REPAIRING CORRUPTED CRYSTAL STRUCTURES WITH GEOMETRY-GUIDED MULTI-BRANCH RELAXATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Crystal repair complements generation: recovering the intended structure from a corrupted input at fixed composition, using the geometry that survives corruption. Every atom is still there; which positions may be trusted is not known in advance. A shared collision projection reads short-range separations in the input and proposes where to start. Five CHGNet refinement routes relax that start under different degrees of freedom and tolerances. A frozen ExtraTrees selector reads observable geometric changes and candidate-level energy, force, and stress, then emits one structure. Clean structures and corruption labels stay hidden at inference and inform matching and subgroup analysis only after every prediction is fixed. On 2,520 corrupted test inputs, multi-branch repair matches 76.35% (1,924/2,520), against 44.01% (1,109/2,520) for relaxation-only and 29.96% (755/2,520) for the unchanged input, the share already inside the matcher tolerance; the 32.34-point gain over relaxation-only is bought with five relaxations against one. Match is highest for atom overlaps and lattice distortions, and lower for coordinate noise and coordination defects.

1 INTRODUCTION

Generative models have broadened access to crystal candidates: DiffCSP predicts periodic structures from composition (Jiao et al., 2023), MatterGen supports property-conditioned design (Zeni et al., 2025), and text-based models emit serialized atomistic structures (Gruver et al., 2024). A corrupted structure offers different evidence. Its lattice and surviving atomic layout constrain the target even where individual coordinates have drifted. We call this setting *crystal repair*: recovering the intended crystal from a corrupted input, holding composition and atom inventory fixed.

Repair must keep what survived and correct the rest. Lattice distortion, coordinate noise, atom overlaps, and local coordination damage disturb different structural relationships. Instruction-guided editing (Ito et al., 2026) and host-conditioned inpainting (Zhong et al., 2025; Reents et al., 2026) show how much partial structure buys. Every atom survives corruption; which positions can be trusted is not known in advance. Three decisions structure the task: how to initialize corrections, which coordinates or lattice degrees of freedom to release, and which resulting candidate to keep.

Multi-branch repair turns each decision into a component. A shared collision projection reads short-range separations in the corrupted structure and proposes where to start. Five CHGNet refinement routes then relax it, each exposing different degrees of freedom. Fixed-cell and restricted-site routes hold chosen parts of the input, while full-cell routes admit coupled lattice and coordinate correction; the routes also differ in convergence tolerance. CHGNet drives every route and reports final energy, force, and stress (Deng et al., 2023).

The last decision is selection. A frozen ExtraTrees selector, fitted once on a Train set of 200 inputs, commits to one of the five outputs. Its features are candidate geometry, displacement from the input, and final energy, force, and stress. Those scores track recovery, so predicted energy stays one signal among several. Generation, selection, and stopping consult only the corrupted structure and its observable measurements, while the paired clean crystal and corruption annotations wait for later evaluation.

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.

Table 1: Representative methods grouped by input constraints and objectives. Fixed composition occurs in both generation and refinement; repair additionally evaluates recovery of the intended crystal from a corrupted structure.

Direction and representative methods	Given information and allowed changes	Primary objective
Generation from composition or properties: CDVAE, DiffCSP, FlowMM, FlowLLM, MatterGen (Xie et al., 2022; Jiao et al., 2023; Miller et al., 2024; Sriram et al., 2024; Zeni et al., 2025)	Composition, property constraints, or a sampled latent state; species may be fixed or generated, depending on the task.	Produce plausible or property-conditioned crystals, including refinement of generated candidates.
Feedback-driven design and editing: Reflexion, Self-Refine; LLMatDesign, MatAgent, MAPPS, LapidaryEngine (Shinn et al., 2023; Madaan et al., 2023; Jia et al., 2024; Takahara et al., 2025; Zhou et al., 2025; Ito et al., 2026)	General feedback mechanisms instantiated with design goals, tools, or editing instructions; edits can change or preserve composition.	Improve task outcomes, optimize material properties, or implement requested structure edits.
Structure-conditioned reconstruction: CHGGen, XtalPaint (Zhong et al., 2025; Reents et al., 2026)	A known host or observed substructure constrains reconstruction of selected missing atomic positions.	Complete partial structures or explore guest configurations consistent with the host.
Physical relaxation and assessment: CHGNet, MatterSim, MACE-MP-0; Matbench Discovery and potential-model error analysis (Deng et al., 2023; Yang et al., 2024; Batatia et al., 2025; Riebesell et al., 2025; Deng et al., 2025)	An initial structure is refined at fixed atom inventory; energies, forces, and stresses assess or guide updates.	Reach a relaxed configuration or assess stability; exact structural recovery is a separate criterion.

We contribute a repair method emitting one structure, combining geometry-conditioned initialization, complementary constrained relaxation, and learned selection. On 2,520 test-set inputs covering 580 material identifiers, it reaches 76.35% Match (1,924/2,520) against 44.01% (1,109/2,520) for relaxation-only and 29.96% (755/2,520) for the unchanged input, which bounds the share already inside the matcher tolerance; the 32.34-point gain over relaxation-only costs five relaxations, up to 860 optimizer steps, against one. Ablations credit two sources: collision projection mostly on atom overlaps, and learned selection over the candidate set mostly on coordinate and coordination damage. Route diversity is not separately identified.

2 Related Work

Existing generation, editing, inpainting, and relaxation methods condition on different inputs and pursue different objectives. Repair starts from the corrupted structure itself and targets one specific crystal, not a class of plausible ones. Table 1 organizes these directions by their available information and objectives. Section 3 makes these choices explicit.

Generation from Composition or Properties. CDVAE generates periodic structures and reconstructs given ones, while DiffCSP predicts lattice parameters and coordinates for a specified composition (Xie et al., 2022; Jiao et al., 2023). FlowMM develops symmetry-aware flow matching (Miller et al., 2024), and MatterGen extends diffusion-based generation to property-conditioned design (Zeni et al., 2025). Generation already refines its own candidates. FlowLLM seeds a flow model with structures sampled from a large language model (LLM) (Sriram et al., 2024). Conditioning ranges from composition and properties to learned latent states. We keep that refinement step but change what it starts from: our five CHGNet refinement routes begin at the damaged structure.

