Crystoria: A Large-Scale Crystal Dataset Linking Thermodynamic and Dynamical Stability for Materials AI

Ke-Xin Pang[1], Ze-Feng Gao[2,3,*], Yan Gao[1,†], and Zhong-Yi Lu[2,3,‡]

[1]State Key Laboratory of Metastable Materials Science and Technology and Hebei Key Laboratory of Microstructural Material Physics, School of Science, Yanshan University, Qinhuangdao 066004, China
[2]School of Physics and Key Laboratory of Quantum State Construction and Manipulation (Ministry of Education), Renmin University of China, Beijing 100872, China
[3]AI-Driven Quantum Materials Research Center, School of Physics, Renmin University of China

[*]Correspondence: zfgao@ruc.edu.cn
[†]Correspondence: yangao9419@ysu.edu.cn
[‡]Correspondence: zlu@ruc.edu.cn

Article type: Data Express Letter

arXiv:2610.38195v1 [cond-mat.mtrl-sci] 28 Oct 2026

Abstract

Large-scale crystal databases have transformed materials discovery by making millions of structures and thermodynamic properties accessible to materials informatics. Yet dynamical stability remains sparsely characterized at a comparable scale because its evaluation is substantially more computationally demanding than conventional structure and energy calculations. This mismatch creates a critical gap between the rapidly expanding crystal-structure space and the physical information available for materials research and machine learning. Here we introduce Crystoria, a large-scale crystal dataset designed to bring dynamical stability to the scale of modern materials data. We construct Crystoria by systematically curating low-energy and metastable inorganic crystals, standardizing their structural identities, and combining machine-learning-accelerated structural relaxation with phonon-based dynamical screening. The resulting dataset contains 2,649,640 crystal structures, spanning 89 elements, 219 of the 230 space groups, and primitive cells containing 1–48 atoms. Among them, 1,793,353 structures (67.68%) are classified as dynamically stable, revealing substantial variation in dynamical stability across chemical and crystallographic space. By bringing dynamical stability to the scale of millions of crystal structures, Crystoria provides a physically grounded foundation for materials artificial intelligence (AI) and enables systematic exploration of the dynamical crystal landscape.

Keywords: crystal structures; thermodynamic stability; dynamical stability; phonons; machine-learning interatomic potentials; materials informatics; scientific datasets

Specifications Table

Table 1. Data specifications for Crystoria.

Item	Specification
Subject	Materials physics
Specific subject area	Crystal thermodynamics and lattice dynamics
Type of data	Atomic structures, hull energies, phonon frequencies and stability classes
Data format	Processed computational data from crystal curation, model relaxation and phonon calculations; Parquet tables with structured atomic coordinates
Data source location	Alexandria, Materials Project, JARVIS, NIMS HeuslerDB and OQMD; some Alexandria and Materials Project data obtained through MatterGen
Data accessibility	Zenodo (version 1.0.0); 10.5281/zenodo.23572554
Related research	None

1. Background & Summary

The rapid growth of high-throughput computation and data-driven methods has fundamentally changed the way crystalline materials are discovered and studied. Large computational resources, including the Materials Project, AFLOW, OQMD, JARVIS, NOMAD, and Materials Cloud, have established a data-centric paradigm in which crystal structures and calculated materials properties can be systematically explored across broad chemical spaces[1–6]. The development of interoperable computational infrastructures and open materials-analysis tools has further enabled the generation, organization, provenance tracking, and reuse of large-scale computational materials data[7–10]. Large-scale thermodynamic data have in turn enabled systematic analyses of inorganic stability and metastability, establishing energy above the convex hull as a central descriptor for computational materials discovery[11–13]. More recently, machine-learning approaches

have shifted materials research from fitting properties of individual compounds toward learning transferable representations and exploring increasingly large structural spaces[14–16]. Large-scale data-driven discovery has now reached a qualitatively different regime: GNoME reported more than two million structures predicted to be stable with respect to previous computational data, while MatterGen demonstrated generative design of diverse inorganic crystals using a large curated corpus of stable structures[17,18]. The resulting expansion of accessible crystal space has made the quality, diversity, and physical content of materials data increasingly important for both computational discovery and materials artificial intelligence (AI)[19–21]. Building large-scale materials datasets with physically meaningful property information is therefore becoming a central requirement for the next generation of data-driven materials research.

Against this rapid expansion of crystal-structure data, efforts to characterize dynamical stability have also progressed, but on a substantially smaller scale. First-principles lattice-dynamical calculations provide a direct computational route to determining whether a crystal is locally stable against collective atomic displacements[22,23]. Early high-throughput work demonstrated the feasibility of systematically evaluating phonon properties for inorganic materials, including complete density-functional perturbation-theory phonon calculations for 1,521 compounds[24]. Dedicated phonon repositories subsequently expanded the accessible collection to approximately 10^4 harmonic phonon calculations, providing a standardized source of dynamical information for inorganic crystals[25]. In parallel, high-throughput studies have used phonon calculations to screen dynamically stable materials within targeted chemical families, demonstrating the value of incorporating lattice dynamics into large-scale materials searches[26,27]. The emergence of machine-learning interatomic potentials (MLIPs) has opened a further route toward scaling dynamical calculations beyond what is practical with direct first-principles evaluation[28–34]. Recent work has begun to systematically benchmark universal MLIPs on phonon properties using approximately 10,000 first-principles phonon calculations, showing that several modern models can reproduce harmonic phonon properties with high accuracy[35]. This progression from direct first-principles calculations to machine-learning-accelerated phonon evaluation has recently enabled dynamical-stability datasets and benchmarks at the 10^5 scale, including PhononBench-MP40 with 46,899 completed phonon records and PhononBench with 133,838 AI-generated crystals evaluated for dynamical stability[36,37]. Nevertheless, these resources remain substantially smaller than modern general-purpose crystal databases, which already contain millions of structures[38,39]. Thus, although the scale of dynamical-stability characterization has increased by orders of magnitude, a pronounced gap remains between the size of available crystal-structure spaces and the number of structures for which systematic dynamical-stability data are available.

Recent advances in machine-learning interatomic potentials have made this scale gap increasingly tractable. Modern universal MLIPs are trained on large and chemically diverse first-principles datasets and use rotationally equivariant representations to improve the description of atomic interactions across broad chemical spaces[30–34,40]. Importantly, their suitability for dynamical calculations is no longer purely prospective. The phonon benchmarks also reveal model-dependent failures that are not necessarily apparent from energy and force tests alone. These differences make direct vibrational comparisons an essential part of assessing a model for large-scale dynamical-stability screening[35,41]. At the level of large-scale materials discovery, the current generation of force fields has also reached high task-level performance. In the Matbench Discovery benchmark, which evaluates machine-learning models in a prospective high-throughput materials-discovery setting, EquFlashV2 achieves a thermodynamic-stability-classification accuracy of approximately 97.8% and an F1 score of 0.929, with a combined performance score of 0.907[42,43]. These results establish that contemporary MLIPs can provide both the atomistic fidelity and computational efficiency needed for large-scale structural relaxation and stability screening. The remaining challenge is therefore not simply the availability of accurate atomistic models, but how their capabilities can be systematically deployed to generate consistent and trustworthy dynamical data across a rapidly expanding crystal-structure space. This requires coordinated treatment of structural relaxation, consistent recognition of equivalent crystal structures, the source and calculation conditions of thermodynamic data, phonon-based classification, near-zero imaginary modes, unsuccessful calculations, and independent higher-fidelity validation. The central question is consequently: how can reliable dynamical-stability data be generated at the scale of modern crystal databases while maintaining physical consistency, reproducibility, and traceability?

Here we address this challenge by introducing Crystoria, a large-scale crystal dataset designed to bring dynamical-stability characterization to the scale of modern materials data. We systematically curate low-energy and metastable inorganic crystal structures from multiple thermodynamic sources, standardize their structural representations and identities, and integrate machine-learning-accelerated structural relaxation with a uniform phonon-based dynamical-stability screening workflow. The phonon force model is assessed against external density-functional-theory (DFT) phonon calculations to test the reliability of its dynamical classifications. Crystoria contains 2,649,640 crystal structures, spanning 89 elements, 219 of the 230 space groups, and primitive cells containing 1–48 atoms. Among them, 1,793,353 structures (67.68%) are classified as dynamically stable, providing dynamical-stability data across broad chemical and crystallographic spaces. By extending systematic dynamical-stability characterization from specialized datasets in the 10^5 regime to the million-structure scale, Crystoria establishes a large-scale, physically curated resource for incorporating dynamical stability into computational materials discovery and materials artificial intelligence.

