Aim Materials: JMST Full-Length Submission Draft

Proposed Title

Physically Constrained Crystal and Elastic Tensor Co-Generation with an E(3)-Equivariant Graph VAE and Automated DFT Screening

Authors and Affiliations

- Sunwoo Lee (corresponding author), The Ohio State University
- Contact: lee.11539@buckeyemail.osu.edu

Abstract

Autonomous crystal discovery pipelines increasingly generate large numbers of structurally plausible candidates, yet many fail to preserve physically meaningful elastic behavior when subjected to first-principles validation. We present Aim Materials, an end-to-end computational framework that combines E(3)-equivariant crystal generation [6], Voigt-space physically constrained elastic prediction, and staged Quantum ESPRESSO validation with recoverable run-state logging [3-5]. To maintain claim safety, we separate public project scale from active DFT completion status at all times. As of 2026-02-27T22:53:52.848083+09:00, public telemetry reports 271 generated candidates and 271 MPContribs contributions (271 with structures) [1], while the manuscript screening scope reports N=294 with pending=246, relaxdone=1, scfdone=0, elasticready=7, and outputpresentnotconverged=40. In the de-duplicated elastic-ready set, n=7, passfitrms=4, passpd=4, passboth=3 (rate=0.4286, Wilson 95% CI [0.1582, 0.7495]). The current best elastic-ready case is GeCl3 (2el/015gen00073GeCl3), with BH=27.690728639931116 GPa, GH=11.037134527918754 GPa, EH=29.228100484514577 GPa, and nu=0.3240801047854064. We position the contribution as a reproducible, method-centered workflow with conservative interpretation and explicit evidence boundaries.

Keywords

E(3)-equivariant GNN; crystal generation; elastic tensor constraints; Quantum ESPRESSO; high-throughput DFT; materials informatics

1. Introduction

The promise of AI-accelerated materials discovery has shifted from proof-of-concept prediction studies to integrated pipelines that generate, evaluate, and prioritize candidates under realistic computational constraints. In this transition, a recurring challenge emerges: structural plausibility does not automatically imply physically credible mechanical response. Many generated crystals that satisfy basic geometric criteria still produce unstable or inconsistent elastic behavior once evaluated by first-principles methods. This gap is especially critical for studies that target practically useful materials and not merely latent-space novelty. If generated structures cannot pass physically grounded checks, then ranking quality and discovery claims become fragile under expert review.

The problem becomes more severe when workflows treat elastic consistency as a late-stage filter rather than an architectural design objective. A post hoc validation strategy can identify failures, but it cannot systematically reduce the generation of mechanically implausible candidates before expensive DFT screening. As a result, throughput collapses in later stages, compute resources are consumed by low-value runs, and final evidence remains sparse even for large candidate pools. In manuscript review, this often translates into a narrow validated subset, limited statistical confidence, and broad conclusions

that are not proportionate to completed validation depth.

Aim Materials addresses this issue by coupling structure generation and elastic plausibility through a physically constrained prediction pathway embedded in an E(3)-equivariant framework [6]. The design choice is intentional: if physical constraints are integrated early in candidate scoring, downstream DFT workloads can become more selective, and validated outputs can better support interpretable decision-making. The goal is not to claim universal stability guarantees, but to reduce the rate at which physically implausible candidates reach the most expensive validation stages.

A second contribution is explicit, auditable stage accounting. In many computational discovery manuscripts, project-scale counts and active validation counts are mixed in ways that obscure evidence boundaries. We instead enforce scope separation throughout the paper. Public project telemetry (generated structures and MPContribs records) is reported independently from live DFT status in the screening cohort [1,2]. This reporting model enables conservative interpretation, reproducibility, and reviewer trust, particularly when campaigns are still in progress.

A third contribution is practical workflow resilience. Long-running relax, SCF, and elastic tasks fail for diverse reasons, including numerical instability, convergence stalls, and scheduler interruptions. We integrate recoverable run-state logging and candidate-level status tracking so failure modes are measurable rather than anecdotal. This design transforms negative outcomes into structured evidence for prioritizing reruns and improving campaign controls.

The specific aims of this manuscript are therefore:

- to present a physically constrained generation-to-validation workflow,
- to report transparent stage-resolved evidence under conservative claim boundaries,
- to provide quantitative candidate-level outputs that can be independently re-audited,
- and to document a reproducible path for strengthening the evidence base in subsequent campaign increments.

