PhononBench:A Large-Scale Phonon-Based Benchmark for Dynamical Stability in Crystal Generation

Xiao-Qi Han[1], Peng-Jie Guo[1], Ze-Feng Gao[1,*], Zhong-Yi Lu[1,*]

[1]School of Physics, Renmin University of China, Beijing, China
*Corresponding authors

Abstract

In recent years, generative artificial intelligence has made significant advances in the design of crystalline materials, giving rise to a variety of approaches based on graph neural networks, diffusion models, and large language models. Existing evaluations commonly follow the stability–uniqueness–novelty (S.U.N.) framework, where "stability" is primarily assessed using thermodynamic criteria, which do not fully capture the dynamical stability essential for a material's practical existence. In fact, dynamical stability is a key determinant of whether a material can be synthesized and persists, with phonon spectrum calculations based on first-principles methods serving as the established standard for its evaluation. However, the high computational cost of such calculations has, until now, prevented large-scale and systematic assessment of dynamical stability in generated crystals. In this work, we introduce Phonon-Bench, the first large-scale benchmark for dynamical stability in AI-generated crystals. Leveraging the recently developed MatterSim interatomic potential, which achieves density-functional-theory (DFT)–level accuracy in phonon predictions across more than 10,000 materials, Phonon-Bench enables efficient large-scale phonon calculations and dynamical-stability analysis for 108,843 crystal structures generated by six leading crystal generation models. PhononBench reveals a widespread limitation of current generative models in ensuring dynamical stability: the average dynamical-stability rate across all generated structures is only 25.83%, with the top-performing model, MatterGen, reaching just 41.0%. Further case studies show that in property-targeted generation—illustrated here by band-gap conditioning with MatterGen—the dynamical-stability rate remains as low as 23.5% even at the optimal band-gap condition of 0.5 eV. In space-group-controlled generation, higher-symmetry crystals exhibit better stability (e.g., cubic systems achieve rates up to 49.2%), yet the average stability across all controlled generations is still only 34.4%. An important additional outcome of this study is the identification of 28,119 crystal structures that are phonon-stable across the entire Brillouin zone, providing a substantial pool of reliable candidates for future materials exploration. By establishing the first large-scale dynamical-stability benchmark, this work systematically highlights the current limitations of crystal generation models and offers essential evaluation criteria and guidance for their future development toward the design and discovery of physically viable materials. All model-generated crystal structures, phonon calculation results, and the high-throughput evaluation workflows developed in PhononBench will be openly released at https://github.com/xqh19970407/PhononBench.

Keywords: crystal generation models, benchmark, dynamical stability

Introduction

In recent years, AI-driven inverse materials design has advanced rapidly and comprehensively, achieving breakthroughs across multiple aspects of crystal materials design [1–4]. Diffusion-based

approaches [5, 6], such as DiffCSP [7], have demonstrated significant improvements in both speed and accuracy for crystal structure prediction compared with traditional DFT-based search workflows. CrystalFlow [8] further enhances efficiency by reformulating the underlying algorithm using flow matching, substantially accelerating inference. For functional materials design, MatterGen [1] exhibits strong performance across key metrics including Stability, Uniqueness, and Novelty, and leverages adapter-based fine-tuning to enable precise generation of various classes of functional materials. The active learning–based workflow for inverse functional materials design (InvDesFlow-AL) [9] has achieved notable success in the design of high-temperature superconductors, together with other generative approaches [10–12] for functional materials, further demonstrating the capability of AI to explore complex material systems. Meanwhile, crystal generation models incorporating space-group control—such as CrystalFormer [13, 14] and DiffCSP++ [15] significantly improve the symmetry and physical plausibility of generated structures. With the rapid development of large language models [16, 17](LLM), methods such as CrystaLLM [18] and FlowLLM [19] integrate natural language into crystal generation, enabling more intuitive and flexible conditioning for materials design.

However, it is essential to recognize the limitations of current progress. Despite the rapid development of crystal generative models, their assessment of material stability has largely focused on thermodynamic stability, typically evaluated using metrics such as E_{hull} (energy above the convex hull) [1, 9, 20, 21]. Yet the synthesizability and practical existence of materials depend not only on thermodynamic stability but, more critically, on dynamical stability—whether a structure resides in a local potential well and can withstand small perturbations without collapsing. Dynamical instability often manifests as imaginary phonon modes [3, 22, 23], which not only directly signal mechanical instability but have also long posed a persistent challenge for DFT practitioners [24]. The origins of imaginary modes are diverse, potentially arising from symmetry breaking, insufficient q-mesh resolution, pseudopotential choices, approximation errors, or intrinsic structural instability [25, 26]. More importantly, because rigorous dynamical stability validation (e.g., full phonon dispersion calculations) is computationally demanding [27], most existing generative models have never performed systematic dynamical stability tests on their generated structures. As a result, these models may produce a large number of structures that appear thermodynamically stable but are in fact dynamically unstable [28], undermining both the reliability and practical utility of their predictions.Therefore, building on the successes of current generative models, establishing systematic and efficient dynamical stability evaluation for generated structures has become a key challenge and an important frontier for guiding computationally generated materials toward experimental synthesis and enhancing the trustworthiness of model predictions.

With the rapid development of universal machine-learning interatomic potentials (uMLIPs) [21, 29–32], a growing number of models have achieved energy and force prediction accuracies that approach or even surpass those of DFT, thereby greatly enhancing computational efficiency in atomistic materials simulations [33–37]. Representative advances include MEGNet [38], which reaches a formation energy accuracy of 21 meV on the GNoME [2], and M3GNet [39], which incorporates atomic coordinates, lattice vectors, and three-body interactions and has become one of the most prominent uMLIPs. Building on M3GNet, MatterSim [29] is pretrained on 17 million first-principles data points, enabling zero-shot generalization across the first 89 elements of the periodic table over temperatures of 0–5000 K and pressures of 0–1000 GPa. Its accuracy in predicting energies, forces, and stresses surpasses previous models such as MACE [40] by roughly an order of magnitude, enabling reliable calculations of lattice dynamical, mechanical, and thermodynamic properties. More importantly, Miguel A. L. Marques et al. conducted a systematic assessment based on phonon calculations for over 10,000 materials (Figure 1) [41]. Their results show that MatterSim achieves phonon-spectrum prediction accuracy comparable to DFT, with average errors even smaller than

the inherent differences between the PBE and PBEsol functionals, and markedly outperforming all other models. Furthermore, in dynamical-stability classification, MatterSim attains a 95% true-positive rate, achieving a level of reliability nearly equivalent to a full DFT workflow while requiring only a tiny fraction of its computational cost. Taken together, these results clearly demonstrate that MatterSim is not merely "good enough," but in fact a genuinely reliable and efficient tool for high-throughput phonon calculations and dynamical-stability analysis.

In this work, we introduce PhononBench and systematically perform phonon calculations for 108,843 crystal structures generated by six commonly used generative models, providing the first large-scale assessment of the dynamical stability of AI-generated materials. PhononBench results reveal that current generative models remain markedly limited in ensuring phonon stability: the average dynamical stability rate across all models is only 25.83%, with the best-performing MatterGen model reaching merely 41.0%. Further case studies on functional-material generation, exemplified by band-gap–conditioned generation with MatterGen, show that although the highest stability is achieved at a target band gap of 0.5 eV, the corresponding dynamical stability rate remains low at 23.5%. In space-group–controlled generation, crystals with higher symmetry exhibit improved phonon stability—for instance, cubic systems reach stability rates of up to 49.2%—yet the overall average stability is still modest at 34.4%. Notably, through this comprehensive evaluation, we newly identify 28,119 crystal structures that are fully phonon-stable across the entire Brillouin zone. These structures constitute a substantial and reliable pool of candidate materials for subsequent materials design and discovery, highlighting both the promise of generative models in expanding the known materials space and their current limitations in guaranteeing dynamical stability.