Feedback-Driven Design and Editing. Reflexion uses linguistic feedback and memory to improve later actions, and Self-Refine alternates feedback with revision (Shinn et al., 2023; Madaan

et al., 2023). What these supply is a decision mechanism, not a rule for editing composition. In inverse design that mechanism searches over compositions. MatAgent proposes them, generates structures, and redirects the search using evaluated properties (Takahara et al., 2025). LLMatDesign instead adds, substitutes, or removes species to approach target properties (Jia et al., 2024). MAPPS couples the same loop to physics tools and scientific feedback (Zhou et al., 2025). Their edits may change composition; ours may not. LapidaryEngine spans both regimes, supporting composition-changing edits and composition-preserving refinement of a distorted crystal (Ito et al., 2026). We follow its constraints, not its supervision: what guides us is observable geometry, never instructions or target properties.

Structure-Conditioned Reconstruction. Crystal inpainting exploits structural information that is already given. CHGGen pairs host-guided inpainting with a foundation interatomic potential to place guest atoms in a known host (Zhong et al., 2025). XtalPaint conditions diffusion on a partial structure and reconstructs missing sites, including hydrogen (Reents et al., 2026). In both, an observed host anchors the reconstruction and every retained position counts as sound. Repair has no such host: all atoms survive corruption, so no site is known to be correct. We borrow the conditioning and discard the anchor, letting each route decide which coordinates may move.

Physical Relaxation and Assessment. Machine-learning interatomic potentials (MLIPs) update an existing crystal without changing its atom inventory. CHGNet supplies the energies, forces, and stresses that drive relaxation (Deng et al., 2023). MatterSim and MACE-MP-0 extend this capability across conditions and chemistries (Yang et al., 2024; Batatia et al., 2025). Physical relaxation is our reference comparator, not our method. Relaxation decides where an optimization stops, not which crystal is recovered. Recovery further depends on the initialization, on which coordinates may move, and on which candidate is returned. Matbench Discovery scores stability prediction rather than structural recovery (Riebesell et al., 2025), and softening analyses record where universal potentials fail on out-of-distribution configurations (Deng et al., 2025). We take the potentials but not their objective: their final measurements become selection features, and recovery is judged against the clean crystal.

Prior work supplies most of the machinery repair needs. It offers generative backbones, tool-mediated agents, structure-conditioned reconstruction, and machine-learning interatomic potentials. Their objectives stay separate: one family searches for new compositions, another completes designated missing sites, and relaxation seeks physically favorable configurations. None of them fixes composition, preserves surviving geometry, and commits to one output chosen from evidence observable at inference. We keep the target fixed to one given crystal, reporting ablations that aggregate over one protocol rather than isolating single modules.

3 METHOD

3.1 PROBLEM SETUP AND BASELINE

We represent an input crystal as $x_0 = (L_0, Z, U_0)$, where $Z = (z_1, \ldots, z_n)$ lists the atomic species and $U_0 \in [0, 1)^{n\times 3}$ holds fractional coordinates for n sites. Fractional coordinates are row vectors. The rows of $L_0 \in \mathbb{R}^{3\times 3}$ are therefore the lattice vectors. The procedure receives the corrupted structure and its formula and returns one structure $\hat{x}$. Composition, site count, and site identities are fixed; coordinates may move, and so may the lattice in selected routes. The paired clean structure $x^\star$ is withheld during repair and used only for later evaluation.

The *identity baseline* returns the input unchanged, $\hat{x}_{\text{base}} = x_0$. It runs no collision projection, no refinement route, and no selector. It measures how often the corrupted input already satisfies the matching criterion, not a module ablation. Relaxation-only instead follows one optimization path under a single choice of movable atoms, cell degrees of freedom, and convergence tolerance. Multi-branch repair keeps physical refinement but explores five such configurations from a shared geometry-conditioned initialization, then returns one output. Selection never touches the withheld clean structure.

3.2 Overview and Running Example

Our running example has two atoms at an unusually short periodic distance, while the surrounding structure stays informative. The collision measurement suggests a separation direction but neither the original coordinates nor the state of the cell. That signal builds one shared initialization (A), feeds the five CHGNet refinement routes (B), and helps a frozen ExtraTrees selector choose one structure (C). Figure 1 shows the workflow.