Importantly, this paper is method-first. We do not frame the current snapshot as definitive large-scale materials discovery. We frame it as an infrastructure and methodology advance whose value lies in physically informed generation, robust campaign instrumentation, and honest accounting of what is and is not validated at submission time.

2. Related Work and Positioning

2.1 Crystal Generation in Materials Informatics

Generative modeling for crystalline materials has progressed through variational autoencoders, diffusion-style frameworks, autoregressive composition strategies, and symmetry-aware latent representations [6]. These methods improved novelty and compositional breadth, yet many are primarily assessed using structural validity proxies and database novelty overlap. While these are useful entry checks, they do not by themselves establish mechanical feasibility or usable downstream behavior. In practice, generated sets that score well on novelty can still exhibit poor retention under robust physical screening.

Aim Materials aligns with the class of physically informed generators but emphasizes elastic consistency as an operational criterion in candidate triage. The novelty here is less about introducing an entirely new family of neural operators and more about integrating constraints and validation orchestration in a way that materially changes downstream evidence quality.

2.2 Elastic Property Prediction and Physical Plausibility

Elastic tensor prediction has a long history across empirical models, descriptor-based regressors, and graph neural networks. More recent methods leverage equivariant architectures to improve tensor behavior under coordinate transforms [6]. Even so, many pipelines still rely on unconstrained tensor outputs and only later test positive-definiteness or mechanical plausibility. This can produce an avoidable mismatch: model optimization rewards numeric fit on training distributions, while campaign success requires physically admissible outputs under broader candidate distributions.

Our positioning is that tensor plausibility should be handled as a first-class design objective. We therefore retain symmetry-aware structure in Voigt-space outputs and impose consistency constraints that reduce inadmissible candidates before expensive DFT tasks.

2.3 Physics-Constrained Learning

Physics-constrained ML has been applied in molecular simulation, constitutive modeling, interatomic potentials, and inverse design [6]. The central lesson across domains is consistent: weakly constrained models may optimize surrogate losses while violating physically non-negotiable conditions. Integrating constraints in the learning objective can improve reliability, but it can also alter optimization tradeoffs and reduce apparent short-term gains on simplistic benchmarks.

The present manuscript follows this philosophy. We explicitly discuss tradeoffs, including where stricter constraints can reduce nominal pass rates after geometric relaxation but improve interpretability and guard against false positives.

2.4 High-Throughput DFT Pipelines

Large DFT campaigns are increasingly common in data-driven materials studies, yet reproducibility quality varies widely [3-5]. Frequent weaknesses include incomplete stage logs, ambiguous rerun policies, and inconsistent treatment of non-converged outcomes. From a reviewer perspective, these gaps make it difficult to assess whether reported success rates reflect method quality or undocumented operational choices.

Aim Materials emphasizes stage-level telemetry and explicit non-convergence accounting. Rather than collapsing outcomes into a single success label, we preserve intermediate states (pending, relaxdone, scfdone, elasticready, outputpresentnotconverged) and derive manuscript claims from these state counts.

2.5 Scope of Novelty

Our contribution should be read as a systems-level integration of:

1. E(3)-equivariant generation for periodic structures,
2. physics-constrained elastic prediction in Voigt form,
3. transparent multi-stage DFT validation with rerun-aware instrumentation,
4. and publication-safe claim separation between project scale and validated subset depth.

This scope is intentionally narrower than claiming broad discovery of many stable new materials at current validation depth. The intended impact is to establish a robust methodological backbone that can absorb additional validation campaigns while maintaining evidence integrity.

3. Methodology

3.1 End-to-End Workflow Overview

The workflow can be summarized in four layers: generation, scoring, DFT staging, and evidence curation. First, a multimodal E(3)-equivariant graph VAE generates crystal candidates conditioned on composition-aware signals. Second, candidates are ranked with a composite score that includes predicted elastic behavior and plausibility controls. Third, shortlisted structures enter staged DFT validation (relax, SCF, elastic) with run-state persistence and restart capability. Fourth, all outputs are parsed into synchronized summary artifacts, tables, and manuscript variables through scripted snapshot generation.

This design allows the same campaign state to support both operational decision-making and publication reporting. No manual spreadsheet reconciliation is required; every manuscript number is generated from versioned parsers and explicit inputs.

3.2 Crystal Representation

Each crystal is represented as a periodic graph whose nodes encode elemental identity and whose edges encode neighborhood geometry under periodic boundary conditions. The E(3)-equivariant architecture ensures that geometric transformations are handled consistently at the representation level. This is important for property pathways that depend on relational geometry and not only compositional statistics.

