PHYSICAL SURVIVAL AND PROVENANCE LIMITS ON WATERMARKING AI-GENERATED CRYSTALS

Anonymous authors
Paper under double-blind review

ABSTRACT

When crystals are generated, they are relaxed and standardised before scientific use. However, this can erase or alter structural evidence of their production. We ask what a structure-only watermark can preserve through this processing, how an auditor should calibrate the surviving evidence, and which provenance claims the resulting decision supports. BASINMARK connects the deposited observation to a collection audit by selecting generated candidates according to keyed structural fingerprints. Exact relaxation to the same basin minimum erases coordinate marks, so the protocol instead selects among alternatives whose fingerprint distinctions may survive deposition. Across the evaluated generators and interatomic potentials, cell-shape bins account for much of the observed fingerprint instability; removing them in two generator populations substantially improves fingerprint-label survival while preserving structural diversity. To ensure rigorous statistical verification, under a key-independent null, counting each unique fingerprint class once provides exact false-positive control. However, because detection operates at the structural class level, collections with identical fingerprints remain indistinguishable—allowing watermark evidence to be copied or reused across datasets. Ultimately, physical processing, fingerprint choices, and structural duplicate handling determine how much watermark evidence actually survives.

1 INTRODUCTION

Generative models propose candidate crystals that undergo relaxation and property screening before scientific use (Zeni et al., 2025). As these structures enter real-world discovery workflows, tracking their production history becomes crucial for reliable reuse (Wilkinson et al., 2016), alongside standard database duplicate checks (Widdowson & Kurlin, 2026) and stability benchmarks (Betala et al., 2025). However, proving provenance introduces a distinct challenge: evidence linking a usable crystal to its origin must remain verifiable even after physical processing (Belhajjame et al., 2013).

In practice, discovery pipelines use learned interatomic potentials ranging from spherical-harmonic architectures like MACE (Batatia et al., 2022) to Clifford algebra-equivariant frameworks like CliffordIP (Polat et al., 2026) or density functional theory (DFT) to relax candidates into stable geometries (Riebesell et al., 2025), followed by symmetry analysis to standardize their crystallographic settings (Togo et al., 2024). Consequently, the final deposited structure is a product of both physical relaxation and crystallographic description (Riebesell et al., 2025). Just as with protein post-processing evaluations, coordinate-based watermarks must survive these refinement steps to remain detectable (Zhang et al., 2025). Because AI-generated crystals are always relaxed before downstream use, any viable watermark must be evaluated after the exact same treatment (Antunes et al., 2024).

Rather than perturbing coordinates directly, selection-based watermarking biases generated outputs during sampling. While protocols like MUSE (Fang et al., 2026) use keyed selection rules, auditors must still determine how much independent evidence selected outputs actually contain—since repeated pseudorandom queries introduce statistical dependencies that compromise false-positive control (Cai et al., 2025). In fingerprint-based schemes, identical structures share the same keyed label (Goldreich et al., 1986), fundamentally limiting the production histories an audit can resolve under the data-processing inequality (Polyanskiy & Wu, 2025). Effective structure-only auditing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

after deposition must therefore address three connected requirements: fingerprint survival through physical relaxation, statistical calibration via independent structural classes, and the exact provenance resolution of the audit metric (Polyanskiy & Wu, 2025).

Here, we introduce BASINMARK, a candidate-selection protocol for *structure-only provenance auditing after deposition*. BASINMARK allows an auditor holding a secret key to evaluate deposited structures without requiring original coordinates, generator states, or trusted side metadata. Using this framework, we examine which structural markers survive relaxation, how repeated fingerprint classes alter statistical confidence, and what source history a fingerprint-based decision can truly resolve. We evaluate these principles across MATTERGEN (Zeni et al., 2025), WYFORMER (Kazeev et al., 2025), CRYSTALFORMER (Cao et al., 2025), and WYCRYST (Zhu et al., 2024), alongside a five-potential relaxation benchmark. Our relaxation experiments uncover a key representation adjustment that dramatically improves watermark survival while preserving population-wide structural diversity. Furthermore, calibration and class-reuse experiments reveal how duplicate fingerprints impact statistical evidence and reshape the resulting provenance guarantees.

This paper presents three primary contributions:

1. *Physical survival:* We prove that exact same-basin relaxation completely erases coordinate perturbations, derive a local stability–diversity bound, and demonstrate a fingerprint representation that significantly boosts survival.
2. *Class-level calibration:* We separate total record counts from distinct structural classes and variance-equivalent sample sizes to construct an exact, deduplicated statistical audit.
3. *Provenance resolution:* We formalize when distinct crystal collections yield identical audit decisions and demonstrate the empirical limits of this resolution through class-reuse and cross-database matching.

The remainder of this paper develops the deposited observation model, candidate selection algorithm, audit calibration framework, and provenance limits. Section S1 provides formal proofs, while Section S2 details representations, dataset populations, and controls.

2 RELATED WORK

Crystal generation and the deposited observation. Generative models propose candidate crystals that undergo physical relaxation and property screening prior to scientific deployment (Zeni et al., 2025), making the relaxed, deposited structure the actual target of interest for downstream auditing (Riebesell et al., 2025). Current generative frameworks range from periodic geometry models like CDVAE (Xie et al., 2021) and DIFFCSP (Jiao et al., 2023) to symmetry-informed architectures such as WYCKOFFDIFF (Kelvinius et al., 2025) and WYFORMER. Because physical relaxation alters candidate geometries during stability assessment (Riebesell et al., 2025), provenance auditing cannot rely on raw generator outputs and must instead track structures through these post-generation refinement steps.

Representations across processing. Structure-only auditing after deposition requires distinguishing genuine physical changes from harmless crystallographic re-descriptions (Thomas et al., 2021). Tools like SWORD (Huang et al., 2026) standardize symmetry-equivalent descriptions, while periodic-set invariants such as average minimum distances (AMD) (Widdowson et al., 2020) and pointwise distance distributions (PDD) (Widdowson & Kurlin, 2026) enable detailed geometric comparisons. Crystallographic canonicalization effectively handles origin shifts and supercell expansions (Lee & Shin, 2023), but feature stability under physical relaxation remains an empirical property of the chosen representation (Antunes et al., 2024). A discrete, keyed audit therefore demands both a robust equivalence convention and a fingerprint label whose survival can be measured under target physical treatments.

Scientific-structure watermarks and selection. Watermarking scientific data generally follows two paradigms: embedding perturbations directly into structural coordinates or selecting among generated outputs. In bio- and molecular modeling, FOLDMARK (Zhang et al., 2025) embeds coordinate marks in protein structures, while MOLMARK (Hu et al., 2025) applies learnable atom-level

watermarking. Subsequent evaluations in FOLDMARK measure coordinate survival after Amber relaxation, Rosetta FastRelax, and loop remodeling, defining watermark persistence explicitly relative to physical post-processing workflows. Conversely, selection-based approaches like MUSE avoid coordinate distortions altogether by using keyed selection rules to sample from candidate batches.

Calibration and provenance resolution. Evaluating selection-based watermarks requires isolating independent units of statistical evidence (Dathathri et al., 2024). While keyed generation is well established in language models (Kirchenbauer et al., 2023; Kuditipudi et al., 2023) and image diffusion noise (Wen et al., 2023), repeated pseudorandom queries introduce complex dependencies that degrade false-positive control (Cai et al., 2025). Furthermore, cryptographic undetectable watermarks (Christ et al., 2024; Gunn et al., 2025), quality-preserving attacks (Zhang et al., 2024), and semantic forgery theoretical bounds (Müller et al., 2025) highlight that statistical detection guarantees do not automatically imply strict provenance guarantees.

