Skepthical review: *A Two-Stage Classification Pipeline for Discovering Thermodynamically Stable and Mechanically Robust ABO_3 Perovskites*

Summary

This manuscript proposes a two-stage, classification-first machine-learning pipeline to screen ABO$3 perovskites for (i) thermodynamic stability and (ii) mechanically "viable/robust" behavior when high-throughput DFT elastic data are sparse and often contaminated by unphysical values. Using a dataset of 1283 computed perovskites, Stage 1 trains a GrE (Sec. 2.3, Sec. 3.2). A Gaussian Process Regressor (GPR) trained on the "clean" elastic subset provides predictive variances that are used as an uncertainty discount in downstream candidate ranking (Sec. 2.3, Sec. 2.5, Sec. 3.5). The paper further includes a hurdle model for band gaps (metallicity classifier + log-gap regressor; Sec. 2.4, Sec. 3.3) and SHAP analyses for interpretability (Sec. 3.4), and reports a Pareto front of 16 screened candidates among 1068 "uncharacterized" compositions (Sec. 3.5).} = 0 vs > 0), evaluated with Leave-One-Cluster-Out (LOCO) cross-validation grouped by A-site element clusters (Sec. 2.1–2.2, Sec. 3.1). Stage 2 reframes unreliable elasticity regression as a mechanical-viability classification problem: a GBC is trained on a small curated subset labeled via heuristic physical filters on $K_{\rm VRH}$ and $G_{\rm VRH}

The core idea—explicitly treating elastic-property artifacts as a first-class modeling problem and using uncertainty-aware screening rather than brittle regression—is compelling and practically relevant. However, several methodological and reporting gaps limit confidence in the mechanical stage and the robustness/reproducibility of the final candidate list: the mechanical viability label is built from very few negative examples (8), its construct validity is unclear (artifact detection vs true instability), probability/uncertainty calibration is not assessed despite probabilities being used as objectives, the Pareto-front extraction and uncertainty penalty are under-specified and potentially sensitive, and key data-provenance and implementation details are missing (Sec. 2–3). Addressing these would substantially strengthen the paper's reliability and impact.

Strengths

- Well-motivated framing around a real bottleneck in HT-DFT datasets: elastic tensors/moduli are often sparse and unreliable, and naïve regression can be misleading (Introduction, Sec. 2.3).
- Conceptually strong "classification-first" mechanical screening strategy that pragmatically avoids over-interpreting noisy $K_{\rm VRH}/G_{\rm VRH}$ targets (Sec. 2.3, Sec. 3.2).
- Generalization-aware validation for thermodynamic stability via LOCO grouped by A-site element, going beyond random splits that can leak chemical similarity (Sec. 2.2, Sec. 3.1).
- Uncertainty-aware screening concept (discounting uncertain predictions) is directionally appropriate for discovery/ranking settings where reliability matters (Sec. 2.5, Sec. 3.5).
- Interpretability via SHAP yields feature attributions that are broadly physically plausible (tolerance-/octahedral-factor strains for stability; density/volume/formation energy and structural descriptors for mechanical viability), providing sanity checks against pure overfitting (Sec. 3.4).
- Clear high-level pipeline narrative and effective use of standard visual tools (SHAP beeswarms, Pareto fronts) to communicate model behavior and candidate selection (Figs. 3–5, Sec. 3.4–3.5).

Major issues

1. **Mechanical Viability Classifier: extreme class imbalance (207 viable vs only 8 unviable) and unclear construct validity of the negative label (Sec. 2.3, Sec. 3.2).** With only 8 negatives, reported accuracy/F1 can be misleading, ROC-AUC estimates are unstable, and the model may be learning "similarity to the curated subset" or detection of calculation artifacts rather than true mechanical instability. Since this classifier and its uncertainty discount directly shape the Pareto front (Sec. 3.5), the evidential basis for the mechanical screening claims is currently weak.

