READING IS NOT VERIFYING: AUDITING RELATION-CONDITIONED EVALUATORS FOR SELECTION AND GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluation scores for relation-conditioned generation may respond to a declared relation yet prefer outputs that violate it. We separate these as relation sensitivity and relation completeness and audit each: one intervention changes the declared relation at fixed output, the other breaks the relation at fixed declaration. In covalent protein–ligand structure prediction, a transferred contact score treats the required bond as a clash and accepts only 1.3% of 4,587 deposited complexes. A public pipeline already ships an exclusion for it; a CASP16 entry demotes predictions failing such checks. Excluding graph-local bonded neighbours raises acceptance to 95.8% while keeping sensitivity to distal clashes, yet withholding the false penalty adds no test that the bond exists: for this capped contact-score family, we prove mask-only exclusion relation-incomplete under large ligand translations. The repaired score prefers a broken bond under displacement on 65 of 272 frozen predictions. Bounded rigid-body search finds bond-violating solutions on 93.0% of tasks under the original score and 30.5% under the repair. Despite that gap, the repair improves controlled guidance, preserving 197 of 200 bonds against 0 of 200 unrepaired. Selection depends on the pool: correct-bond selection rises from 36.5% to 73.9% in mixed six-candidate pools, but a preregistered within-generator test does not establish an effect. Five further panels probe the audit's limits. On HumanEval+ a text-similarity score selects a program with explicit failing-test labels on 36.1% of eligible tasks, and an English–Chinese replication gains 18.5 percentage points in terminology compliance under checker-gated selection. Repairing a mismatch can therefore improve decisions without ensuring the requirement is met, so sensitivity and verification need separate tests before a score drives selection or guidance.

1 INTRODUCTION

A validity checker in wide use accepts only 1.3% of 4,587 deposited covalent protein–ligand complexes. It is not broken: it applies a non-bonded contact rule, and the bond the task requires is a contact. Practitioners have noticed: one public pipeline documents the mismatch and ships an exclusion for it (BioPipelines contributors, 2026), and a CASP16 entry demotes predictions that fail such checks (Morehead et al., 2026). But excluding the pair only withholds the false penalty; it adds no test that the bond formed. That asymmetry is this paper's subject. A score can respond to a stated requirement without checking that the output meets it; once selection or guidance optimizes such a score, the missing check changes which outputs are returned.

A relation-conditioned evaluator can therefore fail in two ways, and neither implies the other: it can fail to *read* the requirement, or fail to *verify* that the output meets it. Reading asks whether changing the declared relation changes the score; verifying asks whether a violating output can tie or outrank a satisfying one.

A minimal example. Let an output Z be a set of edges and require $E = \{A-B\}$. Consider $s(Z, E) = 1/(1 + |Z \setminus E|)$, which penalizes edges outside E. For $Z = \{A-B\}$, changing the requirement to $\{A-C\}$ lowers the score from 1 to $1/2$: it reads the requirement. Yet the empty graph also scores 1 under the original requirement, despite lacking the required edge. A top-scoring

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.

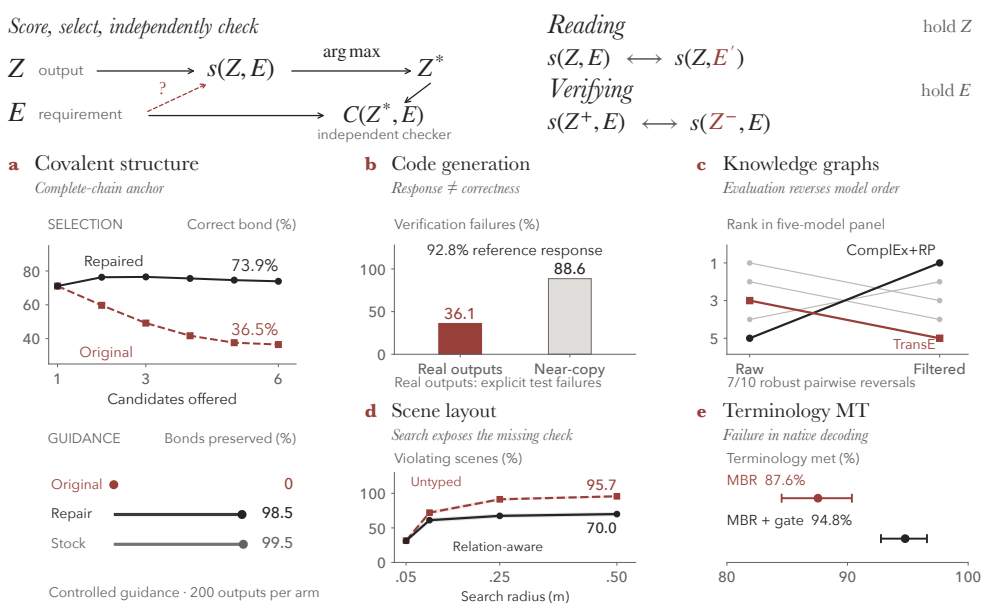

Figure 1: **A molecular case study and cross-domain scope checks.** The score selects Z^*; independent checker C tests it against E. (**a**) Molecular selection and controlled guidance span the full chain. (**b**) Code separates reference response from verification; the real-output bar counts explicit failing-test labels (Appendix A.12). (**c**) KG evaluation reverses model order. (**d**) Scene search exposes a missing placement check. (**e**) MBR can omit required terminology; its gate is a constructive control. Other domains test narrower claims, not the full molecular chain. Intervals are 95% bootstraps; Table 4 maps scope.

output can violate the requirement even when a satisfying alternative exists: exempting a relation from a penalty and checking its presence are different operations.

An audit of decisions. The audit tests for each separately (Figure 1). The **reading test** holds the output fixed and changes the requirement; the **verifying test** holds the requirement fixed and compares satisfying against violating outputs, as judged by a separate compliance checker. We then test whether the failure changes selection on frozen predictions and whether optimization or guidance can reach it. The contribution is this pair of failure modes, an audit that tests for each separately, and the decision consequences we measure. Standard information ordering explains why tuning cannot always recover an ignored requirement; that is not our novelty claim.

Molecules reveal the full mechanism. In mixed six-candidate pools, the repair raises cluster-weighted correct-bond selection from **36.5% to 73.9%**, against **99.9% available in the pools**. It still selects only broken-bond winners on 44 of 292 tasks with a bonded alternative. The scoring choice thus changes decisions, though these controlled transfers do not establish a deployed defect and the preregistered within-generator effect remains unestablished.

Scope. Those sources document the mismatch and show validity checks used in selection; they do not establish a covalent misselection inside any pipeline. The molecular study follows one mechanism through measurement, selection, search and guidance. Code, knowledge graphs, scenes, terminology translation and exploratory reward models supply *cross-domain checks of scope and boundaries*, not five replications of that chain.

2 Two tests: does the score use the requirement, and check it?

The protocol needs an output Z, a requirement E, a score $s(Z, E)$, and a *separate compliance checker* that reads E directly: a distance test, an executable test suite, or another domain-specific procedure. In ungated arms, neither the checker nor the score uses the other's result. Checker-gated

arms are constructive controls: they enforce compliance by design and do not independently validate that endpoint.

Reading varies E at fixed Z; verifying compares satisfying and violating outputs at fixed E. Passing one test need not imply passing the other. Reference substitutions in the text panels probe representation sensitivity, not an isolated change to the task requirement.

For selection, the requirement is simple: whenever a candidate set contains a compliant output, every highest-scoring output should comply. One violating winner disproves this property on that set; finding none does not establish it for unseen outputs. Appendix A gives the formal definition and reporting rules, including unresolved cases.

Evaluation controls. Compliance is determined independently of the audited score. Checker thresholds are frozen before predictions are evaluated, and unresolved cases are excluded rather than assigned a score. We distinguish publicly preregistered tests, protocols frozen without a public timestamp, and post-hoc analyses; their units and status are listed in Appendix Table 11. Intervals are computed within each experiment, not pooled across domains.

Theoretical explanation: a standard information ordering. If the same output needs different decisions under different requirements, an output-only rule cannot choose correctly in both cases. This standard result (Blackwell, 1951) explains the reading obligation; it does not supply the empirical contribution or establish that a score verifies.

Theorem 1 (Information available to the decision rule)**.** *Suppose outputs, requirements, labels and available actions take finitely many values. For a loss L, let $R^\star_Z$ be the smallest expected loss of a rule that observes only the output, and $R^\star_{(Z,E)}$ the smallest when it also observes the requirement; both allow randomized rules. Then $R^\star_{(Z,E)} \le R^\star_Z$. Equality holds exactly when, for each output, one action minimizes expected loss under every requirement that occurs with it. Otherwise $\Delta = R^\star_Z - R^\star_{(Z,E)} > 0$, and every output-only rule has worst-case regret at least Δ relative to the informed optimum.*

Ignoring the requirement is harmless when the same decision remains best. The audit therefore looks for cases needing different decisions, rather than assuming such cases exist. Appendix A.16 supplies the proof, extensions and detection bounds.

3 The repair improves clash scoring but does not verify the bond

For covalent structure prediction, Z contains the atomic coordinates, E names the required protein–ligand bond, and the checker measures the named atoms' distance. The contact score rewards separation; the requirement demands that one pair stay close. The rule was designed for non-bonded pairs, and its released interface cannot represent a declared protein–ligand bond; six of ten screened rules share this gap (Appendix A.15.1). We first isolate the scoring mechanism, then test its decision consequences.

The two scores. Both scores reward separation: a larger value means fewer severe overlaps. They differ only in which atom pairs they inspect. Write Z for a structure—atoms, coordinates, which component each atom belongs to, and the bonds inside each component—and E for the bonds the task declares between components, with $x = (Z, E)$. In the formulas, d_{ij} is the distance between atoms i and j, r_i is an atom's clash radius, and $\pi(i)$ identifies protein (P) or ligand (L). The original score is

$$s_0(x) = \min\Big\{1,\ \min_{\pi(i)=\mathrm{L},\pi(j)=\mathrm{P}} \frac{d_{ij}}{r_i + r_j}\Big\}, \qquad V_\tau(x) = \mathbb{1}[s_0(x) \ge \tau], \quad \tau = 0.75. \tag{1}$$

Each atom-pair distance is divided by the sum of its clash radii. The score uses the smallest ratio, caps it at one, and accepts the structure at 0.75 or above. Both scores below use this cap, including in the proof and optimization experiments. The score cannot distinguish a required bond from an unwanted overlap. For example, a typical required sulfur–carbon bond has ratio $1.81/3.50 = 0.52$, below the cutoff.

The repair removes the required pair and its immediate bonded neighbours from the clash calculation. In graph terms, it scores only atom pairs separated by more than two bonds. Let $\delta_G(i, j)$ denote that graph distance after the declared bond has been added. The repaired score is

$$\mathcal{A}_2(x) = \{(i, j) : \pi(i) = \mathrm{L}, \pi(j) = \mathrm{P}, \delta_G(i, j) > 2\}, \quad s_2(x) = \min\Big\{1, \min_{(i,j)\in\mathcal{A}_2(x)} \frac{d_{ij}}{r_i + r_j}\Big\}. \tag{2}$$

This is the usual connectivity-based exclusion (Eastman et al., 2017), with radii and cutoff unchanged. We call s_0 the *original score* and s_2 the *repaired score*. A separate distance test is the *bond check*. The repaired score measures unwanted clashes once bonded neighbours are set aside; the bond check asks whether the bond formed.

