BOUNDING RETRAINING EQUIVALENCE AND THE DELETION FLOOR IN MATERIALS MACHINE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In materials machine learning, closely related retained structures can sustain accurate property predictions even after removing a specific record, rendering post-deletion prediction error an ambiguous metric for machine unlearning. To resolve this ambiguity, we define the *deletion floor* as the expected target loss under a specified retraining procedure at the deleted request. Standard indistinguishability constraints yield a sharp interval bounding an update's target loss around this baseline reference. Theoretically, a conditional neighbor bound links a low deletion floor directly to retained fit, prediction regularity, and local label agreement, while an exact ridge identity isolates residual fit from the prediction change induced by record deletion. Empirically, controlled redundancy sweeps show an $\approx 8\times$ drop in median normalized retraining loss when one retained relative remains after deletion. Across two distinct fitting regimes in a paired Materials Project study, the lower-floor regime also exhibits a larger prediction change on more than 50% of the shared requests. Systematic comparisons against approximate updates and the original model decouple deliberate target suppression from preserved overall model utility. Consequently, request-level unlearning evaluations should report reference loss, prediction change, and retained utility together, interpreting post-deletion accuracy against what retraining itself leaves behind.

1 INTRODUCTION

Materials-property datasets map chemical compositions and crystal structures directly to thermodynamic targets (Ward et al., 2018; Polat et al., 2026). Because ionic substitutions routinely preserve underlying crystal lattices across varying chemical compositions, physically interpretable relationships naturally form among distinct database entries (Hautier et al., 2011). Recent findings in data duplication reveal that retained related examples can sustain original model behavior even after exact retraining (Ye et al., 2025). Consequently, removing a specific record's training contribution differs fundamentally from eliminating the broader physical behavior represented by that record (Triantafillou et al., 2026).

Standard certified removal frameworks evaluate updated models by comparing their output distributions against references trained without the deleted records (Guo et al., 2019; Sekhari et al., 2021). In practice, however, widespread unlearning membership attacks frequently overestimate privacy guarantees (Hayes et al., 2025), while simple accuracy-gap measurements often conflict with distribution-based forgetting scores (Triantafillou et al., 2024). Resolving what accuracy remains after deletion requires examining individual requests against a clear reference. Without establishing what exact retraining itself would predict for a removed target, continued post-deletion accuracy remains an ambiguous signal of unlearning performance.

To resolve this ambiguity, we study individual deletion requests through the *deletion floor* defined as the expected target loss under a specified retraining procedure. When a model accurately fits retained data and exhibits smooth prediction behavior across neighboring inputs, structural and property redundancies preserve target predictability. This mechanism decouples retained support, reference matching, and deletion sensitivity. Mathematically, a conditional neighbor bound directly connects retained structural support to the deletion floor, while retraining equivalence restricts how far an

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.

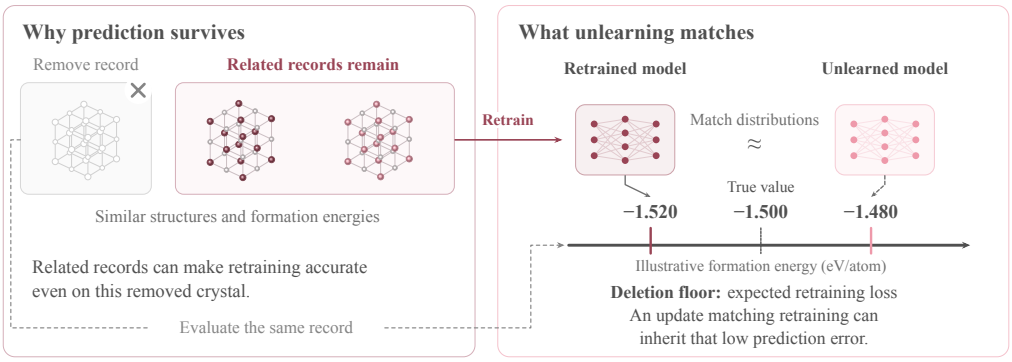


Figure 1: **Related materials can keep a removed record predictable.** With good retained fit and smooth predictions, remaining relatives can keep retraining loss low; an update matching retraining can preserve that accuracy. The dashed path brings the removed record to the formation-energy comparison, where model arrows identify illustrative predictions beside its true value. The deletion floor is expected retraining loss at that record. Distributional certification requires model-law evidence beyond agreement of individual predictions.

update's loss can deviate from this reference. Furthermore, an exact ridge identity isolates residual model fit from the true prediction change induced by record deletion (Figure 1).

We validate these theoretical distinctions through a controlled redundancy sweep alongside 44,344 crystal structures from the Materials Project (MP) (Jain et al., 2013). Exact fixed-feature ridge calculations evaluate the floor and prediction change on identical requests, while approximate updates and unedited baselines expose how target suppression trades off against overall model utility. This work delivers three key contributions:

- **Derives a request-level reference interval:** Formulates a sharp bounded-loss interval that precisely identifies the target losses compatible with approximate retraining equivalence.
- **Establishes a theoretical mechanism for retained predictability:** Proves a conditional neighbor bound linking the deletion floor to retained support, and derives an exact ridge decomposition that separates residual fit from deletion-induced sensitivity.
- **Proposes a robust evaluation protocol:** Demonstrates via synthetic redundancy sweeps and materials benchmarks why reference loss, prediction change, and preserved utility measure distinct properties and must be reported independently.

The remainder of this paper is organized as follows: Section 2 positions our contributions within the unlearning literature, Section 3 presents the mathematical characterization, and Section 4 details experiments across synthetic sweeps, materials fitting regimes, and approximate update procedures. Section 5 outlines the proposed request-level reporting protocol. Appendix A provides mathematical proofs, Appendix B specifies experimental parameters, and Appendix C presents complete operating-point summaries and neural descriptor diagnostics.

2 Related Work

Retraining objectives and evaluation. Retraining establishes the baseline benchmark for eliminating a record's training contribution, with certified removal requiring that an updated model's distribution remain statistically indistinguishable from one trained entirely without that record (Guo et al., 2019; Sekhari et al., 2021). Because exact retraining is computationally expensive, practical approximate algorithms frequently trade theoretical guarantees for efficiency; SCRUB, for instance, balances forgetting objectives against retained utility without providing a distributional certificate (Kurmanji et al., 2023). Consequently, robust evaluation hinges on clarifying both the target objective and the specific properties measured by an evaluation metric. Existing membership inference attacks frequently overestimate privacy protections (Hayes et al., 2025), while simple accuracy-gap

metrics often diverge from distribution-based forgetting scores (Triantafillou et al., 2024). Evaluating post-deletion performance thus requires isolating the exact baseline error left by retraining at individual deletion requests.

Duplication, memorization, and retained support. Data redundancy presents a distinct challenge for unlearning algorithms. Empirical studies by Ye et al. (2025) demonstrate that adversarially duplicated or near-duplicated training instances can preserve original model behavior even after exact retraining. Similarly, Zhao et al. (2024) connect the difficulty of approximate unlearning to data memorization and retain–forget entanglement, showing that low memorization leaves a model's prediction largely insensitive to the inclusion or exclusion of a specific training point.