   Recommendation: Strengthen Sec. 2.3 and Sec. 3.2 with (i) label provenance and (ii) imbalance-appropriate evaluation: (1) Clearly explain how the 8 unviable points were identified and, if available in the source database/workflow, distinguish failed elastic calculations (numerical issues, non-convergence) from physically unstable outcomes (e.g., violated Born criteria, negative eigenvalues of elastic tensor, phonon instabilities). (2) Report per-fold (or aggregated) confusion matrices plus class-wise precision/recall/F1, emphasizing recall on the unviable class; include precision–recall curves or average precision for the minority class. (3) Add robustness/sensitivity analysis suited to tiny negatives: leave-one-negative-out experiments and/or bootstrapped confidence intervals for metrics; show whether model behavior depends on one or two negative points. (4) If possible, expand the negative set systematically (e.g., by using elastic-tensor validity checks, Born stability criteria where tensors exist, flagged failed calculations, or known unstable phases) and re-train/re-evaluate. (5) If expansion is not possible, temper claims in Sec. 4: present the viability classifier explicitly as a preliminary/conservative artifact/instability filter rather than a comprehensive detector of mechanical failure.

2. **Data provenance and curation are under-specified for both the 1283-compound dataset and the 1068 "uncharacterized" screening set (Sec. 2.1, Sec. 2.3, Sec. 2.5).** Given the paper's emphasis on imperfect HT-DFT data, readers need to know which database/workflow produced energies, structures, tilts, and elastic quantities; how duplicates/polymorphs were handled; and how perovskite topology/Glazer tilt systems were assigned. Similarly, the elastic filtering thresholds (e.g., $0 < K_{\rm VRH} < 300$ GPa, $G_{\rm VRH} > 0$) and their impact on retained/excluded points are not justified quantitatively.

Recommendation: Expand Sec. 2.1, Sec. 2.3, and Sec. 2.5 (or add a dedicated data/curation subsection) to include: (1) explicit source database(s) and citations; key DFT workflow details relevant to energies and elastics (functional, pseudopotentials, k-point density, cutoffs, relaxation/stress criteria; or a pointer to the workflow documentation). (2) How duplicates/polymorphs were treated (e.g., lowest E_{hull} retained; how multiple structures per composition are handled). (3) The algorithm/rules mapping space group (or structure) to Glazer tilt labels, and treatment of ambiguous/unknown tilt assignments. (4) For elastics: show distributions/histograms of K_{VRH} and G_{VRH} and report how many points are excluded by each rule (negativity vs upper bound); justify the $K_{\text{VRH}} < 300$ GPa cutoff with references or dataset-driven reasoning, and clarify whether excluded points represent artifacts vs genuine high stiffness. (5) Define precisely how the 1068 "uncharacterized" set is derived from the 1283 total (missing elastic data only? missing experimental reports? something else), and confirm all 1068 satisfy the structural perovskite criterion used in the study.

3. **Uncertainty discount and Pareto-front candidate selection are under-specified and insufficiently validated (Sec. 2.3, Sec. 2.5, Sec. 3.2, Sec. 3.5).** The GPR predictive variance is used as an epistemic uncertainty proxy and combined multiplicatively with classifier probabilities via an ad hoc min–max normalization (Eq. (1)), but the variance is not calibrated/validated (does higher variance actually imply higher error?) and min–max scaling makes the discount dependent on the candidate set and sensitive to outliers. The Pareto-front extraction procedure and any tie-breaking/secondary ranking for Table 1 are described only qualitatively, limiting reproducibility and confidence in the stability of the 16 reported candidates.

   Recommendation: In Sec. 2.3 and Sec. 2.5, fully specify and validate the uncertainty and optimization steps: (1) Provide the exact Pareto-front extraction algorithm (e.g., non-dominated sorting in the 2D objective space) and any subsequent ranking rule used to order/select the 16 materials in Table 1; include pseudocode or a concise step-by-step description. (2) Evaluate uncertainty quality for $K_{\text{VRH}}/G_{\text{VRH}}$ GPRs: calibration/coverage of prediction intervals and/or a binned plot of predicted variance vs absolute error under cross-validation (show that variance correlates with error). (3) Add sensitivity/robustness checks in Sec. 3.5: recompute fronts under bootstrap/CV variability and report how often each candidate appears; test at least one alternative discounting scheme (e.g., quantile-based scaling, clipping, a tunable penalty strength, or a lower-confidence-bound style score). (4) Explicitly define in-text the domain for min/max in Eq. (1) (screened set vs training set) and address outlier sensitivity (e.g., use robust scaling or clipping).