2. Methods

Extending dynamical-stability information to millions of crystals requires each phonon result to refer to a well-defined atomic arrangement. We bring together primitive cells from Alexandria, Materials Project, JARVIS, NIMS HeuslerDB, and OQMD, retaining their original structure and calculation sources (Fig. 1a). Some Alexandria and Materials Project structures and reference energies are obtained through the data collections distributed with MatterGen[18]. The same periodic crystal can appear under different lattice bases, origins, or database entries. The three TiCoSb representations illustrate how such descriptions are reduced to a common primitive-cell representation before repeated structures are removed. This step makes the subsequent energy and phonon calculations refer to distinct crystal arrangements rather than alternative descriptions of the same lattice. Spatial inversion is treated separately because it can reverse handedness; explicitly generated inversion partners are retained. Complete construction procedures are given in Supplementary Methods S1[44].

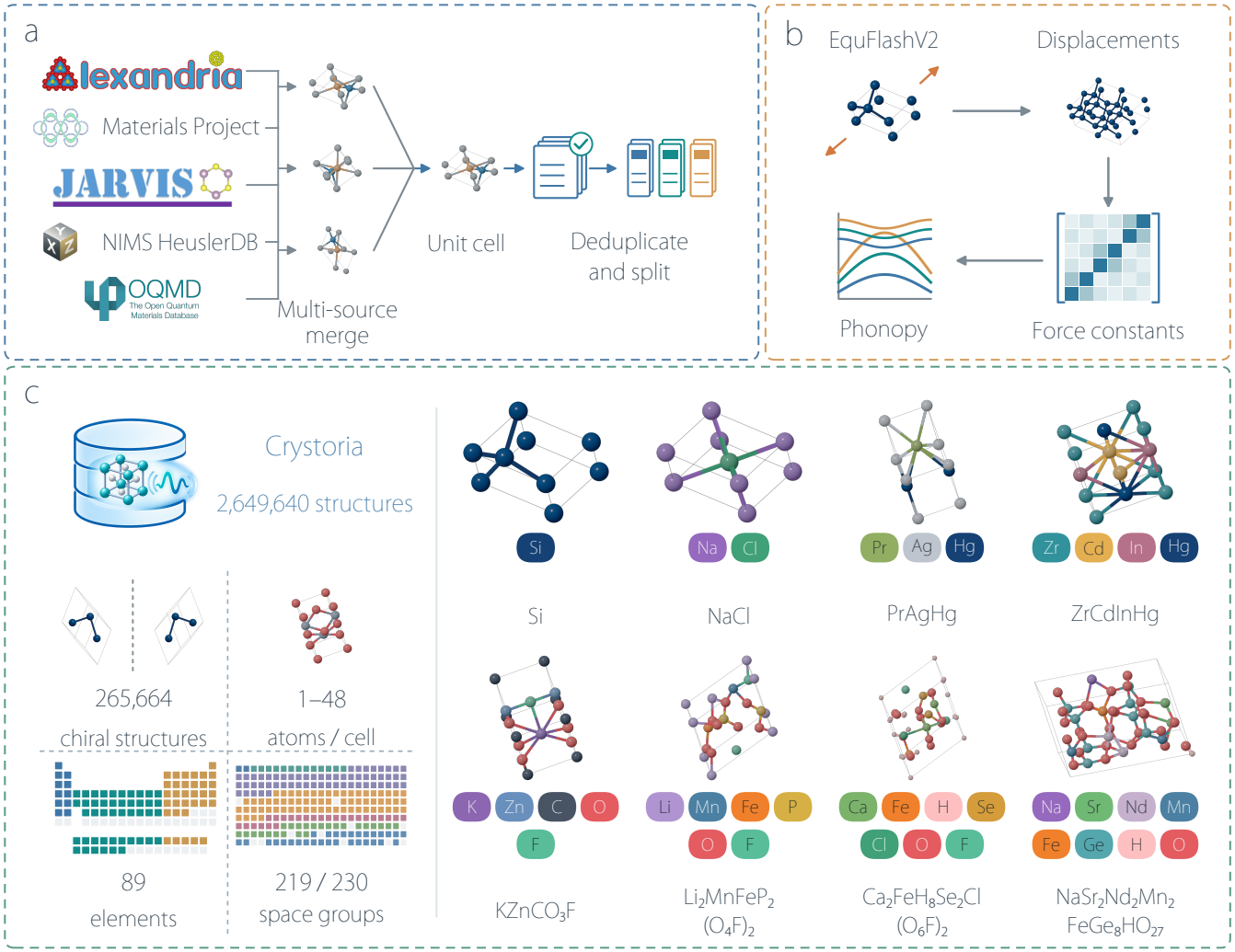


Figure 1. From crystal structures to dynamical-stability data. (a) Five database sources are brought into a common structural description. Equivalent TiCoSb cells represent one crystal; explicit inversion partners are retained and linked to their calculation sources. The source-energy selection is $E_{\mathrm{hull}} \leq 100$ meV atom^{-1}. (b) EquFlashV2 lattice relaxation, with a 0.03 eV Å^{-1} force criterion, followed by 0.01 Å displacements in a $2 \times 2 \times 2$ Si supercell and phonon evaluation on a $12 \times 12 \times 12$ mesh. (c) The 2,649,640 structures include 265,664 identified as chiral in their source cells. Te illustrates inversion-related geometry, and SnO_2 illustrates primitive-cell size. Eight relaxed structures satisfying the stable-class margin span one to eight elements. Filled element keys identify atomic species; space-group tiles distinguish crystal systems. Crystal structures are rendered from their atomic coordinates with MatterViz.

Starting structures are ordered, three-dimensional crystals with 1–48 atoms per primitive cell and source hull energies no greater than 100 meV atom^{-1}. Hull energy measures competition with other phases at the same overall composition, restricting the collection to low-energy and metastable candidates. Where several admissible source values exist, the largest controls selection before relaxation. Primitive-cell reduction and symmetry analysis use pymatgen and spglib[8,45],

with species-preserving structural comparisons.

The selected structures then undergo a common sequence of relaxation and lattice-dynamical evaluation, illustrated by the Si cell and displaced supercell in Fig. 1b. EquFlashV2 relaxes both atomic positions and the cell until the maximum residual force reaches the 0.03 eV $Å^{-1}$ criterion, within a limit of 200 steps[43]. Atomic displacements of 0.01 Å are applied in $2 \times 2 \times 2$ supercells, and the induced forces determine how a displacement of one atom changes the forces on its neighbours. These couplings form the force-constant matrix. Phonopy uses the symmetrized force constants to obtain collective vibrational modes on a non-Γ-centred $12 \times 12 \times 12$ mesh[22]. The calculation therefore proceeds from a relaxed atomic arrangement to its restoring response, rather than inferring dynamical stability from energy alone. An imaginary harmonic mode identifies a collective displacement along which the calculated energy decreases; imaginary frequencies are plotted as negative values in this work.

The minimum sampled frequency is retained as a continuous quantity. Structures are classified as stable for $\omega_{\min} \geq -0.10$ THz, borderline for $-0.30 \leq \omega_{\min} < -0.10$ THz, and unstable for $\omega_{\min} < -0.30$ THz. These margins separate nearly soft modes from pronounced harmonic instabilities. Only structures with converged relaxations and finite phonon results enter the 2,649,640-structure dataset.

Each relaxed parent's EquFlashV2 energy is placed on a fixed reference hull compatible with the Materials Project 2020 (MP2020) corrections[46]. This model-derived hull estimate is distinct from the source values and describes the same atomic arrangement as the phonon frequencies.

3. Data Record

This construction yields 2,649,640 structures, comprising 2,516,445 retained source-derived structures and 133,195 explicitly generated inversion partners, with 31,997,598 atoms in their primitive cells. Among the full collection, 265,664 structures are identified as chiral in their source cells. The paired Te structures in Fig. 1c illustrate why opposite handed geometries are retained: spatial inversion preserves crystal energies and phonon frequencies, but can reverse handedness-dependent properties. Generated partners retain inverted coordinates and inherit their parents' scalar energies and phonon frequencies. Supplementary Figure S1a-c resolves these populations across enantiomorphic pairs and other Sohncke space groups.