Results

Systematic Evaluation of Dynamical Stability in Crystal Generation Models

In this work, we employed 6 crystal generative models—including the LLM-based CrystaLLM [18]; the graph-neural-network–enhanced diffusion models MatterGen [1], DiffCSP [7], and InvDesFlow-AL [9]; the flow-matching model CrystalFlow [8]; and the space-group-controlled CrystalFormer [13] (covering variants trained on different datasets)—to generate a total of 221,000 novel crystalline materials (Figure 1). After generation, we removed duplicates with respect to each model's training set and performed post-processing to correct malformed CIF files. Structural relaxations were then carried out for the remaining crystals, of which 108,843 converged successfully. Subsequently, we conducted full phonon-spectrum calculations on the relaxed structures with MatterSim [29] coupled to Phonopy [42, 43], which identified 28,119 crystals as dynamically stable, yielding a stability ratio of 25.83%. Our systematic evaluation demonstrates that achieving dynamical stability remains a significant challenge for all current crystal generative models.

To ensure fair comparability across different generative models, we adopt the ratio of dynamically stable structures as a unified evaluation metric. Specifically, this ratio is defined as the number of phonon-stable crystals (i.e., those without imaginary phonon modes) divided by the number of successfully relaxed structures. This metric effectively eliminates biases arising from differences in novelty, CIF compliance rates, and relaxation success rates among models. Furthermore, our experiments show that the stability ratio produced by a generative model converges once the sample size exceeds approximately 4,000, with the remaining uncertainty being sufficiently small to avoid affecting the performance ranking reported in this study (Detailed convergence analysis of the dynamical stability rate and a summary of crystal generation and stability statistics are provided in the Supplementary Information). Except for CrystaLLM, all models in the figure exceed this convergence threshold in terms of phonon-calculated samples.

As shown in Figure 1, the three models with the highest dynamical stability are MatterGen (41.0%), InvDesFlow-AL (38.4%), and CrystalFormer (34.4%). All of these models were pre-trained on large-scale, high-quality crystal databases such as Alex20. In contrast, models trained solely on smaller datasets, such as MP20—for example, CrystalFlow—exhibit significantly lower stability ratios (only 16.7%). Notably, the stability ratio of InvDesFlow-AL is approximately 130% higher than that of CrystalFlow. These results indicate that pre-training on large, high-quality datasets can substantially enhance the performance of generative models in predicting crystal dynamical stability. A noteworthy point is that CrystalFlow exhibits the fastest generation speed among all models. As shown in Table 1, a detailed comparison will be provided later.

In terms of model architecture, MatterGen employs a GemNet-based [44] diffusion generative framework, whereas InvDesFlow-AL is built on the DiffCSP-Ab-Initio-Generation approach using EGNN [45]. Both are diffusion-based models, with their performance ranking first and second, further highlighting the notable advantage of diffusion frameworks for crystal generation tasks. At the data representation level, MatterGen uses Cartesian coordinates (atomic positions), while the DiffCSP series employs fractional coordinates. This difference may affect model performance when scaling to supercells: Cartesian coordinates preserve the actual interatomic distances, whereas fractional coordinates change relative to the cell size, introducing a physical inconsistency that could contribute to performance differences. In addition, MatterGen utilizes D3PM [46], specifically designed for discrete data, which may also be an important factor in its superior performance. Moreover, CrystalFormer ranks third and achieves relatively good performance by constraining crystal generation in a manner that aligns better with physical intuition. However, this approach reduces novelty, as the strong constraints make it difficult to generate low-symmetry yet potentially stable materials. A detailed discussion is provided in Figure 2. Detailed statistics on crystal generation, structural relaxation, and dynamical stability for all evaluated generative models are summarized in Table 1 in the Supplementary Materials.

In this test, CrystaLLM achieved a dynamical stability ratio of only 3.0%, ranking last. It should be noted that although the model supports unconditional generation, the generative quality can degrade significantly in the absence of effective prompts due to the inherent next-token prediction mechanism of large language models. To ensure fair comparison, we retained its unconditional generation mode, producing a total of 16,000 crystals. After post-processing, only 2,074 structures were valid, of which 1,951 converged successfully, and ultimately only 58 were dynamically stable. Although the effective test sample size did not reach 4,000, the resulting uncertainty is insufficient to affect the final ranking. These results also indirectly indicate that, for the current task, LLM-based generative methods still lag significantly behind architectures such as graph neural networks.