4. **Potential information leakage / shortcut learning in thermodynamic stability prediction due to inclusion of formation energy features when the label is E_{hull} (Sec. 2.2, Sec. 3.1, Sec. 3.4.1).** Formation energy is thermodynamically closely tied to hull construction; if it is computed from the same DFT energies used to build E_{hull} in the same database, the stability classifier may be learning a near-direct proxy rather than generalizable structure–chemistry relationships, complicating interpretability claims and transferability under LOCO.

   Recommendation: Clarify in Sec. 2.1–2.2 the provenance of formation energy (is it computed from the same total energies and reference states used for the convex hull in the same dataset?). Then add an ablation study in Sec. 3.1: report LOCO performance with and without formation energy (and optionally other highly correlated global energetics features), and comment on how SHAP attributions change. If performance drops substantially, frame conclusions accordingly (the model is powerful but partly leverages database-specific thermodynamic bookkeeping); if performance remains strong, that strengthens claims of learned transferable descriptors.

5. **Probabilities are used directly as screening objectives (stability probability, viability probability) but probability calibration is not assessed (Sec. 2.2, Sec. 2.5, Sec. 3.1, Sec. 3.2).** Miscalibrated probabilities can distort the Pareto geometry and systematically favor certain chemistries (e.g., specific A-site clusters) even if ranking metrics like AUC look acceptable.

   Recommendation: Add calibration diagnostics for both classifiers in Sec. 3.1 and Sec. 3.2: reliability diagrams and summary metrics (Brier score, ECE), ideally also stratified by LOCO fold/A-site cluster to detect chemistry-dependent miscalibration. If miscalibration is non-trivial, apply Platt scaling or isotonic regression (trained properly within the CV scheme) and use calibrated probabilities in Sec. 2.5 screening; report how calibration changes the Pareto front and Table 1 candidates (Sec. 3.5).

6. **Reproducibility/implementation detail is incomplete across modeling stages (Sec. 2.1–2.4, Sec. 3.3).** Several choices materially affect results but are not fully specified: feature preprocessing (scaling/transforms/imputation), categorical encoding (tilt systems/space group/crystal system), hyperparameters and tuning for GBC/GBR/GPR and the hurdle model, class weighting/resampling for imbalance, train/test split protocol and seeds, and how VRH moduli were computed from elastic tensors (if applicable).

   Recommendation: Add a concise but complete implementation subsection (new Sec. 2.6 or expanded Sec. 2.1–2.4) listing: (1) the full feature list and preprocessing steps (scaling, log transforms, missing-value handling); (2) categorical encodings for structural/tilt descriptors and how unknown/ambiguous labels are treated; (3) model hyperparameters and tuning strategy (grid/random/Bayesian search; objective; CV setup), including any class weights; (4) exact split protocols and random seeds for each stage (including hurdle model); and (5) the procedure/software used to compute $K_{\text{VRH}}/G_{\text{VRH}}$ from elastic tensors (or the database field definitions). Provide code and/or processed datasets if possible, or at minimum enough detail to reproduce Table 1.

Minor issues

1. Definition of thermodynamic stability as strictly $E_{\text{hull}} = 0$ eV/atom may be overly restrictive and differs from common screening practice that includes small metastability windows (Sec. 2.1, Sec. 2.2, Sec. 3.1). This choice changes class balance and may exclude experimentally accessible perovskites.

   Recommendation: In Sec. 2.1–2.2, justify $E_{\text{hull}} = 0$ with references and show the E_{hull} distribution (including fractions within e.g. 10, 25, 50, 100 meV/atom). Add a sensitivity analysis in Sec. 3.1 using an alternative label (e.g., $E_{\text{hull}} < 50$ meV/atom) and report effects on performance and candidate selection; if keeping $E_{\text{hull}} = 0$ as primary, explicitly acknowledge the conservativeness in Sec. 4.