Connecting watermarking to deposited crystal evidence. BASINMARK bridges watermark selection with the surviving evidence available after physical deposition. Exact energy relaxation erases coordinate marks that collapse to shared basin minima, leaving discrete fingerprint survival as the key observable. By treating distinct fingerprint classes as independent statistical units, the audit isolates valid evidence while establishing which structural histories remain indistinguishable. Physical relaxation and fingerprint aggregation jointly determine the fundamental provenance limits of deposited crystals.

3 STRUCTURE-ONLY AUDITING AFTER DEPOSITION

3.1 OBSERVATION CHANNEL AND AUDIT OBJECTIVE

A periodic crystal is completely specified by its lattice vectors, atomic species, and fractional coordinates. However, multiple descriptions can refer to the exact same physical material due to origin shifts, basis changes, or supercell expansions (Widdowson et al., 2020). We denote $x \equiv_{\mathrm{crys}} x'$ when two structural descriptions differ only by such crystallographic equivalences including lattice-basis transformations, rigid rotations, point-group operations, identical-site permutations, or finite-index cell re-descriptions.

To account for post-generation processing, we model the observation pipeline using three sequential operations: the relaxation map $\mathcal{R}_0$ maps an input geometry to its local basin minimum under fixed physical conditions; the canonicalization map $\mathcal{C}$ selects a unique representative for each crystallographic equivalence class; and the quantization map Q_ρ records structural geometry at resolution $\rho > 0$. The resulting observation channel is expressed as

$$\mathcal{T}(x) = Q_\rho(\mathcal{C}(\mathcal{R}_0(x))) \,. \tag{1}$$

Here, Equation (1) explicitly separates physical relaxation, crystallographic standardization, and numerical precision. In practice, numerical stopping criteria approximate exact relaxation. Together, these steps determine which structural features survive to reach an auditor, connecting coordinate-level modifications directly to fingerprint-based evidence (Figure 1). For any coordinate-marking function W_k, the immediate question is whether perturbing a candidate within its energy basin leaves any detectable trace after relaxation.

Theorem 1 (Exact-relaxation annihilation)**.** *Fix a deterministic relaxation map taking each basin to one minimum and a distribution of admissible clean inputs X. Suppose $W_k(x)$ remains in the same basin as x for every admissible k, x. Then $\mathcal{T}(W_k(x)) = \mathcal{T}(x)$ pointwise. For each key, the deposited laws of $W_k(X)$ and X agree, so any structure-only detector has equal true-positive and false-positive rates (FPRs) under these two hypotheses.*

The Theorem 1 establishes a fundamental boundary: coordinate perturbations confined within a single energy basin are completely erased once relaxation reaches the basin minimum. When relaxation terminates early under finite force thresholds, residual forces leave the output y at some distance from the true stationary minimum $x^\star$. For a smooth potential $V \in C^2$ whose Hessian is bounded below by $\mu > 0$ (after removing zero modes), this spatial displacement is bounded by $\|y - x^\star\|_2 \le \|\nabla V(y)\|_2/\mu$. While residual forces dictate spatial offsets, whether these offsets alter a structural label ultimately depends on the chosen fingerprint representation.

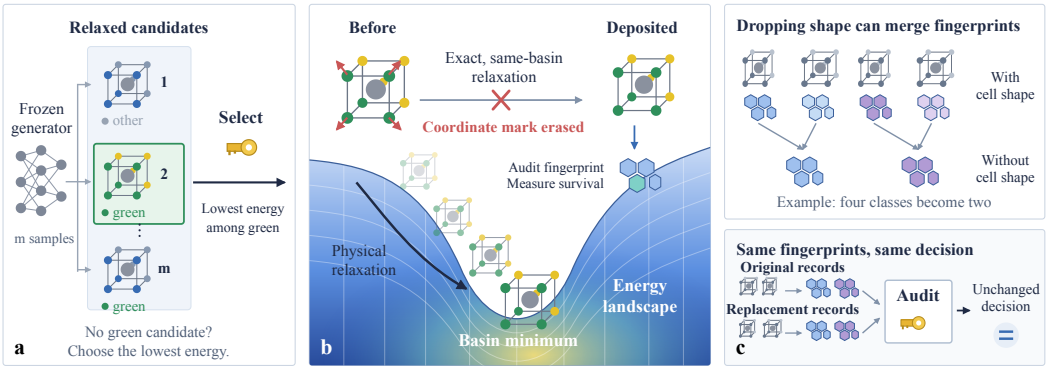

Figure 1. BasinMark: selecting crystals and interpreting their watermark evidence. **a**, The secret key makes some candidate fingerprints eligible ("green"). The producer selects the lowest-energy eligible candidate and falls back to the lowest-energy candidate when the eligible set is empty. **b**, The energy landscape illustrates convergence to a basin minimum, where exact same-basin relaxation erases coordinate perturbations. Comparing fingerprints before and after deposition measures fingerprint survival. **c**, Top: coloured polygon clusters denote distinct fingerprints; removing cell-shape bins merges four illustrated classes into two. Bottom: original and replacement collections supply the same distinct fingerprint set. Under the same key, they have identical green-class counts and audit scores, placing the audit's resolution at the shared classes.

3.2 Candidate Selection and the Fingerprint Audit

To overcome exact-relaxation annihilation, BasinMark shifts the watermarking mechanism from intra-basin coordinate perturbations to candidate selection during sampling. Rather than tracking basin occupancy, the audit operates directly on discrete structural fingerprints. The primitive-cell fingerprint Φ captures space-group symmetry, reduced chemical composition, and a sorted set of species/site-symmetry/orbit-size triples with normalized orbit dimensions, alongside an optional cell-shape block. Merging identical symmetry triples allows distinct structures to share a fingerprint, while the cell-shape block discretizes cell ratios b/a, c/a, and unit cell angles.

Candidate selection uses a keyed pseudorandom function (PRF) (Goldreich et al., 1986) that assigns "green" status whenever $\mathrm{PRF}_k(\Phi) < \gamma$. Under an ideal random-function model, distinct fingerprints map to independent uniform values in $[0, 1)$, making $\gamma \in (0, 1)$ the target green probability. During generation, the producer samples m candidates per batch, ranks eligible green candidates by energy E_i, and selects the lowest-energy candidate $i^\star$, breaking ties with a fixed rule.

Algorithm 1 BasinMark candidate selection and deduplicated class auditing.

procedure SELECT(k, m, γ)
  Generate m candidates independently of k; relax them to $x_1, \ldots, x_m$.
  Compute fingerprints $\Phi(x_i)$ and energies E_i for all candidates.
  $J \leftarrow \{i : \mathrm{PRF}_k(\Phi(x_i)) < \gamma\}$
  if $J = \varnothing$ **then**
    $J \leftarrow \{1, \ldots, m\}$
  end if
  return $x_{i^\star}$, where $i^\star \in \arg\min_{i \in J} E_i$
end procedure
procedure AUDIT(Y, k, γ, α)
  $S \leftarrow \{\Phi(y) : y \in Y\}$; $K \leftarrow |S|$ ▷ Y: deposited collection; deduplicate to unique classes
  $G_K \leftarrow \sum_{f \in S} \mathbf{1}\{\mathrm{PRF}_k(f) < \gamma\}$ ▷ count green classes
  return $\Pr\{\mathrm{Bin}(K, \gamma) \geq G_K\} \leq \alpha$ ▷ reject unmarked null
end procedure

The producer emits one structure per batch using the selection logic in Algorithm 1. Repeated over time, this yields a deposited collection Y. The auditor collects the set S of unique deposited fingerprints, counts $K = |S|$ distinct classes, and rejects the unmarked null hypothesis if the observed green class count G_K is statistically significant.