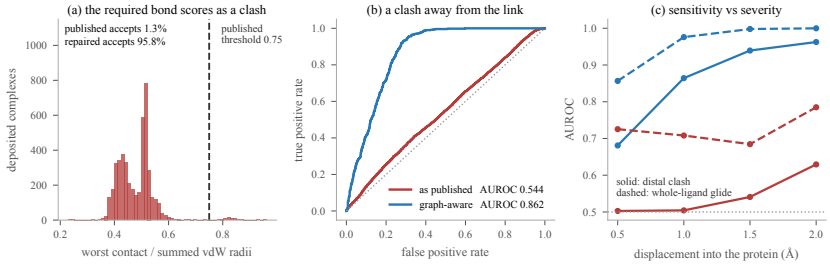

Figure 2: The repair removes the false alarm without losing clash detection. (a) The original rule rejects most deposited covalent complexes because it scores the required bond as a clash; the repair accepts 95.8%. (b–c) When we introduce clashes away from the bond, the repair separates correct from corrupted structures and responds more strongly to more severe damage.

Why changing the cutoff is not enough. A cutoff above the required bond's score rejects it; lowering it enough to accept the bond also accepts overlaps the published rule rejects. No cutoff separates the two, because the score does not know which short distance is required (Appendix A.10).

Proposition 1 (Removing pairs cannot verify the bond)**.** *Assume the allowed outputs include rigidly translating the ligand along the path used in the proof. For any finite set of non-bonded pairs included in the score, the capped distance score attains its maximum at a structure in which the required protein–ligand bond is broken.*

To see why, translate the ligand away from the protein. Every retained contact eventually exceeds its clash threshold, so the clipped score reaches one, while the declared bond distance exceeds the valid range. Theorem 1 concerns information unavailable to a score; Proposition 1 concerns a score that uses the declaration only to exclude contacts. The latter can read the declaration and still attain its maximum without the bond. Both proofs are in Appendix A.16.

This existence result does not estimate frequency; the experiments below test its occurrence in deposited structures, predictions and optimization.

Does the score read the declared bond? The bond check uses a fixed 1.2–2.2 Å window and abstains when an endpoint cannot be identified. Acceptance rises from 1.3% to 95.8% $[95.2, 96.3]$ on the 4,587 deposited complexes. Strict crystallographic-quality filters give the same picture: among the 2,591 complexes that pass both the model-fit and the density/occupancy filter, acceptance is 1.1% against 96.3% (Appendix A.4). The declared bond or its neighbourhood is the closest contact in 98.4% of cases. On structures with introduced distal overlaps, AUROC improves from 0.544 to 0.862: repair still detects unwanted clashes (Figure 2; Appendix A.2).

Does the repaired score check that the bond exists? The prediction cohort contains 292 later targets; 272 frozen Boltz-2 outputs initially satisfy the bond check. For each of these 272, we translate the entire ligand away from the protein along the bond axis in 0.25 Å steps up to 2 Å. This preserves ligand shape and increases the bond distance by exactly the displacement.

A broken-bond position strictly outscored every bonded position on 245 of 272 tasks under the original score and 65 under the repair. Requiring the separate bond check eliminates these constructed

failures. These counts describe the displacement intervention, not the original prediction error rate. Figure 3(**a**) shows one example; Appendix C.3 reports the grid, tie rule and full results. The 272 predictions also initialize the local-search audit; 100 of these tasks enter the generation experiment in Section 5.

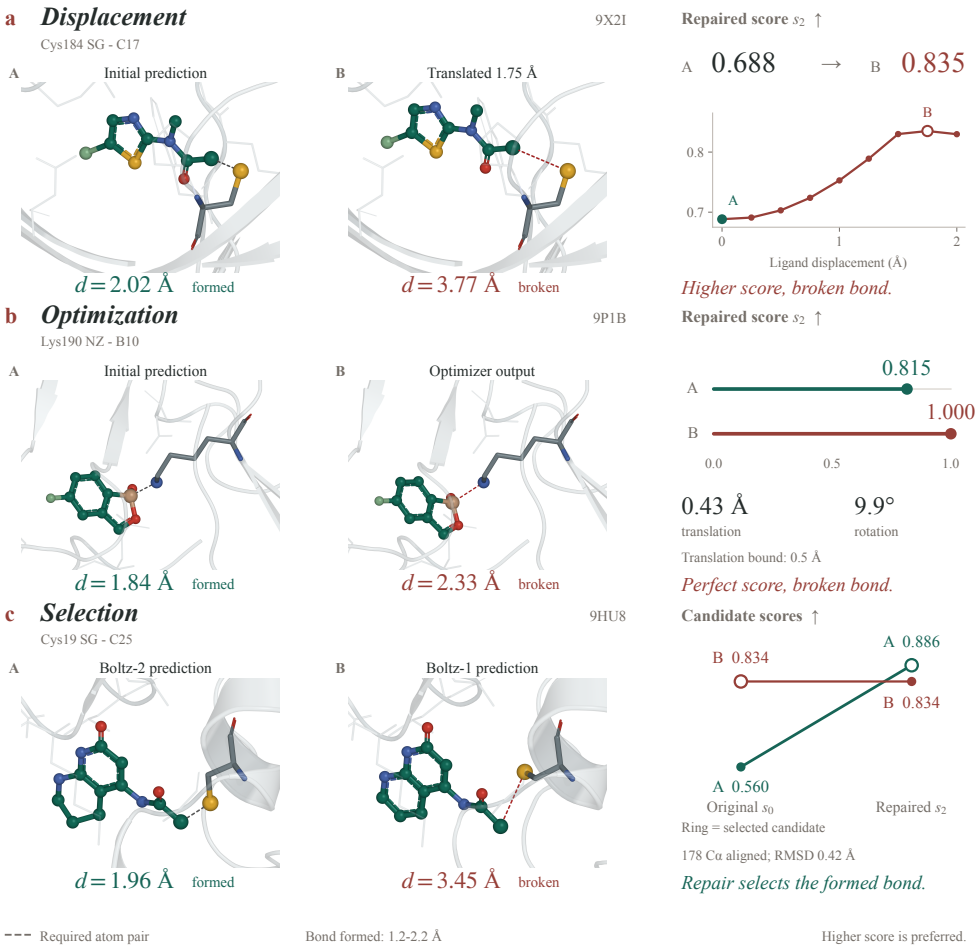

Figure 3: **Clash-score repair helps selection but does not verify the bond.** Each row shows one target, with scores at right. (**a**) Sliding the ligand along the declared bond axis takes the bond out of the covalent window while the score climbs from 0.688 to 0.835. (**b**) A rigid-body search bounded at 0.5 Å reaches the ceiling of 1.000 with the bond outside the valid window. (**c**) Two real predictions of one target, neither moved by us: the original score puts the broken candidate first, the repaired score the formed one.

Can this change which output is selected? The repair removes the dominant false alarm while retaining clash detection, but can still assign its highest score to a broken bond. The next question is whether a candidate pool contains such a structure and whether the score selects it.

4 When does the scoring error change selection?

Selection depends on the candidate pool. We compare mixed-model predictions with repeated samples from one model, then vary the mixture. The 292 tasks form 74 homology clusters at 30% sequence identity and 80% bidirectional coverage. Clusters receive equal weight; paired 95% bootstrap intervals resample clusters, not individual predictions (Appendix A.7).

4.1 CANDIDATES FROM DIFFERENT MODELS EXPOSE THE ERROR

For each of 292 tasks, the candidate pool contains six existing predictions: Boltz-1, Boltz-2, Chai-1, Protenix, and independent Boltz-1 and Chai-1 repeat runs. We ask which candidate each score would select. For every pool size $n = 1, \ldots, 6$, we evaluate all subsets of n different candidates and split exact ties evenly. The success label—whether the named atoms form the required bond—is used only to evaluate the selection. Neither score sees it.

Table 1: Percentage of selections with the required bond. The original score degrades as the pool grows. Repair improves selection but leaves residual failures. The bond–angle score uses the bond check itself, so it is not an independent validation. "Best in pool" counts success whenever a bonded candidate is available. Protein groups receive equal weight.

n	Original s_0	Repaired s_2	Bond–angle	Best in pool
1	71.1	71.1	71.1	71.1
2	59.7	76.3	89.8	90.1
3	49.2	76.5	96.2	96.6
4	41.7	75.6	98.0	98.4
5	37.6	74.6	99.0	99.3
6	36.5	73.9	99.8	99.9

At $n = 6$, correct-bond selection is 36.5% under the original score and 73.9% after repair, a gain of 37.5 points (95% confidence interval $[26.8, 48.3]$). The available candidates permit 99.9% (Table 1). Figure 3(**c**) illustrates a corrected selection.

The repaired score still selects violations. On **44 of 292 tasks**, every candidate tied for the highest repaired score has a broken bond, even though a bonded alternative exists. These are unmodified model outputs. With ties between bonded and broken candidates, ranking errors account for nearly all remaining failures; under 0.1 points come from having no bonded candidate (Appendix Table 31).

A concrete selection consequence. On 7I90, both contact scores select a Boltz-1 prediction whose declared atoms are 5.92 Å apart. Gating the repaired score on the bond instead selects a Chai-1 repeat at 1.48 Å, which also passes the independent pose check. This post-hoc example sits far from the bond threshold; it shows the audited ranking discarding an alternative that satisfies both checks, not a deployed decision or confidence-based selection (Appendix C.3).

Independent pose and angle endpoints distinguish bond-check improvement from broader structural quality (Section 4.4).

4.2 ONE MODEL RARELY SUPPLIES BOTH BONDED AND BROKEN CANDIDATES

With the generator fixed, we draw 25 Boltz-2 samples on each of the same 292 tasks. Before generation we publicly preregistered the cohort and analysis, predicting that more samples would worsen original-score selection (H1) and that repair would improve it (H2).

Neither hypothesis is established. Original-score bond success changes from 88.9% at one sample to 86.7% at 25, with an interval for the decline spanning zero ($[-5.1, 9.6]$ points). Repair reaches 93.1%, a gain of 6.4 points $[-2.0, 15.0]$; pose preservation is also unproven. Appendix Table 29 reports both registered predictions and their outcomes; Figure 5 compares the two candidate sources.

The labels explain the difference: only 51 of the 292 pools contain both bonded and broken candidates, and the original score selects a broken candidate first in 20 of those. Most pools contain only one bond state. This limits the observed selection cost without resolving the score's verification failure.

4.3 CHANGING THE CANDIDATE SOURCE CHANGES THE OUTCOME

At fixed pool size six, replacing Boltz-2 samples with Boltz-1 predictions tests candidate-source sensitivity. It changes both generator and score distributions, so it does not isolate a single causal factor.

Table 2: Changing the candidate mixture changes the consequence of the scoring error. "Mixed pool" is the percentage containing both correct and incorrect bonds; the other columns report correct-bond selection at pool size six. Percentages are cluster-weighted.

Pool	Mixed pool	Original s_0	Repaired s_2	Best in pool	Difference
6 samples, one model	8.9%	88.7	87.7	93.5	-1.0 $[-5.9, 3.6]$
5 samples + 1 weak	49.3%	40.3	76.0	94.4	$+35.7$ $[24.8, 46.9]$
4 samples + 2 weak	52.1%	34.0	70.1	94.2	$+36.0$ $[26.0, 46.7]$
Six frozen model panels	54.8%	36.5	73.9	99.9	$+37.5$ $[26.8, 48.3]$

A single replacement sharply increases conflicting pools and reduces original-score success, even though a bonded candidate usually remains available (Table 2). A distant broken-bond ligand can score well on contact separation. Thus candidate composition determines the cost of the same missing check and warrants a fresh selection audit.