Broader behavioral removal. Removing a specific record's training influence represents a distinct objective from eliminating the overarching behavioral concept or domain knowledge associated with that record (Triantafillou et al., 2026). Correlated-knowledge unlearning evaluates how inference persists through retained structural relationships (Wu et al., 2024), whereas retain–forget entanglement focuses on unwanted collateral damage inflicted on related retained samples during forgetting (Cheng et al., 2026). Additionally, distributional unlearning selects data subsets to excise entire target distributions while leaving complementary distributions intact (Allouah et al., 2026). Together, these formulations highlight the need to distinguish between broader distribution removal and sample-level reference agreement.

Positioning of this work. This work investigates the baseline target loss left by exact retraining at individual deletion requests. By defining the *deletion floor*, we complement existing update algorithms and evaluation metrics by identifying the exact target behavior constrained by reference matching. Building on prior observations of persistence under data duplication and low memorization, we formalize retained support as a theoretical mechanism, connecting it directly to retraining equivalence, local neighbor fit, and ridge leverage. Our bounded-loss transfer theorem delineates the precise loss intervals permitted under indistinguishability, while our exact ridge decomposition decouples retraining accuracy from deletion-induced prediction changes. Paired materials experiments confirm that lower reference loss can coexist with larger prediction sensitivity. Ultimately, these findings establish a reference-based evaluation protocol for regression unlearning, while leaving broader knowledge and distribution removal as distinct objectives.

3 Bounding loss around the retraining reference

3.1 The deletion floor and bounded-loss transfer

We first define the retraining reference following data deletion. Let $\mathcal{D} = ((x_i, y_i))_{i=1}^n$ denote an indexed dataset, where $\mathcal{F}$ specifies requested deletion records and $\mathcal{R}$ contains remaining indices. Explicit indexing keeps identical record copies distinct. Applying a training procedure $\mathcal{A}_{\mathrm{tr}}$ to $\mathcal{R}$ induces a model distribution law Q over training randomness. At a target deletion request (x_f, y_f), let $h(\theta) = \ell(f_\theta(x_f), y_f)$ be a nonnegative measurable loss with finite expectation under Q.

Definition 1 (Deletion floor)**.** The deletion floor is the expected target loss under the specified retraining procedure,

$$\Phi(x_f; \mathcal{F}, \mathcal{D}) = \mathbb{E}_{\theta\sim Q}[h(\theta)].$$

Definition 1 quantifies the expected residual error left at the deleted target by retraining. Under squared loss, a low floor indicates that exact retraining inherently predicts the deleted target accurately. Because this benchmark reflects the specific retraining procedure $\mathcal{A}_{\mathrm{tr}}$, alternative estimators may yield losses above or below this reference.

To evaluate what an update algorithm constrained by retraining equivalence can alter, let P represent its output law on the same measurable model space. Symmetric (ϵ, δ) retraining equivalence requires $P(S) \le e^\epsilon Q(S) + \delta$ and $Q(S) \le e^\epsilon P(S) + \delta$ across all measurable sets S (Guo et al., 2019). By translating standard hypothesis-testing constraints to target loss observables (Kairouz et al., 2015), we establish sharp bounds on the update's expected loss relative to the floor.

Theorem 1 (Sharp bounded-loss transfer)**.** *Let $B > 0$, $\epsilon \ge 0$, $\delta \in [0, 1]$, and let h be a measurable observable taking values in $[0, B]$. If P, Q satisfy symmetric (ϵ, δ) retraining equivalence and $\Phi =$*

$\mathbb{E}_Q[h]$, *then*

$$\max\{0, e^{-\epsilon}(\Phi - \delta B), B - e^{\epsilon}(B - \Phi) - \delta B\} \le \mathbb{E}_P[h] \le \min\{B, e^{\epsilon}\Phi + \delta B, B - e^{-\epsilon}(B - \Phi - \delta B)\}.$$

For each fixed $(\epsilon, \delta, B, \Phi)$ *with* $0 \le \Phi \le B$*, both endpoints are attainable over bounded observables and admissible pairs, converging to* Φ *as* $(\epsilon, \delta) \to (0, 0)$.

Theorem 1 bounds the expected target loss deviation permitted under approximate equivalence. Applying the event inequalities to both the loss h and its complement $B - h$ confirms that both interval endpoints converge tight to Φ as $(\epsilon, \delta) \to (0, 0)$, establishing exact sharpness over bounded observables.

Applying the guarantee to squared error. Because standard squared error is unbounded, we apply Theorem 1 by defining a clipped loss observable $h_B(\theta) = \min\{(f_\theta(x_f) - y_f)^2, B\}$ for a fixed threshold $B > 0$. Measurable predictions yield integrable clipped losses $0 \le h_B \le B$ under both P and Q, anchored by the clipped reference expectation $\Phi_B = \mathbb{E}_Q[h_B]$. While theoretical bounds govern this clipped observable, our empirical analyses report unclipped raw squared loss to describe full retraining behavior. Local target agreement reflects retraining equivalence, but complete distributional equivalence requires model-law validation beyond scalar agreement at a single point.

A target-error threshold. When an unlearning objective requires increasing target error to $\mathbb{E}_P[h] \ge \tau$ (where $0 < \Phi < \tau < B$) under pure $(\epsilon, 0)$ equivalence, both upper bounds in Theorem 1 must simultaneously reach τ. Solving for ϵ establishes the necessary equivalence budget:

$$\epsilon \ge \max\left\{\log \frac{\tau}{\Phi}, \log \frac{B - \Phi}{B - \tau}\right\}. \tag{1}$$

The first term quantifies the required loss increase from Φ to τ, while the second reflects the corresponding reduction in the complement expectation $B - h$. For example, elevating expected unit-bounded loss from $\Phi = 0.1$ to $\tau = 0.5$ requires $\epsilon \ge \log 5$. Equation (1) provides a necessary compatibility condition between target-suppression goals and formal equivalence guarantees; relaxed guarantees with $\delta > 0$ introduce additive slack across the full interval.

3.2 How retained support can keep the floor low

To understand why retraining often leaves deleted targets predictable, consider a retained neighbor (x_r, y_r) near the deleted input x_f under distance metric d. If a retrained model accurately fits x_r, varies smoothly between inputs, and target labels align, the retained example constrains prediction error on the removed target.

Assumption 1 (Uniform retained-neighbor control)**.** For a fixed retained neighbor (x_r, y_r), assume $f_\theta(x_f)$ and $f_\theta(x_r)$ are measurable in θ. For Q-almost every θ, suppose $|f_\theta(x_f) - f_\theta(x_r)| \le \bar{L} d(x_f, x_r)$ and $|f_\theta(x_r) - y_r| \le \bar{\varepsilon}$ for fixed $\bar{L}, \bar{\varepsilon} \ge 0$.