For a batch yielding d distinct fingerprints, the selection mechanism encounters d independent pseudorandom trials. The probability q_d of discovering at least one green candidate in the batch is

$$q_d = 1 - (1-\gamma)^d. \tag{2}$$

Averaging Equation (2) over the distribution of distinct candidate counts D_m gives the expected producer green rate $q = 1 - \mathbb{E}[(1-\gamma)^{D_m}]$. When candidate diversity is maximized ($D_m = m$), this simplifies to $1-(1-\gamma)^m$. Because selection queries depend solely on structural fingerprints, separate batches that generate identical fingerprints share the same pseudorandom label, coupling their audit outcomes.

For instance, with $m = 8$ and $\gamma = 1/2$, a batch of eight unique fingerprints yields a green candidate with high probability ($255/256$), whereas a redundant batch of eight identical fingerprints yields a green candidate with probability $1/2$. Structural diversity provides the marking opportunity, while energy ranking selects the optimal stable candidate. The auditor's task is then to evaluate the statistical evidence present in the distinct classes that survive deposition.

3.3 Distinct Classes Determine Null Calibration

Because fingerprint classes share pseudorandom labels across identical outputs, distinct fingerprint classes form the true independent units of statistical evidence. To analyze false-positive rates under an unmarked null, consider a collection assembled independently of the secret key. Its n total records partition into K distinct fingerprint classes with class sizes n_i and relative weights $w_i = n_i/n$. Under the ideal random-function model, class labels $B_i \sim \mathrm{Bernoulli}(\gamma)$ remain independent across distinct fingerprints.

Proposition 2 (Collision calibration)**.** *The naive green fraction $\widehat{g} = \sum_i w_i B_i$ satisfies*

$$\mathbb{E}\widehat{g} = \gamma, \qquad \mathrm{Var}(\widehat{g}) = \gamma(1-\gamma)\sum_i w_i^2 = \frac{\gamma(1-\gamma)}{n_{\mathrm{eff}}}, \qquad n_{\mathrm{eff}} = \Big(\sum_i w_i^2\Big)^{-1}.$$

After deduplication, $G_K = \sum_{i=1}^K B_i \sim \mathrm{Bin}(K,\gamma)$. Thus K is the independent-unit count of the exact deduplicated test, while n_{eff} is the variance-equivalent sample size of the naive average.

Proposition 2 highlights the risk of naive sample averaging. For example, a dataset of $n = 4$ records split into class sizes $(3,1)$ yields $K = 2$ distinct classes and a variance-equivalent sample size of $n_{\mathrm{eff}} = 1.6$. Duplicating records inflates record counts without adding fresh key queries. The relation $1 \le n_{\mathrm{eff}} \le K \le n$ holds generally, with equality $n_{\mathrm{eff}} = K$ occurring only when all classes are equally populated. Deduplicating the collection assigns exactly one vote to each unique fingerprint, recovering an exact binomial distribution $\mathrm{Bin}(K,\gamma)$ and maintaining strict false-positive control.

3.4 A Local Model of Stability and Diversity

Designing effective fingerprint representations requires balancing structural stability against class collision. If a fingerprint representation maps random inputs to class i with probability v_i, the collision diversity is defined as $D_2 = (\sum_i v_i^2)^{-1}$. Across n independent samples, the expected number of observed unique classes is $\mathbb{E}K_n = \sum_i [1-(1-v_i)^n]$. We connect this macro-level population diversity to micro-level physical stability by modeling fingerprint assignment as a geometric partition. Small relaxation displacements trigger label changes whenever a structure crosses a cell boundary in descriptor space.

Theorem 3 (Local stability–collision-diversity bound)**.** *Let a bounded measurable unit-volume region $\Omega \subset \mathbb{R}^D$ ($D \ge 2$) be partitioned up to null sets into finitely many measurable cells A_i with positive volumes v_i and finite Euclidean perimeters $P(A_i)$. Draw X uniformly on Ω and U uniformly on the unit sphere. Let $p(\varepsilon)$ denote the probability that $X + \varepsilon U$ leaves the cell of X. Defining $\kappa_D = \Gamma(D/2)/(\sqrt{\pi}\,\Gamma((D+1)/2))$, the boundary mismatch probability satisfies*

$$p(\varepsilon) = \frac{\kappa_D}{2}\,\varepsilon \sum_i P(A_i) + o(\varepsilon), \qquad \lim_{\varepsilon \downarrow 0} \frac{p(\varepsilon)}{\varepsilon} \ge \frac{\kappa_D D \omega_D^{1/D}}{2}\, D_2^{1/D}.$$

Theorem 3 establishes a fundamental structural trade-off: high class diversity (D_2) inevitably increases boundary exposure, making fine-grained fingerprints more vulnerable to relaxation shifts.

The derivation links translated-set variation (Galerne, 2011), Euclidean isoperimetry (Maggi, 2012), and convexity.

When coarsening a representation such as by removing sensitive cell-shape bins, the coarse fingerprint $\Phi_c = h \circ \Phi_f$ maps fine labels into merged structural classes. For paired structures (X, Y) evaluated before and after relaxation, the coarse mismatch probability p_c, observed class count $K_c(S)$, and collision diversity $D_{2,c}$ satisfy

$$p_c \leq p_f, \qquad K_c(S) \leq K_f(S), \qquad D_{2,c} \leq D_{2,f}. \tag{3}$$

Therefore, Equation (3) demonstrates why coarsening improves mark survival: every coarse mismatch requires a fine mismatch, but deleting unstable feature blocks drastically reduces label flips while preserving the vast majority of unique structural classes.

3.5 What a Fingerprint-Class Decision Identifies

While deduplication ensures statistical validity, it also dictates what provenance claims an audit can support. Because the detector inspects structural fingerprints rather than raw coordinate histories, it evaluates collections at class-level resolution.

Theorem 4 (Fingerprint-preserving substitution)**.** *Suppose the audit rule depends strictly on the fingerprint multiset, $A_k(x_{1:n}) = a_k(M_\Phi(x_{1:n}))$.*

(i) Pointwise equality: *Any alternative collection $\tilde{x}_{1:n}$ satisfying $M_\Phi(\tilde{x}_{1:n}) = M_\Phi(x_{1:n})$ yields identical audit decisions $A_k(\tilde{x}_{1:n}) = A_k(x_{1:n})$ under every key k.*

(ii) Distributional equivalence: *If M_Φ shares the same probability distribution under two provenance hypotheses $P_{0,k}$ and $P_{1,k}$, the audit rejection probabilities under both hypotheses are identical. This equivalence holds unconditionally for random keys drawn from a common distribution.*

The exact provenance resolution of structure-only auditing delineates with the Theorem 4. For the deduplicated test, equality of the unique fingerprint set S is sufficient to produce identical audit scores, regardless of record counts or individual record origins. Consequently, an adversary can resample or substitute structures within exposed green fingerprint classes to construct new rejecting collections without possessing the secret key. Fingerprint-based watermarking authenticates structural classes rather than individual physical records.