4.4 What the repair improves, and what it does not show

A higher bond-check rate is not the same as a better structure. We therefore measure geometry that neither the repaired score nor the bond check uses: the angles at the two declared link atoms. The band $[73.9°, 153.3°]$ was calibrated to accept 95% of deposited structures and frozen before any prediction label was computed (Appendix C.2). Two things then separate. At pool size six the repair raises the angle-check pass rate from 44.8% to 72.5%, a gain of 27.7 points $[17.4, 38.0]$, so the improvement reaches an endpoint the repair was not built to satisfy. On the 266 tasks (68 clusters) with complete atom correspondence, pose success within 2 Å moves by $+4.0$ points $[-1.1, 9.3]$. The interval contains zero: better overall poses are not established. Within one generator even the angle gain is uncertain, at $+7.1$ points with an interval spanning zero. Appendix Table 30 gives the definitions, counts and paired intervals; these are separate experiments and are not pooled.

5 Can optimization discover the same failure?

The displacement audit constructs a failure along a chosen path. We now ask whether a local optimizer or a generator guided by the score reaches a failure without that prescribed path. The two experiments use different units and protocols, detailed in Appendix E.

Moving and rotating the ligand. Starting from the 272 initially bonded predictions, we maximize each score over ligand translation and rotation, holding the protein and ligand shape fixed. With translation bounded at 0.5 Å and rotation at $30°$, the best structure found fails the bond check on 93.0% of tasks under the original score and 30.5% under the repair. Every invalid structure it selected outscored its own starting point. At a 5 Å bound the repaired score still fails on 34.9%. Adding the bond check removes these failures in all tested searches. Appendix E.1 gives the algorithm and all four translation bounds. Figure 3(**b**) shows a 0.43 Å translation and $9.9°$ rotation that reach score 1.000 with a broken bond.

Guidance during generation. We insert the objectives into Boltz-2's sampling loop on 100 tasks with two fixed seeds per task and the four settings in Table 3. The first two differ only in the eligible clash pairs; the third adds a bond penalty and the fourth retains stock steering. Inputs, masks and controls are specified in Appendix E.2.

All 800 generations produced readable structures. With the original contact score as the isolated guidance objective, 0 of 200 outputs pass the bond check and the median declared distance is 2.73 Å. Pair exclusion restores 197 of 200 bonds (98.5%); Table 3 reports the paired contrast $+97.4$ points $[93.8, 100.0]$. The choice of checked pairs therefore governs this controlled setting. Yet exclusion supplies no bond test: the 197 preserved bonds follow from leaving the pair unscored, not from checking it.

Adding the bond term raises the count from 197 to 199 of 200, and stock Boltz-2 also reaches 199. The failure therefore belongs to the transferred contact objective, not to the generator, and the large

Table 3: What happens when each objective guides generation. The original clash objective breaks every bond when used alone; exclusion restores nearly all. Counts are unweighted; differences are cluster-weighted percentage points with paired 95% bootstrap intervals over 34 clusters.

Generation setting	Bond check passed	Bond distance, median	Difference
1. original clash score	**0/200 (0.0%)**	2.73 Å	—
2. repaired clash score	197/200 (98.5%)	1.65 Å	$+97.4$ $[93.8, 100.0]$ vs 1
3. repaired + bond term	199/200 (99.5%)	1.64 Å	$+1.8$ $[0.0, 5.2]$ vs 2
4. stock Boltz-2 steering	199/200 (99.5%)	1.66 Å	-0.7 $[-4.4, 2.2]$ vs 3

effect is the contrast between the original and the repaired steric objective. The three-task pilot in Appendix C.5 uses a different objective and is reported separately.

6 Cross-domain checks of scope and boundaries

The molecular case supplies the main evidence chain. The following checks ask which parts of the diagnosis transfer and where the conclusions stop; they do not establish the full mechanism in five additional domains. Table 4 separates these roles.

Table 4: Audit evidence and its scope in each domain. The molecular study supplies the complete chain; the other panels test narrower claims. The reported rates are not pooled across domains.

Domain	What the audit establishes	Scope
Covalent structure	Both tests, plus selection, search and guidance	Complete chain
Code generation	Reference-response probe and failing-output selection	Post-hoc response probe
Knowledge graphs	Reading only; raw/filtered FB15k-237 rankings reverse	Reading side only
Scene layout	Both tests; selection and optimization break declared placements	Perturbed layouts
Terminology MT	Reference probe and MBR selection; second-language replication	Different score contexts
Reward models	Both tests; all six tie with or prefer a violation	Exploratory

Code: reference response does not establish correctness. On the original 147-task HumanEval+ cohort, chrF selects an output with explicit failing-test labels on 53 tasks (36.1%). Seven additional timeout-only labels produce the archived 40.8%; all seven match the canonical outputs in a post-hoc functional replay, which does not reproduce the original timing rule. Changing the reference often changes selection, yet verified one-operator mutants tie with or outrank every passing completion on 88.6% of 105 tasks. Reference sensitivity therefore does not certify task correctness (Appendix A.12).

Knowledge graphs: relation handling reverses model order. Across five public FB15k-237 checkpoints, removing other known-true answers from the competitors reverses 7 of 10 model orderings, intervals excluding zero under both rankings. This changes model comparison but stays on the reading side (Appendix A.15.1).

Scene layout: search exposes the missing check. On 847 initially compliant ProcTHOR scenes, local search within a quarter metre breaks a declared placement on 91.4% under a collision score and 67.4% after repair. At the smallest 0.05 m bound, the two arms do not separate (31.5% versus 31.3%); the repair's effect depends on search scale. Among 844 mixed pools of 17 layouts, repair reduces violating selections from 44.2% to 20.5% (paired gap $[19.9, 27.6]$ points). Candidates are perturbations with a compliant original always available; a placement gate enforces the check by construction (Appendix A.13).

Terminology translation: an available correct term is discarded. Of 498 English–French inputs, 472 have a compliant candidate; chrF-based MBR selects a violating one on 36 of those (7.6%). The checker gate recovers these available alternatives by construction. In one case, MBR prefers *froid* (cold temperature) to *rhume commun* (common cold), although the latter is already in the pool. A second-language replication retains 952 of 3,071 English–Chinese segments under frozen rules requiring a declared term and a reference that satisfies it. Its $+18.5$-point gate gap, $[16.1, 20.9]$, is conditional on this filter; source-content overlap with English–French was not measured. Reference-based probes and sample-based MBR remain parallel evidence; terminology compliance is not overall translation quality (Appendix A.14).

Reward models: exploratory paired evidence. Across 427 paired audits from IFEval/IFBench (Zhou et al., 2023; Pyatkin et al., 2025), all six reward models sometimes tie with or prefer a violation. Only three models exceed the required 90% declaration-response rate, fewer than the four prespecified; this keeps the panel exploratory (Appendix A.11).

A boundary: measurement change need not reverse ranking. On 34 ImageNet configurations, switching to ReaL labels produces three point reversals among 561 pairs, but none meets the paired-interval criterion. SQuAD, TriviaQA, WebNLG and GoEmotions also supply bounded nulls (Appendix A.15.1). A common score shift cannot change an ordering; candidate-specific differences must overcome the original gap. This explains why measurement disagreement alone is insufficient, not a predictive law for new domains.

7 Related work

Sets of molecular-validity checks (Buttenschoen et al., 2024; Harris et al., 2023) check generated structures; we audit one scoring rule instead, on a task that declares a bond. Such a set closes the gap only if one check tests that bond. Local bonded exclusion is standard (Eastman et al., 2017). Our contribution is the two-sided audit and its measured consequences, not first recognition of that incompatibility. Nor is exclusion itself the remedy: it withholds a penalty without supplying evidence, so the repaired score stays relation-incomplete (Proposition 1) and still prefers a broken bond on 65 of 272 predictions. Verification needs a positive test; the audit is what shows one is missing.

The reading intervention follows metamorphic testing: change one input factor while holding the rest fixed (Segura et al., 2016). Task-invariant metric audits test output properties (Vyas et al., 2026); we intervene separately on the declaration and on its satisfaction. Reward-model and verifier evaluations fix the prompt and compare outputs — the verifying side alone. Unlike work on imperfect learned reward proxies (Gao et al., 2023), our audit shows the gap in a deterministic score, though neither test isolates its cause (Appendix A.1).

8 Scope, limitations, and conclusion

The audit is most useful when a score is transferred to a task with additional requirements, or promoted from a descriptive metric to a selection or guidance objective. For a new task, test whether the score penalizes a required relation, then whether repair leaves its absence unnoticed, and report the case where no candidate passes the separate check.

The limits affect interpretation. Stock Boltz-2 preserves 199 of 200 bonds, so our guidance result is not a defect claim about its deployed pipeline. Mixed-model selection is post-hoc, the preregistered within-generator effect unestablished, and model confidence unaudited. Gains depend on candidate composition: the three family pairs with Boltz-1 show them ($+26.6$ to $+33.6$ points), and the three without it show none. Bond compliance does not establish pose quality; other domains support only the narrower claims in Table 4. Audit both sides before optimizing a score; field incidence and deployment benefit remain unmeasured.

AI USE STATEMENT

Generative AI tools were used solely to assist with editing and polishing the manuscript, including improving grammar, clarity, concision, organization, and readability. They were not used to formulate the research questions or hypotheses, design the methodology or experiments, generate or analyze data, develop mathematical claims or proofs, or interpret the results. All scientific content, experimental procedures, analyses, results, and conclusions were developed and verified by the authors. The authors reviewed and approved all AI-assisted edits and take full responsibility for the final manuscript.

Ethics Statement. This work does not involve human participants, personal or sensitive data, clinical decision-making, or wet-lab experimentation.

Reproducibility statement. From pinned inputs the anonymous package replays the selection analyses, the bond-axis intervention, the steering table with its cross-machine duplicate check, the rigid-body optimizer table, and every number printed on the case-study figures, whose structures it also ships. Generating new structures needs the frozen cohort seeds and a GPU; PDB and corruption construction is documented in Appendix A.2.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Benoit Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Benchmarking structure-based three-dimensional molecular generative models using genbench3d: ligand conformation quality matters. *arXiv preprint arXiv:2407.04424*, 2024.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0590.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.

Giulia Bianco, Stefano Forli, David S. Goodsell, and Arthur J. Olson. Covalent docking using AutoDock: Two-point attractor and flexible side chain methods. *Protein Science*, 25(1):295–301, 2016. doi: 10.1002/pro.2733.

BioPipelines contributors. BioPipelines tool reference: Analysis, PoseBusters, 2026. URL `https://biopipelines.readthedocs.io/en/latest/tool/analysis/`. Software documentation, accessed September 25, 2026. PoseBusters subsection.

David Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, 1951.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL `https://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data`.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15:3130–3139, 2024.

Vincent B. Chen, W. Bryan Arendall, Jeffrey J. Headd, et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66:12–21, 2010. doi: 10.1107/S0907444909042073.

Yihong Chen, Pasquale Minervini, Sebastian Riedel, and Pontus Stenetorp. Relation prediction as an auxiliary training objective for improving multi-relational graph representations. In *Conference on Automated Knowledge Base Construction*, 2021. URL https://openreview.net/forum?id=Qa3uS3H7-Le.