Bounding the input distance $d(x_f, x_r) \le \rho$ and label gap $|y_f - y_r| \le \eta$, target error decomposes into retained fit error, prediction variation, and label disparity:

$$\begin{aligned} |f_\theta(x_f) - y_f| &\le \underbrace{|f_\theta(x_r) - y_r|}_{\text{retained fit}} + \underbrace{|f_\theta(x_f) - f_\theta(x_r)|}_{\text{prediction variation}} + \underbrace{|y_r - y_f|}_{\text{label gap}} \\ &\le \bar{\varepsilon} + \bar{L}\rho + \eta \equiv C(x_f). \end{aligned} \tag{2}$$

Under Assumption 1, taking expectations bounds the absolute-loss floor by C, while squaring first bounds the squared-loss floor by C^2. Consequently, high retained accuracy, smooth local predictions, and small label gaps jointly guarantee a low deletion floor. For finite neighbor sets, the tightest valid bound holds almost surely across retrained models.

An empirical diagnostic. We construct a request-level surrogate for the theoretical bound via

$$\widehat{C}(x_f) = \widehat{\varepsilon} + \widehat{L} d(x_f, x_r) + |y_f - y_r|,$$

where x_r represents the nearest retained neighbor in standardized descriptor space. Here, $\widehat{\varepsilon}$ is estimated from the original model's 90th-percentile absolute training residual, and $\widehat{L}$ is set to the 99.5th percentile of sampled prediction difference quotients. We designate this empirical ranking FLOORSCORE. Neighbor selection is performed prior to evaluating deletion losses, avoiding direct use of the observed deletion losses in constructing the proxy.

3.3 AN EXACT RIDGE CONNECTION: FIT, LEVERAGE AND DELETION

While neighbor bounds establish sufficient conditions for low floors, ridge regression enables exact calculation of retraining residuals and decouples baseline model fit from deletion sensitivity. Consider a feature matrix $A \in \mathbb{R}^{n\times p}$ with rows $a_i^\top$, target vector $y \in \mathbb{R}^n$, and objective $n^{-1}\|Aw - y\|^2 + \lambda\|w\|^2$ ($n \geq 2$, $\lambda > 0$) minimized by $w_{\mathcal{D}}$. Because row deletion alters mean-loss normalization from $n\lambda$ to $(n-1)\lambda$, we define an adjusted full-data estimator $\widetilde{w}$ using the retained-size penalty coefficient:

$$M = A^\top A + (n-1)\lambda I, \qquad \widetilde{w} = M^{-1}A^\top y, \qquad \widetilde{h}_i = a_i^\top M^{-1} a_i. \tag{3}$$

The adjusted fit $\widetilde{w}$ maintains all training rows while adopting the penalty scaling of the retained system. Applying the Sherman–Morrison identity (Hager, 1989) yields an exact closed-form expression for the retrained residual r_{-i} and corresponding deletion floor Φ_i:

$$r_{-i} := a_i^\top w_{-i} - y_i = \frac{a_i^\top \widetilde{w} - y_i}{1 - \widetilde{h}_i}, \qquad \Phi_i = r_{-i}^2. \tag{4}$$

The denominator remains strictly positive due to the positive-definiteness of M. Equation (4) reveals that retraining error is governed jointly by the adjusted full-data residual $a_i^\top \widetilde{w} - y_i$ and the deleted point's leverage $\widetilde{h}_i$. A small adjusted residual yields a low deletion floor provided leverage remains below unity, whereas high full-data residual error carries over into large retraining loss.

Comparing predictions before and after deletion isolates model sensitivity to record removal. Defining full-model residual $r_i = a_i^\top w_{\mathcal{D}} - y_i$, deletion response is captured by target prediction change $|r_i - r_{-i}|$. Because squared losses discard residual signs, evaluating prediction change requires signed residual tracking. We decompose the total prediction change into penalty rescaling and row removal components:

$$a_i^\top (w_{-i} - w_{\mathcal{D}}) = \underbrace{\lambda a_i^\top M^{-1} w_{\mathcal{D}}}_{\text{penalty rescaling}} + \underbrace{\frac{\widetilde{h}_i \widetilde{r}_i}{1 - \widetilde{h}_i}}_{\text{removing the row}}, \tag{5}$$

where $\widetilde{r}_i = a_i^\top \widetilde{w} - y_i$. Equation (5) demonstrates that a highly accurate model fit can still exhibit strong prediction sensitivity upon record removal, motivating joint measurement of retraining floors and deletion changes.

Why the no-update reference matters. The unedited original model serves as a necessary control because baseline predictions may already align closely with retraining outcomes. To formalize this intuition, consider $n \geq 2$ and differentiable, μ-strongly convex objectives ($\mu > 0$) sharing a regularizer. Assuming per-example gradient norms are bounded by G and model predictions are L_p-Lipschitz in parameters ($G, L_p \geq 0$), unconstrained minimizers satisfy the leave-one-out stability bound:

$$|f_{w_{\mathcal{D}}}(x_f) - f_{w_{\mathcal{R}}}(x_f)| \leq \frac{2GL_p}{\mu n}. \tag{6}$$

The $O(1/n)$ scaling in Equation (6) confirms that small original-to-retrained prediction shifts occur naturally in stable models, highlighting why no-update controls are indispensable when evaluating update agreement against retraining benchmarks.

4 EXPERIMENTS AND RESULTS

Empirical evaluations demonstrate how data redundancy and fitting capacity shape the retraining reference, establishing the benchmark against which approximate unlearning updates and unedited baselines are assessed across matched requests.

4.1 SHARED REFERENCE AND MEASUREMENT PROTOCOL

Evaluating deletion behavior requires tracking both loss relative to the retraining reference and physical prediction shifts. For each single-example deletion request i, we record the squared target loss e_i, its corresponding retraining floor Φ_i, and the normalized error ratio $q_i = e_i / \max(\Phi_i, 10^{-12})$. Predictive utility is measured by test MSE relative to the original model. To isolate distinct prediction shifts, let $\widehat{y}_i^D$, $\widehat{y}_i^R$, and $\widehat{y}_i^U$ denote predictions from the original, retrained, and updated models, respectively. The *deletion-induced prediction change* $|\widehat{y}_i^D - \widehat{y}_i^R|$ measures original model sensitivity to record removal, while the *update-to-reference prediction distance* $|\widehat{y}_i^U - \widehat{y}_i^R|$ measures update deviation from retraining. The scalar loss gap $g_i = |e_i - \Phi_i|$ relates directly to the normalized ratio via

$$g_i = \left|\max(\Phi_i, 10^{-12}) q_i - \Phi_i\right|. \tag{7}$$

The safeguard denominator 10^{-12} ensures numeric stability for near-zero reference losses without altering observed ratios, as all measured ridge floors exceed 1.12×10^{-7} (eV/atom)^2. Thus, Equation (7) unifies target loss comparisons while prediction distances capture directional shifts.