4 Experiments

4.1 Setup and Representation

We evaluate fingerprint survival through relaxation, statistical calibration under class repetition, and provenance resolution under class reuse using relaxation pairs, representation ablations, and generator simulations. Our benchmarks cover MatterGen, WyFormer, CrystalFormer, and WyCryst, using reference populations from MP-20 which are Materials Project (MP) crystals with at most 20 atoms per unit cell (Xie et al., 2021). Unless specified otherwise, experiments use a green-label probability $\gamma = 1/2$, candidate batch size $m = 8$, nominal rejection level $\alpha = 0.001$, and target power of 90%. We define a *flip* as an unequal pre- versus post-treatment fingerprint, and K as the count of distinct fingerprints. Pooled structure pairs capture variation across crystal systems, a shared DFT reference panel supports our five-potential comparison, and generator ablations isolate individual flip-driving components.

Across a 71-prototype metallic library and a 212-structure library spanning all seven crystal systems, the structural fingerprint successfully resists origin shifts, supercell expansions, identical-site reindexing, unimodular basis changes, and rigid rotations. Omitting primitive-cell reduction severely degrades invariance, preserving only 40.8% of labels under anisotropic supercell expansion. By requiring equivalent crystallographic descriptions to map to identical fingerprints, these baseline tests establish the convention for interpreting relaxation-induced changes.

4.2 Default Instability Concentrates in Low-Symmetry Structures

Table 1. In the pooled before/after relaxation pairs, the default fingerprint changes rarely for cubic structures but almost always for monoclinic and triclinic structures. This concentration of flips shows where the representation is most vulnerable to relaxation.

Crystal system	Pairs	Flips	Flip rate (%)
Cubic	2,578	162	6.28
Hexagonal	773	540	69.86
Tetragonal	1,935	1,568	81.03
Trigonal	1,473	1,262	85.68
Orthorhombic	2,119	1,960	92.50
Monoclinic	2,918	2,895	99.21
Triclinic	2,106	2,105	99.95
All	13,902	10,492	75.47

Physical relaxation alters the default structural fingerprint in $10{,}492$ out of $13{,}902$ pairs (75.5%). However, this aggregate rate masks strong crystal-system dependence: flip rates range from just 6.3% in rigid cubic structures to 99.95% in unconstrained triclinic systems (Table 1). This structural variance significantly exceeds the 71.8–76.9% flip range observed across five distinct interatomic potential families reweighted to the same system composition.

Breakdown analysis reveals that binned cell-shape coordinates drive the vast majority of observed instabilities. Across tested potentials, 69.6–78.0% of all fingerprint flips alter only a cell-shape bin, a proportion that rises to 82.9–88.4% among triclinic flips. Omitting the cell-shape block directly targets this primary failure mode.

4.3 Removing Cell Shape Substantially Improves Measured Stability

We compare five fingerprint settings at symmetry tolerance 0.1 Å: no cell shape and cell-bin parameters 0.30, 0.10, 0.05 (default), and 0.02. Figure 2a–b shows relaxation flips and distinct class counts in the relaxed collections. Removing cell shape reduces flip rates from 78.0% to 25.7% for MatterGen and from 67.8% to 4.0% for CrystalFormer. This removes 67.1% and 94.2% of the original flips while reducing distinct class counts by only three and zero, respectively. The endpoint comparison is consistent with Equation (3); different cell-bin widths need not define nested partitions. A quantized AMD control remains sensitive to relaxation, with flip rates of 98.7% and 98.5% for CHGNet and SevenNet at bin width 0.02.

4.4 Distinct Fingerprints Repair the Statistical Audit

Figure 2c compares naive and deduplicated audits across nine released generator collections. The largest difference occurs for WyckoffLLM-vanilla: its 999 records contain 649 distinct fingerprints, and deduplication reduces false rejections from $64/1{,}000$ to $0/1{,}000$ key trials. These are empirical estimates; Proposition 2 provides exact conditional control under the key-independent null.

Figure 2d shows model power when candidate fingerprints are independent and distinct and deposition preserves their labels. The first class counts reaching 90% power are $K = 71, 18, 10, 10$ for candidate budgets $m = 2, 4, 8, 16$. The small decreases between some adjacent K values arise from changes in the exact binomial rejection threshold.

4.5 Observed Classes Transfer Audit Decisions

Class reuse experiments test whether exposing marked outputs allows an adversary to construct a rejecting collection without knowing the key. An adversary observing N selected outputs matches their fingerprints within a candidate corpus and resamples 200 records with replacement. We then measure how often this assembled collection triggers an audit rejection on the unmarked null.

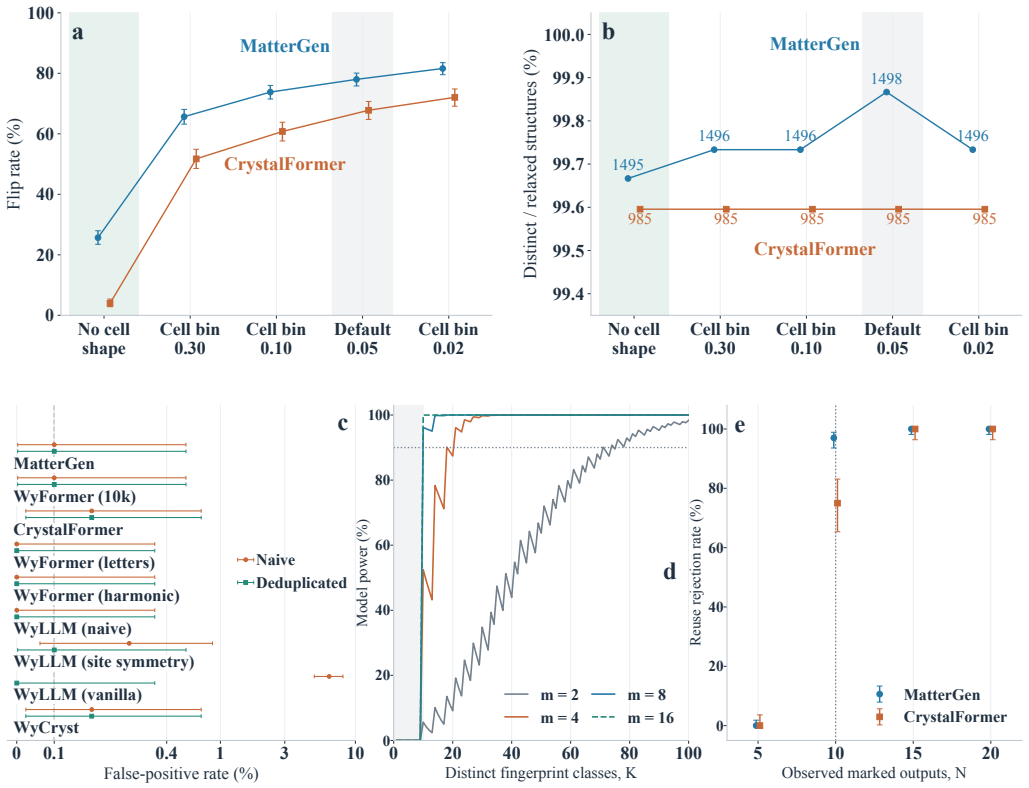

Figure 2. a–b, Removing cell shape reduces relaxation flips while retaining nearly all distinct classes. Shading marks no cell shape (green) and the default bins (gray); lines connect discrete settings. In **b**, labels give class counts K and the vertical range is restricted. **c**, Naive and deduplicated false-positive rates across nine generator collections; the dashed line marks $\alpha = 0.001$. The horizontal scale is linear through 0.4% and logarithmic above it. **d**, Model power for independent distinct candidates and fingerprint-preserving deposition. The dotted line marks 90% power; rejection is impossible in the shaded region $K < 10$. **e**, Rejection after class reuse into 200-record collections. Points are offset for visibility; the vertical guide marks $N = 10$. Panels **c–e** use $\gamma = 1/2$ and $\alpha = 0.001$; **e** uses $m = 8$. Whiskers in **a**, **c**, and **e** are exact two-sided 95% binomial intervals. Sample sizes and numerical results are in Tables 4, 6, and 7. These panels use separate experimental populations.