The same structural summary spans 89 elements, 219 space groups, and primitive cells containing 1-48 atoms. Its eight relaxed examples progress from Si and NaCl through PrAgHg and ZrCdInHg to oxygen-containing compounds with five to eight elements. All eight satisfy the adopted stable-class criterion.

Parquet tables store each atomic arrangement as lattice vectors, species, and fractional coordinates in JavaScript Object Notation (JSON). Each entry links source and relaxed-structure hull energies to the minimum phonon frequency and stability class. Units, calculation conditions, source attribution, and inversion partners accompany these quantities. Supplementary Methods S3 defines the quantities, units, and populations underlying the reported proportions.

Data repository: Zenodo (version 1.0.0). Permanent identifier: 10.5281/zenodo.23572554. Reviewer access: public Zenodo record 23572554.

4. Data Overview

4.1 Diversity of the dataset

Crystoria brings lattice-dynamical information to a chemically broad collection spanning 89 elements. Substantial populations extend across the main-group, transition-metal, and lanthanide regions of the periodic table (Fig. 2a). Oxygen is the most frequent element, occurring in 271,353 structures, while La and Tb each occur in more than 190,000. The chemical coverage therefore includes both light-element compounds and large populations containing heavier metals, placing different atomic environments within the same dynamical-stability study.

At the compound level, this breadth is concentrated in multicomponent crystals. The composition-order distribution peaks at three elements and remains high at four; together, ternary and quaternary compounds contribute 2,400,530 structures (90.60%; Fig. 2b). The population then falls steeply, but 82,193 structures contain five to eight elements. The principal chemical scope of Crystoria is therefore not the elemental or binary limit, but multicomponent lattices in which several atomic species contribute to the collective vibrational response. The higher-order tail extends this coverage to more compositionally complex compounds.

The variety of compositions is accompanied by broad crystallographic coverage: 219 of the 230 space groups and all seven crystal systems are represented (Fig. 2c). The widely varying tile intensities distinguish dense symmetry families from sparsely populated groups. Space group 12 contains the largest population, with 378,575 structures (14.29%), and substantial populations also occupy tetragonal and trigonal groups. The database thus extends well beyond a collection of familiar cubic lattices. Because crystal symmetry constrains the equivalence of atomic sites and displacements, this

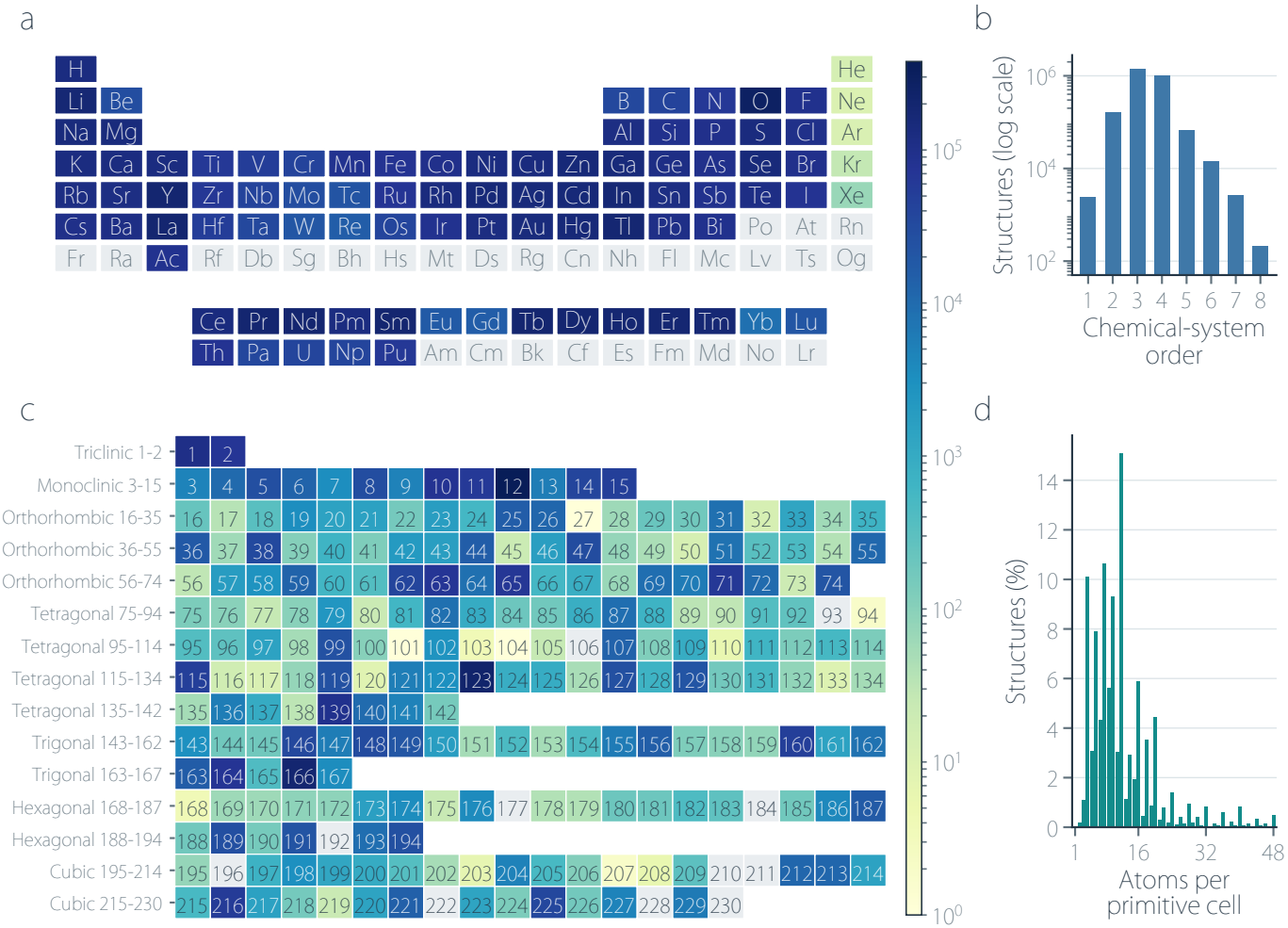


Figure 2. Chemical and structural diversity of Crystoria. All panels describe the same 2,649,640 structures. (a) Number of structures containing each element; a compound contributes to every element it contains. (b) Number of distinct elements per structure, from one to eight, with structure counts on a logarithmic scale. (c) Distribution over the 230 space groups, with 219 represented. These symmetry assignments refer to the source-standardized cells. (d) Number of atoms in the primitive cell over the 1–48-atom domain. Panels (a) and (c) share a logarithmic count scale; gray space-group cells indicate absence. Space-group numbers appear within the cells.

coverage makes lattice-dynamical information available across distinct forms of crystallographic order, not only across different compositions.

This combined chemical and crystallographic diversity is mostly accommodated within compact primitive cells. The cell-size distribution has pronounced peaks at small atom counts, with the largest at 12 atoms (399,001 structures), followed by a less populated tail extending to 48 atoms (Fig. 2d). Overall, 2,418,571 structures (91.28%) contain at most 20 atoms per primitive cell; this fraction reaches 93.00% among ternary and quaternary compounds. Multicomponent chemistry is therefore not confined to large cells. Taken together, the four distributions identify Crystoria as a chemically and crystallographically diverse collection whose main population comprises compact multicomponent crystals, with additional coverage of larger atomic arrangements.

4.2 Stability of the dataset

The source-to-relaxed hull-energy map contains a dense ridge near the equality line, indicating broad agreement between the two energy descriptions, together with a spread on both sides (Fig. 3a). In total, 23.51% of structures differ by more than 20 meV atom^{-1}. These shifts combine changes in atomic geometry with differences in the energy model, corrections, and reference phases. Supplementary Figure S2 resolves median absolute differences by element, space group, and cell size for independent source-derived structures. Re-evaluating the energy of the relaxed geometry therefore matters even when the starting structures have already passed a low-energy selection.

The relaxed-structure distribution nevertheless remains strongly concentrated near the reference hull. Its cumulative

distribution rises rapidly at low energies, with a median of 21.84 meV atom^{-1} and a 95th percentile of 87.93 meV atom^{-1}. In the full collection, 2,591,836 structures (97.82%) remain within 100 meV atom^{-1} of the reference hull. Thus, within the common model description, the expanded structural coverage retains a predominantly low-energy character.