2. LOCO grouping only by A-site element may not fully test compositional generalization because B-site chemistry/oxidation and A–B pairing can dominate stability and mechanical behavior; cluster sizes may also vary substantially (Sec. 2.2, Sec. 3.1).

   Recommendation: Report the distribution of LOCO fold sizes (and positive/negative counts) across the 53 A-site clusters in Sec. 3.1 or Supplementary Information. Briefly discuss limitations of A-only grouping and, if feasible, add a supplementary comparison to at least one alternative grouping (by B-site or by (A,B) pair) to show sensitivity.

3. Screening results need clearer positioning regarding novelty and the incremental value of the second-stage mechanical filter (Sec. 3.5). Several highlighted candidates (e.g., $DyVO_3$, $YCrO_3$) are well-studied; it is unclear what "uncharacterized" means and what baseline (stability-only) would return.

   Recommendation: In Sec. 2.5 and Sec. 3.5, define "uncharacterized" precisely (missing elastic constants only? not in the database? not experimentally reported?). Add a baseline shortlist using stability probability alone (or stability + simple heuristic mechanical filters) and compare overlap with the final Pareto set to quantify how much the mechanical/uncertainty stage changes selection.

4. The role of the hurdle band-gap model is not clearly integrated into the main discovery pipeline (Sec. 2.4, Sec. 3.3, Sec. 3.5, Sec. 4). As written, it reads as disconnected from the final multi-objective screening and Table 1.

   Recommendation: Either integrate the band-gap outputs into screening (as an additional objective/constraint with an application-motivated target range) and reflect this in Sec. 3.5/Table 1, or explicitly reframe it in Sec. 4 as an extensibility demonstration. In Sec. 3.3, add error breakdown by gap ranges and report the effect of metallicity misclassification on MAE.

5. Interpretability section would benefit from more quantitative and less causal phrasing (Sec. 3.4.1–3.4.2). Current discussion sometimes blurs correlation vs causation (e.g., density/formation energy 'indicating' robustness) and does not report variability across folds/background choices for SHAP.

   Recommendation: In Sec. 3.4, add quantitative summaries (e.g., mean $|\mathbf{SHAP}|$ for top features, and rank stability across CV folds), specify SHAP computation details (background dataset, sample size, SHAP variant) in captions, and adjust language to "correlates/associates with" rather than causal claims. Consider adding a few dependence plots in Supplementary Information to illustrate interactions.

6. Figures and tables have several clarity gaps that affect actionability of results: overplotting/metric incompleteness for the band-gap figure; SHAP plot labeling/legibility; and Table 1 missing/unclear uncertainty values (Sec. 3.3–3.5; Figs. 2–5; Table 1).

   Recommendation: Improve Fig. 2 using hexbin/alpha and report both classifier and regressor metrics (and gap-range breakdown). For Figs. 3–4, clarify color semantics for categorical/binary features and add units/transform notes to labels. Populate Table 1's uncertainty-related column with the actual numeric quantity used (e.g., normalized variance or discount factor) and define it in the caption and Sec. 2.5.

7. Eq. (1) is not fully well-posed as written: the reference set for min/max is unspecified, and division-by-zero / out-of-range cases are not addressed (Sec. 2.5 around Eq. (1)).

   Recommendation: In Sec. 2.5, state explicitly whether min/max are computed over the 1068 screened candidates (or another fixed reference), add an epsilon safeguard for max=min, and clip the normalized term to $[\mathbf{0}, \mathbf{1}]$. Also consider robust scaling (quantiles) if min–max is retained.

8. Mechanical-stage terminology and counts are potentially confusing (207 filtered 'training subset' vs 215 labeled set including 8 unviable) (Sec. 2.3).

   Recommendation: Standardize terminology in Sec. 2.3: explicitly define the 215-point labeled set $= \mathbf{207}$ viable $+\mathbf{8}$ unviable, and clarify that GPR is trained only on the 207 viable points while the classifier uses all 215.