Yunmo Chen, William Gantt, Tongfei Chen, Aaron Steven White, and Benjamin Van Durme. A unified view of evaluation metrics for structured prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12868–12882, 2023. doi: 10.18653/v1/2023.emnlp-main.795.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Mostafa Dehghani, Yi Tay, Anurag Arnab, Lucas Beyer, and Ashish Vaswani. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, 2020. doi: 10.18653/v1/2020.acl-main.372.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Justin S. Diamond and Markus A. Lill. Neural SHAKE: Geometric constraints in neural differential equations. *Journal of Cheminformatics*, 17:115, 2025. doi: 10.1186/s13321-025-01053-w.

Peter Eastman, Jason Swails, John D. Chodera, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, 2017. doi: 10.1371/journal.pcbi.1005659.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Mu Gao, Aurélien F. A. Moumbock, Abdul Qaseem, Qing Xu, and Stefan Günther. Covpdb: a high-resolution coverage of the covalent protein–ligand interactome. *Nucleic Acids Research*, 50 (D1):D445–D450, 2022.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In *Proceedings of ACL*, pp. 179–188, 2017. doi: 10.18653/v1/P17-1017.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pp. 11–15, 2008.

Charles Harris, Kieran Didi, Arian R. Jamasb, Chaitanya K. Joshi, Simon V. Mathis, Pietro Liò, and Tom L. Blundell. Posecheck: Generative models for 3d structure-based drug design produce unrealistic poses. In *NeurIPS Generative AI and Biology Workshop*, 2023.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of ACL*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras Wang, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in LLMs. *arXiv preprint arXiv:2410.18451*, 2024.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

Nir London, Rand M. Miller, Shyam Krishnan, et al. Covalent docking of large libraries for the discovery of chemical probes. *Nature Chemical Biology*, 10(12):1066–1072, 2014. doi: 10.1038/nchembio.1666.

Alex Morehead, Jian Liu, Pawan Neupane, Nabin Giri, and Jianlin Cheng. Protein-ligand structure and affinity prediction in CASP16 using a geometric deep learning ensemble and flow matching. *Proteins: Structure, Function, and Bioinformatics*, 94(1):295–301, 2026. doi: 10.1002/prot.26827.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

Curtis G. Northcutt. Benchmarking keras and pytorch pre-trained models, 2019. URL https://github.com/cgnorthcutt/benchmarking-keras-pytorch. Per-image ImageNet top-5 predictions; audited commit 74df938db0f266947abedb2ec9f741af65e2d8ed.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1043–1067, 2024. doi: 10.18653/v1/2024.findings-emnlp.57.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, et al. Boltz-2: towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. *arXiv preprint arXiv:2507.02833*, 2025.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1264.

RCSB Protein Data Bank. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, 51(D1):D488–D508, 2023. doi: 10.1093/nar/gkac1077.

Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You CAN teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BkxSmlBFvr.

Andrea Scarpino, György G. Ferenczy, and György M. Keserű. Comparative evaluation of covalent docking tools. *Journal of Chemical Information and Modeling*, 58:1441–1458, 2018.

Sergio Segura, Gordon Fraser, Ana B. Sanchez, and Antonio Ruiz-Cortes. A survey on metamorphic testing. *IEEE Transactions on Software Engineering*, 42(9):805–824, 2016. doi: 10.1109/TSE.2016.2532875.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler–lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011. URL https://www.jmlr.org/papers/v12/shervashidze11a.html.

Natesh Singh, Philippe Vayer, and Bruno O. Villoutreix. The covalent docking software landscape: Features and applications in drug design. *Briefings in Bioinformatics*, 26(6):bbaf697, 2025. doi: 10.1093/bib/bbaf697.

Hou In Ivan Tam, Hou In Derek Pun, Austin T. Wang, Angel X. Chang, and Manolis Savva. SceneEval: Evaluating semantic coherence in text-conditioned 3D indoor scene synthesis. *arXiv preprint arXiv:2503.14756*, 2025.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. doi: 10.48550/arXiv.physics/0004057.

Akshay Vyas, Angelo Pimienta, and Nicholas Ruozzi. A novel approach to evaluating evaluation metrics for multi-output structured prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26062–26071, 2026. doi: 10.1609/aaai.v40i31.39808.

Ross Wightman. PyTorch Image Models, 2019. URL https://github.com/huggingface/pytorch-image-models.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Haotong Yang, Zhouchen Lin, and Muhan Zhang. Rethinking knowledge graph evaluation under the open-world assumption. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0609.

Tian-Hao Zhang, Jun-Tao Zhu, Zi-Xuan Huang, Jing Xie, Jian-Feng Pei, and Luhua Lai. Benchmarking co-folding methods to predict the structures of covalent protein–ligand complexes. *Acta Pharmacologica Sinica*, 47(5):1374–1384, 2026. doi: 10.1038/s41401-025-01721-5.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

A AUDIT PROTOCOLS AND EVIDENCE MAP

The appendices supply protocols and supporting checks for the main text. Formal derivations are confined to §A.16, §A.15 and §A.6; each states in plain language what it claims before deriving it, and no measured result in this paper depends on following a derivation. A reader checking the experiments can read the protocol and results sections alone. Table 5 follows its argument in order. Each destination states the cohort and analysis relevant to that claim; exploratory extensions and boundary tests are reported separately.

Table 5: Where to verify each main-text claim. Section and table references are linked to the corresponding protocols or results.

Main-text question	Protocol and supporting evidence
Why do the two tests differ? (§2)	Proofs and assumptions: §A.16.
What does the repair fix? (§3)	Cohort and corruptions: §A.2; quality subset: §A.4; displacement grid: §C.3.
When does selection change? (§4)	Candidate-source comparison and preregistration: §B; residual errors and independent endpoints: §C.2; coordinate-level selection cases: Tables 33 and 34.
Can optimization reach the failure? (§5)	Local search: §E.1; generation arms: §E.2. The separate pilot is in §C.5.
Which other domains support which claim? (§6)	Reward models: §A.11; code: §A.12; KG: §A.12; scene optimization: §A.13; terminology MT: §A.14; boundary and null tests: §A.15.1.
When does a scoring error change a ranking?	Decomposition over fourteen panels: §A.15.

Table 11 records analysis status and resampling units. Appendix C contains selection records and exploratory extensions; Appendix D contains supporting molecular measurements. Neither is a separate replication of the main experiments.

Selection requirement used in Section 2. For a candidate set $\mathcal{C}$ containing an output that passes the independent checker C, we call the score *relation-complete on that set* when $\arg\max_{Z \in \mathcal{C}} s(Z, E) \subseteq \{Z : C(Z, E) = 1\}$. One violating maximizer refutes this property on the set. Passing finite probes does not establish it on unseen outputs. Passing the reading and verifying tests need not imply one another; this is a logical claim, not statistical independence. The main text therefore reports the outcomes of the two tests separately.

Reporting statuses. Each test reports COUNTEREXAMPLE, NO COUNTEREXAMPLE FOUND on the outputs actually tested, or UNRESOLVED when an input or a checker result is missing. The second status is a statement about the tested outputs only, never about the score in general. On a counterexample we repair the score and re-run both tests; on an unresolved case we abstain rather than fall back on the unaudited score.

Names used below. A *requirement* is what the task says an output must satisfy. The *reading test* changes that requirement while fixing the output; the *verifying test* fixes the requirement and changes whether the output meets it. We call s_0 the *original score*, s_2 the *repaired score*, and the separate pass/fail distance test the *bond check*. Older artifact filenames use "RIA" for the reading test and "graph-excluded" for the repaired score. These are file labels, not additional concepts. The formal derivation later uses the standard term *fibre* for what the main text calls a *group*: a set of outputs that look identical to the evaluator. The two words mean the same thing, and neither is needed to follow the experiments.

A.1 EXTENDED RELATED WORK

Reward overoptimization can make a learned proxy improve while a gold reward degrades under best-of-n sampling (Gao et al., 2023). Here the verifier structurally omits a declared relation, so task-opposed states can be indistinguishable to a deterministic, noise-free score. Hard-constraint

methods project generation onto a relation-satisfying manifold (Diamond & Lill, 2025); our audit asks whether a supplied score can prefer a relation-violating output during selection or generation.

PoseBusters (Buttenschoen et al., 2024), PoseCheck (Harris et al., 2023), and GenBench3D (Baillif et al., 2024) expose invalid generated poses. Our first failure is the converse: a valid output is rejected because a transferred evaluator discards a task-defining relation. The second persists even when local pair exclusion repairs that measurement. A covalent-docking review places PoseBusters among relevant post-processing tools (Singh et al., 2025). Covalent benchmarks (Gao et al., 2022; Bianco et al., 2016; London et al., 2014; Scarpino et al., 2018) do not audit this transferred validity primitive; predictors report bonded-ligand accuracy separately (Abramson et al., 2024; Zhang et al., 2026). We fix predictions and vary the evaluator across three steric primitives (Table 38).

The reading intervention follows metamorphic testing's paired transformation pattern (Segura et al., 2016); the satisfaction intervention is the addition, since a metric may respond to a declaration yet prefer its violation. Vyas et al. (2026) diagnose approximate metric invariants through an augmented loss, whereas our interventions separately manipulate a declared relation and its realization. Typed-substructure metrics derive scores from output types (Chen et al., 2023); relation erasure removes a required type. Blackwell sufficiency (Blackwell, 1951) and the information bottleneck (Tishby et al., 2000) supply the channel view; we audit and repair one specified evaluation channel. Relative to broader measurement warnings (Dehghani et al., 2021; D'Amour et al., 2022; Raji et al., 2021; Jacobs & Wallach, 2021; Bean et al., 2025), the relation supplies an exact counterexample and executable repair.

A.2 Reference cohort, exclusions and corruption controls

This section supports the deposited-structure and clash-discrimination results in Section 3. Deposited complexes, their corrupted derivatives, and model predictions are separate cohorts; Table 6 records their relationships.

The candidate query and structure retrieval were frozen on 9 August 2026 and required an X-ray PDB entry, resolution at most 2.0 Å, a ligand between 150 and 900 Da, and an explicit protein-anchored `covale` record. Structures with no usable protein–ligand link, unresolved ligand coordinates, or incompatible receptor/ligand extraction were excluded. The archive query returned 8,268 candidates; 4,791 contained a usable protein-anchored link and 4,587 were scored by both evaluators. We retain PDB accessions and component identifiers for every row so that structures can be retrieved independently.

All headline and temporal analyses use this same backbone-link-corrected 4,587-complex cohort.

Recovery is 96.8% for drug-like ligands, 93.2% for cofactors, and 96.4% for modified residues or glycans. Extending the local exclusion from cross-link 1–2/1–3 pairs to 1–4 pairs rescues only 13 additional complexes while more than doubling the median number of hidden pairs, so the main analysis keeps the two-edge rule.

Table 6: Cohort accounting. Rows marked "subset" inherit identifiers from the canonical 4,587-complex cohort; the geometry branch has its own completeness criterion and is used only to calibrate reaction-centre intervals.