To isolate data-level effects, the materials ridge reference refits weights on retained rows while holding feature standardization, RFF kernel bandwidth, and random projections fixed (Rahimi & Recht, 2007). This conditional procedure, denoted $\mathcal{A}_{\mathrm{tr}}(\cdot; T_{\mathcal{D}})$, preserves the request's contribution to learned preprocessing $T_{\mathcal{D}}$. Under deterministic ridge fitting, the squared residual equals the expected loss floor under fixed hyperparameters, with across-request medians reporting central tendencies. Score thresholds are calculated within each diagnostic sample, and request-level distributions accompany aggregate medians.

4.2 RETRAINING RESIDUALS ACROSS CONTROLLED REDUNDANCY SETTINGS

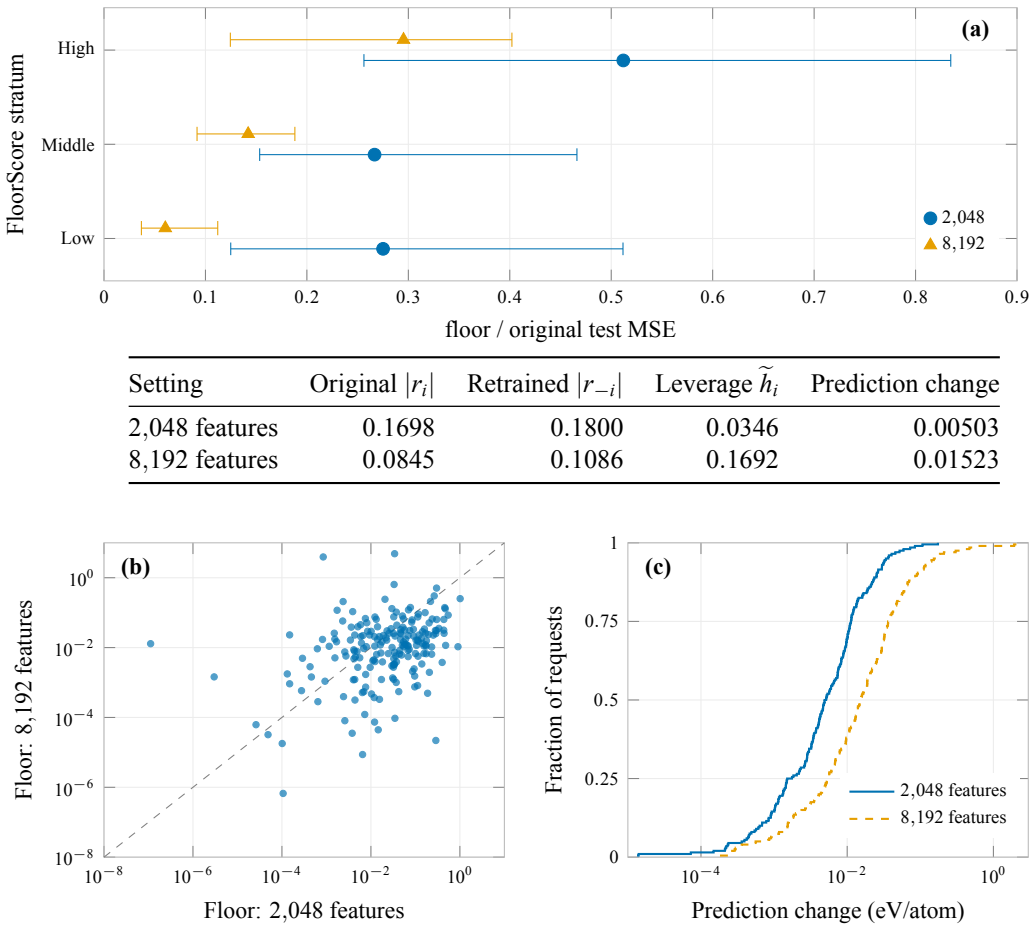
Retained family members substantially reduce retraining loss even after a target record is removed. We evaluate synthetic datasets constructed with 90 prototype families and 160 isolated points in a six-dimensional descriptor space, where labels generated from a smooth random-feature function are corrupted by additive noise. Across family sizes $k \in \{1, 2, 4, 8, 16\}$ and total dataset sizes $90k + 160$, we assess 300 independent trials, deleting one family member and one isolated control separately and normalizing floors by original test MSE.

Transitioning from $k = 1$ (where deletion removes the sole family member) to $k = 2$ (leaving one retained relative) produces a dramatic $7.934\times$ drop in median normalized retraining loss, from 0.326 down to 0.041. Additional family members yield diminishing loss reductions, reaching a median floor of 0.020 ($\approx 2\%$ of test MSE) at $k = 16$. In contrast, isolated controls maintain a high median floor of 0.386 ($\approx 39\%$ of test MSE) throughout the sweep, demonstrating that low retraining loss is localized to requests with retained support.

4.3 MATERIALS FLOORS DEPEND ON THE FITTING REGIME

Deletion floors vary systematically with model capacity and regularization. We analyze $44{,}344$ crystal structures ($35{,}476$ train, $8{,}868$ test) from the MP using 134 structural and chemical descriptors targeting formation energy per atom. Two distinct ridge regimes evaluate exact retained residuals: a $2{,}048$-feature model ($\lambda = 10^{-6}$, 200 diagnostic requests) and an $8{,}192$-feature model ($\lambda = 10^{-8}$, 300 diagnostic requests), coupling feature count and regularization across separate diagnostic samples.

The fitting regime strongly modulates the relationship between the local structural proxy FLOORSCORE and actual retraining loss (Figure 2(a)). In the $2{,}048$-feature setting, low-, middle-, and high-FLOORSCORE requests exhibit median normalized floors of 0.275, 0.267, and 0.512. The $8{,}192$-feature, lower-penalty regime lowers the corresponding medians to 0.0605, 0.142, and 0.295—yielding a $4.88\times$ high-to-low contrast and rank correlation of 0.261—while simultaneously reducing training MSE from 0.0754 to 0.0166 and test MSE from 0.0870 to 0.0704. These results show that FLOORSCORE is a setting-dependent diagnostic whose empirical stratification varies with the fitting regime.



Setting	Original $\lvert r_i \rvert$	Retrained $\lvert r_{-i} \rvert$	Leverage $\widetilde{h}_i$	Prediction change
2,048 features	0.1698	0.1800	0.0346	0.00503
8,192 features	0.0845	0.1086	0.1692	0.01523

Figure 2: **Lower retraining loss can accompany larger prediction changes after deletion.** (a) Median normalized floors across FLOORSCORE strata for the 2,048-feature ($\lambda = 10^{-6}$) and 8,192-feature ($\lambda = 10^{-8}$) fits; error bars are 95% request-bootstrap intervals. Their diagnostic samples contain 200 and 300 requests, respectively, and differ across fits. The embedded table reports medians on a separate paired cohort of 200 shared requests: original and retrained absolute residuals and prediction changes are in eV/atom, while adjusted leverage is dimensionless. (b) Paired squared retraining losses in $(\text{eV/atom})^2$; 132 points fall below equality. (c) Empirical distributions of deletion-induced prediction changes; the 8,192-feature setting has the larger change on 172 requests. Feature count and regularization vary jointly.