Dynamical Stability Analysis of Space-Group-Constrained Crystal Generation

This section analyzes the dynamical stability of crystal structures generated by a space group constrained generative model. We employed CrystalFormer [13] (Alex20) to generate a total of 40,000 crystal structures, with their space-group distribution matched to that of the MP20 training set. After removing structures overlapping with MP20, 8,986 unique crystals were retained for further relaxation and phonon calculations, among which 8,642 successfully completed the phonon calculations. As shown in Figure 2(a), the generated crystals exhibit a distribution across the seven crystal systems that is highly consistent with the Materials Project dataset. Among the 8,642 structures with successful phonon calculations, 2,969 were identified as dynamically stable, corresponding to a stability ratio of approximately 34%. The elemental composition distribution (Figure 2(b)) shows that ternary and quaternary compounds account for more than half of the generated materials, in agreement with the statistical trends observed in the Materials Project.
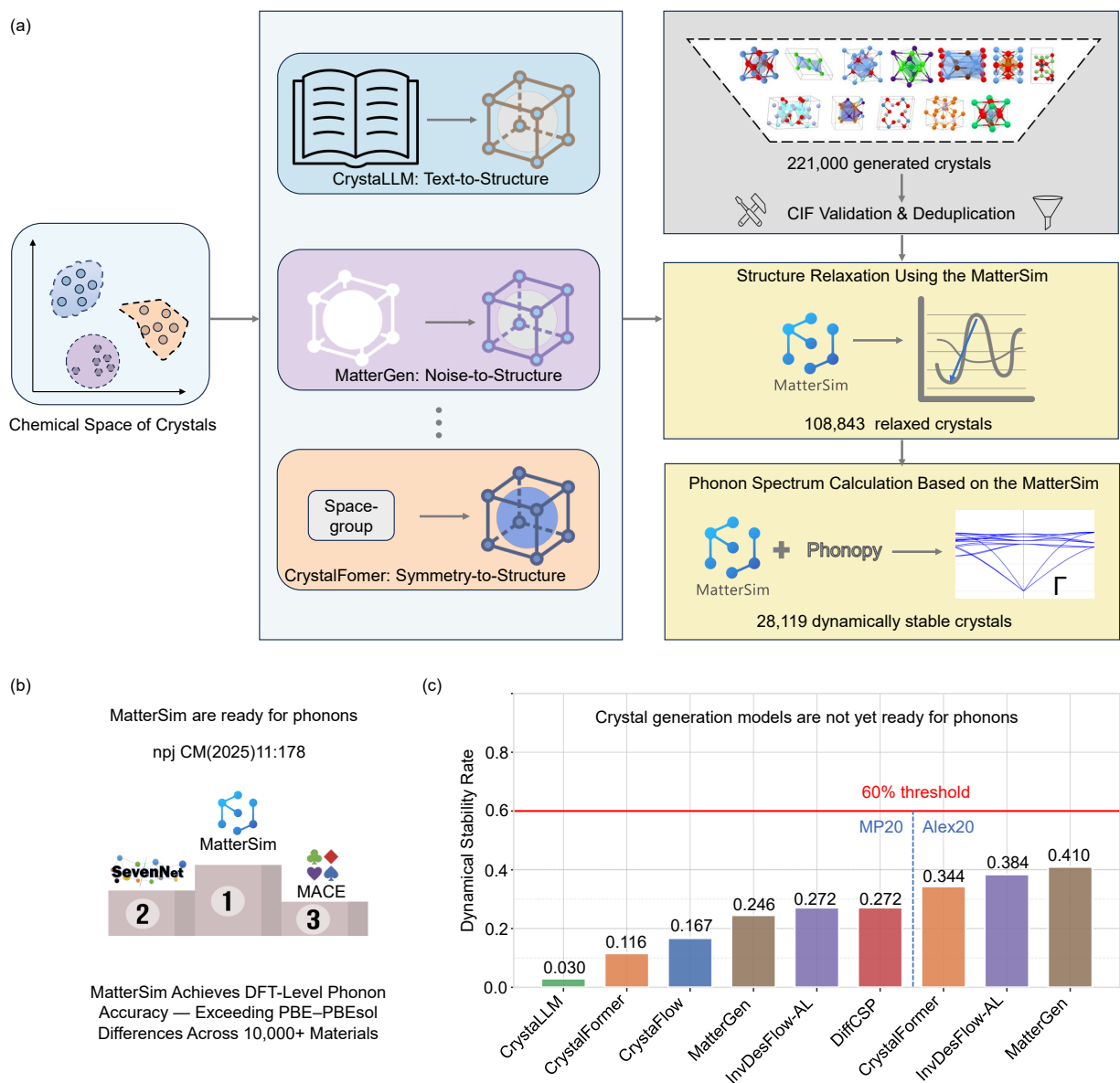
Figure 1: Systematic Evaluation of Dynamical Stability in Crystal Generation Models. (a) Workflow of this study. Eight generative models (CrystaLLM, MatterGen, DiffCSP, InvDesFlow-AL, CrystalFlow, CrystalFormer, etc.) were used to generate a total of 221,000 novel structures. After removing duplicates and post-processing for CIF validity, full phonon-spectrum calculations were performed using MatterSim combined with Phonopy on 108,843 successfully relaxed crystals. In total, 28,119 structures were found to be dynamically stable, corresponding to an overall stability ratio of 25.83%. (b) Based on the systematic phonon-spectrum evaluation of over 10,000 materials by Miguel A. L. Marques et al., we employed MatterSim-v1—which attains DFT-level accuracy with an average error smaller than the difference between PBE and PBEsol functionals—as the unified potential for all subsequent phonon calculations, ensuring consistency in evaluation standards. (c) Key results and model performance comparison. All generated models face challenges in achieving dynamical stability. The top three models are MatterGen (41.0%), InvDesFlow-AL (38.4%), and CrystalFormer (34.4%), all of which benefited from pretraining on large, stable datasets such as Alex20. In contrast, models trained on smaller datasets such as MP20—e.g., CrystalFlow (16.7%)—show significantly lower performance. The LLM-based CrystaLLM (3.0%) performs the worst, highlighting its current disadvantage relative to architectures such as graph neural networks for this task.
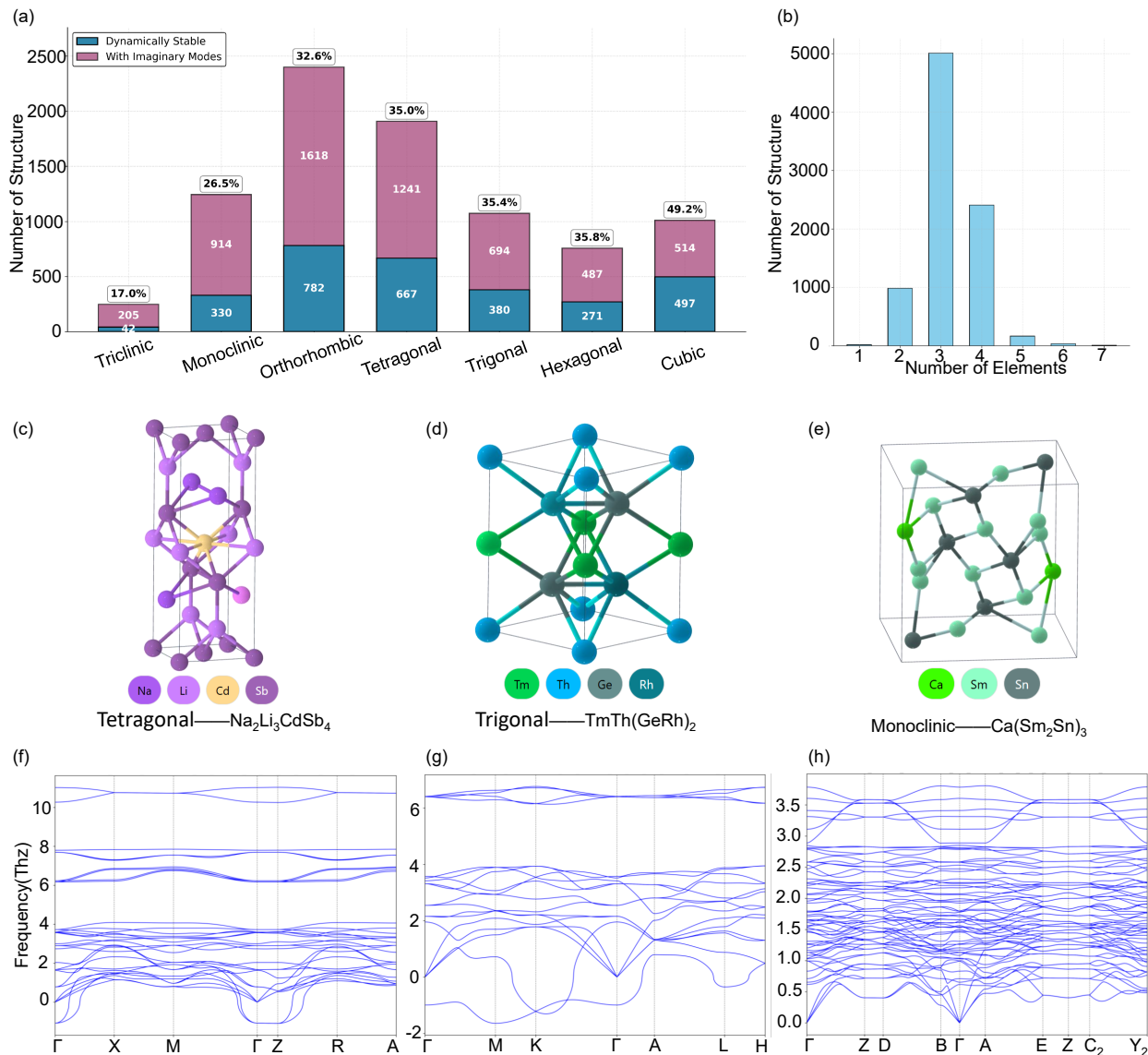
5

Figure 2: Dynamical Stability Analysis of Space-Group-Constrained Crystal Generation. (a) Distribution of the generated crystal structures across the seven crystal systems and their dynamical stability ratios (dark blue: dynamically stable; purple: containing imaginary modes). The cubic system exhibits the highest stability (49.2%), whereas the triclinic system shows the lowest (17%). (b) Distribution of the number of elemental components in the generated materials, where ternary and quaternary compounds account for more than half of the dataset, consistent with the trend observed in MP20. (c) Tetragonal structure: Na_2Li_3CdSb (d) Trigonal structure: $TmTh(GeRh)_2$ (e) Monoclinic structure: $Ca(Sm_2Sn)_3$ (f)–(h) Phonon spectra corresponding to (c)–(e). Na_2Li_3CdSb and $TmTh(GeRh)_2$ exhibit pronounced imaginary phonon modes, indicating dynamical instability, whereas $Ca(Sm_2Sn)_3$ shows no imaginary modes and is thus dynamically stable. These examples illustrate the ability of CrystalFormer to generate materials with complex chemistries, while also highlighting the remaining challenges in ensuring dynamical stability. They underscore the importance of incorporating explicit stability constraints or performing posterior stability screening within the generation pipeline.