9. Sustainability/safety/supply-risk aspects of shortlisted chemistries (e.g., Cr, Rh, heavy rare earths) are not discussed (Sec. 3.5, Sec. 4).

   Recommendation: Add a brief note in Sec. 4 acknowledging that the optimization currently ignores toxicity/criticality and that such constraints could be incorporated as additional objectives/filters in future work.

Very minor issues

1. Minor typographical/formatting inconsistencies (line-break artifacts, inconsistent hyphenation, unit spacing, section heading formatting; punctuation around Eq. (1)) (Sec. 1, Sec. 2.5, Sec. 3.2–3.5).

   Recommendation: Proofread and standardize LaTeX formatting: consistent section numbering/markup, hyphenation ("high-throughput"), unit formatting ("$\mathbf{0\,eV}$"), and punctuation around displayed equations.

2. Acronyms/notation and feature-name definitions are occasionally introduced late or inconsistently (e.g., K/G vs $K_{\mathrm{VRH}}/G_{\mathrm{VRH}}$; VRH/LOCO; 'en_diff'; variance bar notation around Eq. (1)) (Sec. 2.1–2.3, Sec. 3.4).

   Recommendation: Add a short notation pass: define acronyms at first use in each major section, expand feature-name abbreviations on first appearance, and standardize variance notation (e.g., $\bar{\sigma}^2(i)$) consistently in Eq. (1) and surrounding text.

3. Hurdle-model log-transform details are not fully explicit (log base; inverse transform; dimensionful argument) (Sec. 2.4).

   Recommendation: In Sec. 2.4, specify the log base and write the explicit inverse mapping (e.g., $\mathbf{gap} = \exp(\hat{y}) - 1$ for natural log). Optionally note the implicit scaling that makes the argument dimensionless.

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper is predominantly methodological/ML and contains limited formal mathematics (one explicit equation for an uncertainty-penalized score plus a handful of feature/label definitions and inequality-based physical filters). The main analytic check is Eq. (1)'s normalization and bounding behavior, along with consistency of symbols (probabilities, variances, and engineered geometric-strain features).