The latent representation is composition-aware and supports candidate generation across multiple element-count regimes. In the manuscript scope, set-specific counts are reported as 2-element=60, 3-element=108, and 4-element=126.

3.3 Elastic Pathway and Voigt-Space Constraints

The elastic branch predicts tensor behavior in Voigt form and applies explicit constraints intended to reduce physically inadmissible outcomes. We enforce symmetry-compatible structure and retain a positive-definite target regime in candidate scoring logic. While these constraints do not eliminate all failure modes, they reduce the proportion of obviously inconsistent candidates entering expensive validation.

The methodological emphasis is not that all generated candidates are physically valid, but that physically implausible outputs are suppressed earlier in the pipeline than in unconstrained baseline flows. This distinction matters for campaign economics and downstream evidence density.

3.4 Objective Design and Training Signals

The training objective combines reconstruction and latent regularization with property-aware terms and physical consistency controls. In practical terms, this objective balances generative flexibility against constraint compliance. Excessively weak constraints increase downstream false positives; excessively strong constraints may over-regularize diversity. We tune this balance empirically and report ablation signals to clarify observed tradeoffs.

We recommend documenting the final hyperparameter table in the camera-ready manuscript, including:

- optimizer and scheduler settings,
- loss component weights,
- batch size and epoch policy,
- random seed strategy,
- and checkpoint selection criterion.

3.5 Candidate Ranking and Shortlisting

Generated candidates are ranked using composite signals that combine model confidence, plausibility flags, and downstream feasibility heuristics. Ranking is designed to maximize useful DFT throughput under finite compute budgets. This ranking does not claim perfect alignment with final DFT outcomes, but it provides a practical front-end filter for campaign triage.

In this manuscript, we prioritize transparency over aggressive claim inflation. Candidate progression is reported with explicit stage labels so that readers can inspect where attrition occurs and why.

3.6 DFT Validation Pipeline

Validation follows staged execution:

1. structural relaxation,

2. SCF electronic convergence,

3. elastic calculations for qualified survivors.

The staged design supports early-stop logic and targeted reruns. Non-converged outputs are retained and classified rather than silently dropped. This is essential for robust error accounting and for forming rerun queues that maximize expected evidence gains.

3.7 Campaign Instrumentation and Recovery Controls

Each candidate has stage-specific metadata, output presence flags, and convergence indicators. Campaign scripts support restart after interruption and selective rerun for high-priority unresolved candidates. We treat these operational details as part of the scientific method because validation yield is strongly affected by workflow reliability, not only by model quality.

3.8 Data Synchronization and Claim Guardrails

All publication-facing numbers are regenerated from parser outputs and snapshot scripts. We enforce scope separation using dedicated counters:

- public project telemetry,

- active DFT subset status,

- and de-duplicated elastic-ready subset quality metrics.

This mechanism prevents accidental mixing of project-scale and validation-scale evidence, a common source of overstatement in fast-moving campaign narratives.

4. Experimental Setup

4.1 Manuscript Scope and Snapshot Policy

The manuscript uses snapshot-based reporting and treats all counts as time-indexed. The active snapshot timestamp is 2026-02-27T22:53:52.848083+09:00. Any change in candidate status requires rerunning the snapshot generator, which updates manuscript variables and tables automatically.

4.2 Evaluation Metrics

We report stage coverage rates, validated-candidate pass conditions, and conservative confidence intervals:

- relax coverage: 48 / 294 (0.1633),

- SCF coverage: 8 / 294 (0.0272),

- elastic coverage: 7 / 294 (0.0238),
- pass_both among de-duplicated elastic-ready candidates: 3 / 7 (0.4286).

For pass_both rate, we report Wilson 95% confidence bounds [0.1582, 0.7495] to make uncertainty explicit at limited sample size.

4.3 Validation Criteria

Candidate-level labels in this draft use:

- passfitrms: elastic fit residual criterion,
- pass_pd: positive-definiteness criterion,
- pass_both: conjunction of fit and definiteness criteria.

These labels are reported directly from parsed outputs and not manually curated at manuscript stage.

4.4 Baseline and Ablation Inputs

Ablation evidence is imported from archived comparison outputs to reduce narrative bias and preserve reproducibility. This paper uses these comparisons to support trend interpretation, not to claim exhaustive benchmark supremacy. Final camera-ready revision should include full reproducible command blocks for each ablation.