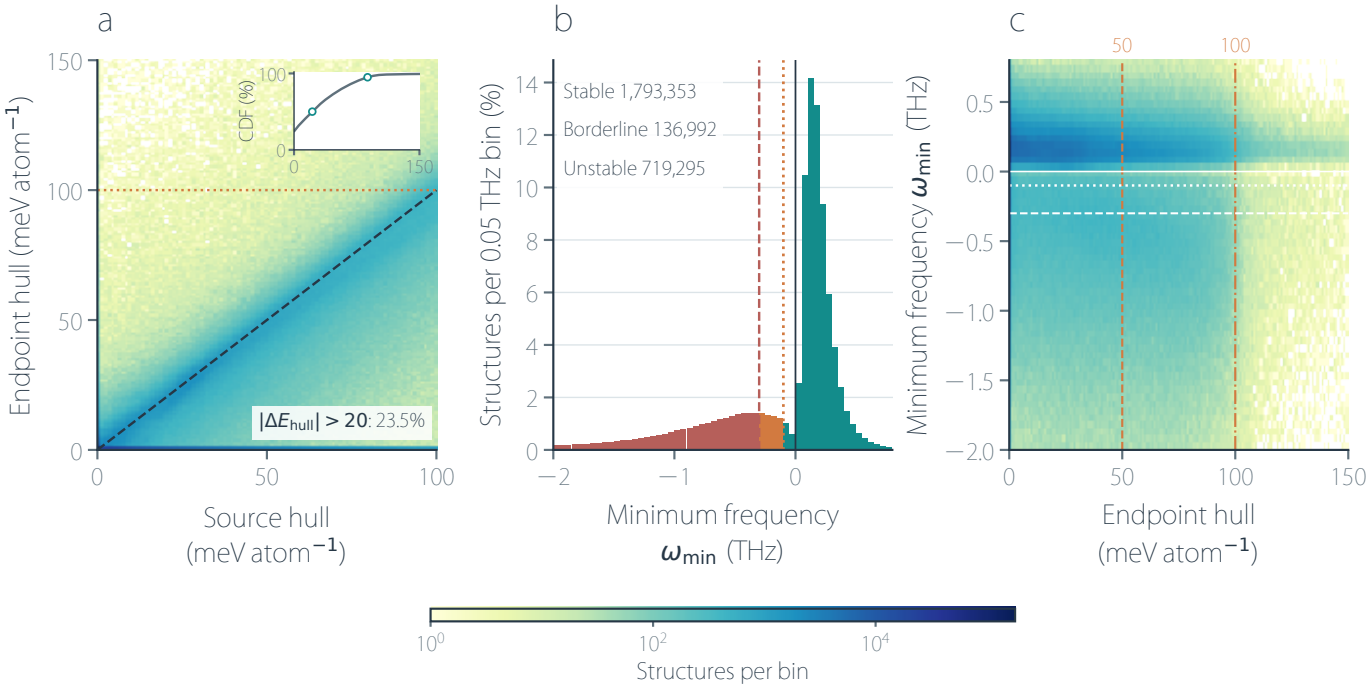


Figure 3. Thermodynamic and dynamical stability landscape of Crystoria. All panels use 2,649,640 structures with converged relaxations and finite phonon results. (a) Source and relaxed-structure hull energies; 18,209 values above 150 meV atom^{-1} are outside the map. The diagonal marks equality; the fraction differing by more than 20 meV atom^{-1} is given within the panel. The inset gives the cumulative distribution of relaxed-structure hull energies, with median and 95th-percentile values of 21.84 and 87.93 meV atom^{-1}. (b) Minimum phonon frequency in 0.05 THz bins, colored by stability class; 129,067 values lie outside the display. Imaginary modes are represented by negative frequencies. (c) Joint energy–frequency distribution containing 2,503,993 structures in the displayed range. Vertical references mark 50 and 100 meV atom^{-1}; horizontal references mark −0.30, −0.10, and 0 THz. Panels (a) and (c) share a logarithmic count scale. The 133,195 generated inversion partners have inherited scalar values. Hull energies are MP2020-aligned EquFlashV2 estimates for the relaxed structures.

The minimum-frequency distribution separates this low-energy population into a sharp peak on the positive-frequency side of zero and a broad tail extending into imaginary frequencies (Fig. 3b). Under the stated margins, 1,793,353 structures (67.68%) are stable, 136,992 (5.17%) are borderline, and 719,295 (27.15%) are unstable. Crystoria therefore provides a million-scale population satisfying a common harmonic-stability criterion, together with nearly soft and unstable lattices that reveal where that criterion is not met. The continuous frequencies distinguish proximity to the stability margin from pronounced imaginary modes.

The unstable population also clarifies why structural relaxation cannot replace a phonon calculation. Although 2,624,127 structures (99.04%) preserve their source space-group type under relaxation, 27.07% of this symmetry-preserving population is dynamically unstable. The full group-to-group and crystal-system comparisons use common symmetry tolerances (Supplementary Figure S4). The broad imaginary-frequency tail therefore does not arise only from crystals that change symmetry. Small residual forces constrain the first derivatives of energy, whereas phonons probe its curvature under collective displacements. A relaxed lattice can retain its symmetry while remaining susceptible to an energy-lowering distortion.

The overall stable fraction also conceals substantial differences between chemical and crystallographic populations. Within Crystoria, 38.51% of oxygen-containing structures satisfy the stable-class criterion, compared with 71.01% of oxygen-free structures. Stable-class fractions are 31.92% in space group 1 and 73.05% in space group 12. The tendency to develop imaginary modes is therefore uneven across the crystal landscape, while stable and unstable structures coexist even within a fixed space-group type. A common phonon calculation resolves differences that composition or symmetry alone cannot determine for an individual crystal.

Thermodynamic and dynamical stability are associated but not interchangeable. A dense positive-frequency band extends over a broad range of hull energies, while imaginary modes remain populated close to the hull (Fig. 3c). Stable-class structures have a lower median hull energy than unstable structures, 16.76 versus 37.65 meV atom^{-1}, consistent

with an overall association between favourable energetics and lattice stability. Yet 187,983 of the 980,560 structures within 10 meV atom^{-1} of the reference hull remain dynamically unstable (19.17%). Even a stringent energetic selection therefore retains many crystals whose atomic arrangement is unstable against a collective distortion.

The converse population is also substantial: 339,576 stable-class structures lie more than 50 meV atom^{-1} above the reference hull. These lattices satisfy the adopted harmonic-stability criterion while remaining energetically unfavourable relative to competing phase combinations. They are metastable candidates with a different physical character from near-hull crystals with imaginary modes. These contrasting populations expose the distinction between a driving force toward competing phases and an unstable displacement of a particular lattice.

Requiring both a relaxed-structure hull energy no greater than 50 meV atom^{-1} and the stable-class phonon margin selects 1,453,777 structures (54.87% of Crystoria). This joint population identifies candidates that combine energetic proximity to competing phases with the adopted harmonic-stability criterion. The value of the database is therefore not only its large stable subset, but its ability to resolve energetic competition and lattice response separately, and then select their intersection across a broad crystal-structure space.

5. Technical Validation

We assess the model-derived phonon classifications against 9,958 nonmagnetic semiconductors with DFT phonons calculated using the Perdew–Burke–Ernzerhof (PBE) approximation[35]. Both calculations use the same fixed DFT structures, supercells, and reciprocal-space points, isolating the predicted vibrational response from differences in model relaxation (Supplementary Methods S2). EquFlashV2 gives 95.11% accuracy and 91.22% balanced accuracy, the mean recall of the two reference classes. The 95% confidence interval for balanced accuracy is 90.36–92.06%, estimated by resampling groups of identical crystal geometries. Supplementary Figure S3 gives the reference populations, class-resolved errors, model comparisons, and uncertainty estimates for both the full and balanced reference sets.

Spatial inversion leaves crystal energies and phonon frequencies unchanged, providing a symmetry-based consistency test. The phonon classes agree in 208 of 209 independently calculated inversion pairs, with a maximum minimum-frequency difference of 0.05444 THz. Supplementary Figure S1d compares the independently calculated energies and phonon frequencies of these pairs.