Stage	n	Relationship / use
Archive query	8,268	resolution, mass, and link query
Usable protein–ligand link	4,791	link and coordinates resolved
Reaction-centre geometry branch	4,742	complete bond/angle geometry
Canonical paired-evaluator cohort	4,587	all headline steric results
Pre-2023 / 2023-onward	3,621 / 966	partition of the same 4,587
Drug-like class	3,000	subset of the same 4,587
Volume-overlap validation	600	deterministic drug-like subset
Corruption parents / derivatives	500 / 4,000	controlled stress test
Post-cutoff model cases	294	separate prediction benchmark

A.2.1 POST-HOC BEST-OF-n AUDIT

This analysis was added after the initial selector experiments and was not preregistered. It uses only the 292 tasks for which all six frozen outputs have resolvable declared endpoints and strict graphs. The panel comprises Boltz-1, Boltz-2, Chai-1, Protenix, and independent Boltz-1 and Chai-1 repeat runs. For each $n = 1, \ldots, 6$, every subset of n distinct candidates is enumerated; each subset and each exact argmax tie receive equal mass. The declared-bond endpoint is computed from the two task-specified atom identities in the output and is never replaced by the nearest contact. Graph-excluded reward contains no bond-distance term, so its bond improvement is an outcome on a separate signal. The joint reward does contain a soft prior on that same distance; its near-perfect hard bond rate is therefore a consistency result, not independent validation.

The numbers of tasks containing zero through six bond-correct candidates are 1, 13, 13, 14, 99, 21, and 131. The within-pool oracle therefore rises from 71.1% under a random singleton to 99.9% with all six candidates. Yet untyped selection falls monotonically from 71.1% to 36.5%, while graph-only selection ends at 73.9% and joint at 99.8%. The joint result shares the endpoint signal; the graph-only gap is the independent measurement result. This is a panel-level selection effect and does not imply an architecture-wide ranking.

The canonical reference analysis contains 4,587 structures: 60 pass the published closest-contact rule and 4,394 pass after local graph exclusion. These counts, rather than a separately filtered plotting cohort, define the reported 1.3% and 95.8% acceptance rates.

A.3 EVALUATOR DESIGN COVERAGE

Table 7 summarizes which requirement each candidate design meets. Lowering the contact threshold is not a correction: it trades the false rejection of bonds for the false acceptance of genuine overlaps and still leaves the bond geometry untested.

Table 7: Coverage of the evaluator designs. "Masked" means the property exists algebraically but has little dynamic range because the bond already determines the minimum; "separate" means the construct is reported as another component rather than folded into the score. R7 is a release-level requirement, met by publishing the annotation snapshot alongside the coordinates.

Requirement	Published	Lower thresh.	Graph-aware	Centre test	Composite
R1: covered-relation acceptance	No	No	Yes	Separate	Yes
R2: uncovered-violation sensitivity	Masked	Weakened	Yes	No	Yes
R3: corruption ordering	Poor	Threshold-dep.	Yes	Bond-local	Yes
R4: out-of-domain consistency	Reference	No	Yes	Separate	Separate
R5: coverage under exclusion	N/A	N/A	No	Yes	Yes
R6: fail-closed integrity	N/A	No	No	Partial	Yes
R7: versioned relations	No	No	No	No	Yes

Table 8: Audit resources and the evaluative claim supported by each.

Resource	n	Purpose
Deposited covalent complexes	4,587	positive-reference acceptance
Drug-like validation subset	600	second intermolecular test
Controlled corruptions	4,000	discrimination and monotonicity
Quality-filtered non-covalent control	324	R4 / collateral damage
Post-cutoff model test set	294	model comparison
Model predictions	1,176	four-model audit
Complete repeat predictions	588	run-to-run replication

A.4 CRYSTALLOGRAPHIC QUALITY SENSITIVITY

We retrieve ligand-instance validation records through the official RCSB PDB Data API (RCSB Protein Data Bank, 2023). The canonical table identifies the chemical component but not a unique non-polymer instance, so we deliberately avoid selecting the best-looking copy. For entries containing

several instances of the target component, every scored instance must pass the stated filter. Metrics are available for 4,316/4,587 entries. "Strict density/occupancy" requires RSCC≥ 0.8, RSR≤ 0.2, average occupancy≥ 0.9, and completeness≥ 0.9 for every target-ligand instance. The model-fit filter is the independent RCSB composite ranking score > 0.5.

Table 9: The headline result under conservative crystallographic quality filters. Wilson 95% intervals are shown in brackets.

Subset	n	Published pass	Relation-aware pass
Canonical cohort	4,587	60 (1.3%)	4,394 (95.8%)
Validation metrics available	4,316	55 (1.3%)	4,131 (95.7%)
All instances: model fit > 0.5	3,013	43 (1.4%)	2,895 (96.1%)
All instances: strict density/occupancy	3,142	33 (1.1%)	3,026 (96.3%)
Both strict filters	2,591	28 (1.1%)	2,495 (96.3%)

The strict subset contains 116 relation-aware failures. These are not the lower-quality tail: failed versus passed structures have median minimum RSCC 0.970 versus 0.951, median maximum RSR 0.069 versus 0.077, median minimum occupancy 1.0 in both groups, and median minimum model-fit ranking 0.823 versus 0.742. The class-stratified strict-subset pass rates are 98.0% for 1,857 drug-like structures, 92.9% for 1,042 cofactors, and 97.9% for 243 modified residues or glycans. Thus filtering on experimental fit strengthens rather than erases the central result; the lower cofactor rate remains a chemistry/connectivity issue for residual analysis.

A.5 Exhaustive audit of strict-subset residual failures

We inspected the exact worst remaining pair for every one of the 116 failures in the strict density/occupancy subset. Table 10 is a descriptive triage and nothing more. The categories are assigned by fixed rules on that one pair—"borderline" means a normalized distance in $[0.65, 0.75)$, "severe" below 0.65, and the HEC entry is the specific pattern HEC.CAC–Cys.SG at no more than 2.3 Å—so they name what the geometry looks like, not what the chemistry is. In particular the HEC category asserts a resemblance to a second haem thioether, not the existence of one. We have no independent evidence that those 34 files are missing a link, and we convert none of these pairs into bonds. Each structure contributes exactly one row, so the counts describe worst pairs rather than structures, and no category enters any denominator reported in the paper. Treating any of them as ground truth would be the same error the paper objects to—inferring a relation from the distance the evaluator is supposed to test.

Table 10: Exhaustive pair-level triage of the 116 high-quality residual failures.

Worst-pair category	Count
Borderline polar contact	42
HEC second-thioether pattern absent from the snapshot	34
Severe other contact	19
Borderline other contact	15
Metal-centred contact	4
Post-snapshot connectivity revision	2

The last two cases are 2ZO5 and 6P43. Their currently served RCSB files have a 12 August 2026 revision whose changed categories include `struct_conn`; the 9 August snapshot's worst pair is now explicitly declared. Running the same implementation on the current files excludes both newly declared pairs (2ZO5 then passes; 6P43 exposes a different undeclared short contact). This check distinguishes an algorithmic mapping defect from mutable archive metadata. More broadly, the concentration of HEC residuals shows why increasing the graph exclusion radius is not the right repair: it would not create a missing second protein–ligand edge. The safe behavior is to retain the contact as a warning, surface a possible missing-relation diagnostic, and require a curated relation before exclusion.

The backbone-link defect was found by remeasuring twelve complexes with an independent parser. In entries containing modified residues, the initial selector could choose a peptide edge such as

FME–ARG or GLU–TYS. The corrected selector requires the apparent ligand partner not to belong to the polymer and otherwise falls through to a genuine protein–ligand edge. Regression tests cover both observed cases; all three slices and the merged result reproduced byte-for-byte.

A.6 Reference-oracle deficit

Erasure can create a measurable penalty on correct references before any model is run.

Proposition 2 (Reference-oracle deficit)**.** *For a benchmark $\mathcal{B} = \{x_n^\star\}_{n=1}^N$ of specification-faithful references and an evaluator $s \in [0,1]$ whose ideal score is one, define $D_{\mathcal{B}}(s) = 1 - N^{-1}\sum_n s(x_n^\star)$. This deficit is computable without model outputs. For a binary gate it is the empirical reference rejection rate. For raw ranking, if a target is exchangeable among k known correct alternatives that all precede every incorrect one, its expected reciprocal rank is H_k/k, giving expected deficit $1 - \mathbb{E}_k[H_k/k]$.*

In words: score the benchmark's own correct answers and measure how far the evaluator falls short of a perfect mark. No model output enters, so the quantity describes the instrument rather than any system being judged. For a pass/fail gate it is just the fraction of correct references the gate rejects. For a ranking metric, the H_k/k term is the score the metric returns when it cannot tell k equally correct answers apart and so orders them arbitrarily; a metric that erased the relation is in exactly that position.

On our covalent references the published gate has deficit 98.7%. On FB15k-237 the exchangeability calculation predicts raw MRR 0.3680, against 0.3683 for a separately measured hash-ordered truth-separating oracle. These are pre-model instrument diagnostics, not upper bounds on arbitrary predictors. A model may score higher by ordering the particular test target first or, in the molecular case, by failing to form the relation that the metric erased. A high reference-oracle deficit therefore indicates that correct references are penalized by the evaluation rule.

A.7 Statistical units and analysis status

Use the experiment-specific units in Table 11: the selection analyses in Section 4 use 74 clusters, while the 100-task steering cohort in Section 5 uses 34. The 75-cluster model cohort below precedes the complete-case restriction; its intervals must not be substituted for the selection intervals.

Proportions use Wilson intervals. Corruption AUROCs use percentile bootstrap intervals resampled by parent crystal structure so that eight derivatives of one complex are never treated as independent. Model intervals resample 75 MMseqs2 homology clusters with replacement and retain all structures within a sampled cluster. The primary clustering uses 30% sequence identity and 80% bidirectional coverage; thresholds of 50% and 90% give 85 and 101 clusters. Run-to-run changes use paired cluster bootstrap intervals.

The model-selection stress test uses 20,000 random 37/38-cluster splits with seed 20260824. Models are selected on one side by either published validity or a passing bond check, and evaluated only on the held-out side. The published criterion's continuous worst-contact ratio breaks pass-rate ties. Reported percentile ranges are the central 95% of repeated split outcomes, not confidence intervals from independent repetitions.

A.8 Software and compute

Seed reproducibility is process-level. Boltz reproduces a sample from `-seed` only when the prediction starts in a fresh process. Calling the predictor repeatedly inside one process advances the random state, so the second and later calls diverge from a single-call run with the same seed. We measured this while testing whether the checkpoint could be loaded once and reused across arms: the first call in a warm process reproduced the reference structure byte-for-byte, and all seven later calls did not. Every run reported here therefore uses one process per prediction, which is also what keeps the paired-seed comparison between arms meaningful. Anyone reproducing these runs should do the same.

The structure audit used Python 3.9, gemmi 0.7.5, RDKit 2025.09.2, PoseBusters 0.6.5, MMseqs2 18-8cc5c, NumPy, SciPy, and Matplotlib. Predictions used `boltz` 2.2.1 and `chai_lab` 0.6.1 on

Table 11: Data and analysis status for the main experiments. "Frozen" means the protocol was hashed before the outcome existed; "preregistered" adds a public timestamp.