Checked items

1. ✔ **Geometric strain feature definitions** (Sec. 2.1, p.3 (feature engineering))
   - **Claim:** Defines tolerance-factor strain and octahedral-factor strain as absolute deviations from ideal values: $\tau_{\text{strain}} = |\tau - \mathbf{1.0}|$ and $\mu_{\text{strain}} = |\mu - \mathbf{0.57}|$.
   - **Checks:** symbol/definition consistency, dimensional/units sanity
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** τ and μ are dimensionless geometric factors., Absolute value is the standard scalar absolute value.
   - **Notes:** Both strain features are dimensionless and nonnegative by construction; notation is consistent within Sec. 2.1 and later interpretability discussion (Sec. 3.4.1).
2. ✔ **Thermodynamic stability label definition** (Sec. 2.1–2.2, p.3)
   - **Claim:** Treats stability as a binary label: stable iff $E_{\text{hull}} = \mathbf{0\ eV/atom}$; metastable iff $E_{\text{hull}} > \mathbf{0\ eV/atom}$.
   - **Checks:** definition consistency, units sanity
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** E_{hull} is reported in $\mathbf{eV/atom}$ and is nonnegative in the dataset as used.
   - **Notes:** Internally consistent as stated. The paper does not mention any numerical tolerance around zero; that is an implementation choice, not an algebraic inconsistency.
3. ✔ **Mechanical-property physical filter** (Sec. 2.3, p.3)
   - **Claim:** Defines a physically filtered subset via $\mathbf{0} < K_{\text{VRH}} < \mathbf{300\ GPa}$ and $G_{\text{VRH}} > \mathbf{0}$.
   - **Checks:** inequality logic, units/dimensional sanity
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** K_{VRH} and G_{VRH} are expressed in $\mathbf{GPa}$.
   - **Notes:** Inequalities are well-formed and unit-consistent (upper bound includes unit $\mathbf{GPa}$).
4. ✔ **Consistency of mechanical subset sizes (207 vs 215)** (Sec. 2.3, p.3–4 and Sec. 3.2, p.6)
   - **Claim:** Uses 207 physically consistent samples and 8 unphysical/unstable samples (total 215) for the viability classifier; uses the 207 filtered samples for GPR.
   - **Checks:** internal consistency of stated sets, symbol/definition consistency
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** The 'full 215-sample subset' comprises the 207 passing the filter plus 8 failing it.
   - **Notes:** Counts reconcile ($\mathbf{207 + 8 = 215}$) and are described consistently in Sec. 3.2. Wording in Sec. 2.3 could be clearer but does not create a contradiction.
5. ⚠ **Predictive variance used as uncertainty** (Sec. 2.3, p.4)
   - **Claim:** Defines GPR predictive variance σ^2 as the uncertainty measure; later uses an averaged predictive variance $\bar{\sigma}^2(i)$ across the two modulus models (K_{VRH} and G_{VRH}).
   - **Checks:** symbol/definition consistency
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
   - **Assumptions/inputs:** Each GPR provides a predictive variance for its target.
   - **Notes:** The concept is consistent, but the paper does not specify exactly how $\bar{\sigma}^2(i)$ is computed from the two variances (mean, weighted mean, sum, etc.), which is needed for a fully specified analytic definition.
6. ⚠ **Uncertainty-penalized viability score** (Eq. (1), Sec. 2.5, p.5)
   - **Claim:** Defines $S_{\text{penalized}}(i) = P_{\text{viability}}(i) \times \left[1 - \frac{\bar{\sigma}^2(i) - \min(\bar{\sigma}^2)}{\max(\bar{\sigma}^2) - \min(\bar{\sigma}^2)}\right]$ to down-weight high-uncertainty candidates.
   - **Checks:** algebraic form, range/bounding sanity, well-posedness
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** $P_{\text{viability}}(i) \in [\mathbf{0}, \mathbf{1}]$. $\mathbf{min}(\bar{\sigma}^2)$ and $\mathbf{max}(\bar{\sigma}^2)$ are computed over a set that contains $\bar{\sigma}^2(i)$ for each candidate being scored. $\mathbf{max}(\bar{\sigma}^2) > \mathbf{min}(\bar{\sigma}^2)$.
   - **Notes:** Algebra matches a standard min–max normalization followed by $\mathbf{1 - normalized}$. If $\bar{\sigma}^2(i)$ lies within $[\mathbf{min}, \mathbf{max}]$, the factor lies in $[\mathbf{0}, \mathbf{1}]$ and $S_{\text{penalized}}(i) \in [\mathbf{0}, \mathbf{1}]$. However, the text does not define the reference set for min/max, does not address the degenerate case $\mathbf{max} = \mathbf{min}$, and does not state any clipping if values fall outside range (e.g., when reusing min/max across different sets). Notation around overbars/superscripts is slightly inconsistent typographically.

7. ✔ **Consistency between Eq. (1) and Table 1 penalty interpretation** (Eq. (1), p.5 and Table 1, p.11)
    - **Claim:** Table 1's 'Uncertainty Penalty' values correspond to the normalized variance term, and Penalized Viability equals Unpenalized Viability × (1 − penalty).
    - **Checks:** symbol/definition consistency, algebraic relationship sanity
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** The 'Uncertainty Penalty' column is $(\bar{\sigma}^2(i) - \min)/(\max - \min)$.
    - **Notes:** The naming aligns with Eq. (1): penalty≈ 0.201 implies multiplier≈ 0.799, consistent with Penalized Viability≈ 0.9999 given Unpenalized Viability≈ 0.9999 (qualitatively consistent without reproducing numerics).
8. ⚠ **Hurdle model transform and back-transform** (Sec. 2.4, p.4 and Sec. 3.3, p.7)
    - **Claim:** Uses a classifier for metal vs non-metal; for non-metals regresses y = \log(1 + \text{band_gap}) and converts back to band gap in eV.
    - **Checks:** definition completeness, units/dimensional sanity
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** log denotes a specific base (not stated). Back-transform is consistent with the chosen log base (e.g., $\exp(\hat{y}) - 1$ for natural log).
    - **Notes:** The logical structure is consistent, but the inverse mapping is not written explicitly and the log base is not specified. Also, \log(1 + \text{band_gap}) uses a dimensionful band gap (in eV), which is a mild dimensional-formalism issue unless an implicit normalization is stated.