6. Usage Notes

Crystoria can support the training and fine-tuning of crystal-generation models using structures that satisfy the stated harmonic-stability criterion. Conditional generation could use the relaxed-structure hull estimate and minimum phonon frequency as distinct targets, linking phase competition and lattice response in the design objective.

Stable, borderline, and unstable examples can also support predictors that guide generation or rank candidates. Such predictors could use continuous frequencies to distinguish nearly soft lattices from pronounced harmonic instabilities.

Temperature-dependent studies involve further calculations of vibrational free energies and, where needed, anharmonic effects[22]. Magnetic configurations have not been systematically enumerated and must be specified for magnetic-state comparisons.

Data Availability

Crystoria comprises the 2,649,640 structures and associated energetic and dynamical quantities described in Data Record. Repository: Zenodo (version 1.0.0). Persistent identifier: 10.5281/zenodo.23572554. Newly generated Crystoria data are released under CC BY 4.0; third-party data retain their source-specific terms and attribution.

Code Availability

Structure-curation, phonon-calculation and analysis programs: GitHub release v1.0.0. Persistent archival identifier: 10.5281/zenodo.23572554. Original Crystoria method code is released under the MIT License. Third-party code and the separately identified GGNN compatibility patches retain their original licenses, including CC BY-NC-SA 4.0 for GGNN.

Supplementary Information

The accompanying Supplementary Information[44] contains the construction methods, reliability tests, physical-quantity definitions, and Supplementary Figures S1-S4.

References

1. Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).

2. Curtarolo, S. *et al.* AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
3. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
4. Choudhary, K. *et al.* The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).
5. Draxl, C. & Scheffler, M. The NOMAD laboratory: From data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).
6. Talirz, L. *et al.* Materials cloud, a platform for open computational science. *Sci. Data* **7**, 299 (2020).
7. Pizzi, G., Cepellotti, N., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: Automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).
8. Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
9. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: Status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
10. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: Recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
11. Sun, W. *et al.* The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2**, e1600225 (2016).
12. Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **4**, eaaq0148 (2018).
13. Bartel, C. J. *et al.* A critical examination of compound stability predictions from machine-learned formation energies. *npj Comput. Mater.* **6**, 97 (2020).
14. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
15. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
16. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83 (2019).
17. Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
18. Zeni, C. *et al.* A generative model for inorganic materials design. *Nature* **639**, 624–632 (2025).
19. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
20. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
21. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 15004 (2016).
22. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
23. Gonze, X. & Lee, C. Dynamical matrices, born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory. *Phys. Rev. B* **55**, 10355–10368 (1997).
24. Petretto, G. *et al.* High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 180065 (2018).
25. Togo, A. MDR phonon calculation database. (2025) doi:10.48505/nims.4197.
26. Lee, H., Hegde, V. I., Wolverton, C. & Xia, Y. Accelerating high-throughput phonon calculations via machine learning universal potentials. *Mater. Today Phys.* **53**, 101688 (2025).
27. Xiao, E. & Tadano, T. High-throughput computational screening of heusler compounds with phonon considerations for enhanced material discovery. *Acta Mater.* **297**, 121312 (2025).
28. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
29. Bartok, A. P., Payne, M. C., Kondor, R. & Csanyi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
30. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
31. Deng, B. *et al.* CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).

32. Batatia, I., Kovacs, D. P., Simm, G. N. C., Ortner, C. & Csanyi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. in *Advances in neural information processing systems* vol. 35 11423–11436 (2022).
33. Liao, Y.-L., Wood, B. M., Das, A. & Smidt, T. E. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. in *International conference on learning representations* (2024).
34. Yang, H. *et al.* MatterSim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967* https://doi.org/10.48550/arXiv.2405.04967 (2024) doi:10.48550/arXiv.2405.04967.
35. Loew, A., Sun, D., Wang, H.-C., Botti, S. & Marques, M. A. L. Universal machine learning interatomic potentials are ready for phonons. *npj Comput. Mater.* **11**, 178 (2025).
36. Li, W.-K., Gao, Z.-F. & Lu, Z.-Y. PhononBench-MP40: A spectrum-resolved benchmark dataset for phonon stability. *Chin. Phys. B* **35**, 086301 (2026).
37. Han, X.-Q., Guo, P.-J., Gao, Z.-F., Li, W.-K. & Lu, Z.-Y. PhononBench: A large-scale benchmark reveals the dynamical stability challenge in AI-generated crystals. *AI for Science* https://doi.org/10.1088/3050-287X/ae9ee4 (2026) doi:10.1088/3050-287X/ae9ee4.
38. Schmidt, J. *et al.* Improving machine-learning models in materials science through large datasets. *Mater. Today Phys.* **48**, 101560 (2024).
39. Cavignac, T. *et al.* AI-driven expansion and application of the alexandria database. *J. Phys. Mater.* **9**, 025014 (2026).
40. Wood, B. M. *et al.* UMA: A family of universal models for atoms. in *Advances in neural information processing systems* (2025).
41. Deng, B. *et al.* Systematic softening in universal machine learning interatomic potentials. *npj Comput. Mater.* **11**, 9 (2025).
42. Riebesell, J. *et al.* A framework to evaluate machine learning crystal stability predictions. *Nat. Mach. Intell.* **7**, 836–847 (2025).
43. Cho, H. EquFlashV2 matbench submissions. (2026) doi:10.6084/m9.figshare.32630910.
44. Pang, K.-X., Gao, Z.-F., Gao, Y. & Lu, Z.-Y. Supplementary information for Crystoria: A large-scale crystal dataset linking thermodynamic and dynamical stability for materials AI. (2026).
45. Togo, A., Shinohara, K. & Tanaka, I. Spglib: A software library for crystal symmetry search. *Sci. Technol. Adv. Mater. Methods* **4**, 2384822 (2024).
46. Wang, A. *et al.* A framework for quantifying uncertainty in DFT energy corrections. *Sci. Rep.* **11**, 15496 (2021).

Author Contributions

Ke-Xin Pang performed the main calculations and drafted the manuscript. Ke-Xin Pang and Yan Gao conceived the study. Ze-Feng Gao, Yan Gao, and Zhong-Yi Lu validated the results, supervised the research, and acquired funding. All authors discussed the results and revised the manuscript.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 12304202, 62476278, and 12434009), the Beijing Natural Science Foundation (Grant Nos. Z260002 and Z250005), the National Key R&D Program of China (Grant No. 2024YFA1408601), the Hebei Natural Science Foundation (Grant No. A2023203007), and the Science Research Project of Hebei Education Department (Grant No. BJK2024085).

Disclosure of AI Use

The authors used AI for language polishing of the manuscript.

Supplementary Information

Crystoria: A Large-Scale Crystal Dataset Linking Thermodynamic and Dynamical Stability for Materials AI

Ke-Xin Pang[1], Ze-Feng Gao[2,3,*], Yan Gao[1,†], and Zhong-Yi Lu[2,3,‡]

Supplementary Methods S1. Crystal Structures and Lattice Dynamics

S1.1 Crystal collection and physical domain

Crystoria contains 2,649,640 ordered, three-dimensional crystal structures, with 31,997,598 atoms in their primitive cells. The selected cells contain 1–48 atoms, have supported atomic species with $Z \leq 100$, and have minimum interatomic separations of at least 0.50 Å. Source attribution and the conditions of the original thermodynamic calculations accompany each structure.

The contributing collections are Alexandria, Materials Project, JARVIS, NIMS HeuslerDB, and OQMD. Some Alexandria and Materials Project data are obtained through the collections distributed with MatterGen, including its DFT relaxation and energy processing[1]. A crystal appearing in several collections remains associated with its original sources.

Energy above the convex hull measures competition with other phases at the same overall composition. The source-energy selection requires every retained admissible estimate to be no greater than 100 meV atom^{-1}. The minimum and maximum admissible source estimates are retained separately, and the maximum controls the initial selection. This cutoff is applied to source estimates before relaxation. The source-to-relaxed comparison in Fig. 3 uses the source hull energy associated with each structure's original calculation.

S1.2 Equivalent cells and independent comparisons

Different lattice bases, origins, or atomic orderings can describe the same periodic crystal. Primitive-cell reduction and symmetry analysis use pymatgen and spglib[2,3]. Initial symmetry processing uses a 0.01 Å positional tolerance and a 5° angular tolerance, with fractional coordinates rounded to six decimal places.