Experiment	Data and unit	Analysis status
Reference and corruption audits	4,587 structures; 4,000 corruptions	corpus frozen; construction prespecified
Bond-axis orbit and local search	272 bond-valid predictions; task	grid and protocol frozen
Steering	100 tasks × 2 replicates × 4 arms; 34 clusters	protocol and hypotheses frozen
Heterogeneous selection and composition	292 tasks × 6 panels; 74 clusters	post-hoc on frozen outputs
Within-generator selection	292 tasks × 25 samples; 74 clusters	preregistered and timestamped
Reaction-centre endpoint	292 tasks; 74 clusters	band frozen before scoring
Reward-model paired audit	82 pairs; 28 constraint types	exploratory; model panel fixed before scoring
Six-model reward-model extension	427 pairs; 53 families; 352 mixed pools	frozen before scoring; one interpretation gate failed
KG and code panels	9 checkpoints; 31 code archives	auxiliary cross-domain audits
Code reading test	147 and 105 tasks × 31 models	post-hoc; metric and cohort taken unchanged from the registered run
Scene local-optimizer audit	847 scenes; 31,455 movable objects	frozen and hashed 2026-09-24T14:14:57Z, before any optimization
Scene layout selection	844 pools × 17 layouts	frozen and hashed 2026-09-24T15:05:29Z, before any candidate
Terminology MT	498 segments × 34 candidates	frozen and hashed 2026-09-24T14:56:52Z; one deviation logged
Terminology MT reference arm	478 segments × 35	frozen before computation; post-hoc to the main protocol
Terminology MT held-out pair	952 English–Chinese segments × 34	addendum frozen and hashed 2026-09-24T19:02:33Z, before any output

one NVIDIA A10G host with 30 GB system RAM. Inputs were processed in chunks of twenty with one data-loading worker and resume-on-completion logic; completion was determined by output counts rather than process exit status.

A.9 Relation-integrity stress test

Four executable fixtures perturb only the declared task relation while holding coordinates fixed. With the true link omitted, or with a link whose nodes do not exist, no pair is excluded and the physical bond remains visible as a failing contact. For a doubly attached ligand, omitting the second link likewise leaves that bond visible. The adversarial case replaces the true Cys SG anchor with Cys CB: the graph mask can then hide the local contact, but the independent relation schema check rejects the unseen anchor class and the composite decision fails. These four cases distinguish the observed behavior of the composite evaluator from that of the graph-aware steric component alone; four fixtures establish the distinction, not a guarantee over all adversarial relations. Together with the existing invariance, corruption, and task-complementarity tests and relation-erasure tests of Appendix A.12, the recorded regression suites pass in their specified environments.

The canonical controlled analysis uses 500 parent structures and 4,000 corruptions. Its AUROC is 0.544 for the untyped score and 0.862 for the graph-aware score. The scored corruption record, rather than a manually relabeled plot, is the source of these values.

A.10 THE COST OF RETUNING AN UNTYPED THRESHOLD

Proof of the cutoff claim in Section 3. The declared bond is included in the minimum, hence $s_0(x) \leq b < \tau$ and the structure fails regardless of every other contact. Lowering τ enough to admit the bond applies to all pairs because the score contains no bond type; the same change necessarily admits non-bonded overlaps in the stated interval. □

Within the published ratio family the cost of ignoring the theorem is explicit.

The main text states this argument in plain language; the proof is recorded here to keep the information tradeoff explicit.

A.11 REWARD MODELS: PAIRED READING AND VERIFYING TESTS

Section 6 reports the six-model extension on 427 pairs. The smaller initial panel below explains construction and feasibility; Table 13 gives the main-text model results and Table 14 the post-hoc joint diagnostic. The unmet reading gate keeps the extension exploratory.

This experiment asks whether a learned reward model changes score in the registered direction under an executable instruction intervention while still failing to prefer satisfaction under the fixed instruction. It is a Stage-A feasibility study, not a preregistered confirmatory experiment. The protocol draft, generator recipe, three-model panel, revisions, and analysis code were recorded before reward-model outcomes were inspected. However, the prospective full-study gate required at least 100 retained pairs and a mixed-validity decision endpoint. Stage A retained 82 pairs and did not run that decision endpoint. We therefore report a scoped finite-domain audit rather than a second full instantiation.

Construction. The carrier Z is response text and the relation E is an executable output constraint. The manifest draws from familiar IFEval families (Zhou et al., 2023) and held-out IFBench families (Pyatkin et al., 2025). Qwen2.5-7B-Instruct (Yang et al., 2024) generates four candidates for each of 259 prompts under a fixed decoding recipe (1,036 responses). Exact checkers accept 439 responses, covering 138 prompts. For each eligible prompt, the first accepted response becomes Z^+. The reading arm replaces only the constraint with one whose checker rejects the same Z^+; the satisfaction arm keeps the original constraint and applies the first deterministic single-character insertion or deletion whose checker accepts Z^+ and rejects Z^-. We retain only exact one-constraint flips. This yields 82 paired audits: 46 from 14 IFEval constraint types and 36 from 14 IFBench types.

Scores and analysis. The frozen panel contains Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024), Llama-3-OffsetBias-RM-8B (Park et al., 2024), and Llama-3.1-Tulu-3-8B-RM (Lambert et al., 2024). Revisions are pinned, inference uses bfloat16 and a common 4,096-token left-truncation policy, and all models score the same four arms. Four arms from one pair exceed the limit under every tokenizer; no model receives a different arm set. Reading accuracy is $\mathbb{1}[s(Z^+, E^+) > s(Z^+, E^-)]$. Verifying failure is $\mathbb{1}[s(Z^+, E^+) \leq s(Z^-, E^+)]$; a tie counts as failure because the violating response remains a maximizer. Intervals are 10,000 percentile bootstraps over instruction type within stratum (seed 20260920).

Table 12: Both sides of the reward-model audit. Values are percentages with 95% instruction-family bootstrap intervals. "Sat. fail" means the violating response ties or outranks its satisfying near-copy.

Reward model	Constraint stratum	n	Decl. response	Verifying fail
Skywork	IFEval seen	46	100.0 [100.0, 100.0]	8.7 [1.9, 18.9]
Skywork	IFBench OOD	36	94.4 [80.6, 100.0]	25.0 [10.8, 40.0]
OffsetBias	IFEval seen	46	100.0 [100.0, 100.0]	13.0 [0.0, 30.4]
OffsetBias	IFBench OOD	36	94.4 [84.6, 100.0]	22.2 [8.2, 38.1]
Tulu 3	IFEval seen	46	100.0 [100.0, 100.0]	0.0 [0.0, 0.0]
Tulu 3	IFBench OOD	36	97.2 [90.3, 100.0]	5.6 [0.0, 14.3]

Across both strata, Skywork strictly prefers the violation on 13/82 pairs; OffsetBias does so on 12/82 and ties on two more; Tulu does so on 2/82. Twenty-one distinct pairs affect at least one

reward model, five affect at least two, and one affects all three. The shared case is a formatting constraint: deleting a single required newline makes the response more natural while making the exact checker reject it, and every reward model raises its score. The result is not that reward models generally ignore instructions—their operational declaration-response rate is 94.4–100%—but that a response in the registered direction does not certify a preference for satisfaction. The stronger OOD effect resolves for Skywork and OffsetBias but not Tulu, so model dependence is part of the finding.

Frozen six-model extension. We next froze a larger experiment before any reward-model outcome from the extension was observed. The sampling frame contains all 541 IFEval and 300 IFBench prompts. Qwen2.5-7B-Instruct and OLMo-2-1124-7B-Instruct each supply four candidates per prompt under fixed seeds and decoding settings, for 6,728 candidates. A reading intervention holds a checker-verified satisfying response fixed and appends an incompatible executable constraint. A verifying intervention holds the original declaration fixed and applies the first checker-verified edit in a fixed enumeration that flips exactly one constraint. Neither edit is selected using reward scores. The resulting archive has 427 paired audits across 53 constraint families. Of 841 eight-candidate pools, 352 contain both strictly valid and invalid responses. All six frozen reward models scored the same 8,009 requests; no model, family, or generator was removed after scoring.

Primary estimates macro-average over constraint family. Intervals are 10,000 percentile family bootstraps; a tie counts as a verifying failure. The six one-sided verifying tests remain positive after Benjamini–Hochberg correction (all $q \leq 0.00030$). Table 13 is not pooled with the Stage-A feasibility estimates above.

Table 13: Frozen six-model reward audit. Values are percentages except gains, which are percentage points; brackets give 95% family-bootstrap intervals. Checker-first gain is executable-constraint validity minus reward-model-alone validity at $n = 8$ in the 352 mixed-validity pools.

Reward model	Decl. response	Verifying fail	Checker-first gain
ArmoRM 8B	88.3 [81.5, 94.2]	9.9 [5.8, 15.1]	+41.6 [35.3, 48.0]
Eurus 7B	92.8 [87.6, 96.8]	15.3 [8.4, 23.0]	+50.0 [43.6, 56.6]
OffsetBias 8B	95.3 [90.5, 98.6]	19.3 [11.7, 28.2]	+40.7 [34.4, 47.2]
Skywork Gemma 27B	65.7 [57.2, 73.7]	20.0 [12.8, 27.9]	+36.9 [30.7, 43.5]
Skywork Llama 8B	68.1 [59.3, 76.7]	25.3 [17.8, 33.5]	+36.1 [30.0, 42.3]
Tulu 3 8B	96.6 [91.8, 99.7]	2.4 [0.6, 4.7]	+43.9 [37.4, 50.5]

The five pre-outcome interpretation gates were: at least 300 pairs and 30 families (427 and 53; pass); at least 200 mixed pools (352; pass); the frozen declaration-response rate above 90% for at least four of six models (three; **fail**); a corrected satisfaction-failure interval excluding zero for at least four models (six; pass); and an $n = 8$ checker-first gain whose lower bound exceeds five points for at least three models (six; pass). Because the conjunction fails, we retain the exploratory finite-domain label. Checker-first validity is 100% by construction in these mixed pools because each contains a valid candidate; its gain does not establish better semantic answer quality or deployment performance.

Joint outcome of the two operational flags (post-hoc). The gate analysis above reports the two sides as separate marginals. Because both per-pair flags are already defined by the frozen script (`declaration_correct` and `satisfaction_failure`), they can be cross-tabulated without changing either definition; ties continue to count as verifying failures. This cross-tabulation was not preregistered and does not enter any gate: the five gates and their outcomes stand exactly as recorded above. It reports whether the two registered flags co-occur on the same model and pair; it is not an additional interpretation gate.

Over the $6 \times 427 = 2{,}562$ model-by-pair cells, 10.3% have the registered declaration response and still rank the violation at least as high. Per model, this is 13.1–15.2% of response-positive pairs, except Tulu 3 at 2.4%. Another 9.7% are response-negative while passing the verifying test, so the two flags are not redundant. The response flag compares absolute rewards across two instructions. A prompt-dependent additive shift can change that flag while leaving every within-instruction preference unchanged; because we do not calibrate score levels across instructions, this table describes the frozen models' operational responses rather than proving semantic understanding.

Table 14: Post-hoc cross-tabulation of the two audit sides on the same model and the same pair. Not preregistered; no gate depends on it. Cell percentages are of all 2,562 model-by-pair cells.

	Verifying passes	Verifying fails
Decl. response +	1,958 (76.4%)	**265 (10.3%)**
Decl. response −	249 (9.7%)	90 (3.5%)

A.12 Code verification and knowledge-graph reading audits

Both domains here declare a relation that a score can ignore, and in both the declaration can be checked by something other than the score. In code generation the requirement is executable: the task's own tests either pass or fail, so compliance is settled by running them, while the audited score compares program *text* to a reference implementation. In knowledge-graph evaluation the declared relation is the query's relation type, and the audited score is a ranking metric whose competitor set may contain answers that are themselves true. Each panel therefore has a score on one side and an independent checker on the other, and neither consults the other.