Figure 2e shows no rejections at $N = 5$, rejection rates of 97% and 75% at $N = 10$ for MATTERGEN and CRYSTALFORMER, and 100% in the tested trials at $N = 15$ and 20. Infeasible trials count as non-rejections, and resampling permits repeated records, including observed structures.

The threshold depends on distinct classes, not the number of records. At $\gamma = 1/2$, ten distinct green classes suffice for rejection because $2^{-10} \leq 0.001$, whereas $2^{-9} > 0.001$. Observing $N = 10$ outputs does not guarantee ten distinct green classes. Guided reuse can therefore transfer the audit signal outside the key-independent null without establishing the origin of individual records.

This resolution limit extends across databases. Matching 27,000 Open Quantum Materials Database (OQMD) records (Saal et al., 2013) against a marked MP-20 set isolates 72 records spanning 67 all-green classes. Auditing the full pool yields an uninformative $p = 0.2444$, whereas auditing the matched subset yields $p = 2^{-67}$. Adaptive matching transfers audit detection across catalogues, proving that structure-only watermarks authenticate fingerprint classes rather than individual physical origins.

5 CONCLUSION

BasinMark demonstrates that proving data provenance depends on which features are preserved during deposition, how many unique classes survive, and whether our metrics can distinguish different processing histories. Under exact same-basin relaxation, coordinate perturbations are erased. However, removing cell-shape bins substantially improves the survival of the discrete fingerprint labels used by the audit while retaining almost all observed classes in our test populations. We calibrate this evidence against our null model by counting each unique class only once. Ultimately, substitution and reuse experiments reveal that this auditing approach resolves down to the broader fingerprint class. If strict, individual record-level authentication is required, signed metadata provides a more suitable solution (National Institute of Standards and Technology, 2023).

LIMITATIONS

Exact annihilation assumes a common basin minimum. Geometric and force bounds are local, and representation gains are population-specific. Calibration requires a key-independent collection, ideal random-function labels, and a pseudorandom PRF. Power also assumes independent selection labels and the specified deposition model. Neural controls use stored outputs. The classical check supplies end-to-end deposition evidence. Separate survival and calibration studies leave joint power unmeasured. Reuse transfers class-level signal.

REPRODUCIBILITY STATEMENT

Implementation, data sources, computational environments, results, and execution scripts are available at `https://anonymous.4open.science/r/crystals-watermarking-iclr/` with Lean 4 formalizations and a coverage map to machine-check the core theoretical proofs.

ETHICS STATEMENT

The study uses generated structures and materials-property records and involves no human participants or personal data. A positive fingerprint audit provides evidence at fingerprint-class resolution. Claims of unique authorship, independent physical origin, or authorisation to reuse a structure require additional evidence and an evaluation designed for those assurances.

USE OF AI STATEMENT

Generative AI supported literature search, drafting and editing, figure and code preparation, data analysis, mathematical formulation, and proof checking. All were reviewed and verified by the authors, who take responsibility for the final content.

REFERENCES

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

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436, 2022.

Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, et al. Prov-dm: The prov data model. *W3C Recommendation*, 14:15–16, 2013.

Siddharth Betala, Samuel P Gleason, Ali Ramlaoui, Andy Xu, Georgia Channing, Daniel Levy, Clémentine Fourrier, Nikita Kazeev, Chaitanya K Joshi, Sékou-Oumar Kaba, et al. Lemat-genbench: A unified evaluation framework for crystal generative models. *arXiv preprint arXiv:2512.04562*, 2025.

T Tony Cai, Xiang Li, Qi Long, Weijie J Su, and Garrett G Wen. Optimal detection for language watermarks with pseudorandom collision. *arXiv preprint arXiv:2510.22007*, 2025.

Zhendong Cao, Xiaoshan Luo, Jian Lv, and Lei Wang. Space group informed transformer for crystalline materials generation. *Science Bulletin*, 2025.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1125–1139. PMLR, 2024.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

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.

Liancheng Fang, Aiwei Liu, Henry Peng Zou, Yankai Chen, Hengrui Zhang, Zhongfen Deng, and Philip Yu. Muse: Model-agnostic tabular watermarking via multi-sample selection. In *International Conference on Learning Representations*, volume 2026, pp. 16541–16574, 2026.

Bruno Galerne. Computation of the perimeter of measurable sets via their covariogram. applications to random sets. *Image Analysis & Stereology*, 30(1):39–51, 2011.

Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. *Journal of the ACM (JACM)*, 33(4):792–807, 1986.

Samuel Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *International Conference on Learning Representations*, volume 2025, pp. 6612–6637, 2025.

Runwen Hu, Peilin Chen, Keyan Ding, and Shiqi Wang. Molmark: Safeguarding molecular structures through learnable atom-level watermarking. *arXiv preprint arXiv:2508.17702*, 2025.

Yuyao Huang, Wei Nong, Shuya Yamazaki, Martin Hoffmann Petersen, Jianghai Wang, Ruiming Zhu, and Kedar Hippalgaonkar. Sword: Symmetry and wyckoff-sequence of ordered and disordered crystals. *arXiv preprint arXiv:2604.17994*, 2026.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.

Nikita Kazeev, Wei Nong, Ignat Romanov, Ruiming Zhu, Andrey Ustyuzhanin, Shuya Yamazaki, and Kedar Hippalgaonkar. Wyckoff transformer: Generation of symmetric crystals. *arXiv preprint arXiv:2503.02407*, 2025.

Filip Ekström Kelvinius, Oskar B Andersson, Abhijith S Parackal, Dong Qian, Rickard Armiento, and Fredrik Lindsten. Wyckoffdiff–a generative diffusion model for crystal symmetry. *arXiv preprint arXiv:2502.06485*, 2025.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International conference on machine learning*, pp. 17061–17084. PMLR, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

In-Ho Lee and Seoleun Shin. Crystal structure search with principal invariants. *Computer Physics Communications*, 292:108889, 2023.

Francesco Maggi. Sets of finite perimeter and geometric variational problems. *Cambridge studies in advanced mathematics*, 135, 2012.

Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20937–20946. IEEE, 2025.

National Institute of Standards and Technology. Digital signature standard (DSS). Technical Report FIPS 186-5, National Institute of Standards and Technology, 2023.

Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of chemical theory and computation*, 20(11):4857–4868, 2024.

Can Polat, Erchin Serpedin, Mustafa Kurban, and Hasan Kurban. Cliffordip: Clifford algebra equivariant interatomic potentials for heterogeneous catalysis. *npj Computational Materials*, 2026.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025.

Janosh Riebesell, Rhys EA 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.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65(11):1501–1509, 2013.

John C Thomas, Anirudh Raju Natarajan, and Anton Van der Ven. Comparing crystal structures with symmetry and geometry. *npj Computational Materials*, 7(1):164, 2021.