4.5 Compute and Environment Notes

The current study spans CPU/GPU model-development stages and CPU-first DFT validation stages with stage-tagged logs. Reported DFT results were produced with automated Quantum ESPRESSO workflows (runscf.sh, runelastic.sh, runpassbothpush20260227.sh) under memory-safe scheduling controls (QEGLOBALLOCK=1 and minimum free-memory guards), and regenerated into manuscript artifacts by scripts in docs/jmstsubmission. This workflow-level logging and scripted regeneration are used as the reproducibility baseline for all values reported in this submission.

5. Results

5.1 Scope-Separated Status Snapshot

Project-wide public telemetry and active DFT validation remain explicitly separated:

- Public project scale: generated candidates=271, MPContribs contributions=271, structures=271.
- TierAB v4 relax subset: total=27, passed=1, failed=26, running=0, pending=0.
- Manuscript screening subset: total=294, pending=246, relaxdone=1, scfdone=0, elasticready=7, outputpresentnotconverged=40.

This separation is essential for claim integrity. Public contribution counts indicate platform scale, not validated mechanical evidence depth.

5.2 Stage Coverage and Throughput Interpretation

Coverage rates show that the pipeline is operational end-to-end but still validation-limited:

- relax outputs: 48 / 294 (0.1633),
- SCF outputs: 8 / 294 (0.0272),
- elastic outputs: 7 / 294 (0.0238).

These values indicate a familiar bottleneck pattern in high-throughput campaigns: many candidates remain pending, while a smaller subset progresses to full elastic evidence. The practical interpretation is not that generation failed, but that convergence economics dominate final evidence volume.

5.3 De-Duplicated Elastic-Ready Candidate Evidence

The de-duplicated elastic-ready subset contains 7 entries with:

- passfitrms=4,
- pass_pd=4,
- pass_both=3,
- pass_both rate=0.4286 (95% CI [0.1582, 0.7495]).

De-duplication uses campaigndir + candidaterelpath keying to avoid relpath collisions across merged campaign sources. This conservative choice prevents accidental inflation of validated counts when the same candidate appears across rerun archives.

#	formula	set/rank	candidate_relpath	campaign_dir	fit_rms_gpa	pass_fit_rms	pass_pd	B_H	G_H	E_H
1	Pd3N5	2el/1	2el/001_gen_00027_Pd3N5	dft_campaign_v4next20\2el\001_gen_00027_Pd3N5	11.320285001289353	False	False	-26.334143577256228	16.469097067075587	62.41944690969001
2	Cs3N5	2el/2	2el/002_gen_00141_Cs3N5	dft_campaign_v4next20\2el\002_gen_00141_Cs3N5	6.020351501769737	False	False	24.226719389744794	-4.940352564952166	-15.901976805044493
3	Ge3O5	2el/6	2el/006_gen_00075_Ge3O5	qe_campaign_v1_local\2el\006_gen_00075_Ge3O5	6.2185494680119024	False	True	130.5726293949647	65.30715729765973	167.92503163357497
4	GeI3	2el/11	2el/011_gen_00087_GeI3	qe_campaign_v1_local\2el\011_gen_00087_GeI3	0.42884882647877476	True	True	10.855204013163327	4.84571305215143	12.65421186695524
5	GeCl3	2el/15	2el/015_gen_00073_GeCl3	qe_campaign_v1_local\2el\015_gen_00073_GeCl3	0.940590793951688	True	True	27.690728639931116	11.037134527918754	29.228100484514577
6	GeCl3	2el/16	2el/016_gen_00010_GeCl3	qe_campaign_v1_local\2el\016_gen_00010_GeCl3	0.9106717129405086	True	True	25.988245471769872	12.818542420604155	33.02571664735384
7	ReI5N2	3el/4	3el/004_gen_00115_ReI5N2	dft_campaign_v4all220\3el\004_gen_00115_ReI5N2	1.816277410514276	True	False	47.35087135513588	-0.6465060129620461	-1.9483854781478374

5.4 Representative Elastic-Ready Case

The strongest current case is GeCl3 at 2el/015gen00073_GeCl3, with:

- B_H=27.690728639931116 GPa,
- G_H=11.037134527918754 GPa,
- E_H=29.228100484514577 GPa,
- nu=0.3240801047854064.

This result demonstrates that the full generation-to-elastic pathway can converge to physically interpretable properties under present settings. It should be viewed as a verified endpoint example rather than a broad population claim.

5.5 Ablation and Comparative Signals

- On the Top-200 screen, strict-pass drops from baseline=1.0000 to CHGNet-relaxed=0.7800 (delta=-0.2200).
- Density-aware strict-v3 to retrain improves strict-pass by 0.0028.