The dynamical stability ratio varies considerably among different crystal systems: cubic structures show the highest stability ratio (49.2%), followed by hexagonal (35.8%), trigonal (35.4%), tetragonal (35.0%), orthorhombic (32.6%), and monoclinic systems (26.5%), while triclinic structures exhibit

the lowest stability ratio at only 17%. These results suggest a possible correlation between structural symmetry and dynamical stability: high-symmetry crystal systems such as cubic tend to exhibit higher stability ratios, whereas low-symmetry systems such as triclinic show markedly reduced stability. This trend may be related to the comparatively smoother potential energy landscapes associated with high-symmetry structures. Moreover, the elemental distribution of the generated materials closely matches that of the real dataset, indicating that the model achieves good coverage of chemical compositional diversity. Nevertheless, there remains room for improvement in the dynamical stability of the generated structures. It is also noteworthy that, relative to generative models without space-group constraints—such as InvDesFlow-AL [9] and DiffCSP [7]—the space-group–restricted model exhibits significantly reduced novelty, reflecting the inherent limitations imposed by symmetry constraints on the generative space.

Figures 2(c)–(e) present three representative crystal structures generated by CrystalFormer: (c) a tetragonal Na_2Li_3CdSb compound, (d) a trigonal $TmTh(GeRh)_2$ compound, and (e) a monoclinic $Ca(Sm_2Sn)_3$ compound. Their corresponding phonon spectra are shown in Figures 2(f)–(h). The phonon results indicate that both Na_2Li_3CdSb and $TmTh(GeRh)_2$ exhibit pronounced imaginary frequencies across multiple phonon branches, signaling strong dynamical instabilities and suggesting that these structures may reside near saddle points or shallow extrema on the potential energy landscape. In contrast, the phonon spectrum of $Ca(Sm_2Sn)_3$ shows no imaginary modes throughout the Brillouin zone, demonstrating robust dynamical stability. The extensive imaginary modes observed in the first two structures may originate from unfavorable bonding configurations, such as excessively short interatomic distances leading to strong repulsive interactions, or local coordination environments that cannot be dynamically sustained. These case studies clearly illustrate the dual nature of the current generative model: while it can construct chemically complex structures that relax to energetically reasonable configurations, many of these structures still face significant dynamical challenges. This highlights that energy-based criteria alone are insufficient for crystal generation tasks and underscores the necessity of incorporating explicit dynamical stability constraints or performing systematic post-generation screening.

Dynamical Stability Analysis of Property-Constrained Crystal Generation

This section presents functional materials generation using the band-gap–conditioned MatterGen model. To ensure statistical significance, we predetermined that each band-gap condition should yield at least ∼4,000 unique crystals for stability evaluation. In total, 56,000 crystals were generated: 16,000 samples for $E_g = 1.5$ eV, and 10,000 samples each for $E_g = 0.5$ eV, 2.5 eV, 3.5 eV, and 4.5 eV. After removing structures overlapping with the MP20 training set, 33,210 crystals were subjected to phonon calculations to assess their dynamical stability. As shown in Fig. 3(a), the dynamical stability exhibits a clear dependence on the target band-gap conditions. The $E_g = 4.5$ eV condition shows the lowest stability: only 617 out of 5,340 evaluated samples are dynamically stable (11.6%). In contrast, the $E_g = 0.5$ eV condition yields the highest stability, with 1,523 stable structures among 6,524 samples (23.5%). The stability ratios for the remaining conditions are as follows: 15.3% for $E_g = 1.5$ eV (1,448 stable out of 9,478), and 13.3% for both $E_g = 2.5$ eV and 3.5 eV (816 stable out of 6,133 and 763 stable out of 5,735, respectively). Overall, the dynamical stability in band-gap–conditioned functional materials generation remains relatively low. Even with the advanced fine-tuned MatterGen framework, the overall stability ratio reaches only 15.6%. This indicates that relying on conventional first-principles packages such as Quantum ESPRESSO (QE) or the Vienna *Ab initio* Simulation Package (VASP) for subsequent phonon calculations would incur substantial computational cost, posing challenges for practical large-scale deployment.

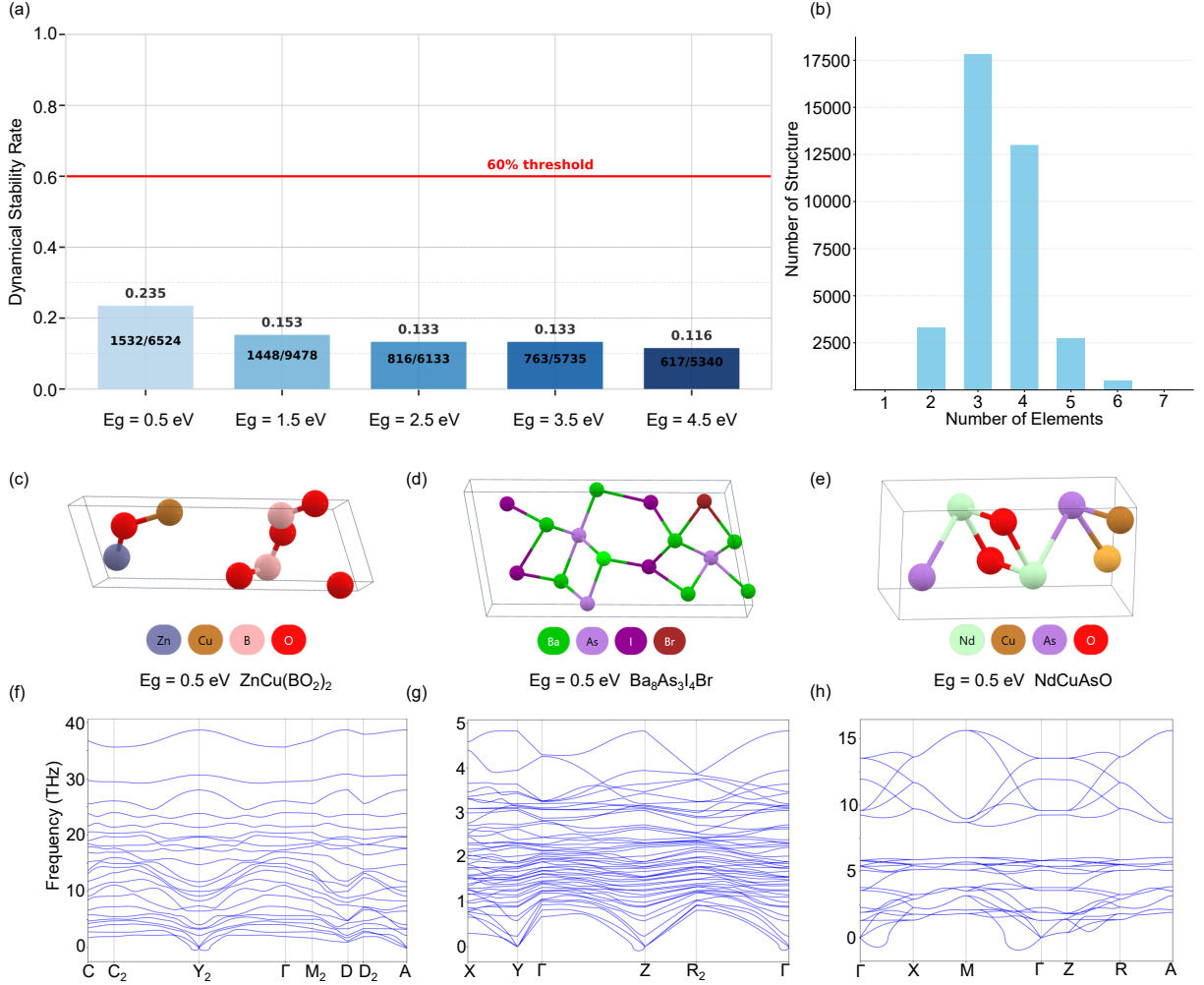From the elemental composition distribution (Fig. 3(b)), we observe that, under the band-