This section supports the code and KG results in Section 6. The code experiments below use distinct cohorts: 147 tasks for selection among unmodified completions, and 105 eligible tasks for controlled edits. KG protocols begin at Section A.12. Model-order diagnostics supplement these tests; they do not add a missing audit side.

The code panel's satisfaction question. We test whether a reference-similarity score selects a failing program when a passing one is available. The archive is also used for the model-order diagnostic below, but this experiment is a verifying test. Its protocol was hashed before the statistic was computed (`code_satisfaction_20260917/PROTOCOL.md`). A task enters the cohort when its 31 temperature-zero completions include at least one that passes the HumanEval+ tests and at least one with a non-pass archive label. Of 164 tasks, 147 qualify; 17 have no archive-labelled passing completion, and none has all completions labelled passing. For each cohort task we take the completion with the highest blind score and ask whether it passes. Under sentence chrF against the canonical implementation it has a recorded non-pass label on 60 of 147 tasks (40.8%, bootstrap interval $[32.7, 49.0]$ over tasks, 10,000 replicates, seed 20260917); under chrF recomputed on the parsed function body, on 50 (34.0%, $[26.5, 42.2]$). Selecting by chrF from subsets of size $1, 2, 4, 8, 16, 31$ yields archive-labelled pass rates of 27.3, 34.2, 41.3, 47.5, 53.9 and 59.2% of the time, against 100% available by construction.

The original observational analysis used archive labels without re-execution. Its non-pass category includes timeouts; it is not a pure functional-error rate. The post-hoc status audit below keeps that analysis intact and separates the labels relevant to the semantic interpretation.

Timeout-label audit and functional replay. Among the 60 selected non-passing outputs, 53 have an explicit `fail` in at least one of `base_status` or `plus_status`; three of these also have a timeout in the other suite. The remaining seven have `base_status=timeout` and `plus_status=pass`. Keeping the original 147 tasks and their selected outputs fixed gives 53/147 (36.1%) with explicit failing-test records. No timeout is automatically relabelled as correct, and the original archive is not overwritten.

For these seven timeout-only winners, we replay all saved base and plus inputs from HumanEvalPlus-v0.1.10, comparing exact outputs with the canonical function on separate deep copies. Each task runs in a separate Python subprocess with a 30-second overall timeout; the inspected functions have no imports or external I/O. All seven match every expected output (Table 15). This checks functional behaviour on the saved inputs; it does not reproduce EvalPlus's original per-test timing policy, certify unseen inputs, or re-evaluate the other 5,077 completions. Consequently the original 40.8% remains an archive non-pass rate, while the main text distinguishes its explicit-failure component. The reference-switch and metric-comparison rates below retain their original archive labels and the same qualification.

The controlled version. A second protocol, hashed before any mutant existed (`code_intervention_20260917/PROTOCOL.md`), replaces observation with a con-

Table 15: Post-hoc replay of all seven chrF-selected outputs whose sole recorded failure was a timeout. Every saved input matches the canonical output.

HumanEval task	Model	Base inputs	Plus inputs
35	CodeGen-6B	2	938
46	GPT-4-1106-preview	4	132
64	GPT-4-1106-preview	7	992
70	GPT-4-1106-preview	9	975
72	CodeGen-6B	6	691
85	GPT-4-1106-preview	4	994
98	CodeGen-2B	7	1000

trolled edit. For each task, we scan the released canonical solution from the top and apply the first eligible operator from a fixed list of thirteen single-token substitutions. The prompt remains unchanged, and the score does not choose the operator or location. We execute both programs on the task's own inputs in a network-disabled subprocess with a five-second timeout. The canonical program defines the expected outputs; the task's tolerance applies to floating-point results.

Of 164 tasks, 26 admit no operator, 19 yield a mutant that still passes, and 14 lack a passing panel completion for comparison. Every canonical program passes, leaving 105 tasks. On 93 (88.6%, bootstrap interval $[81.9, 94.3]$; 10,000 replicates; seed 20260917), chrF scores the broken mutant at least as high as the best passing completion, with median margin 0.47. In all 93, it also outranks every other panel completion. The construction is intentionally simple: a one-character near-copy maximizes reference similarity whether or not it runs. The observational panel above provides the complementary result on unmodified model completions. A post-freeze correction to the subset sampler changed only the best-of-n curve, not this primary endpoint (deviation D9).

Is the failure chrF's, or reference similarity's? Character n-grams are not a credible code verifier by themselves, so we repeat all three endpoints with the cohorts, text pairing, and operators fixed while changing only the score (`code_metric_robustness_20260918/PROTOCOL.md`; post-hoc, not preregistered). CodeBLEU is the stronger comparison because two of its four components encode syntax and data flow. Before reading anything new, the pipeline reproduces the registered chrF value on all 5,084 completions, and the registered `prompt_prefix_match` flag on all of them. The panel's `functional` label agrees with `base_status` and `plus_status` both passing on every task of every archive. Table 16 shows the same pattern for CodeBLEU on all three endpoints; the edit ratio performs worse. On the 105 broken near-copies, CodeBLEU's syntax-match component is exactly 1.0 for 98 of them and its data-flow component is exactly 1.0 for 96, both with median 1.0: the structure-aware components therefore report a perfect match for programs that fail their own tests. A single-token edit can change computation without changing the parse tree or def-use edges, so reweighting these views cannot separate the two states.

Table 16: The code panel under three reference-similarity scores. Cohorts, text pairing and operators are those of the registered runs; only the score changes. Intervals are percentile bootstrap over tasks, 10,000 replicates, seed 20260917. The first two rows use the original archive status, including timeouts (Section A.12); they are not pure semantic-error rates. chrF reproduces the registered values exactly.

Endpoint	chrF	CodeBLEU	Edit ratio
Top-ranked completion has archive non-pass label (n=147)	40.8 [32.7, 49.0]	41.5 [33.3, 49.7]	61.9 [53.7, 69.4]
Selected completion has archive pass label (n=147)	59.2 [51.0, 67.3]	58.5 [50.3, 66.7]	38.1 [30.6, 46.3]
Broken near-copy outranks every passing completion (n=105)	88.6 [81.9, 94.3]	90.5 [84.8, 96.2]	95.2 [90.5, 99.0]

Does the score read the declaration? The registered code endpoints are verifying-side. We add a reading test (`code_reading_test_20260924/`; post-hoc, not preregistered) using the cohort, text pairing, tie-breaking and metric of the registered run unchanged; only the reference changes. Before anything new is read, the pipeline reproduces the registered chrF value on all 5,084 completions exactly (maximum absolute difference 0) and the `prompt_prefix_match` flag on all of them.

Two substitutions are made, with the candidate programs untouched. First, each task's reference is replaced by each of the other 163 canonical solutions. Over the resulting $147 \times 163 = 23{,}961$ pairs the selected completion changes on 92.8%, and the top-ranked program has an archive non-pass label on 66.9% of them against 40.8% under the task's own reference. The score is therefore sensitive to reference substitution; this does not isolate sensitivity to the task requirement. Second, the reference is replaced by the single-operator mutant of itself used above, a program already verified to fail the task's tests. The mutants are rebuilt from the frozen operator list and reproduce the registered `mutant_chrf` on all 105 tasks exactly. Here the selection is unchanged on 101 of 105 tasks (96.2%), and the archive non-pass rate moves from 41.0% to 43.8%, a paired change of $+2.9$ points (95% CI $[0.0, 6.7]$; 10,000 replicates, seed 20260917).

The two substitutions differ in both surface change and program behaviour. An unrelated reference changes much of the text, whereas a one-operator mutant changes the reference implementation's behaviour with little textual change. The original task requirement stays fixed in both experiments. Thus these are reference-representation probes, not controlled interventions on the requirement itself. Their decision changes establish reference sensitivity without proving that the score reads task semantics. The verified failing-output comparisons supply the separate satisfaction-side evidence. Neither result establishes the pattern for learned code metrics.

Is the requirement itself completely verified? The code panel treats passing the released HumanEval+ suite as satisfying the task, which makes that suite the endpoint's oracle. The rule we apply to scores applies to it too: a finite suite licenses nothing outside the inputs it runs. We audited it after the registered experiments, so what follows is exploratory and labelled as such (`code_verifier_blindspot_20260918/PROTOCOL.md`). Probes are built from each task's own recorded calls by replacing one argument at a time, and then in pairs, with the structural boundaries of its type—empty, singleton, repeated, zero, two, negative, retyped, case-folded. An input is kept only when the task's own EvalPlus contract admits it and the reference returns a value on it. A disagreement therefore cannot be blamed on an input the task never promised to handle. Outputs are compared by value, so an integer and a float of equal value agree. Of the 138 mutants above, the released base tests reject 116 and the extended inputs reject three more, leaving 19 that the complete released suite accepts. At most 400 in-contract probes per task falsify one of them: on `HumanEval/94` the mutant returns `None` where the reference returns 2 on the input `[2]`, having agreed with the reference on all 1,009 released inputs for that task. The other 18 are undecided, which is not a claim that they are equivalent. The same search applied to sampled rather than constructed programs finds nothing. Across the 31 archives, 1,245 completions are accepted by both halves of the released suite, 889 of them distinct. No in-contract probe separates any of them from its reference. The verifier's gap is thus reachable by construction and unobserved under sampling—the shape the structural chapters report. A finite pass therefore records that no counterexample was found on those tested inputs; it gives no guarantee on untested inputs.

Two supporting domains, two different roles. On FB15k-237, raw and filtered MRR reverse the ordering of nine controlled checkpoints in all matched-seed comparisons, giving a declaration-side relation-availability witness. On HumanEval+, unmodified model completions and a separate controlled mutant experiment provide satisfaction-side evidence for reference-similarity scores. The code model-order reversals are supporting diagnostics, not a paired reading intervention. Neither panel is a prevalence estimate.

Table 17: Evidence supplied by each domain. A dash means that this work does not run that intervention in the domain. Supporting panels are not pooled into a prevalence estimate.

Domain	Declaration-side evidence	Satisfaction-side evidence	Role
Covalent structure	paired reading intervention	registered bond-axis intervention	principal full instantiation
Reward models	paired constraint replacement	checker-verified minimal edit	exploratory finite two-sided audit
Knowledge graphs	paired raw/filtered relation availability	—	declaration-side witness
Code generation	post-hoc paired reference substitutions	real candidates plus controlled mutant	auxiliary two-sided audit

The main text reports the public-checkpoint KG comparison and the code verification tests. The controlled KG replication and code model-order diagnostic provide additional evidence: On FB15k-237 we froze the protocol before training nine checkpoints (DistMult, TransE, ComplEx at three seeds). At every seed raw MRR orders them TransE $>$ DistMult $>$ ComplEx and relation-aware filtered MRR gives the exact reverse. All 9 matched-seed cross-architecture comparisons reverse with paired test-triple intervals excluding zero, and 25 of 27 seed pairings are robust. In code generation the relation is an executable requirement. Over all 31 complete temperature-zero HumanEval archives in the official EvalPlus release (Liu et al., 2023), sentence chrF against the canonical implementation reverses against HumanEval+ test passing for 27 model pairs. Four of those pairs (five models) also survive chrF recomputed on the parsed target function body.

`analysis/erasure/` contains the domain-agnostic material. An evaluator there is any callable `score(carrier, relations)`, where the first argument carries whatever the domain uses to place things and the second is the declared relation set; nothing in the module names an atom, a bond or a radius.