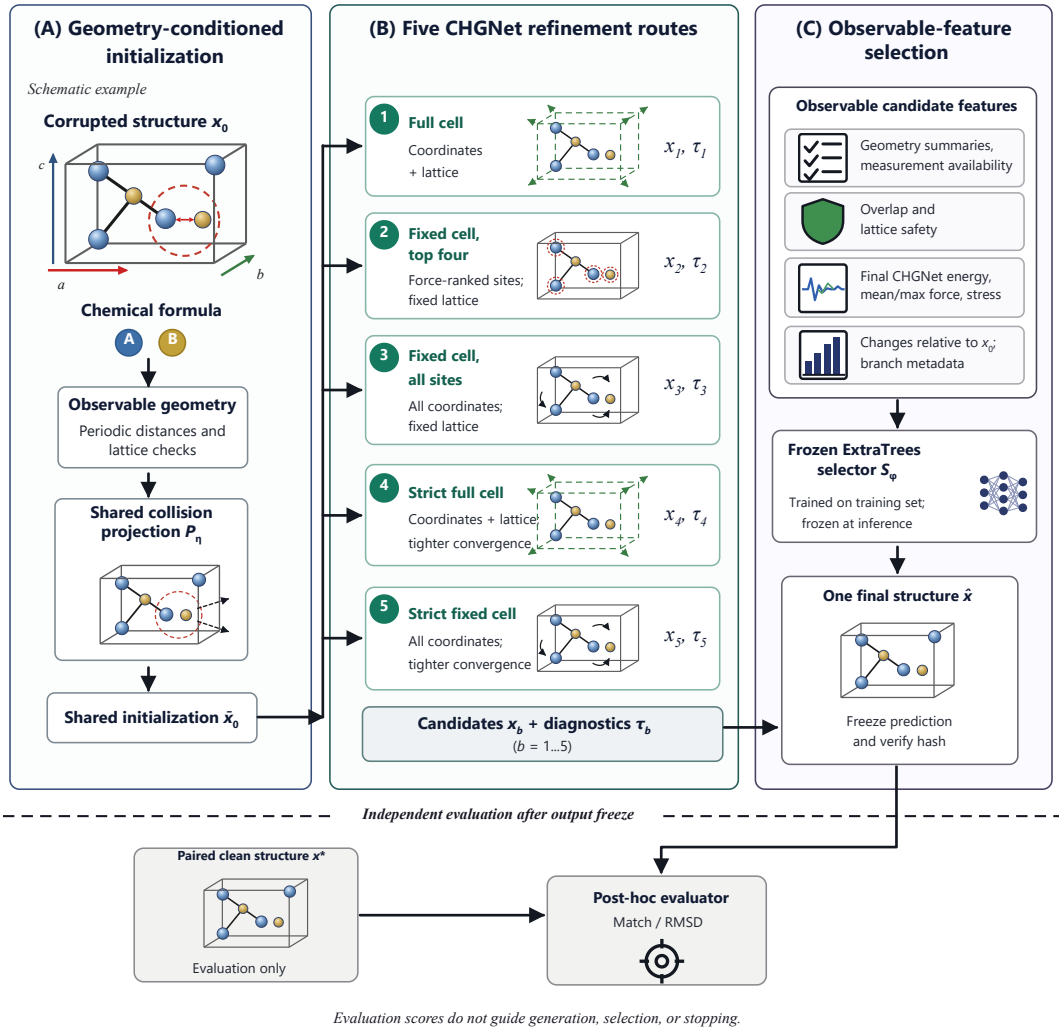

Figure 1: Workflow of multi-branch crystal repair. (A) Observable geometry guides a shared collision-projection initialization. (B) Five CHGNet refinement routes produce candidate structures and diagnostic summaries. (C) A frozen ExtraTrees selector uses candidate geometry, final-state physical measurements, and branch indicators to choose one final structure. Evaluation is post hoc and does not guide generation, selection, or stopping. RMSD denotes structure-matched root-mean-square deviation. Crystal miniatures illustrate the procedure schematically.

3.3 Geometry-Conditioned Candidate Generation

Observable geometry and shared initialization. Distances use pymatgen (Ong et al., 2013) on the current lattice, never clean coordinates; its periodic neighbor machinery searches a finite image shell, not all of $\mathbb{Z}^3$. For distinct sites i and j with fractional coordinates u_i and u_j, the minimum

periodic separation is

$$d_{ij}(x_0) = \min_{\mathbf{m}\in\mathbb{Z}^3} \|(u_i - u_j + \mathbf{m})L_0\|_2 , \tag{1}$$

We flag collisions with the cutoff $\delta_{ij} = \max\{0.55\,\text{Å}, 0.55(r_i + r_j)\}$, built from covalent radii r_i and r_j. We correct the closest colliding pair only, moving either endpoint along its observed separation direction. An input with several collisions gets one correction. Proposed distances mix distance statistics with covalent-radius priors, keep the atom inventory, and bound displacement by 4.0 Å. A fixed score combines overlap, lattice and local-distance anomalies, edit magnitude, a chemical-distance prior, and symmetry. We accept an improving proposal or keep the input, giving

$$\tilde{x}_0 = P_\eta(x_0), \tag{2}$$

where P_η is a bounded discrete proposal search with fixed settings η, not an idempotent projection. One proposal is chosen before any route relaxes. All five routes then share that initialization. In the running example, the pair direction narrows the search. Appendix B.1 gives the distances, score, and wrapping.

Complementary refinement routes. Let $\mathcal{B} = \{1, \ldots, 5\}$ index the five CHGNet refinement routes in Table 2. With CHGNet parameters θ, route-specific configuration λ_b, and refinement operator $\mathcal{R}_b$, each route produces

$$(x_b, \tau_b) = \mathcal{R}_b(\tilde{x}_0; \theta, \lambda_b), \qquad b \in \mathcal{B}, \tag{3}$$

where x_b is the final candidate and τ_b its diagnostic summary: input and candidate geometry summaries, final energy, mean and maximum force magnitudes, scaled stress norm (Appendix E), and any refinement error. CHGNet supplies the physical predictions (Deng et al., 2023). The route set is shared across corruption families, never chosen from the true corruption label.

Table 2: Five refinement routes sharing the same initialization obtained by collision projection of the corrupted structure. Step counts are maximum optimizer budgets, $f_{\max}$ is the force-convergence tolerance in eV/Å. Site-selection and cell constraints differ across branches.

Branch	Movable degrees of freedom	Max. steps	$f_{\max}$
Full cell	All coordinates and lattice	200	0.05
Fixed cell, top four	Coordinates of up to four force-ranked sites	80	0.10
Fixed cell, all sites	All coordinates; fixed lattice	80	0.10
Strict full cell	All coordinates and lattice	300	0.03
Strict fixed cell	All coordinates; fixed lattice	200	0.05

The force-ranked route measures forces at the shared initialization, fixes the up-to-four highest-force sites as its editable set, and moves only those sites. The all-site route relaxes every coordinate with the cell fixed, while both full-cell routes admit strain correction; the stricter variants use lower tolerances and larger budgets. In the running example, the routes separate the pair with and without cell adjustment. Fixed-cell and restricted-site constraints preserve selected degrees of freedom; full-cell refinement does not guarantee that every undamaged coordinate survives. Each route stops at its force tolerance or step cap, never at a hidden match.

3.4 Frozen Candidate Selection

For each route, the frozen ExtraTrees selector constructs

$$\mathbf{v}_b = \Phi(x_0, x_b, \tau_b, b) \in \mathbb{R}^{33}, \qquad p_b = S_\phi(\mathbf{v}_b), \qquad \hat{b} = \arg\max_{b\in\mathcal{B}} p_b, \quad \hat{x} = x_{\hat{b}}, \tag{4}$$

where S_ϕ is an ExtraTrees classifier (Geurts et al., 2006) with fitted parameters ϕ frozen after fitting, and p_b estimates the match probability. Balanced class weights rescale p_b monotonically, so the argmax over the five routes is unchanged and no calibration is needed. The route index enters $\mathbf{v}_b$ only through five one-hot indicators, although Φ takes it explicitly. τ_b supplies the physical measurements, their availability, and geometry summaries; the rest come from x_b and x_0. Appendix B.2 enumerates the 33 features: 21 geometry, physics and availability features, seven geometric-change

features, and five route indicators. Parsing, atom-count, and species-order checks stay outside this vector.

We train the classifier on the Train set of 200 inputs, each with five branch candidates, yielding 1,000 candidate-level training records. Each binary label indicates whether a training candidate matches its paired clean structure. Five-fold cross-validation is performed within the Train set, grouping rows by sample identifier so that all five candidates from one input remain in the same fold; the final classifier is trained on all 1,000 records and then frozen before test-set evaluation. At inference it sees only the input, diagnostic summaries, and route index; clean coordinates, target-derived descriptors, matching scores, and corruption annotations are excluded. Ties follow the fixed order in Table 2. In the running example, residual overlap, displacement from the input, and final-state measurements decide the returned route.

3.5 Inference and Evaluation Boundary

We fix the final prediction for each input retained by the inclusion rules of Appendix A before reference-based assessment. An independent evaluator scores each prediction against its clean counterpart. Neither Match nor RMSD guides generation, selection, or stopping.