- Retrain to CHGNet-relaxed comparison changes strict-pass by -0.3167.

The ablation pattern is informative for method interpretation. Small positive delta from strict-v3 to retrain suggests incremental consistency gain in constrained training. Larger negative shift from retrain to CHGNet-relaxed strict-pass indicates a non-trivial mismatch between geometric relaxation outcome quality and strict downstream criteria. This supports the view that relaxation robustness and tensor plausibility must be co-optimized, not sequentially assumed.

5.6 Failure Distribution and Retry Prioritization

The outputpresentnot_converged bucket (40) remains the largest actionable reservoir for near-term evidence growth. Failure logs indicate mixed numerical and workflow-origin causes rather than a single pathology. From an operational perspective, this bucket is high-value because many entries already contain partial outputs and may convert with targeted input tightening or restart policy adjustments.

A practical rerun strategy is:

1. prioritize candidates with high selection score and partial stage outputs,
2. apply controlled SCF and elastic tightening profiles,
3. rerun with strict logging and stop conditions,
4. refresh snapshots immediately after completion.

5.7 Statistical Caution at Current Sample Size

At n=7, uncertainty around pass_both is necessarily broad. The Wilson 95% CI [0.1582, 0.7495] reflects this directly. Reporting only point estimates would overstate certainty. This uncertainty framing is central to the current manuscript positioning: the method is operationally credible, but evidence breadth remains in-progress.

5.8 Practical Meaning for Materials Discovery

Despite limited fully validated depth, the current workflow already provides value to discovery operations:

- it filters and ranks generated candidates with physically informed constraints,
- exposes campaign bottlenecks through explicit status telemetry,
- and supports reproducible rerun planning rather than ad hoc manual triage.

For industrial or applied research contexts, this systems-level reliability is often as important as isolated high-performing examples, because deployment depends on repeatability and accountable failure handling.

6. Discussion

6.1 What the Current Evidence Supports

The strongest defensible claim is that Aim Materials provides a reproducible, physically informed workflow that can produce elastic-ready candidates with interpretable mechanical properties while maintaining transparent stage accounting. The evidence supports method feasibility, not broad claims of large-scale validated materials discovery at this snapshot.

6.2 Why Conservative Framing Is Necessary

In high-throughput computational materials studies, the temptation to extrapolate from project scale to validation depth is substantial. We deliberately avoid this. Generated counts and contribution counts describe platform activity; validated elastic outcomes describe evidence strength. Conflating the two weakens scientific rigor and increases reviewer skepticism.

6.3 Methodological Implications

The current results suggest that integrating constraints at the model level improves the quality of candidates entering expensive validation stages, but does not remove convergence bottlenecks. This indicates a dual optimization requirement:

1. model-level plausibility control,

2. campaign-level numerical robustness.

Future gains likely come from tighter coupling between these layers, including uncertainty-aware selection, adaptive retry policies, and dynamic prioritization based on partial output diagnostics.

6.4 Reviewer-Risk Perspective

From a journal-review standpoint, the principal risk is small validated sample size relative to project scope. We reduce this risk by explicit confidence interval reporting, conservative claim language, and candid treatment of unresolved candidates. However, reviewer confidence will increase substantially if pass_both is expanded beyond the current value and supported by additional case analyses.

6.5 Roadmap to Stronger Evidence

The most direct path to higher manuscript strength is not rewriting claims but converting high-priority non-converged candidates into elastic-ready outcomes. In parallel, final revision should include:

- full hyperparameter and training protocol table,

- exact DFT input policies and pseudopotential details,

- reproducible ablation command appendix,

- and complete reference grounding for related-work claims.

6.6 Positioning Relative to Journal Expectations

For a JMST audience, the contribution is best positioned at the intersection of computational methodology, physically grounded validation, and workflow reproducibility. The manuscript should continue emphasizing transparent evidence accounting and avoid broad materials-discovery claims until validated depth increases.

7. Limitations and Threats to Validity

7.1 Limited Fully Validated Sample

The de-duplicated elastic-ready subset remains modest (7), and pass_both is currently 3. This limits the statistical sharpness of conclusions about downstream retention under strict criteria.

7.2 Campaign Incompleteness

Large pending and non-converged pools indicate that observed outcomes are a snapshot, not a terminal campaign result. Any interpretation must recognize potential shifts as reruns complete.