Only structures with the same composition are compared. Candidate pairs are selected using reduced lattice lengths, angles, volume, and periodic interatomic distances. The allowed differences are 0.001 for sorted logarithmic lattice lengths and sorted absolute angle cosines, and 0.003 for logarithmic volume. Distance percentiles at 0, 10, 25, 50, 75, 90, and 100% are normalized by $(V/N)^{1/3}$, where V is cell volume and N is atom count; their comparison window is 0.0025.

Equivalence requires a direct, species-preserving correspondence between periodic atomic positions. Relative lattice and normalized atomic-position tolerances are 10^{-4}, with a 0.01° angle tolerance. The comparison permits proper rotations but not reflection, lattice scaling, additional primitive reduction, or supercell matching. Representative selection does not depend on the phonon class or other property values.

S1.3 Relaxed structures and thermodynamic quantities

Both the cell and atomic positions are relaxed with the same EquFlashV2 interatomic potential, using a maximum-force criterion of 0.03 eV Å^{-1} and at most 200 steps. The model is the 45-million-parameter version compared with density-functional theory (DFT) in Figure S3.

For the relaxed structures, EquFlashV2 single-point energies are corrected consistently with the Materials Project 2020 (MP2020) scheme and compared with a fixed reference phase diagram[4]. Both signed and nonnegative distances from this reference hull are retained as outputs of the common model calculation. Source and relaxed-structure values remain attached to the atomic arrangements from which they were obtained. Their differences combine relaxation with changes in the interatomic potential, energy corrections, and reference phases. Figure S2 shows how these differences vary with composition, symmetry, and cell size.

S1.4 Phonon calculations and stability margins

The harmonic response is calculated with Phonopy using atomic displacements of 0.01 Å in $2 \times 2 \times 2$ supercells[5]. The mean force is removed from each displaced supercell, and the reconstructed force constants are symmetrized. Force changes are evaluated without separately subtracting the forces of the undisplaced supercell.

Frequencies are evaluated on a non-Γ-centred $12 \times 12 \times 12$ reciprocal-space mesh. We retain the uncorrected minimum over this mesh as a continuous quantity, with imaginary modes represented by negative values. A negative harmonic mode identifies an atomic displacement along which the calculated energy decreases near the relaxed structure.

The reported classes are stable for $\omega_{\min} \geq -0.10$ THz, borderline for $-0.30 \leq \omega_{\min} < -0.10$ THz, and unstable for $\omega_{\min} < -0.30$ THz. These margins separate pronounced harmonic instabilities from nearly soft modes.

All structures included in the dataset have finite phonon results and converged parent relaxations.

S1.5 Inversion partners and chirality

Spatial inversion generates the missing handed partner through $\mathbf{x} \mapsto -\mathbf{x}$ (mod 1), with the lattice and species order retained. Existing partners are connected rather than added again. The chirality assessment uses consistent source-cell symmetry assignments within the 65 Sohncke space-group types.

The dataset contains 133,195 explicitly generated inversion partners. They inherit the parent energy and phonon results as parity-even scalar quantities, which are unchanged by spatial inversion. Handedness-dependent quantities, such as optical-rotation sign or topological charge, are not inherited. Chirality counts refer to the source-cell symmetry, while phonon properties refer to the relaxed arrangements. Independently calculated inversion pairs provide a separate consistency test in Supplementary Methods S2.2 and Figure S1.

Supplementary Methods S2. Reliability and Physical Interpretation

S2.1 Crystal symmetry before and after relaxation

Space-group types are determined independently from the source and relaxed atomic coordinates using spglib, with a common positional tolerance of 0.01 Å and an angular tolerance of 5°. All 2,649,640 structures enter this comparison, including generated inversion partners.

The space-group type is unchanged for 2,624,127 structures (99.04%) and changes for 25,513 (0.96%). Agreement ranges from 95.82% for triclinic source structures to 99.95% for cubic source structures. Figure S4 gives the complete correspondence over space groups 1–230 and the population and agreement within each crystal system. Among structures retaining their space-group type, 710,222 (27.07%) are dynamically unstable. Symmetry retention and harmonic stability therefore describe different aspects of the relaxed lattice.

S2.2 Inversion symmetry

Crystal energies and harmonic phonon frequencies are unchanged by spatial inversion. Among 209 independently calculated inversion pairs, 208 have the same phonon class. The largest minimum-frequency difference is 0.05444 THz. Figure S1 shows the corresponding symmetry populations and independent pair comparisons.

S2.3 Comparison with DFT phonons

The external reference comprises 9,958 nonmagnetic semiconductors with DFT phonon calculations using the Perdew–Burke–Ernzerhof (PBE) approximation: 8,189 stable and 1,769 unstable[6]. EquFlashV2 and MatterSim forces are evaluated on identical 0.01 Å displacements of the fixed DFT structures. The original supercell matrices and commensurate reciprocal-space points are retained. Mean-force removal and force-constant reconstruction are common to both models, and no model relaxation is applied.

The reference uses a numerical tolerance of −35 K for acoustic modes at Γ. This reference convention differs from the frequency margins and fixed reciprocal-space mesh used for Crystoria.

EquFlashV2 gives 95.11% accuracy and 91.22% balanced accuracy on the full reference set. Balanced accuracy is the mean recall of the stable and unstable classes, avoiding dominance by the more numerous stable structures. Stable recall is 97.25%, unstable recall is 85.19%, and stable precision is 96.81%.

The 95% confidence interval for balanced accuracy is 90.36–92.06%, obtained by resampling groups of identical crystal geometries. Differences between the two potentials are assessed with a paired exact McNemar test. Figure S3 also includes a balanced set of 2,688 structures with agreeing PBE and PBEsol classifications, and reports the model differences and their uncertainty. Overlap with the training data of either potential remains undetermined.

Supplementary Methods S3. Physical Quantities and Use of the Data

S3.1 Structural and energetic information

Each structure is described by its lattice vectors, atomic species, and fractional coordinates, together with composition, primitive-cell atom count, and source symmetry. The machine-readable tables use Parquet, with atomic structures represented in JavaScript Object Notation (JSON).

Thermodynamic information distinguishes source hull energies from the estimate for the relaxed structure. Dynamical quantities comprise the sampled minimum frequency and its stability class. Energies are reported per atom and frequencies in THz, with source and calculation conditions accompanying the values.

The chemical, symmetry-resolved, and energy-conditioned proportions in the main text use the complete 2,649,640-structure population. Chemical and symmetry-resolved populations use the source compositions and space groups shown in Figure 2; energy windows use hull estimates for the relaxed structures. Generated inversion partners remain included in these population statistics.