Candidate geometry and final-state physical measurements serve as selection signals, whereas Match and RMSD measure recovery against the reference structures. We report them separately. Failed or missing predictions are not counted as matches.

4 Experiments

We evaluate multi-branch repair on paired corrupted and clean structures spanning lattice distortion, coordinate noise, atom overlap, and coordination defects. The main comparators are the identity baseline and relaxation-only. Family and paired-outcome analyses locate gains and regressions, and module ablations test the projection, the five routes, and the selector.

4.1 Experimental Setup

4.1.1 Datasets

Our data come from MP-20 crystals through the processed `data/basic` release of Crystal-Text-LLM (Gruver et al., 2024). MP-20 derives from the Materials Project (Jain et al., 2013). Clean sources are the release's 9,046 test-split records, each paired with a synthetically corrupted structure. Appendix A gives the source corpus, perturbation settings, and inclusion criteria.

The Train set contains 200 inputs drawn from the same clean-source pool, each producing five candidates and 1,000 labelled candidate-level training records. Five-fold cross-validation is performed within the Train set, and the final selector is frozen before evaluation on the test set. Material-level disjointness between the Train set and the 2,520 test inputs is unverified; this provenance limitation is documented in Appendix D. The test set holds 2,520 corrupted inputs over 580 material identifiers. Forty of these identifiers also occur in a historical screening collection, corresponding to 360 test-set records. Appendix E reports a sensitivity analysis excluding this overlap. Throughout, N denotes corrupted–clean pairs, not unique crystals.

Table 3: Training and test data for crystal repair. Each Train-set input yields five candidate records for ExtraTrees training; these records are not independent crystal inputs.

Data role	Inputs	Candidate/evaluation records
Train set	200	1,000 candidate rows
Test set	2,520	2,520 final outputs

4.1.2 Compared Methods and Inference Settings

The *identity baseline* (row "Baseline" in the tables) returns the corrupted structure x_0 unchanged and measures how often the input already satisfies the matching criterion. *Relaxation-only* applies

the full-cell CHGNet route to the unmodified corrupted structure, using at most 200 steps and a maximum-force threshold of 0.05 eV/Å; it uses neither collision projection nor learned selection. *Multi-branch repair* applies the shared collision projection and the five CHGNet refinement routes in Table 2, then selects one output with the frozen selector. All comparisons share the input records, the one-output requirement, and the StructureMatcher configuration. Neither candidate generation nor selection accesses the paired clean structure. Appendix B gives configuration details.

4.1.3 Evaluation Metrics

Match is the primary recovery metric. For N evaluated records and M successful structure matches, it equals $100M/N$. We use pymatgen StructureMatcher with `ltol=0.2`, `stol=0.3`, and `angle_tol=5.0`. Match therefore measures agreement under fixed tolerances, not equality of serialized coordinates. RMSD describes geometric deviation on the finite matched-output subset, with its contributor count reported separately. Parseability, composition preservation, and Validity complement these metrics but do not establish target recovery or first-principles stability. Following the structural-validity convention used in crystal-generation evaluation (Jiao et al., 2023), we count an output as valid when it parses as a pymatgen structure and the minimum-image distance between each pair of distinct sites exceeds 0.5 Å. The Validity rate is the fraction of the N evaluated outputs satisfying this predicate; it does not use the clean structure, composition agreement, CHGNet convergence, energy, force, or stress thresholds. Appendix C gives the metric definitions and aggregation scope.

4.2 Main Results: Single-Output Structural Recovery

Table 4 reports the main comparison. Multi-branch repair matches 1,924 of 2,520 inputs (76.35%), compared with 755 (29.96%) for the identity baseline, an absolute improvement of 46.39 percentage points. Nearly a third of corrupted inputs already lie inside the matcher tolerance, and 46.03% do on the lattice family, so part of the recoverable headroom is a tolerance effect rather than recovered geometry.

Table 4: Single-output recovery on the test set compared with the unchanged input. Matched counts and rates use the same N. Δ is the absolute improvement in percentage points.

Method	N	Matched	Match (%)	Δ
Baseline	2,520	755	29.96	–
Relaxation-only	2,520	1,109	44.01	+14.05
Multi-branch repair (ours)	2,520	1,924	**76.35**	+46.39

Relaxation-only is the stronger comparator at 44.01%, but the comparison is not compute-matched. The five CHGNet refinement routes allow up to 860 optimizer steps (200 + 80 + 80 + 300 + 200 in Table 2), against 200 for relaxation-only.

4.3 Performance Across Defect Families

Table 5: Defect-family recovery on the test set under the unchanged matcher. Baseline and repair cells give matched counts followed by Match (%). Δ denotes percentage-point improvement calculated from the counts. Each family contains 630 records.

Defect family	N	Baseline: M (%)	Repair: M (%)	Δ
Coordinate noise	630	227 (36.03)	393 (62.38)	+26.35
Coordination defect	630	198 (31.43)	391 (62.06)	+30.63
Lattice distortion	630	290 (46.03)	522 (82.86)	+36.83
Atom overlap	630	40 (6.35)	618 (98.10)	+91.75

Table 5 shows that overlap repair is the most reliable family, at 98.10%. Constructed separations of 0.3–0.7 Åfall below the collision cutoff $\delta_{ij} \geq 0.55$ Åfor all but the lightest pairs, so this family is

nearly separable by construction and should not be read as evidence that collision geometry transfers. Lattice recovery reaches 82.86%, and coordinate and coordination repair 62.38% and 62.06%. Family rates do not control for corruption severity.

4.4 Paired Recovery and Error Analysis

A *gain* is a transition from non-match to match, a *harm* the reverse, and *unchanged* preservation of the match status. Table 6 accounts for all test-set inputs. The test set contains 1,218 gains and 49 harms, yielding a net increase of 1,169 matches. The 49 harms affect 6.49% of the 755 already-matching inputs.

Table 6: Paired changes in binary match status relative to the baseline. Gain, harm, and unchanged sum to N in each row; gain minus harm equals the increase in matched count. Unchanged denotes the match status, not necessarily an unchanged structure.

Scope	N	Gain	Harm	Unchanged
Test set	2,520	1,218	49	1,253

On the test set, parseability and composition preservation are both 100%, while Validity is 98.02%. Enabling projection raises Validity from 73.69% to 98.02%, and about 50 outputs of the complete method fail the 0.5 Åminimum-separation predicate. Unparseable entries are excluded at construction (Appendix A), so parseability and composition are interface-level checks rather than discriminative recovery results. Validity also separates from Match (76.35%): a well-formed output with no sub-threshold periodic separation need not recover the intended structure. Matched-output RMSD is conditional: its average is not an all-record reduction in geometric error.