4.4 LOWER FLOORS CAN COEXIST WITH LARGER DELETION CHANGES

Evaluating identical deletion requests reveals that higher model accuracy does not imply lower deletion sensitivity. We compare both fitting regimes on a shared set of 200 deletion requests using identical train/test splits.

The paired medians embedded in Figure 2 show that the 8,192-feature setting has smaller original and retrained residuals but larger adjusted leverage and deletion-induced prediction change. Compared with the 2,048-feature baseline, it achieves a lower retraining floor on 132 of 200 requests while exhibiting a larger prediction change on 172 requests, dropping median squared retraining loss from 0.0324 to 0.0118 $(\text{eV/atom})^2$ while increasing median prediction change from 0.00503 to 0.0152 eV/atom (Figure 2(b)–(c)).

On 104 individual requests, the 8,192-feature, lower-penalty regime predicts removed targets more accurately after retraining despite moving predictions farther from the original fits. This decoupling

highlights that retraining floors evaluate post-deletion accuracy, whereas prediction changes quantify target sensitivity.

4.5 TARGET LOSS AND RETAINED UTILITY AT FIXED OPERATING POINTS

Approximate unlearning interventions reveal a trade-off between target suppression and preserved overall model utility. Using the shared cohort of 200 requests in the $2{,}048$-feature setting ($\lambda = 10^{-6}$), we compare approximate updates against exact retraining and the unedited no-update control (Table 1).

Evaluated updates start from independent copies of the original model under fixed hyperparameter settings: NegGrad+ (Kurmanji et al., 2023) (30 updates combining target loss ascent with retain loss descent), SCRUB-style regression (Kurmanji et al., 2023) (40 alternating ascent/teacher-matching updates), Newton-style removal (1 damped Hessian step), retain fine-tuning (60 steps), random relabeling (40 steps), and gradient ascent (30 target ascent steps), all using learning rate 0.010 and retained minibatches ≤ 256.

Table 1: **Increasing target error can also reduce predictive utility.** A target-loss ratio above one indicates greater error than retraining, and a test-MSE ratio above one indicates worse utility than the original model. Gradient ascent raises both ratios. Several other updates remain near the no-update control. Entries are medians over 200 shared requests in full-corpus ridge with $2{,}048$ features and $\lambda = 10^{-6}$ at the fixed update settings.

Procedure	Target loss / floor	Test MSE ratio
Retrain (reference)	1.000	1.000
No update	0.932	1.000
Retain fine-tuning	0.930	1.000
Newton-style update	0.934	1.000
Random relabeling	0.928	1.000
SCRUB-style regression	1.141	1.001
NegGrad+	1.645	1.013
Gradient ascent	3.081	1.083

The unedited control exhibits a median target-loss/floor ratio of 0.932 and reference distance of 0.00503 eV/atom. Retain fine-tuning, Newton-style updates, and random relabeling closely mirror the control's loss ratio and test utility. In contrast, gradient ascent forces target loss to $3.081\times$ the reference while increasing test MSE by 8.3% ($1.083\times$ ratio). These results emphasize that target loss inflation alone does not signify successful unlearning unless evaluated alongside reference agreement and retained utility.

5 EVALUATION PROTOCOL AND SCOPE

Unlearning evaluations must decouple post-deletion accuracy from record sensitivity: the deletion floor quantifies what a specified retraining reference predicts, whereas prediction change measures model sensitivity to record removal. Because higher post-retraining accuracy can coexist with larger prediction shifts, reporting both properties is necessary to contextualize unlearning performance.

A request-level reporting protocol. Evaluations should explicitly define the target deletion objective, training reference, and any refitted preprocessing steps. For each request, benchmarks should report original, retrained, and updated target losses, deletion-induced prediction changes, update-to-reference prediction distances (when available), and retained test utility, alongside an unedited original model as a no-update control. Retaining signed residuals enables exact recovery of prediction shifts under squared loss. Comparing predictions across original, retrained, and updated models isolates true update effects beyond baseline reference agreement, whereas proving formal retraining equivalence requires distributional certification beyond scalar point agreement.

Interpreting operating points. Evaluating approximate updates requires measuring target loss shifts against both reference floors and preserved utility. An update satisfying retraining equivalence naturally remains accurate when the reference deletion floor is low. Conversely, target-suppression

interventions elevate loss above the floor, requiring separate utility accounting; for example, gradient ascent forces target loss higher while degrading overall test utility. Because squared loss discards residual signs and near-zero reference floors produce artificially inflated loss ratios, evaluations should report full request-level distributions and explicit loss normalizations alongside aggregate medians.

Scope and limitations. These findings reflect specific experimental and theoretical scopes. In controlled sweeps, global redundancy scales jointly with total dataset size, while in materials studies, feature dimension and regularization vary together under fixed descriptors, preprocessing, and model architectures. Crystal graph neural network diagnostics provide exploratory observations per request, and FLOORSCORE serves as a setting-dependent proxy whose empirical rank correlation varies with model capacity. Additionally, leave-one-out stability bounds assume uniform gradient control. Broader behavioral removal, comparative hyperparameter tuning, distributional certification, and end-to-end refitting of learned representations remain distinct objectives requiring expanded frameworks. Within this scope, our protocol provides transparent accounting of retraining references and update behavior.

6 CONCLUSION

The deletion floor establishes an explicit baseline for evaluating regression unlearning by measuring the residual target error left by exact retraining. Retraining equivalence restricts updated models around this reference, retained structural support maintains low reference loss, and ridge leverage explains how accurate retraining can remain highly sensitive to record removal. Across synthetic redundancy sweeps and materials benchmarks, empirical measurements confirm that lower reference loss frequently coexists with larger prediction sensitivity. Consequently, continued post-deletion accuracy can only be interpreted relative to reference floors and target objectives. Standardizing unlearning evaluation requires reporting retraining predictions, record removal shifts, and retained model utility together.

REPRODUCIBILITY STATEMENT

The accompanying repository (https://anonymous.4open.science/r/materials-unlearning-iclr/) contains implementation details, data provenance, environment information, result records, and verification scripts. It also includes Lean 4 formalizations verifying bounded-loss transfer and sharpness, endpoint convergence, target-error thresholds, retained-neighbor bounds, ridge identities, and conditional leave-one-out stability.

ETHICS STATEMENT

This study uses synthetic data and public materials-property records, involving no human subjects or personal data. Demonstrating that retained structural support preserves target accuracy after deletion is a technical observation and should not be used to infer compliance with specific legal or regulatory data removal mandates. Applications seeking broader behavioral concept removal require evaluations tailored to that specific objective.

USE OF AI STATEMENT

Generative AI tools assisted with literature review, text polishing, code development, and data analysis. The submitting authors retain full responsibility for the final text, theoretical proofs, cited sources, code, and reported empirical findings.

REFERENCES

Youssef Allouah, Rachid Guerraoui, and Sanmi Koyejo. Distributional machine unlearning via selective data removal. In *International Conference on Learning Representations*, volume 2026, pp. 119258–119289, 2026.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine unlearning under retain-forget entanglement. *arXiv preprint arXiv:2603.26569*, 2026.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1979.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

William W Hager. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989.