Limitations

- The provided PDF contains very few explicit equations/derivations; most methods are described verbally, limiting the scope of algebraic verification.
- Key domain-specific quantities (e.g., explicit formulas for Goldschmidt tolerance factor τ and octahedral factor μ) are referenced but not defined in the document, so they cannot be audited symbolically here.
- No detailed derivations are shown for the uncertainty variance aggregation ($\bar{\sigma}^2$) or the multi-objective/Pareto procedure, so only definitional consistency can be checked.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

All 14 automated numeric consistency checks passed, including: (i) no-skill baseline matching the stated stable fraction ($0.131 = 13.1\%$), (ii) implied counts from percentages being close to integers for $N = 1283$, (iii) exact subset and LOCO fold identities ($207 + 8 = 215$; $52 + 1 = 53$), (iv) algebraic range sanity for the penalty normalization term, (v) Table 1 penalized-viability recomputations matching within tight tolerances, and (vi) Table 1 rank order matching the stated stability+penalized-viability scoring rule.

Checked items

1. ✔ **C1** (Page 3 (Methods 2.2) and Page 5/6 (Results 3.1; Figure 1 caption))
    - **Claim:** No-skill baseline of 0.131 corresponds to 13.1% frequency of stable compounds in the dataset.
    - **Checks:** percent_to_fraction_consistency
    - **Verdict:** PASS
    - **Notes:** Computed $13.1/100 = 0.131$ and matched the reported baseline.
2. ✔ **C2** (Page 3 (Methods 2.2) and Page 5 (Results 3.1))
    - **Claim:** Only 13.1% of the 1283 compounds are stable.
    - **Checks:** count_from_percentage
    - **Verdict:** PASS
    - **Notes:** Implied stable count = $1283 \times 0.131 = 168.073$, which is 0.073 away from the nearest integer (168).
3. ✔ **C3** (Page 3 (Methods 2.3) and Page 6 (Results 3.2))
    - **Claim:** Elastic constants available for only 16.8% of compounds in dataset of 1283.
    - **Checks:** count_from_percentage
    - **Verdict:** PASS
    - **Notes:** Implied elastic-constants count = $1283 \times 0.168 = 215.544$, which is 0.456 away from the nearest integer (216).
4. ✔ **C4** (Page 3 (Methods 2.3) and Page 4/6 (Methods/Results 3.2))
    - **Claim:** Mechanical subset sizes: 'physically filtered subset of 207 materials' and 'full 215-sample subset' with '8 unphysical or unstable ones'.
    - **Checks:** parts_sum_to_total
    - **Verdict:** PASS
    - **Notes:** Verified $207 + 8 = 215$ exactly.