4.5 Module Ablations and Relaxation Control

We ablate collision projection, route diversity, and learned selection under the same protocol. *Without collision projection* passes the unmodified corrupted structure to the five routes, retaining the frozen selector. *Without branch diversity* retains projection but returns one full-cell route. *Without learned selection* retains the five candidates and replaces the selector with a fixed score over force, stress, and energy per atom, taking the lowest value. Relaxation-only removes all three, serving as a baseline rather than a single-module ablation. The Full row of Table 7 and the Full column of Table 8 refer to the complete combination, which is the reference for each removal.

Table 7: Module ablations on the test-set stratum ($N = 2,520$). Δ is the Match difference from Full, in percentage points. Parse, Comp., and Validity are percentages over the reported input scope. RMSD is averaged over finite matched outputs, with the contributor count in parentheses. The complete method provides the reference for each removal; relaxation-only is a baseline.

Method	Match	Δ vs. Full	Parse / Comp. / Validity	RMSD (valid N)
Full method	1,924/2,520 (76.35)	–	100 / 100 / 98.02	0.02747 (1,924)
Without collision projection	1,337/2,520 (53.06)	-23.29	100 / 100 / 73.69	0.03470 (1,337)
Without branch diversity	1,640/2,520 (65.08)	-11.27	100 / 100 / 97.54	0.02795 (1,640)
Without learned selection	1,640/2,520 (65.08)	-11.27	100 / 100 / 97.90	0.02929 (1,640)
Relaxation-only	1,109/2,520 (44.01)	-32.34	100 / 100 / 71.31	0.03377 (1,109)

Removing collision projection reduces test-set Match from 76.35% to 53.06% (Table 7), a 23.29-point decrease. Both other controls return exactly 1,640 matched inputs (65.08%): one keeps a single route, the other keeps five under a fixed score, so neither isolates a component on its own. The 11.27 points require the frozen ExtraTrees selector acting on the diverse candidate set, and route diversity alone is not separately identified. The single-route control keeps that route with projection and reaches 65.08%, while relaxation-only relaxes it without projection and reaches 44.01%. The 21.07-point gap therefore isolates initialization at matched compute and with no selector, up to the development-based choice of that single route. Appendix E reports exploratory paired-bootstrap ranges; because the resampling units and cluster sizes differ, these ranges are used as descriptive robustness checks rather than significance intervals.

Table 8 reveals distinct failure patterns. Removing projection reduces overlap Match from 98.10% to 10.95%; relaxation-only reaches 7.30% there, against 90.16% for the single-route control with projection. Coordinate recovery falls from 62.38% to 42.38% and 43.65% under the two controls, and coordination from 62.06% to 46.35% and 46.67%. Lattice recovery varies least, from 82.86% to 80.63–82.86%. On the lattice family the two agree exactly at 522, plausibly because the detector rarely fires there and the projection is the identity. The coordinate and coordination rows differ (393 vs 376 and 391 vs 370), so it does fire on some non-overlap inputs, which bounded rejection during construction cannot fully exclude; we report no per-family firing rates that would settle this. Two controls fall slightly below relaxation-only here, 513 and 508 against 514, so neither the extra routes nor the selector helps uniformly. Initialization matters most for overlap repair, and learned selection matters most elsewhere. We report no per-route match rate and no share of inputs with a matching candidate, so 76.35% cannot be set against this candidate set's ceiling.

Table 8: Test-set damage-family Match counts. Columns report Baseline, Full, and the four module controls; each family contains $N = 630$ records.

Family	Identity M (%)	Full M (%)	w/o proj.	w/o branch	w/o select.	Relax. only
Lattice	290 (46.03)	522 (82.86)	522 (82.86)	513 (81.43)	508 (80.63)	514 (81.59)
Coordinate	227 (36.03)	393 (62.38)	376 (59.68)	267 (42.38)	275 (43.65)	261 (41.43)
Overlap	40 (6.35)	618 (98.10)	69 (10.95)	568 (90.16)	563 (89.37)	46 (7.30)
Coordination	198 (31.43)	391 (62.06)	370 (58.73)	292 (46.35)	294 (46.67)	288 (45.71)

Disabling projection also shifts the candidate distribution seen by the frozen selector, so its effect includes that interaction. One route cuts the candidate budget, so no control is compute-matched. The ablations show practical value under the evaluated protocol, not a full separation of interactions.

5 Conclusion and Future Work

On the test set, multi-branch repair reaches 76.35% Match (1,924/2,520), against 44.01% (1,109/2,520) for relaxation-only and 29.96% (755/2,520) for the unchanged-input baseline. Removing collision projection costs 23.29 percentage points, concentrated on atom overlap, while the two 65.08% controls show that the observed 11.27-point gap is associated with learned selection over a diverse candidate set rather than separately identifying route diversity. These findings are scoped to one MP-20-derived corpus, one pretrained CHGNet potential, and four synthetic single-family corruption types; material-level disjointness between the 200-input Train set and the test set remains unverified, the complete method uses up to 860 optimizer steps versus 200 for relaxation-only, and the current study does not report an external repair comparator or the per-input candidate ceiling. Future work will evaluate budget-matched search on source-disjoint materials and experimentally observed defects, and will use first-principles calculations to assess the physical stability of repaired candidates.

AI Use Statement

Generative artificial intelligence (AI) tools were used only to assist manuscript preparation, language polishing, and code-oriented workflow planning. They were not used as a source of scientific ground truth. All experimental results, metrics, tables, and conclusions reported in this paper are produced by the authors' computational pipeline and are subject to the validation procedures described above.

Ethics Statement

This study uses synthetically corrupted inorganic crystal structures derived from publicly available materials data and involves no human participants, personal data, or biological subjects. The repaired structures are computational candidates; scientific or engineering use should be preceded by appropriate first-principles, experimental, and expert materials-science validation.

REFERENCES

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, et al. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163:184110, 2025. doi: 10.1063/5.0297006.

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, 2023. doi: 10.1038/s42256-023-00716-3.

Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A. Persson, and Gerbrand Ceder. Systematic softening in universal machine learning interatomic potentials. *npj Computational Materials*, 11:9, 2025. doi: 10.1038/s41524-024-01500-6.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *International Conference on Learning Representations*, Vienna, Austria, 2024. OpenReview.net. doi: 10.48550/arXiv.2402.04379.

Yusei Ito, Yuta Suzuki, Tomoya Murata, and Masaki Adachi. Lapidaryengine: Fully conversational crystal generation. *arXiv preprint arXiv:2606.14215*, abs/2606.14215, 2026. doi: 10.48550/arXiv.2606.14215.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1:011002, 2013. doi: 10.1063/1.4812323.