Geoffroy Hautier, Chris Fischer, Virginie Ehrlacher, Anubhav Jain, and Gerbrand Ceder. Data mined ionic substitutions for the discovery of new compounds. *Inorganic chemistry*, 50(2):656–663, 2011.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 497–519. IEEE, 2025.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pp. 1376–1385. PMLR, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

Can Polat, Mustafa Kurban, Erchin Serpedin, and Hasan Kurban. Quantumcanvas: a multimodal benchmark for learning two-body quantum interactions. *Machine Learning: Science and Technology*, 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, et al. Are we making progress in unlearning? findings from the first neurips unlearning competition. *arXiv preprint arXiv:2406.09073*, 2024.

Eleni Triantafillou, Ahmed Imtiaz Humayun, Monica Ribero, Alexander Matt Turner, Michael C Mozer, and Georgios Kaissis. Is your algorithm unlearning or untraining? *arXiv preprint arXiv:2604.07962*, 2026.

Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

Ruihan Wu, Chhavi Yadav, Russ Salakhutdinov, and Kamalika Chaudhuri. Evaluating deep unlearning in large language models. *arXiv preprint arXiv:2410.15153*, 2024.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

Dayong Ye, Tianqing Zhu, Jiayang Li, Kun Gao, Bo Liu, Leo Yu Zhang, Wanlei Zhou, and Yang Zhang. Data duplication: A novel {Multi-Purpose} attack paradigm in machine unlearning. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6399–6418, 2025.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Barbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37:12293–12333, 2024.

A PROOFS OF THE MAIN CHARACTERIZATION

A.1 PROOF OF THEOREM 1: BOUNDED-LOSS TRANSFER

The bounded-loss transfer theorem establishes how symmetric (ϵ, δ) retraining equivalence restricts expected loss shifts relative to the retraining reference floor Φ.

Proof of Theorem 1. Applying hypothesis-testing constraints on bounded observables (Kairouz et al., 2015), let $a = e^{\epsilon}$. For any measurable loss $h \in [0, B]$, the layer-cake representation combined with distributional indistinguishability $P(S) \leq aQ(S) + \delta$ yields

$$\mathbb{E}_P[h] = \int_0^B P(h > t)\, dt \leq a \int_0^B Q(h > t)\, dt + \delta B = a\Phi + \delta B.$$

Reversing the roles of P and Q gives the lower bound $\mathbb{E}_P[h] \geq a^{-1}(\Phi - \delta B)$. Applying these identical constraints to the complementary bounded observable $B - h$ establishes

$$\mathbb{E}_P[h] \geq B - a(B - \Phi) - \delta B, \qquad \mathbb{E}_P[h] \leq B - a^{-1}(B - \Phi - \delta B).$$

Intersecting these four inequalities with the physical loss range $0 \leq \mathbb{E}_P[h] \leq B$ completes the proof of the transfer interval.

To prove sharpness, normalize by B and define $q = \Phi/B$. Consider a two-point sample space with observable $h = B\mathbf{1}_S$, setting $Q(S) = q$ and $P(S) = p$. Symmetric (ϵ, δ)-indistinguishability reduces to the four joint constraints

$$p \leq aq + \delta, \quad q \leq ap + \delta, \quad 1 - p \leq a(1 - q) + \delta, \quad 1 - q \leq a(1 - p) + \delta,$$

subject to $0 \leq p \leq 1$. The feasible interval for p corresponds exactly to the normalized upper and lower bounds of Theorem 1. Feasibility is guaranteed because $p = q$ satisfies all constraints for valid parameters. Endpoint attainability holds across all $0 \leq \Phi \leq B$, including extreme points $q = 0$ and $q = 1$. As $(\epsilon, \delta) \to (0, 0)$, all candidate bounds converge tight to Φ, establishing endpoint convergence. □

Justification of Equation (1)**.** Under exact $(\epsilon, 0)$ equivalence ($\delta = 0$), requiring updated target error to reach $\mathbb{E}_P[h] \geq \tau$ (where $0 < \Phi < \tau < B$) forces both upper bounds in Theorem 1 to be at least τ, namely $e^{\epsilon}\Phi \geq \tau$ and $B - e^{-\epsilon}(B - \Phi) \geq \tau$. Taking logarithms yields the two necessary constraints in Equation (1), defining the minimum equivalence budget ϵ compatible with the target error objective.

A.2 JUSTIFICATION OF THE RETAINED-NEIGHBOR BOUND

Justification of Equation (2)*.* Under uniform retained-neighbor control (Assumption 1), target error at deletion request x_f decomposes into retained neighbor fit, local prediction variation, and label disparity. For Q-almost every retrained predictor θ, triangle inequality yields

$$|f_\theta(x_f) - y_f| \leq |f_\theta(x_f) - f_\theta(x_r)| + |f_\theta(x_r) - y_r| + |y_r - y_f| \leq \bar{L}\rho + \bar{\varepsilon} + \eta \equiv C(x_f).$$

Because $C(x_f)$ is deterministic and nonnegative, taking expectations bounds absolute loss by $\mathbb{E}_Q[|f_\theta(x_f) - y_f|] \leq C(x_f)$, while squaring first yields squared loss bound $\mathbb{E}_Q[(f_\theta(x_f) - y_f)^2] \leq C(x_f)^2$. For clipped squared loss, expectation is bounded by $\min\{C(x_f)^2, B\}$. For any finite collection of valid retained neighbors, intersecting their measure-one validity sets preserves a measure-one domain where the minimum bound remains valid. □

A.3 DERIVATION OF THE RIDGE IDENTITIES

Derivation of Equations (4) *and* (5)*.* Define the retained normal matrix $H_{-i} = A_{-i}^\top A_{-i} + (n - 1)\lambda I = M - a_i a_i^\top$. For $n \geq 2$ and $\lambda > 0$, both H_{-i} and M are positive definite. Setting $t = a_i^\top H_{-i}^{-1} a_i \geq 0$, Sherman–Morrison rank-one update formulas yield adjusted leverage $\widetilde{h}_i = t/(1 + t) < 1$, ensuring positive leverage denominators $1 - \widetilde{h}_i > 0$.

The stationarity condition for retained ridge fitting is

$$(M - a_i a_i^\top) w_{-i} = A^\top y - a_i y_i.$$

Multiplying by M^{-1} and isolating the weight vector gives $w_{-i} = \widetilde{w} + M^{-1} a_i (a_i^\top w_{-i} - y_i)$. Taking the inner product with $a_i^\top$, subtracting target y_i, and solving for residual $r_{-i} = a_i^\top w_{-i} - y_i$ yields the exact leave-one-out residual identity in Equation (4).