5. ✔ **C5** (Page 3 (Methods 2.3))
    - **Claim:** Physical filter criteria are $0 < K_{\text{VRH}} < 300$ GPa and $G_{\text{VRH}} > 0$; verify inequality bounds are internally consistent (strict) and not contradictory.
    - **Checks:** inequality_sanity_check
    - **Verdict:** PASS
    - **Notes:** Strict bounds are feasible: $0 < K_{\text{VRH}} < 300$ has a non-empty interval; $G_{\text{VRH}} > 0$ is also feasible.
6. ✔ **C6** (Page 5 (Eq. 1 in Methods 2.5))
    - **Claim:** Penalty normalization term: $(\bar{\sigma}^2(i) - \min(\bar{\sigma}^2))/(\max(\bar{\sigma}^2) - \min(\bar{\sigma}^2))$ should be in $[0, 1]$ when $\bar{\sigma}^2(i)$ is within $[\min, \max]$; thus multiplicative factor $(1 - ...)$ should be in $[0, 1]$.
    - **Checks:** formula_range_check
    - **Verdict:** PASS
    - **Notes:** Synthetic unit tests ($\min = 2$, $\max = 5$; $\sigma = 2, 3.5, 5$) confirmed normalized term $\in [0, 1]$, $(1 - \text{normalized}) \in [0, 1]$, and $S_{\text{penalized}} \leq P_{\text{viability}}$.
7. ✔ **C7** (Page 11 (Table 1))
    - **Claim:** Penalized viability should equal Unpenalized viability probability multiplied by $(1 - \text{uncertainty penalty})$.
    - **Checks:** recompute_from_table_formula
    - **Verdict:** PASS
    - **Notes:** Recomputed $0.9999 \times (1 - 0.201) = 0.7989201$ vs reported 0.799 (rounding-consistent).
8. ✔ **C8** (Page 11 (Table 1))
    - **Claim:** Penalized viability should equal Unpenalized viability probability multiplied by $(1 - \text{uncertainty penalty})$ (Rank 2).
    - **Checks:** recompute_from_table_formula
    - **Verdict:** PASS
    - **Notes:** Recomputed $0.9999 \times (1 - 0.187) = 0.8129187$ vs reported 0.813 (rounding-consistent).
9. ✔ **C9** (Page 11 (Table 1))
    - **Claim:** Penalized viability should equal Unpenalized viability probability multiplied by $(1 - \text{uncertainty penalty})$ (Rank 3).
    - **Checks:** recompute_from_table_formula
    - **Verdict:** PASS
    - **Notes:** Recomputed $0.9999 \times (1 - 0.187) = 0.8129187$ vs reported 0.813 (rounding-consistent).
10. ✔ **C10** (Page 11 (Table 1))
    - **Claim:** Penalized viability should equal Unpenalized viability probability multiplied by $(1 - \text{uncertainty penalty})$ (Rank 4).
    - **Checks:** recompute_from_table_formula
    - **Verdict:** PASS
    - **Notes:** Recomputed $0.9999 \times (1 - 0.312) = 0.6879312$ vs reported 0.688 (rounding-consistent).
11. ✔ **C11** (Page 11 (Table 1))
    - **Claim:** Penalized viability should equal Unpenalized viability probability multiplied by $(1 - \text{uncertainty penalty})$ (Rank 5).
    - **Checks:** recompute_from_table_formula
    - **Verdict:** PASS
    - **Notes:** Recomputed $0.9999 \times (1 - 0.387) = 0.6129387$ vs reported 0.613 (rounding-consistent).
12. ✔ **C12** (Page 11 (Table 1 caption + table columns))
    - **Claim:** Ranking criterion: 'ranked by the sum of their predicted thermodynamic stability probability and penalized mechanical viability score' should match the listed rank order for top 5.
    - **Checks:** ranking_by_computed_score
    - **Verdict:** PASS
    - **Notes:** Computed sums (stability+penalized viability): rank1=1.520; rank2=1.506; rank3=1.452; rank4=1.431; rank5=1.357, confirming non-increasing order.
13. ✔ **C13** (Page 3 (Methods 2.2))
    - **Claim:** LOCO CV: data partitioned into 53 clusters; during each fold trained on 52 clusters and validated on 1 held-out cluster.
    - **Checks:** fold_count_consistency
    - **Verdict:** PASS
    - **Notes:** Verified $52 + 1 = 53$.
14. ✔ **C14** (Page 4 (Methods 2.4) and Page 7 (Results 3.3))
    - **Claim:** Metals fraction: 48.6% of materials are metals (band gap = 0 eV).

- **Checks:** complement_percentage
- **Verdict:** PASS
- **Notes:** Computed complement: non-metals = $100 - 48.6 = 51.4\%$.

Limitations

- Only parsed text content was available; no underlying datasets, fold-wise results, or prediction outputs are included, preventing verification of reported ML metrics and Pareto-optimality claims.
- Figures are present as images; per instructions, no numeric extraction from plots/pixels was attempted, limiting checks to numbers explicitly stated in text/tables.
- Some statements (e.g., penalty normalization using min/max variance) depend on hidden intermediate quantities not reported in the PDF, so only algebraic sanity checks or table-level recomputations are feasible.