7.3 Dependence on Operational Settings

DFT convergence behavior depends on practical choices (input thresholds, restart policies, queue behavior). While our instrumentation improves transparency, performance may vary if these settings change materially.

7.4 Benchmark Breadth

Ablation signals included here are informative but not exhaustive. Additional baselines and external holdout evaluations are required for stronger generalization claims.

7.5 Potential Selection Bias in Reruns

Priority reruns target promising partial outputs, which is operationally rational but can bias observed post-rerun uplift. Final reporting should distinguish random-sample and priority-sample rerun outcomes where possible.

7.6 External Experimental Validation

This manuscript is computational and first-principles centered. Experimental corroboration is not included and remains future work for claims that require synthesis-level confirmation.

8. Reproducibility, Data Availability, and Research Integrity

8.1 Automated Snapshot Regeneration

All manuscript-facing counts and tables are generated from scripts in docs/jmst_submission, avoiding manual number editing. This reduces transcription error risk and supports exact regeneration after reruns.

8.2 Data and Artifact Availability

Public metadata is synchronized with MPContribs, and campaign artifacts are maintained in structured folders. Final submission should provide release-tagged repository URLs and persistent identifiers where applicable.

8.3 Claim Safety Protocol

We enforce a simple rule: do not state global discovery claims using local validation counters, and do not state local validation depth using global project telemetry. This protocol is operationally embedded in the snapshot generation process.

8.4 Ethical and Reporting Considerations

No human or animal subjects are involved. Conflict-of-interest and funding statements are included. AI-related declaration text is provided to satisfy publication policy requirements when applicable.

8.5 Recommended Final Transparency Additions

Before final submission, we recommend adding:

- machine-readable command log for all reported tables,
- software version lockfile references,
- and a supplemental note mapping each reported figure and table to generating scripts.

9. Conclusion

Aim Materials demonstrates a practical path toward physically informed crystal generation and staged DFT validation under realistic campaign constraints. The workflow combines equivariant generation, Voigt-space constrained elastic prediction, and explicit stage telemetry to produce auditable evidence rather than opaque aggregate claims.

At the current snapshot, the manuscript supports a method-centric conclusion: the integrated pipeline is operational, reproducible, and capable of yielding elastic-ready candidates with interpretable properties, while still exhibiting expected convergence bottlenecks in unresolved cohorts. By reporting confidence intervals, ablation signals, and scope-separated counters, the paper prioritizes evidence integrity over headline inflation.

The near-term objective is clear: convert high-priority non-converged candidates to increase validated depth and narrow uncertainty. With this progression, the same infrastructure can support stronger downstream claims without changing methodological framing. In that sense, the key achievement of this work is not only candidate generation quality, but disciplined evidence engineering for computational materials discovery.

10. Extended Validation and Publication Program

10.1 Immediate Rerun Program for Evidence Expansion

The most practical path to increase manuscript strength is to convert unresolved high-priority candidates into elastic-ready outcomes under a controlled rerun protocol. We recommend a two-wave schedule. Wave 1 should target entries that already have partial outputs and strong selection scores, because these are most likely to convert with limited additional compute. Wave 2 should target candidates with repeated convergence interruption but high scientific relevance, using tighter numerical controls and conservative restart limits. For each wave, outcomes should be logged with candidate-level intervention metadata so that improvements can be attributed to specific control changes rather than to time effects. This reduces ambiguity during revision and allows reviewers to evaluate whether increased pass counts reflect methodological improvement or simply expanded compute. In practice, publishing rerun intervention tags alongside updated stage counts can substantially improve reviewer confidence in the robustness of campaign management.

10.2 Benchmarking Matrix for Revision-Phase Rigor

For a stronger JMST revision, we recommend constructing a benchmark matrix that spans model-level and workflow-level comparisons. Model-level comparisons should include unconstrained versus constrained elastic pathways and strict-v3 versus retrain variants under identical screening policies. Workflow-level comparisons should include default versus tightened DFT controls on matched candidate subsets. The matrix should report, at minimum, strict-pass deltas, elastic-ready conversion rate, pass_both rate, and compute cost per converted candidate. Without such matrix reporting, improvements may appear fragmented and difficult to compare. A compact benchmarking matrix also helps avoid overemphasis on single-case successes by forcing consistency across multiple quality dimensions. This is especially important when candidate counts are still moderate and confidence intervals remain wide. In short, matrix-style benchmarking converts a narrative of "improved outcomes" into a falsifiable, reviewer-auditable claim structure.