For prediction change, comparing full minimizer $(M+\lambda I) w_{\mathcal{D}} = A^\top y$ with adjusted fit $M\widetilde{w} = A^\top y$ gives parameter shift $\widetilde{w} - w_{\mathcal{D}} = \lambda M^{-1} w_{\mathcal{D}}$. Subtracting predictions yields

$$a_i^\top (w_{-i} - w_{\mathcal{D}}) = a_i^\top (\widetilde{w} - w_{\mathcal{D}}) + (r_{-i} - \widetilde{r}_i) = \lambda a_i^\top M^{-1} w_{\mathcal{D}} + \frac{\widetilde{h}_i \widetilde{r}_i}{1 - \widetilde{h}_i},$$

proving Equation (5). □

Under an unnormalized penalty $\alpha > 0$, full matrix $M = A^\top A + \alpha I$ uses unadjusted full fit $w_{\mathcal{D}}$ and standard leverage $h_i = a_i^\top M^{-1} a_i$, yielding residual ratio $r_{-i} = r_i/(1 - h_i)$ and squared loss ratio $r_i^2/\Phi_i = (1 - h_i)^2$. When penalty λ is defined per sample mean loss, adjusted matrix M provides the exact sample-size normalization correction.

A.4 Leave-one-out stability: statement and proof

Leave-one-out stability bounds quantify how record deletion impacts parameter estimation in strongly convex objectives (Bousquet & Elisseeff, 2002). Let $n \geq 2$, let $\ell_i(w)$ be example losses, and let $r(w)$ be a differentiable regularizer. Objective functions $F_{\mathcal{D}}(w) = \frac{1}{n} \sum_{i \in \mathcal{D}} \ell_i(w) + r(w)$ and $F_{\mathcal{R}}(w) = \frac{1}{n-1} \sum_{i \in \mathcal{R}} \ell_i(w) + r(w)$ are μ-strongly convex ($\mu > 0$) with minimizers $w_{\mathcal{D}}, w_{\mathcal{R}}$. Assuming loss gradients are bounded $\|\nabla \ell_i(w)\| \leq G$ and predictions are L_p-Lipschitz in parameters, parameter shift satisfies $\|w_{\mathcal{D}} - w_{\mathcal{R}}\| \leq 2G/(\mu n)$, proving Equation (6).

Proof of the leave-one-out stability bound. Stationarity of the retained objective gives $\nabla r(w_{\mathcal{R}}) = -(n-1)^{-1} \sum_{i \in \mathcal{R}} \nabla \ell_i(w_{\mathcal{R}})$. Substituting into the full gradient $\nabla F_{\mathcal{D}}(w_{\mathcal{R}})$ yields

$$\nabla F_{\mathcal{D}}(w_{\mathcal{R}}) = \frac{1}{n} \left(\nabla \ell_f(w_{\mathcal{R}}) - \frac{1}{n-1} \sum_{i \in \mathcal{R}} \nabla \ell_i(w_{\mathcal{R}}) \right), \qquad \|\nabla F_{\mathcal{D}}(w_{\mathcal{R}})\| \leq \frac{2G}{n}.$$

Strong convexity of $F_{\mathcal{D}}$ at stationarity $\nabla F_{\mathcal{D}}(w_{\mathcal{D}}) = 0$ implies

$$\mu \|w_{\mathcal{R}} - w_{\mathcal{D}}\|^2 \leq \langle \nabla F_{\mathcal{D}}(w_{\mathcal{R}}), w_{\mathcal{R}} - w_{\mathcal{D}} \rangle \leq \frac{2G}{n} \|w_{\mathcal{R}} - w_{\mathcal{D}}\|.$$

Dividing by $\|w_{\mathcal{R}} - w_{\mathcal{D}}\|$ yields $\|w_{\mathcal{R}} - w_{\mathcal{D}}\| \leq 2G/(\mu n)$. Applying prediction Lipschitz continuity $|f_{w_{\mathcal{D}}}(x_f) - f_{w_{\mathcal{R}}}(x_f)| \leq L_p \|w_{\mathcal{D}} - w_{\mathcal{R}}\|$ completes the proof of Equation (6). □

The $O(1/n)$ asymptotic decay relies on uniform gradient bounds G and positive lower bound $\mu > 0$. For mean-squared loss, $\mu = 2\lambda$ provides strong convexity, though unrestricted regression lacks global gradient bounds. Thus, algorithmic stability provides a conditional theoretical mechanism for low prediction sensitivity, whereas exact ridge identities provide unconstrained exact calculations.

B Experimental details

B.1 Controlled family construction

To systematically analyze the impact of structural redundancy on deletion floors, we construct a synthetic benchmark in a six-dimensional descriptor space using 300 cosine Random Fourier Features (RFFs). Projection frequencies are drawn from a standard Gaussian distribution $N(0, I)$, phases are sampled uniformly over $[0, 2\pi)$, and latent target function weights are initialized from a Gaussian distribution with standard deviation 0.30. Target evaluations share a fixed latent function and a held-out set of $1{,}500$ noiseless test points across all sweep evaluations.

Each generated dataset contains 90 prototype family centers (coordinate standard deviation 1.5) with within-family perturbations (standard deviation 0.05) alongside 160 isolated background points, yielding labels corrupted by additive Gaussian noise with standard deviation 0.02. Model training uses ridge regression with regularization parameter $\lambda = 10^{-4}$ under mean-squared error. Across family sizes $k \in \{1, 2, 4, 8, 16\}$ (scaling total dataset size as $90k + 160$), we evaluate 300 independently generated dataset draws, refitting ridge weights after deleting a family member and an isolated control separately. The resulting family-size sweep is summarized in Figure 3.

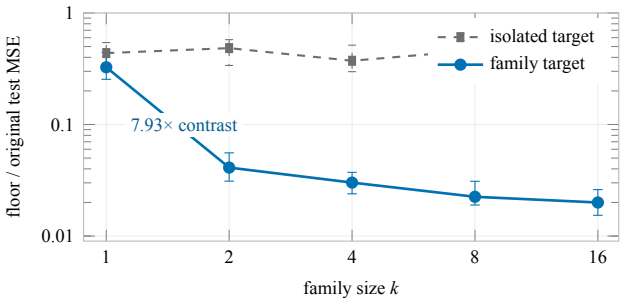


Figure 3: **Retained relatives can keep a deleted family member predictable.** Increasing family size from $k = 1$ to $k = 2$ leaves one relative after deletion and sharply lowers the median normalized retraining loss. Isolated controls remain higher throughout the sweep. Loss is divided by each draw's original-model test MSE, and error bars are 95% bootstrap intervals over 300 generated datasets and targets per family size. All 90 families grow together, coupling redundancy with total training size.

B.2 Materials representation and reference

The real-world evaluation uses 44,344 crystal structures extracted from the MP export, filtered by maximum site count $n_{\text{sites}} \leq 20$, energy above hull ≤ 0.08 eV/atom, and formation energy ≤ 2 eV/atom. Structures are randomly partitioned into 35,476 training entries and 8,868 test entries, targeting formation energy per atom in eV/atom.