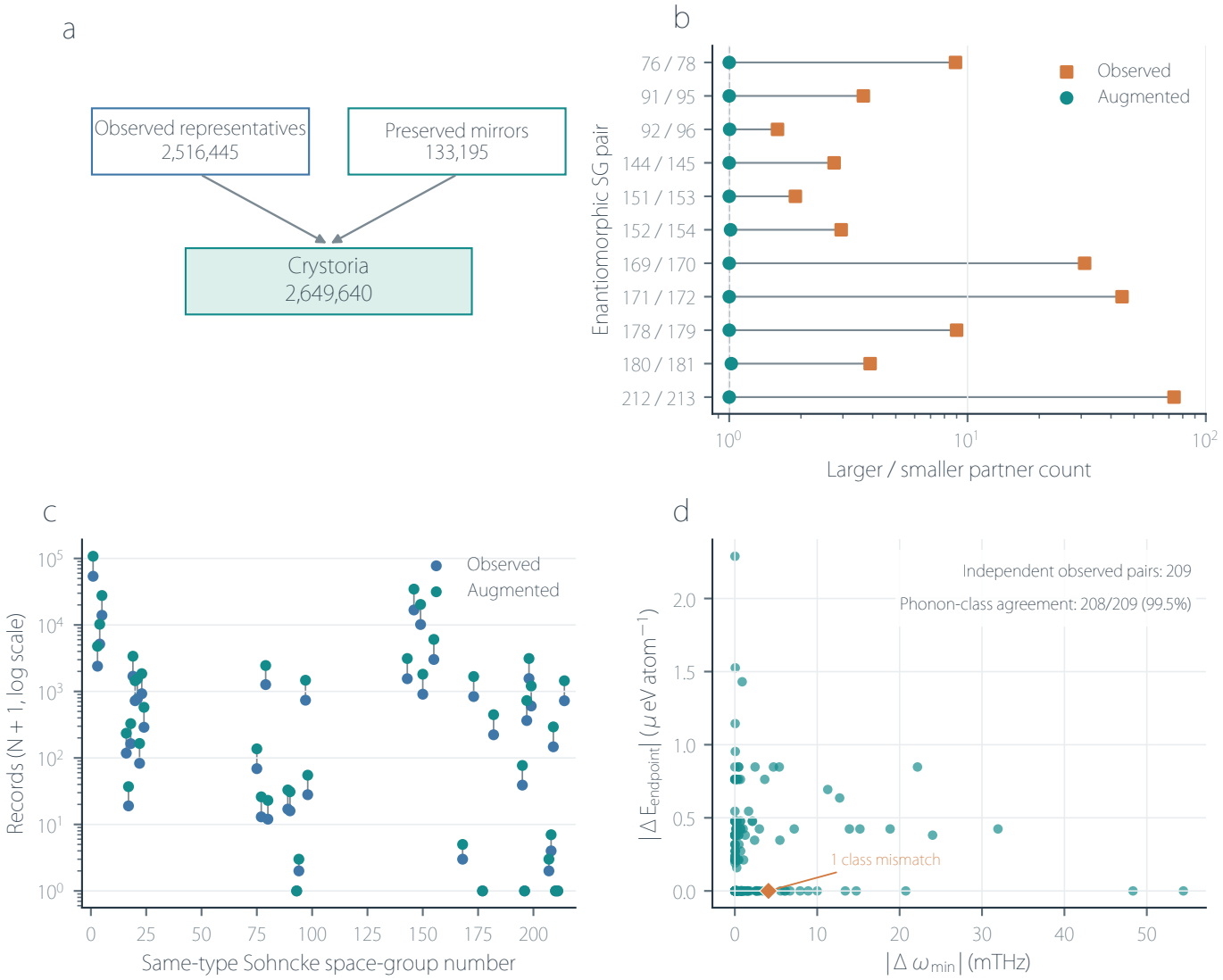


Figure S1. Distinct crystal structures and inversion partners. All panels describe the same 2,649,640 structures. (a) The population comprises 2,516,445 source-derived structures and 133,195 generated inversion partners. (b) Relative populations of the two partners in each of the 11 enantiomorphic space-group pairs, before and after adding missing inversion partners. Ratios use observed counts; a ratio is omitted if its denominator is zero. (c) Populations of the other 43 Sohncke space-group types, with 38 represented. Counts are plotted as $N+1$ on a logarithmic scale. Symmetry assignments in (b,c) refer to the source structures. (d) Differences in relaxed energy and minimum phonon frequency for 209 independently calculated inversion pairs. The class agrees in 208 pairs (99.5%). Generated partners inherit scalar values from their calculation parents. Structural comparison criteria are given in Supplementary Methods S1.2.

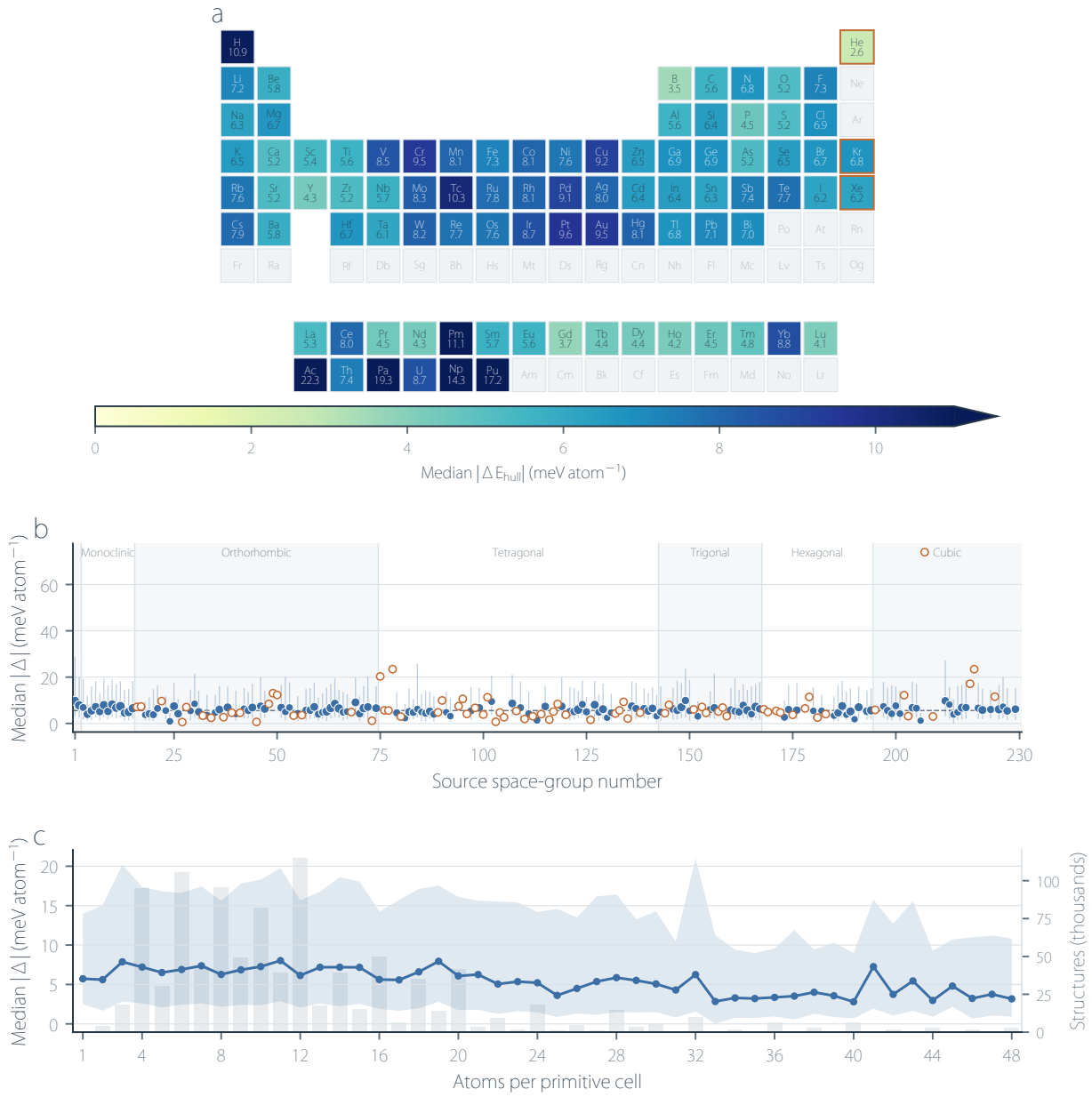


Figure S2. Comparison of source and relaxed-structure hull energies. Absolute differences compare the selected source energy above the convex hull with the nonnegative MP2020-aligned EquFlashV2 estimate for the relaxed structure. The comparison includes 993,629 independent source-derived structures, excluding generated inversion partners and cases using a substitute reference-hull estimate. (a) Median absolute difference for each of 87 represented elements. A compound contributes to each constituent element; orange borders indicate fewer than 100 structures. (b) Median difference for the 217 represented source space groups, with interquartile ranges. Filled points represent at least 100 structures and open points fewer than 100. (c) Median difference as a function of primitive-cell atom count, with interquartile ranges shaded and sample counts in gray. Differences reflect changes in geometry, potential, energy corrections, and reference phases.