Shuyi Jia, Chao Zhang, and Victor Fung. LLMatDesign: Autonomous materials discovery with large language models. arXiv preprint arXiv:2406.13163, 2024.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17464–17497, New Orleans, LA, USA, 2023. Curran Associates, Inc. doi: 10.52202/075280-0767.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, abs/2303.17651, 2023. doi: 10.48550/arXiv.2303.17651.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. Flowmm: Generating materials with riemannian flow matching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35664–35686, Vienna, Austria, 2024. PMLR.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028.

Timo Reents, Arianna Cantarella, Marnik Bercx, Pietro Bonfà, and Giovanni Pizzi. Score-based diffusion models for accurate crystal-structure inpainting and reconstruction of hydrogen positions. *npj Computational Materials*, 12:203, 2026. doi: 10.1038/s41524-026-02090-1.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, 2025. doi: 10.1038/s42256-025-01055-1.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, abs/2303.11366, 2023. doi: 10.48550/arXiv.2303.11366.

Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M. Wood. Flowllm: Flow matching for material generation with large language models as base distributions. In *Advances in Neural Information Processing Systems*, volume 37, Vancouver, Canada, 2024. Curran Associates, Inc. doi: 10.52202/079017-1464.

Izumi Takahara, Teruyasu Mizoguchi, and Bang Liu. Accelerated inorganic materials design with generative ai agents. *Cell Reports Physical Science*, 6:103019, 2025. doi: 10.1016/j.xcrp.2025.103019.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, Online, 2022. OpenReview.net. doi: 10.48550/arXiv.2110.06197.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Daniel Z"ugner, Andrew Fowler, Tian Xie, et al. Mattersim: A deep learning atomistic model across elements, temperatures, and pressures. *arXiv preprint arXiv:2405.04967*, abs/2405.04967, 2024. doi: 10.48550/arXiv.2405.04967.

Claudio Zeni, Robert Pinsler, Daniel Zugner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbe, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025. doi: 10.1038/s41586-025-08628-5.

Peichen Zhong, Xinzhe Dai, Bowen Deng, Gerbrand Ceder, and Kristin A. Persson. Crystal structure prediction with host-guided inpainting generation and foundation potentials. *Materials Horizons*, 12:9669–9678, 2025. doi: 10.1039/D5MH00774G.

Lianhao Zhou, Hongyi Ling, Keqiang Yan, Kaiji Zhao, Xiaoning Qian, Raymundo Arróyave, Xiaofeng Qian, and Shuiwang Ji. Toward greater autonomy in materials discovery agents: Unifying planning, physics, and scientists. arXiv preprint arXiv:2506.05616, 2025.

A Evaluation Data and Coverage

A.1 Source corpus and paired inputs

Clean structures come from MP-20, a Materials Project-derived collection (Jain et al., 2013), through the `data/basic` release of Crystal-Text-LLM (Gruver et al., 2024). The *repository source corpus* holds 45,229 records: 27,136 training, 9,047 validation, and 9,046 test. Construction uses the 9,046 test-split entries as *held-out source records* and keeps source identifiers for traceability. Each entry carries metadata and a CIF, parsed by pymatgen into the paired clean target.

The repository counts differ from the 2,520 records evaluated here: one source crystal can contribute several corrupted records, and test-split membership does not by itself establish source-level independence.

Each record pairs a corrupted structure x_0 with a clean counterpart $x^\star$. Only `sample_id`, `formula`, and `before_structure` reach the repair procedure; the identifier links records to outputs without supplying a clean structure. Corruption labels, construction metadata, and the clean structure are withheld from repair.

A.2 Controlled corruptions and inclusion rules

One controlled perturbation is applied to each clean source crystal, preserving elemental species and atom ordering where the operation permits. Four corruption families are used:

- **Lattice distortion**: change one or more lattice lengths, shift a lattice angle, or apply independent anisotropic length scales; fractional coordinates stay fixed.
- **Coordinate noise**: add zero-mean Gaussian noise with standard deviation $\sigma = 0.02$ to fractional coordinates and wrap them into $[0, 1)$.
- **Atom overlap**: select two atoms and move one toward an anchor atom so that their separation lies between 0.3 and 0.7 Å.
- **Coordination defect**: choose a coordination center and perturb one or two nearest neighbors by pulling them away, pushing them closer, or twisting the local shell. Displacement factors are retained in the construction metadata.

The base construction uses random seed 42. Entries that fail parsing, exceed the atom-count filter, or cannot produce a valid perturbation are excluded and tracked as non-evaluable outcomes. Construction also rejects invalid lattices and unintended collisions outside the atom-overlap family, with bounded sampling attempts.

B Supplementary Inference Configuration

B.1 Shared initialization and refinement routes

Collision projection runs once per input and its output is shared by all five CHGNet refinement routes; the restricted-site route picks its editable set once, from forces at initialization. Table 9 gives branch names and bounded optimizer budgets.

Table 9: Branch identifiers and reported refinement budgets. $f_{\max}$ is the force-convergence tolerance in eV/Å. The configurations agree with Table 2; inference-time branch choice does not use the true corruption family.

Branch identifier	Paper name	Max. steps	$f_{\max}$
`full`	Full cell	200	0.05
`fixed_top4`	Fixed cell, top four	80	0.10
`fixed_all`	Fixed cell, all sites	80	0.10
`strict_full`	Strict full cell	300	0.03
`fixed_all_strict`	Strict fixed cell	200	0.05

Collision proposal search. For each distinct pair, δ_{ij} is the collision cutoff of Section 3. The overlap penalty is $O(x) = \sum_{i<j}[\max(0, 1 - d_{ij}(x)/\delta_{ij})]^2$. Covalent radii come from ASE, with an element radius and then 0.85 Å as fallbacks. If $O(x_0) = 0$, initialization leaves the corrupted structure unchanged. Otherwise the closest pair fixes a periodic separation direction, and we move either endpoint to: 1.05, 1.30, and 1.60 times its cutoff; the 0.05, 0.15, 0.25, 0.40, and 0.60 quantiles of other pair distances in the corrupted structure; the first two neighbor distances of either endpoint; and 0.92, 1.00, and 1.08 times their covalent-radius sum. We keep distances above the current separation, deduplicate at six decimals, and reject proposals whose separation change exceeds 4.0 Å. Proposals preserve lattice lengths, angles, and species order. Fractional coordinates are wrapped periodically, with Cartesian displacement capped at 4.0 Å. The two caps differ: one bounds the separation change, the other Cartesian displacement.