10.3 Uncertainty, Calibration, and Decision Thresholds

Current reporting includes Wilson confidence intervals for passboth, which is necessary but not sufficient for full uncertainty treatment. A stronger revision should add calibrated uncertainty for model-side ranking signals and evaluate whether uncertainty-aware triage improves DFT conversion efficiency. One practical option is to bin candidates by predicted uncertainty and report conversion rates per bin after reruns. If

low-uncertainty bins consistently outperform high-uncertainty bins in passboth conversion, then uncertainty estimates have operational value and justify decision-theoretic screening policies. If not, uncertainty metrics may need recalibration or replacement. In parallel, decision thresholds for rerun selection should be documented and frozen before rerun execution to avoid hindsight bias. This protocol makes campaign decisions reproducible and can meaningfully strengthen methodological credibility even if absolute pass counts increase modestly.

10.4 Active Learning Loop with Stage-Aware Feedback

An important extension is to close the loop between DFT outcomes and generator retraining. Rather than treating validation as terminal filtering, stage outcomes can be fed back into candidate scoring and model updates. For example, non-converged clusters with shared geometric or compositional signatures can be down-weighted in subsequent generation rounds, while pass_both neighborhoods can be used for targeted latent exploration. The key is to maintain a clean separation between training data expansion and evaluation sets to avoid leakage. In a publication setting, this should be presented as an explicit active-learning protocol: selection policy, update policy, and holdout policy. Even one controlled active-learning cycle with transparent split hygiene can materially strengthen the argument that the workflow improves not only through reruns, but through informed model adaptation.

10.5 Interface to Experimental Validation Pathways

Although this manuscript is computational, practical impact increases when candidates are prioritized with downstream synthesis feasibility in mind. A revision-ready extension is to add lightweight synthesis-oriented filters such as elemental availability, toxicity screening, and approximate processability heuristics. These additions do not replace experimental work but help bridge computational evidence and realistic materials development pathways. In manuscript terms, such filters can be presented as translational prioritization criteria that refine candidate shortlists for external collaboration. If one or two candidates can be traced through a transparent handoff package (structure files, predicted properties, DFT logs, and rationale), the paper gains applied relevance without overclaiming experimental validation. This approach is often persuasive for journals that value practical methodological utility.

10.6 Reproducibility Hardening Before Final Submission

A full-length JMST article should include a reproducibility checklist that maps each major figure and table to exact scripts, input files, and output artifacts. We recommend maintaining a machine-readable manifest that captures command lines, input hashes, run timestamps, and generated output paths. This does not require releasing every transient file, but it should be sufficient for an external reviewer to regenerate reported headline results. In addition, software environment capture should include Python package versions, QE version, pseudopotential provenance, and campaign script revisions. If these items are documented prior to final upload, the manuscript transitions from a strong narrative to a reproducible scientific record. Given the pace of campaign updates, this hardening step is likely the highest leverage action for preserving trust as numeric values evolve during ongoing reruns.

10.7 Suggested Revision Milestones

To operationalize the above program, we propose milestone-gated revision targets:

1. Milestone R1: pass_both increased to at least 5 with refreshed confidence interval reporting.
2. Milestone R2: benchmark matrix completed with model and workflow deltas.
3. Milestone R3: uncertainty-aware triage analysis included.
4. Milestone R4: reproducibility manifest and finalized references integrated.

These milestones enable phased strengthening without destabilizing the manuscript core. They also provide a clear response framework if reviewers request deeper evidence during peer review.

10.8 Expected Outcome of the Upgrade Path

If the milestone program is executed, the manuscript will move from a feasibility-dominant evidence profile to a quantitatively reinforced profile with narrower uncertainty and clearer benchmark context. The key expected changes are: higher validated-depth confidence, more defensible comparative claims, and stronger reproducibility traceability. In turn, reviewer concerns are likely to shift from "insufficient evidence breadth" toward higher-level questions on generalization and transferability, which are more favorable discussion spaces for a methods paper. This transition is precisely the purpose of the upgrade plan: not to alter the core scientific claim, but to support that claim with a denser and more statistically interpretable evidence base.

Declarations

CRediT authorship contribution statement

Sunwoo Lee: Conceptualization, methodology, software, formal analysis, data curation, writing - original draft, writing - review and editing, visualization.

Conflict of interest

The author declares no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

The author acknowledges The Ohio State University for institutional support and local CPU/GPU workstation resources used for model training and DFT workflow execution.