Atsushi Togo, Kohei Shinohara, and Isao Tanaka. Spglib: a software library for crystal symmetry search. *Science and Technology of Advanced Materials: Methods*, 4(1):2384822, 2024.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Daniel Widdowson, Marco Mosca, Angeles Pulido, Vitaliy Kurlin, and A Cooper. Average minimum distances of periodic point sets. *arXiv preprint arXiv:2009.02488*, 2020.

Daniel E Widdowson and Vitaliy A Kurlin. Pointwise distance distributions for detecting near-duplicates in large materials databases. *SIAM Journal on Applied Mathematics*, 86(3):898–918, 2026.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):160018, 2016.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Hanlin Zhang, Benjamin L Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: impossibility of strong watermarking for language models. In *Forty-first International Conference on Machine Learning*, 2024.

Zaixi Zhang, Ruofan Jin, Guangxue Xu, Xiaotong Wang, Marinka Zitnik, Le Cong, and Mengdi Wang. Foldmark: Safeguarding protein structure generative models with distributional and evolutionary watermarking. *BioRxiv*, pp. 2024–10, 2025.

Ruiming Zhu, Wei Nong, Shuya Yamazaki, and Kedar Hippalgaonkar. Wycryst: Wyckoff inorganic crystal generator framework. *Matter*, 7(10):3469–3488, 2024.

S1 FORMAL SETTING AND PROOFS

Formal setting. A crystal description is specified by $x = (L, \{(s_i, f_i)\})$, comprising lattice vectors $L \in \mathbb{R}^{3\times3}$, atomic species s_i, and fractional coordinates $f_i \in [0,1)^3$. Unimodular basis transformations, origin shifts, proper rigid rotations, point-group operations, identical-site permutations, and finite-index cell re-descriptions define the crystallographic equivalence relation $\equiv_{\mathrm{crys}}$. Ideal canonicalization maps satisfy $\mathcal{C}(x) = \mathcal{C}(x')$ if and only if $x \equiv_{\mathrm{crys}} x'$, which practical tolerance-based implementations approximate. All functions and observation statistics are assumed to be measurable on their respective spaces.

Structural fingerprints take values in a countable label space. Under an ideal random-function model, distinct query inputs receive independent uniform values in $[0,1)$, and thresholding at $\gamma \in (0,1)$ assigns independent $\mathrm{Bernoulli}(\gamma)$ green statuses. The implemented pseudorandom function (PRF) satisfies this model under standard cryptographic pseudorandomness assumptions.

Proof of Theorem 1. Fix a key k and an admissible structural input x. Because the watermarking map W_k preserves basin occupancy, both $W_k(x)$ and x reside within the same energy basin, whose unique stationary minimum is $\mathcal{R}_0(x)$. Consequently,

$$\mathcal{R}_0(W_k(x)) = \mathcal{R}_0(x), \qquad Q_\rho(\mathcal{C}(\mathcal{R}_0(W_k(x)))) = Q_\rho(\mathcal{C}(\mathcal{R}_0(x))).$$

The second equality establishes that $\mathcal{T}(W_k(x)) = \mathcal{T}(x)$ pointwise. For any admissible random input X, this pointwise equivalence implies that the observed representations $\mathcal{T}(W_k(X))$ and $\mathcal{T}(X)$ are almost surely identical, yielding identical observation distributions. Applying any measurable detector to these matching laws produces equal rejection probabilities under both hypotheses. The result extends immediately to randomized detectors whose internal randomness is independent of the input. (This pointwise result assumes exact basin relaxation; finite force stopping criteria are analyzed separately below.) □

Proof of Proposition 2. Condition on a nonempty deposited collection selected independently of the key and marked observations. Let $f_1, \ldots, f_K$ denote its K distinct structural fingerprints with multiplicities $n_1, \ldots, n_K$ and relative weights $w_i = n_i/n$ (such that $\sum_i w_i = 1$). Under the ideal random-function model, the distinct class statuses $B_i \sim \mathrm{Bernoulli}(\gamma)$ are mutually independent. By linearity of expectation and independence, the naive green sample average satisfies

$$\mathbb{E}\widehat{g} = \sum_{i=1}^{K} w_i\gamma = \gamma, \qquad \mathrm{Var}(\widehat{g}) = \sum_{i=1}^{K} w_i^2 \,\mathrm{Var}(B_i) = \gamma(1-\gamma)\sum_{i=1}^{K} w_i^2.$$

Equating this variance expression to $\gamma(1-\gamma)/n_{\mathrm{eff}}$ yields the variance-equivalent effective sample size $n_{\mathrm{eff}} = (\sum_i w_i^2)^{-1}$. Deduplicating the collection evaluates each distinct label B_i exactly once, so their sum $G_K = \sum_{i=1}^{K} B_i$ follows a binomial distribution $\mathrm{Bin}(K,\gamma)$.

For $Z \sim \mathrm{Bin}(K,\gamma)$, define the rejection threshold $c_K = \min\{c \in \{0,\ldots,K+1\} : \Pr(Z \geq c) \leq \alpha\}$. The deduplicated audit rejects the null hypothesis if and only if $G_K \geq c_K$, guaranteeing a false-positive rate bounded by α. (Setting $c_K = K+1$ indicates that rejection is impossible for the given parameters). Because discrete tail probabilities make the test conservative, conditioning and taking expectations preserves this upper bound for random collections independent of the PRF. Finally, $\sum_i w_i^2 \leq 1$ implies $n_{\mathrm{eff}} \geq 1$, while the Cauchy–Schwarz inequality yields $1 = (\sum_i w_i)^2 \leq K\sum_i w_i^2$, establishing $n_{\mathrm{eff}} \leq K \leq n$. □

Candidate selection and occupancy. For a key-independent candidate batch containing d distinct fingerprints, the probability that all d independent candidates evaluate as non-green is $(1-\gamma)^d$. The selection rule emits a green candidate whenever at least one is present, falling back to an unconstrained choice only when all statuses are zero. Conditioning on the random distinct candidate count D_m and taking expectations yields the marginal producer green probability $q = 1 - \mathbb{E}[(1-\gamma)^{D_m}]$. This marginal calculation accounts for queries shared across batches.

For a population partitioned into fingerprint classes with probabilities v_i, let I_i indicate whether class i is observed at least once in n independent draws. The distinct class count is $K_n = \sum_i I_i$, with expectation $\mathbb{E}I_i = 1 - (1-v_i)^n$. Linearity of expectation yields the expected occupancy $\mathbb{E}K_n = \sum_i [1-(1-v_i)^n]$, which extends to countably many classes by monotone convergence.