Figure 3: Dynamical Stability Analysis of Property-Constrained Crystal Generation. (a) Dynamical stability ratios of generated materials under different band-gap constraints. Among the 33,210 structures subjected to phonon analysis, the $E_g = 4.5$ eV condition yields the lowest stability (11.6%), whereas $E_g = 0.5$ eV yields the highest (23.5%). The stabilities for the other settings are 15.3% for $E_g = 1.5$ eV and 13.3% for both $E_g = 2.5$ eV and $E_g = 3.5$ eV. Overall, the stability rate under band-gap conditioning remains low (15.6%), implying that subsequent phonon validation with QE or VASP incurs considerable computational cost and poses challenges for large-scale applications. (b) Distribution of the number of elemental components in the generated materials. (c)–(e) Three representative crystal structures generated by MatterGen under the $E_g = 0.5$ eV constraint: $ZnCu(BO_2)_2$, $Ba_8As_3I_4Br$, and NdCuAsO. (f)–(h) Phonon spectra corresponding to panels (c)–(e). All three structures exhibit pronounced imaginary (negative) frequencies across multiple phonon branches, indicating strong dynamical instabilities in their current configurations.

gap constraint, ternary and quaternary compounds account for more than 50% of the generated materials. This trend closely aligns with the statistical patterns found in the Materials Project database, indicating that the model maintains a reasonable exploration of chemical space in controlled-generation tasks. In addition, Fig. 3(c)–(e) illustrates three representative crystal structures generated by MatterGen—$ZnCu(BO_2)_2$, $Ba_8As_3I_4Br$, and NdCuAsO—with a target band gap of $E_g = 0.5$ eV. Their corresponding phonon spectra are shown in Fig. 3(f)–(h). The phonon calculations reveal pronounced imaginary (negative) frequencies across multiple phonon branches for all three materials, indicating strong dynamical instabilities in their current structural configurations.
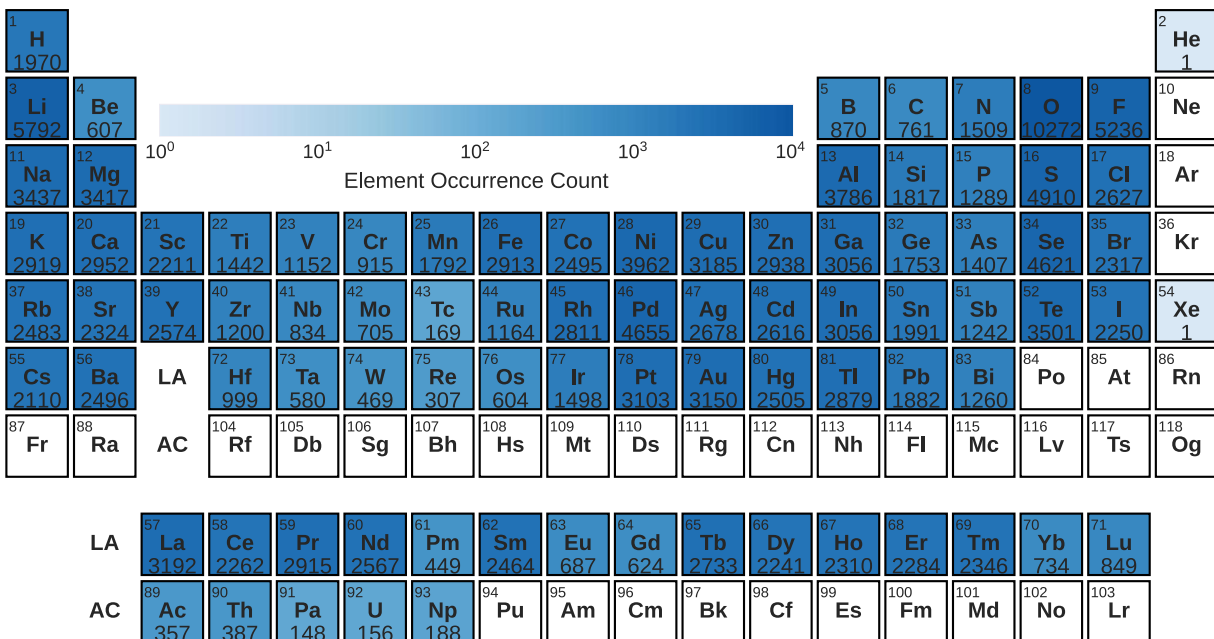
8

Figure 4: Elemental distribution heatmap of dynamically stable crystals. The heatmap illustrates the chemical element distribution of 28,119 newly discovered crystal structures with phonon (dynamical) stability generated by the crystal generation model. Each element is colored according to its occurrence frequency in the stable crystal set, with darker colors indicating higher frequencies. The analysis shows that oxygen (O) is the most prevalent element (10,272 occurrences), followed by lithium (Li, 5,792) and fluorine (F, 5,236), whereas noble gas elements appear only rarely. This distribution clearly indicates that current generative models tend to predict stable compounds containing chemically active elements, in good agreement with established chemical intuition.

To facilitate further analyses of structural reliability and potential properties, we will release all phonon calculations and optimized structures for the full set of 33,210 generated materials, providing a comprehensive benchmark for subsequent validation and methodological developments. The statistics of structural relaxation and dynamical stability stratified by different band-gap ranges are summarized in Table 2 in the Supplementary Materials.

Dynamically Stable Crystal Structures

In this study, we conducted a systematic evaluation of the dynamical stability of generated crystal materials and identified 28,119 dynamically stable crystal structures. This discovery provides a rich pool of candidate systems for materials science research and substantially expands the database of known stable materials. Although current crystal generation models are not explicitly optimized for phonon stability during training, they demonstrate a capability for novel materials design that significantly surpasses traditional manual design approaches. Through elemental statistical analysis (Fig. 4), we observe clear regularities in the elemental distribution of these stable crystals. Oxygen (O) appears most frequently, with 10,272 occurrences, followed by lithium (Li) with 5,792 occurrences, and fluorine (F) with 5,236 occurrences. In contrast, noble gas elements are rarely present in dynamically stable crystals, consistent with their chemical inertness. These trends indicate that the generative model effectively captures fundamental chemical principles governing real materials. All evaluated crystal structures and their corresponding phonon calculation details will be fully open-sourced, providing a transparent and verifiable data foundation for the community. The validated

9

dynamically stable crystals constitute a reliable candidate set for exploring novel functional materials, enabling density functional theory practitioners to focus on functional property investigations without concerns about basic dynamical stability. The open availability of this dataset is expected to accelerate functional materials discovery and advance computational materials science toward higher accuracy and efficiency. The elemental distribution of the 28,119 dynamically stable crystals is shown in Figure 2 of the Supplementary Materials.