For a proposal x relative to the current corrupted structure x_0, the initialization score is

$$J(x; x_0) = 8O(x) + 1.5L(x) + C(x) + 0.05A(x; x_0) + 1.8H(x) - 0.05G(x; x_0). \quad (5)$$

Here L penalizes an ill-conditioned lattice and extreme volume per site, C local neighbor-distance outliers, A edit magnitude and the changed-site fraction, H the log-distance discrepancy from the covalent-radius sum, and G relative changes in symmetry-operation counts at spatial tolerances 0.03, 0.05, 0.10, and 0.20 Å with a 5° angle tolerance, each normalized by the initial count or one, whichever is larger. The terms use only the corrupted structure and the proposal. We reject proposals that increase O, take the lowest-scoring strict improvement over the unchanged input, and otherwise retain it. One pass yields the shared starting structure.

B.2 Selection features and information access

The selector receives the 33 ordered features of Table 10. Displacement features compare final and input fractional coordinates after periodic wrapping, converted with the input lattice. A changed site moves more than 10^{-3} Å; a large displacement exceeds 0.2 Å. Lattice features average absolute relative length and absolute angle changes. Initialization and refinement status are auxiliary diagnostics, not classifier features. Table 11 separates this interface from evaluation.

Table 10: The 33 scalar selector features. Physical measurements are taken at candidate endpoints; geometry changes are relative to the input. Interface checks on parsing and atom identity are separate from the feature vector.

Feature group	Count	Ordered contents
Measurement availability	2	Fixed-cell indicator; availability of candidate physical measurements.
Candidate geometry	5	Overlap, lattice, and local-distance penalties; minimum separation; volume per site.
Candidate physics	6	Energy per atom; mean and maximum force; scaled stress norm; fixed public score; stress-to-mean-force ratio.
Input geometry	5	Input overlap, lattice, and local-distance penalties; minimum separation; volume per site.
Derived summaries	3	Candidate-minus-input volume per site and minimum separation; sum of candidate mean force and scaled stress norm.
Geometric changes	7	Changed-site count; mean, maximum, and root-mean-square displacement; mean absolute relative lattice-length change; mean absolute angle change; large-displacement site count.
Branch identity	5	One-hot indicators in Table 2 order.

Supervision and frozen decision rule. We train an `ExtraTreesClassifier` with 240 estimators, a minimum of two samples per leaf, balanced class weights, and random seed 20260924. The Train set contains 200 inputs, each producing five candidates and 1,000 labelled candidate-level training records; each label records whether that candidate matches its paired clean structure. Five-fold cross-validation is performed within the Train set, grouping rows by sample identifier so that all five candidates from one input remain in the same fold; the final classifier is trained on all 1,000 records. No cross-validated selection accuracy is reported, nor is the label prevalence among the

Table 11: Information roles during repair and post-hoc evaluation. The boundary concerns the current evaluation sample; it is not a claim that selector training or pretrained-potential training uses no supervision.

Stage	Available information	Information withheld
Input and projection	Corrupted structure, formula, observable collision geometry	Clean coordinates, target-specific repair operations, corruption labels
CHGNet refinement	Corrupted structure after collision projection, fixed branch constraints and budgets, predicted physical quantities	Paired clean structure and its matching scores
Frozen selection	Candidate geometry, final-state physical measurements, geometric changes, branch indicators	Clean-structure fingerprints, answer-revealing retrieval, Match and RMSD against the clean structure
Independent evaluation	Frozen prediction and paired clean structure; labels for subgroup analysis	No return path to alter candidate generation, selection, or stopping

1,000 rows. The effective sample, 200 inputs for 33 features, is small. Inference takes the largest predicted matching probability, breaking ties in branch order: full, fixed top four, fixed all, strict full, strict fixed. These matching labels supervise selector training; test-set targets reach only the evaluator.

Missing measurements and interface checks. Missing or non-finite features are encoded as -999 with a measurement-availability indicator. A CHGNet refinement error returns the geometry-conditioned initialization and exposes the failure as an explicit candidate status. Parsing, site count, and ordered species are interface checks, not classifier features. Candidate identifier sets must match the public input before selection, and the frozen decision rule operates on all five candidate states, including fallbacks.

C Metric Definitions and Supplementary Outcomes

C.1 Fixed matching protocol and denominators

Let $q \in \{1, \dots, N\}$ index records in one declared scope, and let $\mathcal{M}$ be the fixed StructureMatcher predicate. Define the binary statuses of the baseline and repaired outputs as

$$c_q = \mathcal{M}(x_{0,q}, x_q^\star), \qquad r_q = \mathcal{M}(\hat{x}_q, x_q^\star), \tag{6}$$

where $x_{0,q}$, $\hat{x}_q$, and $x_q^\star$ are the input, selected repair, and clean counterpart for record q. The matched counts are $M_{\text{base}} = \sum_q c_q$ and $M_{\text{ours}} = \sum_q r_q$. For either method, $\text{Match} = 100M/N$, and the absolute improvement is $100(M_{\text{ours}} - M_{\text{base}})/N$ percentage points. Improvements come from integer counts before rounding, not from separately rounded percentages.

Validity follows the structural component of the validity metric used by DiffCSP (Jiao et al., 2023), rather than a stability or recovery criterion. For an evaluated output y_q, define

$$v_q = \mathbf{1}\{y_q \text{ parses and } d_{ij}(y_q) > 0.5\,\text{Å for every } i < j\},$$

where d_{ij} is the minimum-image distance defined in Section 3. This is implemented by `Structure.is_valid(tol=0.5)`; a single-site structure satisfies the pairwise condition vacuously. Thus $\text{Validity} = 100 \sum_q v_q/N$; parse failures count as invalid, while composition preservation and CHGNet energy–force–stress measurements are reported separately. This predicate only excludes sub-threshold periodic separations and should not be interpreted as a first-principles stability certificate.

The fixed matcher settings are:

```
ltol=0.2, stol=0.3, angle_tol=5.0
primitive_cell=False, scale=True, attempt_supercell=False
```

Matching therefore means agreement under this tolerance-based protocol, not coordinate-wise identity or an independently established physical-stability certificate.

The scope size N, matched count M, and contributor count N_{RMSD} have different meanings. In these tables RMSD is averaged over the matched outputs of each row. Here N_{RMSD} equals that row's matched count, so the summary covers the matched subset only. Normalization and aggregation beyond this are not supplied. Unmatched outputs get no error, and no all-record reduction is claimed.