Proof of Theorem 3. Let $|A|$ denote Lebesgue volume, $A \triangle B$ symmetric difference, and $A - h = \{x : x + h \in A\}$. Because translation preserves volume, $|A \setminus (A - h)| = |A \triangle (A - h)|/2$. Given that $|\Omega| = 1$ and its constituent cells partition Ω up to null sets, the label mismatch probability under random displacement εU is

$$p(\varepsilon) = \sum_i \mathbb{E}_U |A_i \setminus (A_i - \varepsilon U)| = \frac{1}{2} \sum_i \mathbb{E}_U |A_i \triangle (A_i - \varepsilon U)|.$$

This total includes exits through the boundary of Ω. For a cell A_i with finite perimeter and a fixed unit vector u, the translation limit formula (Galerne, 2011) states

$$\lim_{\varepsilon \downarrow 0} \frac{|A_i \triangle (A_i - \varepsilon u)|}{\varepsilon} = \int_{\partial^* A_i} |u \cdot \nu_i(z)| \, \mathrm{d}\mathcal{H}^{D-1}(z),$$

where $\partial^* A_i$ is the reduced boundary, ν_i is the outward unit normal, and $\mathcal{H}^{D-1}$ is the Hausdorff area measure. The uniform bound $|A_i \triangle (A_i - \varepsilon u)| \le \varepsilon P(A_i)$ enables dominated convergence. By rotational symmetry, $\mathbb{E}|U \cdot \nu|$ is constant for all unit normals ν. The first component of a uniform direction vector U on the unit sphere in $\mathbb{R}^D$ has density proportional to $(1 - t^2)^{(D-3)/2}$ on $(-1, 1)$. Integrating $|t|$ yields $\kappa_D = \mathbb{E}|U \cdot \nu| = \Gamma(D/2)/(\sqrt{\pi}\Gamma((D+1)/2))$. Exchanging the finite sum and limit gives

$$\lim_{\varepsilon \downarrow 0} \frac{p(\varepsilon)}{\varepsilon} = \frac{\kappa_D}{2} \sum_i P(A_i),$$

establishing the first-order asymptotic expansion. Applying Euclidean isoperimetry (Maggi, 2012) to each cell gives $P(A_i) \ge D \omega_D^{1/D} v_i^{1-1/D}$. To relate these cell volumes to collision diversity D_2, define a discrete random variable I with $\Pr(I = i) = v_i$. Applying Jensen's inequality to the convex function $t \mapsto t^{-1/D}$ on $(0, \infty)$ yields

$$\sum_i v_i^{1-1/D} = \mathbb{E}[v_I^{-1/D}] \ge (\mathbb{E} v_I)^{-1/D} = \Big(\sum_i v_i^2 \Big)^{-1/D} = D_2^{1/D}.$$

Combining the perimeter inequality with the asymptotic limit completes the lower bound. □

Justification of deterministic coarsening. Let $\Phi_c = h \circ \Phi_f$ denote a coarsened fingerprinting map. Equality of fine labels implies equality of coarse labels, meaning every coarse mismatch event is a subset of the fine mismatch event on the same joint population ($p_c \le p_f$). Applying h to the fine labels of any collection S yields its coarse labels, so $K_c(S) \le K_f(S)$. For fine class probabilities $v_{f,\ell}$, the mass of a coarse class a is $v_{c,a} = \sum_{\ell : h(\ell) = a} v_{f,\ell}$. Squaring and summing gives

$$\sum_a v_{c,a}^2 = \sum_a \Big(\sum_{\ell : h(\ell) = a} v_{f,\ell} \Big)^2 \ge \sum_\ell v_{f,\ell}^2.$$

Inverting this sum proves $D_{2,c} \le D_{2,f}$. Non-negative cross-terms extend this result to countable classes. All comparisons assume a fixed input distribution; modifying canonicalization tolerances or sample filters defines a separate population context.

Proof of Theorem 4. To prove part (i), substitute equal fingerprint multisets into the deterministic audit decision rule:

$$A_k(\tilde{x}_{1:n}) = a_k(M_\Phi(\tilde{x}_{1:n})) = a_k(M_\Phi(x_{1:n})) = A_k(x_{1:n}).$$

For part (ii), let $\mathbf{X}$ denote a random structure collection and B a measurable set of audit decisions. For fixed key k and hypothesis $j \in \{0, 1\}$,

$$P_{j,k}\{A_k(\mathbf{X}) \in B\} = P_{j,k}\{M_\Phi(\mathbf{X}) \in a_k^{-1}(B)\}.$$

Because a_k is measurable, the inverse image $a_k^{-1}(B)$ is an admissible event. Equal multiset distributions under $P_{0,k}$ and $P_{1,k}$ imply identical probabilities for every B, proving equality of audit outcome distributions. Setting $B = \{1\}$ confirms equal rejection probabilities. Integrating conditional probabilities against a key distribution ν extends this guarantee to random keys. For the deduplicated audit, the unique fingerprint set S determines both K and G_K under a fixed key, yielding identical binomial scores across collections sharing the same unique structural classes. □

Finite force tolerance. Let $V \in C^2(U)$ be an potential energy surface defined on a convex neighborhood U of a local minimum $x^\star$, with zero modes removed and Hessian lower-bounded by $\nabla^2 V \succeq \mu I$ ($\mu > 0$). For any $y \in U$ and displacement $d = y - x^\star$, the line segment connecting $x^\star$ and y lies entirely in U. Since $\nabla V(x^\star) = 0$, integrating along this segment yields

$$d^\top \nabla V(y) = \int_0^1 d^\top \nabla^2 V(x^\star + td) d \, \mathrm{d}t \geq \mu \|d\|_2^2.$$

Applying the Cauchy–Schwarz inequality bounds the left side by $\|d\|_2 \|\nabla V(y)\|_2$. Dividing by $\mu\|d\|_2$ proves the displacement bound $\|y - x^\star\|_2 \leq \|\nabla V(y)\|_2/\mu$ for $d \neq 0$ (the case $d = 0$ is trivial). For N atoms in fixed Cartesian coordinates with maximum per-atom force f, $\|\nabla V(y)\|_2^2 \leq Nf^2$. This inequality bounds spatial offsets near a minimum; translating this offset into label survival requires evaluating the margin between y and nearby fingerprint boundaries.

S2 Additional Experimental Details

S2.1 Fingerprint Construction and Audit Settings

The primitive fingerprint reduces input structures to standard primitive cells, then concatenates space-group numbers, reduced chemical compositions, and sorted sets of (element, site-symmetry symbol, orbit-size) triples. Orbit sizes are normalized by their greatest common divisor, and distinct orbits sharing identical triples are merged. The cell-shape block discretizes cell length ratios b/a, c/a, and unit angles via $\mathrm{round}(r/\delta)\delta$ using nearest-even tie breaking. The default bin size $\delta = 0.05$ creates step sizes of 0.05 for ratios and $0.05°$ for angles. Omitting this block yields the no-cell-shape fingerprint variant.

Default audit parameters are symmetry tolerance $\tau = 0.1$ Å, green probability $\gamma = 1/2$, candidate batch size $m = 8$, nominal level $\alpha = 10^{-3}$, and target power 90%. The audit retains one entry per unique fingerprint and evaluates the exact upper binomial tail. Unmarked collection calibration uses $1{,}000$ pseudorandom keys, while marked controls evaluate 200 keys over fixed structure collections.

Controlled libraries. The metallic library contains 71 prototype structures across Ag, Al, Au, Cu, Ni, Pd, and Pt, spanning five space groups (85% cubic) evaluated using an embedded-atom potential. The broad library comprises 212 generated structures across 25 space groups in all seven crystal systems (with at most 32 atoms per cell, 47% of which are low-symmetry triclinic, monoclinic, or orthorhombic), evaluated using a Lennard–Jones potential. Both benchmark sets serve as controlled representation tests for crystallographic invariance.

Finite-tolerance check. The finite-tolerance experiment evaluates 40 real prototype structures relaxed to 5×10^{-3} eV Å^{-1} force tolerance and deposited at precision $\rho = 10^{-5}$ Å. Across perturbation amplitudes of 0.01–0.30 Å, marked and unmarked relaxed outputs were bit-identical in 0 of 40 cases, confirming that residual numerical differences persist under standard physical stopping thresholds.