Summary of Model Architectures and Inference Performance

Table 1: Comparison of inference throughput for different generative models. To account for differences in maximum batch size due to model architecture and memory usage, speeds are normalized to an equivalent batch size of 200, enabling a fair and consistent evaluation of computational efficiency. The last column indicates the normalized generation speed in crystals per minute.

Model Name	Generated Items	Total Time (h)	Speed (items/min)	Eq. Speed @200 ↑
CrystalFlow [8]	16,000	0.64	416.6	333.3
InvDesFlow-AL [9]	25,000	1.7	240.0	48.0
DiffCSP [7]	16,000	2.3	113.5	45.4
CrystaLLM [18]	30,000	9.4	52.9	42.3
MatterGen [1]	16,000	20.2	13.2	13.2
CrystalFormer [13]	20,000	44.3	6.0	12.0

This section presents a systematic comparison of generation speed across different models. As shown in Table 2, we report each model's throughput for a single generation task (note: "Generated Items" refers to the number of structures used in this speed test, not the final number used for phonon calculations). To ensure fair comparison across architectures and memory usage, the generation speeds are normalized to an equivalent batch size of 200. All generation tasks, except for CrystalFormer, were performed on a single NVIDIA RTX 4090 GPU. The normalized inference throughput clearly reveals substantial differences in generation efficiency among the models. CrystalFlow demonstrates the highest performance, with an equivalent generation speed of 333.3 crystals per minute, significantly outperforming all other models. InvDesFlow-AL and DiffCSP form the second tier, with normalized speeds of 48.0 and 45.4 crystals per minute, respectively. CrystaLLM achieves 42.3 crystals per minute. MatterGen and CrystalFormer exhibit the lowest throughput, with equivalent speeds of only 13.2 and 12.0 crystals per minute; the slower speed of CrystalFormer is likely due to the use of its CPU version in our tests. In summary, while maintaining both novelty and stability of the generated materials, CrystalFlow demonstrates outstanding generation efficiency, exceeding the slowest model by more than an order of magnitude. This superior throughput provides a critical advantage for large-scale virtual screening and iterative optimization of crystal structures using generative models.

As shown in Table 2, the compared models encompass two mainstream architectures: GNN-based crystal generative models and Transformer-based sequence models, with parameter counts ranging widely from lightweight 4.8M to large-scale 53.7M, reflecting diverse model capacities and design philosophies. In terms of architectural distribution, GNNs dominate, consistent with their natural suitability for modeling the periodic graph structures of crystals. Among them, MatterGen, with 53.7M parameters, is the largest model in this comparison, reflecting its design focus on high expressive capacity. CrystalFlow, another representative GNN model, has a moderate size of 20.9M parameters, while InvDesFlow-AL and DiffCSP share a compact GNN architecture with 12.3M parameters each. Regarding Transformer-based architectures, CrystaLLM (Small) is a 26.1M-

Table 2: Overview of the model architectures and their parameter counts. The compared models include both GNN-based crystal generative models and Transformer-based sequence models, covering a wide range of model capacities from lightweight (4.8M parameters) to large-scale (53.7M parameters).

Model Name	Training Set	Architecture Type	Parameters ↓
CrystalFormer	MP20	Transformer	4.8M
CrystalFormer	Alex20	Transformer	4.8M
InvDesFlow-AL	Alex20	GNN	12.3M
DiffCSP	MP20	GNN	12.3M
CrystalFlow	MP20	GNN	20.9M
CrystaLLM	MP20	Transformer	26.1M
MatterGen	MP20	GNN	53.7M
MatterGen	Alex20	GNN	53.7M

parameter sequence model, exceeding the size of most GNN models. In contrast, CrystalFormer, also a Transformer, is the most lightweight model in this comparison with only 4.8M parameters.

Discussion

In practical applications of crystal generative models, different architectures exhibit notable differences in terms of novel material discovery and computational friendliness. For instance, the symmetry-constrained CrystalFormer shows a significant reduction in novelty: in one generation experiment, although 20,000 crystals were produced, only about 5,000 unique chemical formulas were obtained. This indicates that exploring new chemical spaces with such models requires generating a substantially larger number of samples to obtain sufficient novel materials. In contrast, the large language model-based CrystaLLM frequently encounters file parsing failures during generation, with failure rates reaching up to 90% in severe cases, likely related to prompt design, highlighting practical limitations of this approach. Conventional equivariant graph neural network models such as MatterGen, DiffCSP, and InvDesFlow-AL achieve better novelty, but the generated structures often exhibit significantly reduced symmetry, which can lead to increased computational difficulty or instability in subsequent DFT calculations. Overall, there is a trade-off between novel material discovery and computational tractability across different generative architectures, and model selection should be guided by the specific research objectives.

Methods

To reliably assess the dynamical stability of the crystal structures generated in this work, we computed their phonon spectra using the MatterSim-v1 [29] universal machine-learning interatomic potential. Although recent uMLIPs have demonstrated near-DFT accuracy in energies, forces, and structural relaxations, their performance on second-order response properties such as phonons has long lacked systematic validation. A recent large-scale benchmark [41] conducted by Miguel A. L. Marques and co-workers demonstrated that, among seven state-of-the-art uMLIPs, MatterSim-v1 delivers the highest accuracy across key phonon-related properties—including phonon frequencies, phonon DOS, free energy, and heat capacity—with errors even smaller than those arising from the choice of different DFT functionals (e.g., PBE versus PBEsol). Moreover, it achieves an accuracy of approximately 95% in classifying dynamical stability. Traditional DFT phonon calculations require extensive force-constant evaluations and are therefore prohibitively expensive for high-throughput screening at the scale of tens of thousands of materials. Given the validated accuracy and efficiency

of MatterSim-v1, we adopt this model to perform large-scale phonon and dynamical-stability evaluations of the generated structures, enabling reliable and efficient stability screening at unprecedented scale. Generating target functional materials is a crucial step in the inverse design of materials. However, prior to achieving desired attributes, generative models must ensure that synthesized materials inherently exhibit fundamental crystalline characteristics, including periodicity, symmetry, interatomic interactions, and chemically reasonable stoichiometry. To this end, we propose a pretrained crystal generation model. In the following, we will introduce the data representation, model architecture, and training method required for this model.

We developed a high-throughput phonon calculation workflow based on Phonopy and the MatterSim-v1 universal machine learning interatomic potential. Crystal structure files (CIF, POSCAR/CONTCAR, etc.) were first converted to PHONOPYATOMS objects using a custom batch script, and $2 \times 2 \times 2$ supercells were generated to produce compressed `phonopy.yaml.bz2` input files. Phonon calculations were then performed using MatterSim-v1 for both geometry relaxation and force constant evaluation. Specifically, initial structures were reconstructed from the reference Phonopy files and optimized with the FIRE algorithm while preserving crystal symmetry (force convergence criterion: 0.005 eV/Å). Displaced supercells (0.01 Å) were generated by Phonopy, and atomic forces were computed using MatterSim-v1, corrected for translational drift, and used to construct and symmetrize the force constant matrices. High-symmetry paths were automatically generated using SEEKPATH, and phonon band structures were obtained via Fourier interpolation. Dynamical stability was assessed by checking for imaginary modes (threshold $< -1 \times 10^{-3}$ THz). This high-throughput workflow enables the dynamical stability assessment of tens of thousands of structures, providing an efficient and reliable basis for evaluating phonon properties of generative-model-derived crystals.