Each crystal structure is parameterized by 134 features: 94 elemental composition fractions, 8 global geometric attributes, and a 32-bin pair-distance radial histogram spanning 0–12 Å. Standardized feature vectors x map through a cosine RFF kernel $\phi(x) = \sqrt{2/p}\cos(Wx + b)$, where weights $W_{ij} \sim N(0, \gamma^2)$ and bias $b_j \sim \text{Unif}[0, 2\pi)$. Kernel bandwidth γ is set to the inverse median non-zero Euclidean distance across 4,000 randomly sampled training pairs. Feature standardization, bandwidth parameters, and RFF projections remain fixed during deletion, ensuring that exact retraining refits linear weights under identical mean-loss penalties. The two fitting regimes and their train/test errors are summarized in Table 2.

Table 2: **The higher-feature, lower-penalty setting fits both training and test data better.** Both settings use the same train/test partition, and MSE is reported in (eV/atom)^2. The table compares fitting regimes with feature count and regularization coupled. Diagnostic requests are sampled separately; the paired comparison uses the same 200 targets.

Features p	λ	Diagnostic requests	Train MSE	Test MSE
2,048	10^{-6}	200	0.0753784	0.0869666
8,192	10^{-8}	300	0.0166086	0.0703859

B.3 Request sampling and uncertainty

Diagnostic deletion requests are sampled uniformly without replacement from the training set, reserving a shared cohort of 200 requests for paired-regime and operating-point evaluations.

The empirical proxy FloorScore identifies the nearest retained structural neighbor in standardized descriptor space. Its regularity estimate $\widehat{L}$ corresponds to the 99.5th percentile of prediction difference quotients across 20,000 sampled pairs, while residual estimate $\widehat{\varepsilon}$ matches the original

model's 90th-percentile absolute training residual. Score tertiles are computed independently within each diagnostic sample. Statistical uncertainty is quantified via 95% percentile bootstrap confidence intervals (Efron, 1979), drawing 10,000 resamples for ridge strata and 2,000 resamples for synthetic, operating-point, and neural diagnostics.

B.4 APPROXIMATE-UPDATE SETTINGS

Approximate unlearning procedures initialize from independent copies or refits of the original full-data model, using training data for unlearning updates and reserved held-out data exclusively for test utility evaluation.

Table 3: **Fixed update settings define the operating points being compared.** Step counts and objectives specify each intervention's strength; all gradient methods use learning rate 10^{-2} and retain minibatches of at most 256 points. The results characterize the selected interventions at these fixed strengths; comparative tuning lies outside this experiment.

Procedure	Steps	Additional setting
Gradient ascent	30	Target squared-loss ascent
NegGrad+ (Kurmanji et al., 2023)	30	Forget/retain mixing $\alpha = 0.5$
Retain fine-tuning	60	Retain squared-loss descent
SCRUB-style regression	40	20 ascent, 20 retain; weight 0.3
Newton-style update	1	Hessian damping 10^{-2}
Random relabeling	40	Replacement target drawn from retained labels

Algorithm variants adapt standard unlearning formulations to regression tasks: SCRUB-style regression alternates target loss ascent with retain-task and teacher-prediction matching under squared loss (Kurmanji et al., 2023); random relabeling replaces the target label with a randomly sampled retained label during retain fine-tuning; and Newton-style removal adapts Hessian-based unlearning (Guo et al., 2019) via a damped retained Hessian correction scaled by the forget fraction. Utility impact is reported as median test MSE relative to the original model.

C SUPPLEMENTARY RESULTS

C.1 COMPLETE OPERATING-POINT SUMMARIES

Evaluating approximate unlearning procedures across shared deletion requests isolates differences between target loss shifts, baseline reference agreement, and retained model utility.

Table 4: **Target-error increases can move an update farther from the retraining reference.** Gradient ascent has the largest median absolute loss gap and test-error ratio among these operating points. A gap near zero indicates similar target losses; a target-loss ratio above one indicates greater error than retraining. Entries summarize 200 shared requests in 2,048-feature ridge, with brackets giving 95% bootstrap intervals. Gap values are in units of 10^{-3} (eV/atom)^2, and test ratios use original-model test MSE. The retrain gap and its interval are below 10^{-9} in the displayed units.

Procedure	Loss gap $\times 10^3$	Target loss / floor	Test ratio
Retrain (reference)	$< 10^{-9}$	1.000 [1.000, 1.000]	1.000
No update	1.72 [1.07, 2.93]	0.932 [0.922, 0.939]	1.000
Retain fine-tuning	1.83 [1.23, 2.86]	0.930 [0.918, 0.938]	1.000
SCRUB-style regression	3.65 [2.28, 5.38]	1.141 [1.132, 1.156]	1.001
NegGrad+	18.85 [11.58, 24.24]	1.645 [1.631, 1.659]	1.013
Gradient ascent	62.58 [37.40, 78.31]	3.081 [3.048, 3.116]	1.083
Newton-style update	1.66 [1.01, 2.82]	0.934 [0.925, 0.941]	1.000
Random relabeling	1.92 [1.28, 2.58]	0.928 [0.909, 0.936]	1.000

The absolute loss gap $g_i = |e_i - \Phi_i|$ and target loss ratio $q_i = e_i/\Phi_i$ provide complementary perspectives on reference agreement across all eight procedures evaluated on the shared 200-request cohort (Table 4, with hyperparameter configurations specified in Table 3). While overlapping bootstrap intervals indicate statistical similarity across several conservative updates, rigorous verification of algorithmic equivalence requires dedicated distribution-level testing beyond scalar interval comparisons.

C.2 EXPLORATORY NEURAL DESCRIPTOR DIAGNOSTIC

To examine whether retained-support patterns extend beyond the fixed-feature kernel models, we evaluate single-seed retraining losses from a Crystal Graph Convolutional Neural Network (CGCNN) (Xie & Grossman, 2018) on the same crystal population. The 134-dimensional descriptors are used only to define the retained-neighbor diagnostics. The CGCNN architecture comprises three message-passing layers of width 64, processing graphs with up to 12 neighbors within an 8 Å cutoff and 41 Gaussian distance features. Model training proceeds for 60 epochs using the Adam optimizer (Kingma & Ba, 2014) (batch size 128, learning rate 10^{-2}, zero weight decay), achieving a baseline test MSE of 0.0029175 $(\text{eV/atom})^2$.

For each deleted target, we identify its nearest retained descriptor neighbor under fixed standardization and measure rank associations using Spearman correlation without fitting parameters to deletion losses.

Table 5: **Neighbor label gaps show a stronger association than descriptor distance.** Larger label gaps are positively associated with higher neural retraining loss. Descriptor distance has little rank association, and its interval includes zero. Entries are Spearman correlations over 200 requests with 95% request-bootstrap intervals conditional on one training seed. The correlations quantify association within this sample.

Diagnostic	Correlation	95% interval
Nearest retained descriptor distance	0.0367	$[-0.0913,\ 0.1700]$
Label gap to that neighbor	0.2128	$[0.0729,\ 0.3464]$

Within this single-seed diagnostic, neighbor label disparity shows a stronger positive association with post-deletion loss than raw descriptor distance (Table 5). Descriptor distance has little rank association ($\rho = 0.0367$, with a confidence interval spanning zero), whereas the target label gap shows a positive association ($\rho = 0.2128$). This pattern is consistent with local property agreement contributing to retraining predictability in this neural-model setting.