S2.2 Model-Implied Audit Sizes

To estimate required audit sizes, assume m distinct candidate fingerprints per batch, giving a producer green probability $q = 1 - (1 - \gamma)^m$. Suppose fingerprint flip events occur with probability p during relaxation, independently of producer selection. Following a flip, a fresh independent PRF query yields a green label with probability γ. The effective green probability q_p among relaxed deposited outputs is

$$q_p = \underbrace{(1-p)q}_{\text{preserved fingerprints}} + \underbrace{p\gamma}_{\text{fresh-query labels}}.$$

Scanning positive integers K identifies the minimum distinct class count required to reach 90% model power under the exact binomial test (Table 2).

Figure 2d uses $p = 0$ and computes power as $\Pr\{\mathrm{Bin}(K, q) \geq c_K\}$, where c_K is the smallest integer with $\Pr\{\mathrm{Bin}(K, 1/2) \geq c_K\} \leq 0.001$. If no count permits rejection, power is zero.

Table 2. Required distinct class count K to reach 90% model power under measured relaxation flip rates ($m = 8$, $\gamma = 1/2$, $\alpha = 0.001$). Instability in low-symmetry crystal systems dramatically increases required audit sample sizes.

Crystal system	K needed	Potential	K needed
Cubic	14	SevenNet	241
Hexagonal	210	ORB-v3	252
Tetragonal	539	MACE-MPA-0	272
Trigonal	944	MatterSim	281
Orthorhombic	3,455	CHGNet	362
Monoclinic	312,571		
Triclinic	$> 600{,}000$		

S2.3 Relaxation Panels and Matching

The pooled dataset contains 13,902 before/after relaxation pairs stratified by crystal system. A separate potential benchmark evaluates 2,000 pairs across five interatomic potential families: SevenNet (Park et al., 2024), Orb (Rhodes et al., 2025), MACE (Batatia et al., 2022), MatterSim (Yang et al., 2024), and CHGNet (Deng et al., 2023). Reweighting flip rates to match the pooled crystal-system composition establishes a unified reference benchmark (Table 3).

Table 3. Fingerprint flip rates and median root-mean-square (RMS) displacements relative to DFT references across five interatomic potentials, reweighted to a common crystal-system composition.

Potential	Pairs	Flip rate	Matched distance
SevenNet	2,000	0.718	0.0136
ORB-v3	2,000	0.726	0.0141
MACE-MPA-0	2,000	0.735	0.0163
MatterSim	2,000	0.738	0.0171
CHGNet	2,000	0.769	0.0236

Matched distance denotes the dimensionless median normalized RMS displacement $(\mathrm{Vol}/N)^{-1/3}$ among successfully matched structures.

S2.4 Fingerprint Coarsening and Population Definitions

Table 4 gives the counts underlying Figure 2a–b at symmetry tolerance $\tau = 0.1$ Å. Flip rates use valid raw/relaxed pairs; distinctness uses valid relaxed structures. Panel a intervals are recomputed from the flip counts using the exact two-sided Clopper–Pearson method. Table 5 retains the separate symmetry-tolerance sweep.

Table 4. Fingerprint settings at $\tau = 0.1$ Å for Figure 2a–b. K/n counts distinct classes among valid relaxed structures; flip denominators count valid pairs.

	MatterGen		CrystalFormer	
Cell shape	Flips/pairs	K/n	Flips/pairs	K/n
Omitted	385/1,500	1,495/1,500	39/984	985/989
0.30	985/1,500	1,496/1,500	509/984	985/989
0.10	1,107/1,500	1,496/1,500	598/984	985/989
0.05 (default)	1,170/1,500	1,498/1,500	667/984	985/989
0.02	1,224/1,500	1,496/1,500	709/984	985/989

Table 5. Impact of symmetry tolerance τ on no-cell-shape fingerprint flips and distinct class ratios (K/n_{relaxed}) across generator populations.

	MatterGen			**CrystalFormer**		
τ (Å)	Flips/pairs	Rate (%)	K/n_{relaxed}	Flips/pairs	Rate (%)	K/n_{relaxed}
0.01	584/1,500	38.9	1,497/1,500	16/991	1.6	987/991
0.05	590/1,500	39.3	1,495/1,500	29/991	2.9	987/991
0.10	385/1,500	25.7	1,495/1,500	39/984	4.0	985/989
0.30	228/1,500	15.2	1,494/1,500	45/957	4.7	954/958

S2.5 Quantized Continuous-Descriptor Control

A quantized Average Minimum Distance (AMD) fingerprint (length 100, augmented with reduced composition) passes all crystallographic invariance checks. Evaluated on 2,000 pairs using CHGNet and SevenNet at bin width 0.02, quantized AMD yields flip rates of 0.987 and 0.985, demonstrating that continuous descriptor discretization remains sensitive to relaxation shifts unless bin widths are scaled appropriately.

S2.6 Calibration Across Released Generators

Table 6 provides the numerical companion to Figure 2c and includes two reference populations. In the figure, WyLLM abbreviates WyckoffLLM; WyFormer (10k) denotes the 9,995 valid records from the nominal 10,000-output release. The shortened WyFormer and WyLLM variant names refer to the corresponding DiffCSP++ collections.

Table 6. Numerical results for Figure 2c and two reference populations, using 1,000 keys ($\alpha = 0.001$, $\gamma = 1/2$). n denotes records and K distinct fingerprint classes.

Generator	n	K	naive FPR	dedup. FPR
MatterGen	10,000	9,876	0.0010	0.0010
WyFormer/DiffCSP++	9,995	9,912	0.0010	0.0010
CrystalFormer	989	985	0.0020	0.0020
WyFormer-letters	957	956	0.0000	0.0000
WyFormer-harmonic	983	983	0.0000	0.0000
WyckoffLLM-naive	980	979	0.0000	0.0000
WyckoffLLM-site-sym.	999	998	0.0030	0.0010
WyckoffLLM-vanilla	999	649	**0.0640**	**0.0000**
WyCryst	994	994	0.0020	0.0020
MP-20 training set	27,136	27,049	0.0010	0.0020
Experimental-crystal calibration population	5,000	5,000	0.0020	0.0020

S2.7 Marked, Population, and End-to-End Controls

All six released energy-bearing generator collections achieve 100% detection rates across 200 marking keys. Under unmarked keys, false-positive rates remain ≤ 0.005. Population controls on clean reference sets (MP-20 with 27,049 distinct fingerprints and a 5,000-structure experimental-crystal calibration population) yield zero false detections. Synthetic baseline generators (repeating substitution and random search) are detected at rate 1.000.

S2.8 Reuse Simulation and Cross-Database Matching

Table 7 gives the counts for Figure 2e. Each trial assembles 200 records by resampling matches to exposed classes with replacement; observed structures may be reused. Infeasible trials count as non-rejections. The plotted intervals use the exact two-sided Clopper–Pearson method.

Table 7. Class-reuse rejections/trials at $\alpha = 0.001$, $\gamma = 1/2$, and $m = 8$. N counts observed marked outputs.

Observed N	**MatterGen**	**CrystalFormer**
5	0/200	0/100
10	194/200	75/100
15	200/200	100/100
20	200/200	100/100

In cross-database matching, selecting 72 OQMD records matching 67 green fingerprint classes from a marked MP-20 collection yields an adaptively selected subset score of $p = 2^{-67}$ (versus $p = 0.2444$ across the full database), demonstrating that fingerprint-based watermarks authenticate structural classes rather than isolated physical records.