Data availability

The crystal data are available from the Materials Project database via the web interface at https://materialsproject.org or the API at https://api.materialsproject.org.

Code availability

All data and code are publicly available at https://github.com/xqh19970407/PhononBench.

References

[1] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

[2] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[3] Xiao-Qi Han, Zhenfeng Ouyang, Peng-Jie Guo, Hao Sun, Ze-Feng Gao, and Zhong-Yi Lu. InvDesFlow: An AI-driven materials inverse design workflow to explore possible high-temperature superconductors. *Chin. Phys. Lett.*, 42(4):047301, 2025.

[4] Xiao-Qi Han, Xin-De Wang, Meng-Yuan Xu, Zhen Feng, Bo-Wen Yao, Peng-Jie Guo, Ze-Feng Gao, and Zhong-Yi Lu. AI-Driven Inverse Design of Materials: Past, Present, and Future. *Chinese Physics Letters*, 42(2):027403, 2025.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[6] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *ICLR*, 2021.

[7] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal Structure Prediction by Joint Equivariant Diffusion on Lattices and Fractional Coordinates. In *Workshop on "Machine Learning for Materials" ICLR 2023*, 2023.

[8] Xiaoshan Luo, Zhenyu Wang, Qingchang Wang, Xuechen Shao, Jian Lv, Lei Wang, Yanchao Wang, and Yanming Ma. Crystalflow: a flow-based generative model for crystalline materials. *Nature Communications*, 16(1):9267, 2025.

[9] Xiao-Qi Han, Peng-Jie Guo, Ze-Feng Gao, Hao Sun, and Zhong-Yi Lu. Invdesflow-al: active learning-based workflow for inverse design of functional materials. *npj Computational Materials*, 11(1):364, 2025.

[10] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S Jaakkola. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. In *International Conference on Learning Representations*, 2021.

[11] Cai-Yuan Ye, Hong-Ming Weng, and Quan-Sheng Wu. Con-CDVAE: A method for the conditional generation of crystal structures. *Computational Materials Today*, 1:100003, May 2024.

[12] Xiaoshan Luo, Zhenyu Wang, Pengyue Gao, Jian Lv, Yanchao Wang, Changfeng Chen, and Yanming Ma. Deep learning generative model for crystal structure prediction. *npj Computational Materials*, 10(1):254, 2024.

[13] Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation. *Science Bulletin*, 70(21):3522–3533, 2025.

[14] Zhendong Cao and Lei Wang. CrystalFormer-RL: Reinforcement Fine-Tuning for Materials Design. *arXiv preprint arXiv:2504.02367*, 2025.

[15] Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Representation Learning*, volume 2024, pages 6836–6853, 2024.

[16] OpenAI, Josh Achiam, and Steven Adler. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024.

[17] DeepSeek-AI. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[18] Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024.

[19] Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M. Wood. Flowllm: Flow matching for material generation with large language models as base distributions, 2024.

[20] Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, 2025.

[21] Duo Zhang, Xinzijian Liu, Xiangyu Zhang, et al. DPA-2: a large atomic model as a multi-task learner. *npj Computational Materials*, 10(1):293, 2024.

[22] Ioanna Pallikara, Prakriti Kayastha, Jonathan M Skelton, and Lucy D Whalley. The physical significance of imaginary phonon modes in crystals. *Electronic Structure*, 4(3):033002, jul 2022.

[23] Zhenfeng Ouyang, Bo-Wen Yao, Xiao-Qi Han, Peng-Jie Guo, Ze-Feng Gao, and Zhong-Yi Lu. High-temperature superconductivity in li_2auh_6 mediated by strong electron-phonon coupling under ambient pressure. *Phys. Rev. B*, 111:L140501, Apr 2025.

[24] Stefano Baroni, Stefano de Gironcoli, Andrea Dal Corso, and Paolo Giannozzi. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.*, 73:515–562, Jul 2001.

[25] Atsushi Togo and Isao Tanaka. First principles phonon calculations in materials science. *Scripta Materialia*, 108:1–5, 2015.

[26] Stefano Baroni, Stefano de Gironcoli, Andrea Dal Corso, and Paolo Giannozzi. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.*, 73:515–562, Jul 2001.

[27] Jesús Carrete, Wu Li, Natalio Mingo, Shidong Wang, and Stefano Curtarolo. Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X*, 4:011019, Feb 2014.

[28] Anyang Peng, Xinzijian Liu, Ming-Yu Guo, Linfeng Zhang, and Han Wang. The openlam challenges: Lam crystal philately competition. *Machine Learning: Science and Technology*, 6(2):020701, jun 2025.

[29] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. Mattersim: A deep learning atomistic model across elements, temperatures and pressures, 2024.

[30] Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *J. Chem. Theory Comput.*, 20(11):4857–4868, 2024.

[31] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

[32] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022.

[33] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *International Conference on Learning Representations (ICLR)*, 2024.

[34] Saro Passaro and C Lawrence Zitnick. Reducing SO(3) Convolutions to SO(2) for Efficient Equivariant GNNs. In *International Conference on Machine Learning (ICML)*, 2023.

[35] Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential, 2024.

[36] Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale, 2025.

[37] Chao Shen, Xiaoqi Han, Heng Cai, Tong Chen, Yu Kang, Peichen Pan, Xiangyang Ji, Chang-Yu Hsieh, Yafeng Deng, and Tingjun Hou. Improving the reliability of language model-predicted structures as docking targets through geometric graph learning. *Journal of Medicinal Chemistry*, 68(2):1956–1969, 2025.

[38] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

[39] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

[40] Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[41] Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials*, 11(1):178, 2025.

[42] Atsushi Togo, Laurent Chaput, Terumasa Tadano, and Isao Tanaka. Implementation strategies in phonopy and phono3py. *J. Phys. Condens. Matter*, 35(35):353001, 2023.

[43] Atsushi Togo. First-principles phonon calculations with phonopy and phono3py. *J. Phys. Soc. Jpn.*, 92(1):012001, 2023.

[44] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6790–6802. Curran Associates, Inc., 2021.

[45] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 18–24 Jul 2021.

[46] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

Corresponding authors: Correspondence and requests for materials should be addressed to Ze-Feng Gao (zfgao@ruc.edu.cn) and Zhong-Yi Lu (zlu@ruc.edu.cn).

Competing interests: The authors declare no competing interests.

