soc. Jpn.* **92** 012001
[33] Yang H, Hu C, Zhou Y, Liu X, Shi Y, Li J, Li G, Chen Z, Chen S, Zeni C, Horton M, Pinsler R, Fowler A, Zügner D, Xie T, Smith J, Sun L, Wang Q, Kong L, Liu C, Hao H and Lu Z 2024 arXiv:2405.04967
[34] Han X Q, Gao Z F, Li W K, Guo P J and Lu Z Y 2025 arXiv:2512.21227
[35] Hinuma Y, Pizzi G, Kumagai Y, Oba F and Tanaka I 2017 *Comput. Mater. Sci.* **128** 140
[36] Togo A, Shinohara K and Tanaka I 2024 *Sci. Technol. Adv. Mater.: Methods* **4** 2384822
[37] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[38] Chen C, Ye W, Zuo Y, Zheng C and Ong S P 2019 *Chem. Mater.* **31** 3564
[39] Choudhary K and DeCost B 2021 *npj Comput. Mater.* **7** 185
[40] Dunn A, Wang Q, Ganose A, Dopp D and Jain A 2020 *npj Comput. Mater.* **6** 138
[41] Riebesell J, Goodall R E A, Benner P, Chiang Y, Deng B, Ceder G, Asta M, Lee A A, Jain A and Persson K A 2025 *Nat. Mach. Intell.* **7** 836
[42] Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, Riviere M, Tran K, Heras-Domingo J, Ho C, Hu W, Palizhati A, Sriram A, Wood B, Yoon J, Parikh D, Zitnick C L and Ulissi Z 2021 *ACS Catal.* **11** 6059
[43] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 *npj Comput. Mater.* **2** 16028
[44] Ward L, Dunn A, Faghaninia A, Zimmermann N E R, Bajaj S, Wang Q, Montoya J, Chen J, Bystrom K, Dylla M, Chard K, Asta M, Persson K, Snyder G J and Foster I 2018 *Comput. Mater. Sci.* **152** 60
[45] Schütt K T, Sauceda H E, Kindermans P J, Tkatchenko A and Müller K R 2018 *J. Chem. Phys.* **148** 241722
[46] Chen C and Ong S P 2022 *Nat. Comput. Sci.* **2** 718
[47] Deng B, Zhong P, Jun K, Riebesell J, Han K, Bartel C J and Ceder G 2023 *Nat. Mach. Intell.* **5** 1031
[48] Batzner S, Musaelian A, Sun L, Geiger M, Mailoa J P, Kornbluth M, Molinari N, Smidt T E and Kozinsky B 2022 *Nat. Commun.* **13** 2453
[49] Batatia I, Kovács D P, Simm G N C, Ortner C and Csányi G 2022 arXiv:2206.07697
[50] Batatia I, Benner P, Chiang Y, Elena A M, Kovács D P, Riebesell J, Advincula X R, Asta M, Avaylon M, Baldwin W J, Berger F, *et al.* 2025 *J. Chem. Phys.* **163** 172501
[51] Loew A, Sun D, Wang H C, Botti S and Marques M A L 2025 *npj Comput. Mater.* **11** 178