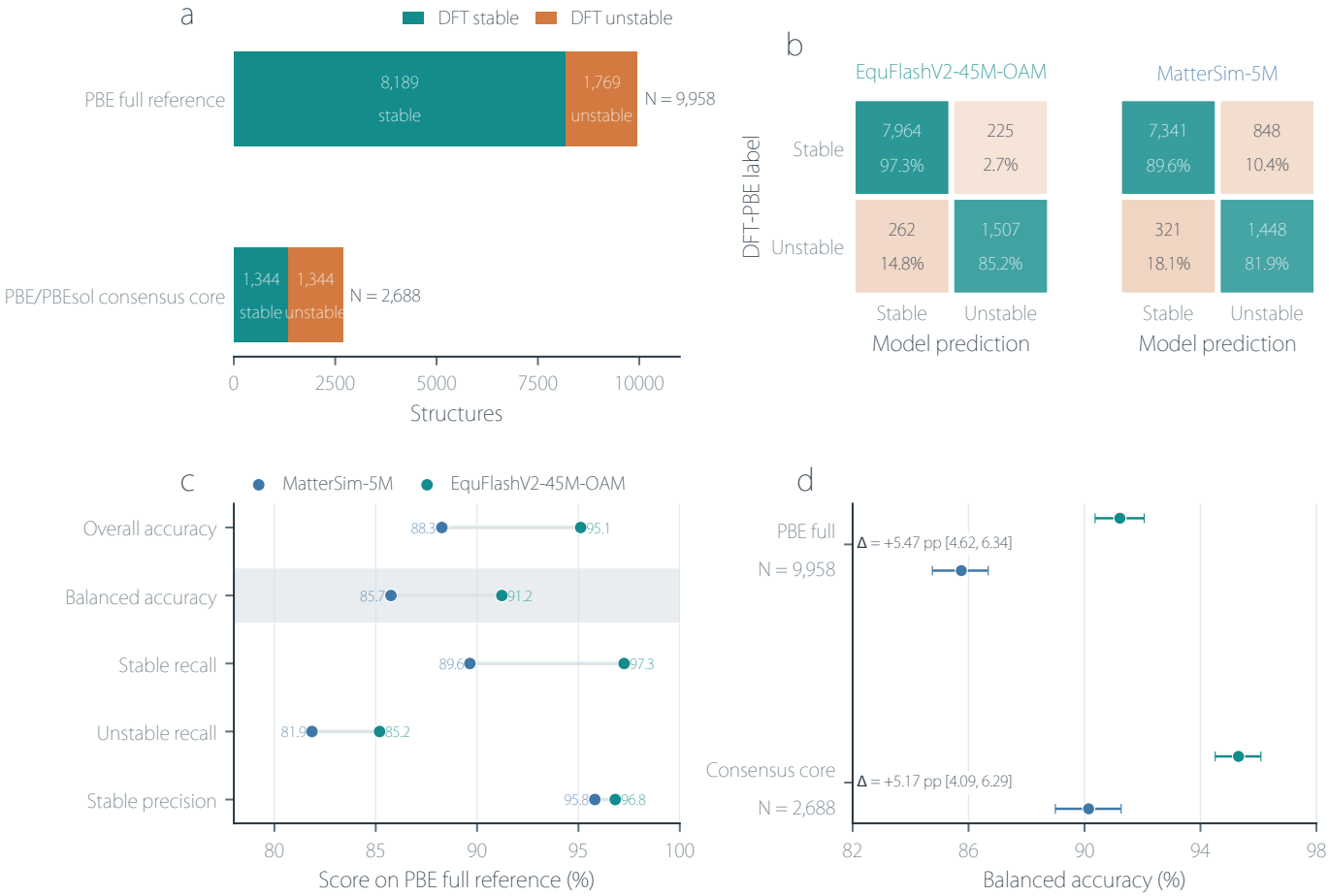


Figure S3. Phonon stability classifications compared with DFT. (a) Stable and unstable populations in the complete PBE reference set of 9,958 structures (8,189 stable; 1,769 unstable) and the balanced PBE/PBEsol-consensus set of 2,688 structures. (b) Confusion matrices normalized within each reference class, with structure counts shown. (c) Classification measures on the complete reference set. EquFlashV2 gives 95.11% accuracy and 91.22% balanced accuracy, compared with 88.26% and 85.75% for MatterSim. (d) Balanced accuracy and 95% confidence intervals obtained by resampling groups of identical crystal geometries. EquFlashV2 exceeds MatterSim by 5.47 percentage points on the full set (95% interval 4.62–6.34; paired exact McNemar $p = 2.93 \times 10^{-116}$) and 5.17 points on the balanced set (4.09–6.29; $p = 9.43 \times 10^{-21}$). Both potentials use identical 0.01 Å displacements of fixed DFT structures, the source supercell matrices and commensurate wavevectors, mean-force removal, and common force-constant reconstruction. Reference classes and predictions use the source −35 K acoustic-mode tolerance at Γ.

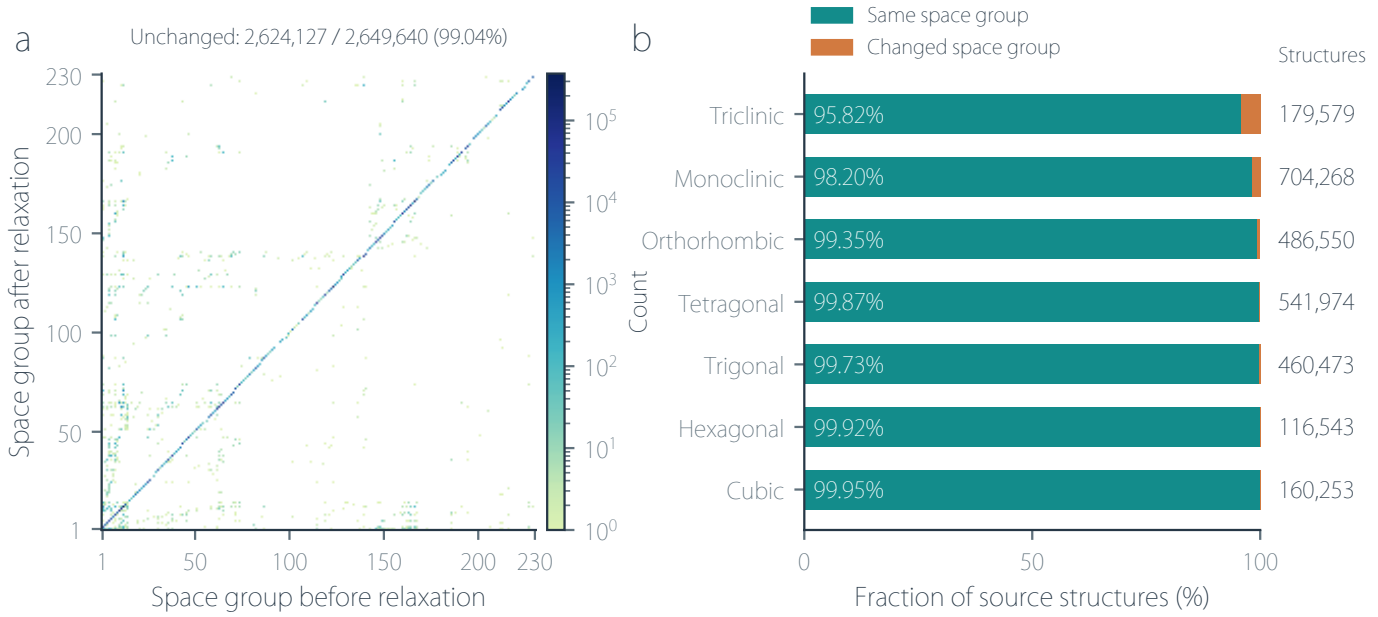


Figure S4. Space-group correspondence before and after relaxation. Both symmetries are determined independently for all 2,649,640 structures, including generated inversion partners, using a common 0.01 Å positional tolerance and 5° angular tolerance. (a) Full correspondence between source and relaxed space-group types 1–230. Color gives the structure count on a logarithmic scale; white cells contain no structures. The diagonal contains the 2,624,127 structures (99.04%) retaining their space-group type; the remaining 25,513 (0.96%) lie off the diagonal. (b) Fractions retaining or changing space-group type within each source crystal system. Percentages inside the bars indicate agreement, and numbers at right give each system's full population. Matching space-group types do not require identical lattice parameters or atomic positions.

Supplementary References

1. Zeni, C. *et al.* A generative model for inorganic materials design. *Nature* **639**, 624–632 (2025).
2. Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
3. Togo, A., Shinohara, K. & Tanaka, I. Spglib: A software library for crystal symmetry search. *Sci. Technol. Adv. Mater. Methods* **4**, 2384822 (2024).
4. Wang, A. *et al.* A framework for quantifying uncertainty in DFT energy corrections. *Sci. Rep.* **11**, 15496 (2021).
5. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
6. Loew, A., Sun, D., Wang, H.-C., Botti, S. & Marques, M. A. L. Universal machine learning interatomic potentials are ready for phonons. *npj Comput. Mater.* **11**, 178 (2025).