C.2 Complete paired-status accounting

The four status counts follow from the reported baseline matches, gains, and harms. Specifically,

$$n_{11} = \sum_q c_q r_q = M_{\mathrm{base}} - n_{10}, \qquad n_{01} = \sum_q (1 - c_q) r_q, \tag{7}$$

$$n_{10} = \sum_q c_q (1 - r_q), \qquad n_{00} = N - n_{11} - n_{01} - n_{10}, \tag{8}$$

where n_{11} denotes both methods matching, n_{01} a gain, n_{10} a harm, and n_{00} neither method matching. Consequently, $n_{11} + n_{00}$ is the unchanged match-status count and $n_{01} - n_{10} = M_{\mathrm{ours}} - M_{\mathrm{base}}$. Table 12 is an algebraic decomposition of Table 6, not a new run or an independent replication.

Table 12: Paired-status counts derived from the reported matched, gain, and harm totals. The four status columns sum to N. Both-match and neither-match values are computed by the identities above; they are not additional measurements.

Scope	N	Both match	Repair only	Identity only	Neither matches
Test set	2,520	706	1,218	49	547

D Evaluation Boundary and Reproducibility Scope

Fixing the prediction set before clean structures are used separates inference from reference-based assessment. Integrity checks verify unique sample identifiers, one final structure per input, schema validity, and unchanged output content via a digest. The stated checks were satisfied for the test-set outputs and establish output integrity and the specified information-access boundary. Training-data independence, absence of pretraining exposure, and exhaustive runtime isolation require separate evidence.

The supplementary configuration specifies the collision search, feature vector, training-label construction, fitted classifier, and missing-measurement policy. Reproduction additionally requires versioned source/split manifests and collection-to-source mappings, severity-specific corruption parameters, checkpoint and dependency hashes, historical candidate-cache provenance, and complete failure/retry and compute accounting.

E Module Ablation Protocol and Sensitivity

Multi-branch repair combines collision projection, five CHGNet routes, and the selector trained on the Train set; the ablation runs use the same 2,520 test-set inputs, StructureMatcher, and one-final-structure protocol. The no-projection control disables only the shared projection. The no-diversity control fixes the full-cell 200-step route chosen during development, not per input. The no-learned-selection control retains all five candidates and takes the lowest-scoring one with the fixed score

$$s_b = \mathrm{mean_force}_b + 0.15\ \mathrm{stress_norm}_b + 0.02\ \mathrm{energy_per_atom}_b, \tag{9}$$

Mean force and energy per atom enter this score as numerical values in eV/Å and eV/atom, respectively. CHGNet predicts a stress tensor $\boldsymbol{\sigma}_b^{\mathrm{CHGNet}}$ in GPa. The implemented calculator sets `stress_weight=0.1`, so the stress feature is

$$\mathrm{stress_norm}_b = 0.1 \left\| \boldsymbol{\sigma}_b^{\mathrm{CHGNet}} / \mathrm{GPa} \right\|_{\mathrm{F}},$$

where $\| \cdot \|_{\mathrm{F}}$ is the Frobenius norm. The coefficients in Eq. (9) are empirical weights on these numerical features, not coefficients defining a physical energy. This stress scaling is also used

by the shared relaxation calculator; it differs from the standard GPa-to-eV/Å^3 conversion factor $1/160.2177$. The reported results retain the implemented scaling, and sensitivity to the standard conversion has not been evaluated. Ties are resolved in the order `full`, `fixed_top4`, `fixed_all`, `strict_full`, `fixed_all_strict`. Relaxation-only applies the full-cell 200-step, $f_{\max} = 0.05$ CHGNet route to the unmodified corrupted structure, without collision projection, alternative routes, or learned selection. It is the full-cell candidate computed under the no-projection control on the same structure; no second run is made.

All five conditions produce exactly one output per input. Table 13 summarizes the test-set Match results and supporting values; it restates the same measurements with explicit intervention definitions and is not an independent replication run. Parseability and composition use all records in scope, while RMSD uses finite matched outputs with the displayed contributor count.

Table 13: Supplementary module-ablation summary on the test-set stratum. The values duplicate the main-text ablation result for reproducibility and include the exact intervention definitions.

Method	Match	Δ Full	Parse / Comp. Validity	RMSD valid N
Full method	1,924/2,520 (76.35)	–	100 / 100 / 98.02	0.02747 (1,924)
Without collision projection	1,337/2,520 (53.06)	-23.29	100 / 100 / 73.69	0.03470 (1,337)
Without branch diversity	1,640/2,520 (65.08)	-11.27	100 / 100 / 97.54	0.02795 (1,640)
Without learned selection	1,640/2,520 (65.08)	-11.27	100 / 100 / 97.90	0.02929 (1,640)
Relaxation-only	1,109/2,520 (44.01)	-32.34	100 / 100 / 71.31	0.03377 (1,109)

The 2,520 test records contain repeated corruption instances grouped under 580 sample identifiers. An exploratory paired bootstrap resamples these identifier groups within three fixed dataset-membership strata: 40 identifiers with four instances each, 500 with four each, and 40 with nine each. All instances of a sampled identifier and all method outcomes are kept together. With 20,000 replicates and seed 20260919, the 95% percentile ranges for the Match difference (ablation minus Full) are $[-26.59, -20.00]$ pp for no projection, $[-12.78, -9.80]$ pp for no branch diversity, $[-12.74, -9.84]$ pp for no learned selection, and $[-35.60, -29.09]$ pp for relaxation-only.

A separate collection-level sensitivity analysis resamples the 17 test collections with replacement and computes each replicate from the summed matched counts and summed record counts. With the same replicate count and seed, its corresponding ranges are $[-24.56, -22.65]$, $[-12.82, -9.60]$, $[-12.67, -9.74]$, and $[-33.88, -30.91]$ pp. These narrower projection and relaxation-only ranges reflect a different resampling unit: entire collections with shared identifiers, rather than identifier groups. Neither collection count nor record count alone determines the interval width. We use both analyses as descriptive sensitivity checks; source-level independence and exchangeability within the resampling strata are necessary for a population-level confidence interpretation. Identifier grouping alone does not establish source-structure deduplication or independent material-level replication.

The 2,520 test inputs correspond to 580 sample identifiers; 40 identifiers also occur in the earlier screening collection and account for 360 test records. Excluding them leaves multi-branch repair at 1,624/2,160 = 75.19%. Backing out the excluded stratum gives 300 of 360 matched, or 83.33%, so removed records are easier than retained ones. On the retained records, the Match losses relative to Full are 22.78, 11.30, 11.39, and 32.04 percentage points for no projection, no branch diversity, no learned selection, and relaxation-only, respectively. This post-hoc exclusion checks sensitivity to the identified screening overlap; it does not establish independence from all development data.