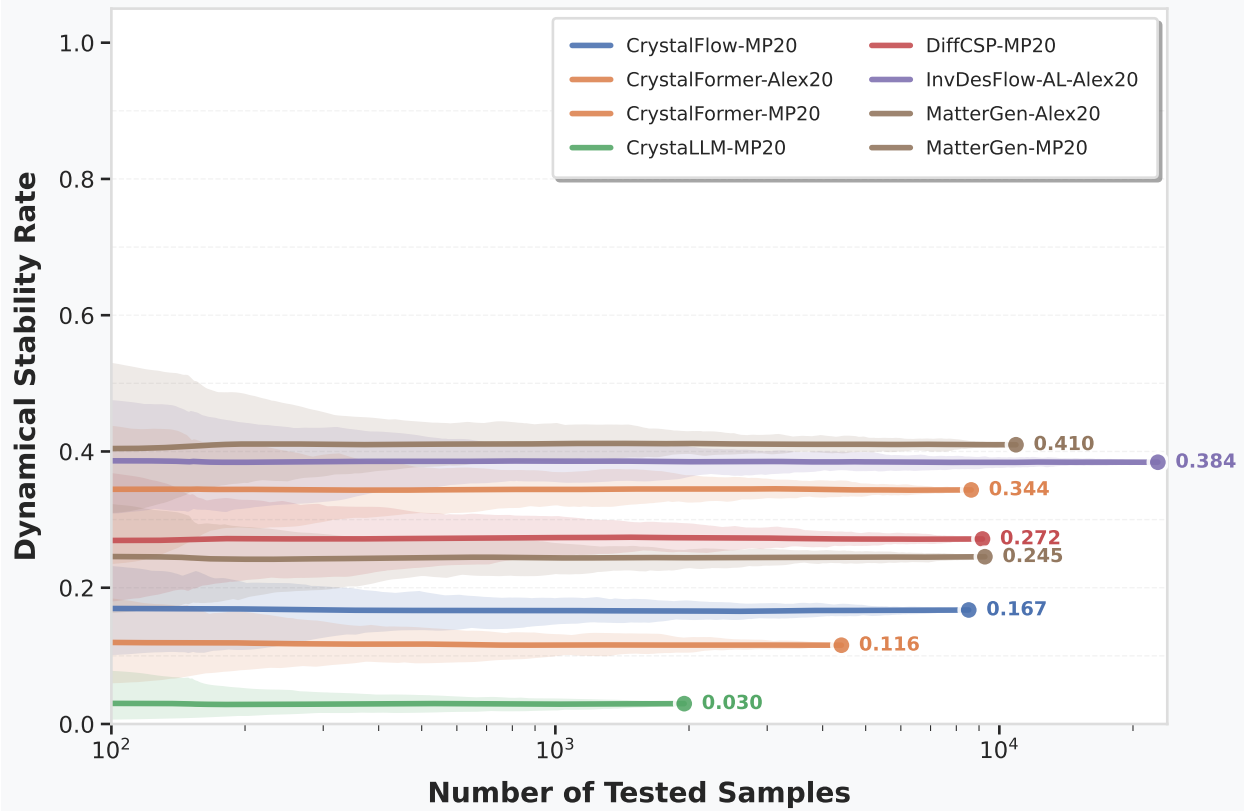**Supplementary information:** The supplementary information is attached.

Figure 5: Convergence of the dynamical stability rate for various crystal generative models as a function of the number of tested samples. The figure demonstrates that stability estimates converge as sample size increases, with errors becoming negligible above 4,000 samples, ensuring robust and fair model ranking.

1 Convergence Analysis of Dynamical Stability Rate

Figure 5 illustrates the convergence of the dynamical stability rate for various crystal generative models as a function of the number of tested samples. This indicates that when the sample size exceeds approximately 4,000, the estimation error of the stability rate becomes sufficiently small, and further increasing the number of samples has negligible impact on the ranking of the models. Although the effective test sample size for CrystaLLM is slightly below this threshold, the error range is still insufficient to alter its final ranking. This analysis provides quantitative support for the comparison of dynamical stability in the main text, ensuring the reliability and fairness of the model performance evaluation.

2 Summary of Crystal Generation and Dynamical Stability Statistics

Table 3 summarizes the statistics of crystal generation and stability for the nine crystal generative models evaluated in this study. For each model, the table lists the number of structures that successfully converged during relaxation, the number of dynamically stable crystals identified through phonon calculations, the number of crystals successfully generated via the input scripts, the count of unique CIF files after removing duplicates, and the total number of crystals originally generated. This comprehensive overview provides a quantitative comparison of the performance

of different models in terms of generation success, structural relaxation, and dynamical stability, offering a valuable reference for further analysis and discussion in the main text.

Table 3: Summary of crystal generation and stability statistics for different models.

Model	Relaxed	Dynamically Stable	Input Script Success	Unique CIFs	Total Generated
CrystalFlow-MP20	8,533	1,428	8,852	9,952	16,000
CrystalFormer-Alex20	8,642	2,969	8,807	8,986	40,000
CrystalFormer-MP20	4,408	510	4,990	5,143	20,000
CrystaLLM-MP20	1,951	58	2,074	2,074	16,000
DiffCSP-MP20	9,163	2,488	9,959	10,000	16,000
InvDesFlow-AL-MP20	8,000	2176	-	-	-
InvDesFlow-AL-Alex20	22,755	8,743	24,997	25,000	30,000
MatterGen-Alex20	10,902	4,469	11,829	11,829	16,000
MatterGen-MP20	9,279	2,278	10,000	10,000	16,000

In Table 3, "Total Generated" denotes the total number of crystals produced by each model. To ensure feasibility for subsequent relaxation and phonon calculations, the generated crystals were first filtered to remove duplicates and invalid CIF files. If the resulting set exceeded 10,000 crystals, a random subset of 10,000 was selected for relaxation; otherwise, all available crystals were used for phonon calculations. Notably, InvDesFlow-AL-Alex20 and MatterGen-Alex20 represent special test cases with substantially larger generation counts, designed to investigate the sample size required for the convergence of dynamical stability estimates. Experiments showed that stability ratios converge reliably with approximately 4,000 samples, making further increases unnecessary for subsequent analyses. Since InvDesFlow-AL-MP20 and DiffCSP-MP20 share the same model architecture and MP20 training set, no systematic difference in data distribution is expected. The dynamical stability ratio of InvDesFlow-AL-MP20 reported in the main text is therefore estimated by randomly sampling 8,000 structures from the 9,163 DiffCSP-MP20 samples with completed dynamical stability evaluations and averaging over five independent trials, which matches the sample size while reducing statistical fluctuations and ensuring a fair and robust comparison between models.

3 Crystal Relaxation and Dynamical Stability Statistics Across Bandgap Ranges

Table 4 summarizes the statistics of crystal relaxation and dynamical stability across different bandgap ranges. For each bandgap category, the table lists the number of structures that successfully converged during relaxation, the number of dynamically stable crystals identified via phonon calculations, and the total number of crystals generated. This overview provides a quantitative reference for analyzing the relationship between bandgap and dynamical stability.

Table 4: Statistics of crystal relaxation and dynamical stability for different bandgap ranges.

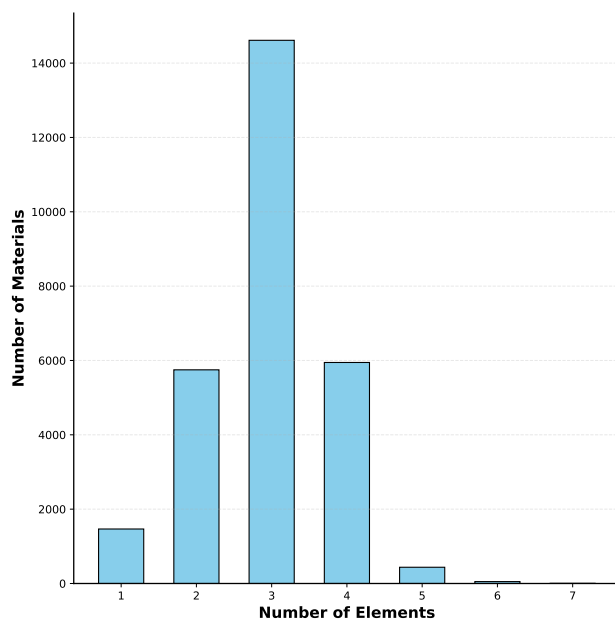| Bandgap Range	Relaxed	Dynamically Stable	Total Generated
$E_g = 0.5$ eV	6,524	1,532	10,000
$E_g = 1.5$ eV	9,478	1,448	16,000
$E_g = 2.5$ eV	6,133	816	10,000
$E_g = 3.5$ eV	5,735	763	10,000
$E_g = 4.5$ eV	5,340	617	10,000

Figure 6: Elemental distribution of 28,119 dynamically stable crystals