Data and code availability

All manuscript-facing artifacts used for this submission (snapshot JSON, candidate tables, status tables, figures, and manuscript-generation scripts) are provided in the submission bundle and can be regenerated from the included scripts under docs/jmst_submission. MPContribs synchronization outputs and campaign-level parsed tables used by this manuscript are included as structured exports in the same bundle.

Data statement

The datasets generated and analyzed during this study are available as submission-attached structured artifacts (including resultssnapshot.json, statusbyset.csv, tablevalidatedcandidatesclean.csv, and related generation scripts). Additional persistent public identifiers can be provided in a post-acceptance repository release.

Appendix A. Figures and Tables Planned for Camera-Ready Submission

A.1 Core Figures

- Figure 1: Overall Aim Materials workflow from generation to DFT stages.

- Figure 2: Model architecture and physically constrained elastic pathway.
- Figure 3: Stage-status distribution by element-count set.
- Figure 4: Elastic fit RMS distribution and pass and fail overlay for validated candidates.
- Figure 5: Failure taxonomy and rerun-priority map.

A.2 Core Tables

- Table 1: Training and model hyperparameters.
- Table 2: DFT protocol settings (pseudopotentials, cutoffs, k-point policy, convergence criteria).
- Table 3: De-duplicated validated elastic-ready candidates with mechanical properties.
- Table 4: Ablation comparison summary and strict-pass deltas.
- Table 5: High-priority non-converged candidates and retry metadata.

A.3 Optional Supplementary Items

- Supplementary S1: Full candidate-level status export.
- Supplementary S2: Reproducibility command list and environment lock information.
- Supplementary S3: Additional case studies for top-ranked unresolved candidates after reruns.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used OpenAI Codex (GPT-5 based coding assistance) for drafting support, language refinement, and manuscript artifact automation. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

[1] A. Jain et al., Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials 1(1) (2013) 011002. https://doi.org/10.1063/1.4812323

[2] S.P. Ong et al., Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Computational Materials Science 68 (2013) 314--319. https://doi.org/10.1016/j.commatsci.2012.10.028

[3] P. Giannozzi et al., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, Journal of Physics: Condensed Matter 21(39) (2009) 395502. https://doi.org/10.1088/0953-8984/21/39/395502

[4] P. Giannozzi et al., Advanced capabilities for materials modelling with Quantum ESPRESSO, Journal of Physics: Condensed Matter 29(46) (2017) 465901. https://doi.org/10.1088/1361-648X/aa8f79

[5] P. Giannozzi et al., Quantum ESPRESSO toward the exascale, The Journal of Chemical Physics 152(15) (2020) 154105. https://doi.org/10.1063/5.0005082

[6] V.G. Satorras et al., E(n) Equivariant Graph Neural Networks, Proceedings of the 38th International Conference on Machine Learning, pp. 9323--9332, 2021.

[7] T. Xie et al., Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, Physical Review Letters 120(14) (2018) 145301. https://doi.org/10.1103/physrevlett.120.145301

[8] C. Chen et al., Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals, Chemistry of Materials 31(9) (2019) 3564--3572. https://doi.org/10.1021/acs.chemmater.9b01294

[9] B. Deng et al., CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nature Machine Intelligence 5(9) (2023) 1031--1041. https://doi.org/10.1038/s42256-023-00716-3

[10] A. Dunn et al., Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm, npj Computational Materials 6(1) (2020) 138. https://doi.org/10.1038/s41524-020-00406-3

[11] P. Huck et al., User applications driven by the community contribution framework MPContribs in the Materials Project, Concurrency and Computation: Practice and Experience 28(7) (2015) 1982--1993. https://doi.org/10.1002/cpe.3698

[12] S. Batzner et al., E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nature Communications 13(1) (2022) 2453. https://doi.org/10.1038/s41467-022-29939-5

[13] H.J. Monkhorst et al., Special points for Brillouin-zone integrations, Physical Review B 13(12) (1976) 5188--5192. https://doi.org/10.1103/physrevb.13.5188

[14] J.P. Perdew et al., Generalized Gradient Approximation Made Simple, Physical Review Letters 77(18) (1996) 3865--3868. https://doi.org/10.1103/physrevlett.77.3865

[15] A. Hjorth Larsen et al., The atomic simulation environment -- a Python library for working with atoms, Journal of Physics: Condensed Matter 29(27) (2017) 273002. https://doi.org/10.1088/1361-648x/aa680e

[16] A.M. Ganose et al., Atomate2: modular workflows for materials science, Digital Discovery 4(7) (2025) 1944--1973. https://doi.org/10.1039/d5dd00019j