A real ML benchmark. `kg_filtered_audit.py` downloads checksum-pinned FB15k-237 from the official RotatE repository and reconstructs the raw/filtered distinction of Bordes et al. (2013). The union of the train, validation, and test facts defines the known typed-relation set. For each of the 20,466 test triples we evaluate both head and tail prediction, giving 40,932 queries. A diagnostic oracle assigns every known true triple a score above every unknown triple; a fixed SHA-256 ordering makes scores unique within the true class. This is not a trained model and uses test labels intentionally: it asks whether the evaluator gives its own perfect truth separator the score such an output should receive.

Table 18: The same relation-erasure mechanism on the real FB15k-237 benchmark. The oracle perfectly separates all known truths from all unknowns. Filtered ranking consumes the known relation set; raw ranking erases it.

Evaluation	MRR	Hits@1	Hits@3	Hits@10
Raw	0.3683	0.2540	0.4178	0.6031
Filtered	1.0000	1.0000	1.0000	1.0000

Multiple known answers occur in 34,528/40,932 queries (84.35%); the mean is 236.371 and the maximum is 4,364. Thus the gap is not a hand-selected triple. It is also not a claim that current knowledge-graph papers use raw ranking: filtered ranking is the field's established correction. It is an independent precedent for the paper's central prescription—the evaluator must consume the task's relation set. Unknown but true facts remain outside that set, an open-world limitation analyzed by Yang et al. (2022).

Controlled multi-seed replication. After the public-checkpoint reversal was known, we froze a prospective-to-outcome replication protocol before launching new training. We train DistMult, TransE, and ComplEx at seeds 20260911–20260913 under one shared reciprocal-relation pipeline: 128-dimensional embeddings, uniform tail corruption with 32 negatives, batch size 4,096, Adam, evaluation every five epochs, and early stopping after four non-improving validation checks. Learning rates are 0.01 for DistMult and ComplEx and 0.001 for TransE; all other training and evaluation choices are fixed across seeds. The best filtered-validation checkpoint occurs at epoch 10 for all nine runs. This is a controlled replication motivated by the discovery, not independent confirmation of the original hypothesis.

For matched seeds, DistMult-minus-TransE changes from -0.01789, -0.01431, and -0.01110 under raw evaluation to $+0.00338$, $+0.00835$, and $+0.01213$ under filtered evaluation. All six corresponding paired-test-triple 95% bootstrap intervals exclude zero (10,000 replicates); the exact intervals are in Table 19. Adding the prospectively frozen ComplEx panel makes all 9/9 matched-seed cross-architecture comparisons reverse robustly. Considering all seed pairings, 27/27 cross-architecture pairs reverse in point estimate and 25/27 are robust; the two remaining filtered intervals touch zero without reversing direction. None of the 9 within-architecture seed comparisons reverses. These counts test panel stability, not intrinsic architecture superiority. Checkpoints, traces, per-query ranks, configurations, and hashes are included in the anonymous artifact.

Table 19: Nine independently trained checkpoints under the frozen shared recipe. Raw and filtered columns are test MRR on the same 40,932 queries.

Model	Seed	Raw MRR	Filtered MRR
DistMult	20260911	0.1340	0.2270
DistMult	20260912	0.1366	0.2291
DistMult	20260913	0.1392	0.2322
TransE	20260911	0.1519	0.2236
TransE	20260912	0.1509	0.2208
TransE	20260913	0.1503	0.2201
ComplEx	20260911	0.1310	0.2348
ComplEx	20260912	0.1298	0.2333
ComplEx	20260913	0.1314	0.2377

The two non-robust cases are not heterogeneous failures: both compare DistMult seed 20260913 with ComplEx seed 20260911 or 20260912. Their raw DistMult advantages are robust (0.83 and 0.94 MRR points), but the filtered ComplEx advantages are only 0.26 and 0.11 points, with intervals $[-0.56, 0.03]$ and $[-0.40, 0.18]$ points. Thus uncertainty comes from two near-ties on the filtered side; neither comparison changes back to the raw ordering. We report 25/27 rather than converting these directional near-ties into successes.

Relation-availability monotonicity. Across these nine controlled checkpoints and the five public checkpoints below, we additionally vary the fraction of known true alternatives supplied to the evaluator over $\{0, .25, .5, .75, 1\}$. Under the frozen missing-at-random channel, the raw evaluator is invariant for all 14 checkpoints, whereas exact expected filtered MRR is monotone nondecreasing for all 14. The aggregate model order first changes between .25 and .5 and changes substantially by full relation availability. This is a mechanism model rather than observed annotation missingness; it validates RIA response and corruption ordering, not field prevalence.

Public-checkpoint replay. We evaluate five released learned checkpoints on the same ordered queries: ComplEx with the relation-prediction auxiliary objective (Chen et al., 2021), plus LibKGE TransE, DistMult, ComplEx, and RESCAL (Ruffinelli et al., 2020). Checkpoint and split files are checksum-pinned. Entity and relation mappings are reconstructed according to each release, all 14,541 candidates are scored in float32, and no model is retrained. The filtered results reproduce the sources' rounded reports. The original ComplEx+RP–TransE comparison was frozen before TransE ranks were computed; the three-checkpoint panel extension is explicitly exploratory.

Table 20: Raw and relation-aware filtered MRR for five public FB15k-237 checkpoints on the same 40,932 queries. Parentheses give rank within the panel.

Checkpoint	Raw MRR	Filtered MRR
ComplEx+RP	0.1746 (5)	0.3886 (1)
TransE	0.1817 (3)	0.3128 (5)
DistMult	0.1838 (2)	0.3428 (4)
ComplEx-LibKGE	0.1855 (1)	0.3477 (3)
RESCAL	0.1772 (4)	0.3559 (2)

Before computing TransE ranks, we froze a paired test-triple bootstrap that keeps each head/tail pair together (10,000 replicates; seed 20260910). The ComplEx-minus-TransE MRR difference is -0.00714 under raw evaluation, with 95% percentile interval $[-0.00863, -0.00563]$, and $+0.07580$ under filtered evaluation, with interval $[0.07302, 0.07860]$. This supports a stable evaluation- induced decision change on this test set. It does not estimate training-seed variation: each model is represented by one checkpoint, trained with a different pipeline, so no architecture-wide superiority claim follows.

For the exploratory five-checkpoint panel, raw and filtered MRR reverse 7 of the 10 pairwise orders (Kendall distance 7). In every reversed pair, both raw and filtered 95% paired-test-triple bootstrap intervals exclude zero. For example, ComplEx-LibKGE minus RESCAL changes from $+0.00826$ ([0.00700,0.00949]) to -0.00822 ([-0.01045,-0.00591]); DistMult minus RESCAL changes from

$+0.00654$ ([0.00531,0.00782]) to -0.01318 ([-0.01543,-0.01093]). These intervals characterize this test set and these frozen artifacts, not training-seed variability. Full pairwise intervals, checkpoint hashes, Hits@k, and per-query ranks are retained with the artifact.

Partial relation channel. After observing the endpoint reversal but before computing intermediate values, we froze an exploratory channel that independently retains each score-relevant known fact with probability p. If $m = r_{\text{raw}} - r_{\text{filtered}}$ such facts outrank the target, its remaining rank is $r_{\text{filtered}} + U$, $U \sim \text{Binomial}(m, 1-p)$. Exact expectation over this distribution gives Table 21. This is a declared missing-at-random mechanism, not measured annotation missingness.

Table 21: Expected public-model MRR under partial known-relation availability.

Availability	ComplEx+RP	TransE	Difference
0%	0.1746	0.1817	-0.0071
25%	0.1884	0.1932	-0.0047
50%	0.2086	0.2091	-0.0005
75%	0.2432	0.2346	0.0085
100%	0.3886	0.3128	0.0758

The preferred model changes between the 50% and 75% grid points. We do not interpolate a crossover or attach training-seed uncertainty.

Code-generation hard witness. Before downloading generations or outcomes, we froze an outcome-independent panel rule over the official EvalPlus v0.1.0 release: retain every complete temperature-zero HumanEval archive. This gives 31 systems, one generated program for each of 164 HumanEval+ tasks (Liu et al., 2023). Sentence chrF against the canonical implementation is the projected score; functional success requires both base and plus tests to pass. Across all model pairs, 27 reversals have paired-task 95% bootstrap intervals excluding zero under both scores (10,000 replicates).

The preregistered prefix diagnostic then exposed a serialization mismatch: most open-model archives exactly contain the v0.1.10 prompt prefix, whereas several API-model archives use older formatting. We therefore froze a strictly harder post-outcome safeguard before calculating it. The generated and canonical target functions are parsed, docstrings and formatting are removed, and chrF is recomputed over their AST-body dumps without reading tests. A pair is promoted only if it reverses robustly under both projected scores. Four pairs involving five models remain (Table 22); thus the result does not depend on prompt prefix formatting. The primary 27-pair count is reported for completeness, but only these four pairs support the main claim.

Table 22: Code-generation differences in percentage points with paired-task 95% intervals. Each row robustly reverses under raw chrF, AST-body chrF, and functional HumanEval+ testing.

A vs. B	chrF	AST-chrF	Test pass
CodeGen2-1B vs. SCB-7B	$+4.1$ [0.1,8.0]	$+9.6$ [3.8,15.3]	-14.6 [-20.7,-9.1]
CodeGen2-3B vs. SCB-7B	$+3.9$ [0.3,7.6]	$+8.4$ [3.3,13.5]	-7.9 [-12.8,-3.0]
SantaCoder vs. SCB-7B	$+6.6$ [2.9,10.1]	$+13.7$ [8.2,19.0]	-7.9 [-12.8,-3.0]
SCB-7B vs. Vicuna-7B	-6.7 [-10.0,-3.2]	-23.7 [-28.7,-18.5]	$+9.8$ [4.3,15.2]

A.13 A SECOND SYSTEM WHOSE OPTIMIZER FOLLOWS THE SCORE

The covalent local-optimizer audit showed that maximizing a relation-blind steric score drives a rigid ligand to bond-invalid optima. This section asks the same question of an unrelated system with its own declared relations. The protocol was frozen before any optimization ran (`scene_optimizer_audit_20260924/PROTOCOL.md`, SHA-256 `61889cd1...8b02ec6`, 2026-09-24T14:14:57Z; runner hashed at 14:15:49Z, results produced afterwards), and it declared in advance that a null control would be reported in the same place as a positive primary result. It was, and it is reported below.

Setup. Scenes are the 1,000 houses of the pinned ProcTHOR-10K validation split already used in Appendix A.15.1. Boxes, the 1 mm penetration test and the declared parent-child hierarchy are imported unchanged from that audit, and are never inferred from geometry. A declared placement counts as satisfied when the child keeps at least half of its horizontal footprint over the parent and the vertical gap stays within 0.10 m; on the unmodified corpus 99.6% of the 41,660 declared pairs meet the first condition and 100% the second, and the thresholds were fixed from that input distribution before freezing. The cohort is the **847** scenes in which every declared placement is satisfied at baseline, mirroring the covalent cohort of predictions whose declared bond is already valid. The **31,455** movable objects are the leaves of the hierarchy, so a moved object never carries children. Each is translated as a rigid body with $\|\Delta\|_\infty \le R$, everything else fixed; rotation is withheld as the more conservative choice.

Result. Table 23 reports the fraction of cohort scenes whose optimum breaks at least one declared placement. Under the relation-blind objective a local search of a quarter of a metre suffices to break a placement in 91.4% of scenes, and half a metre in 95.7%. As in the covalent audit, "broken" and "broken and strictly better scoring" coincided in every cell: all 10,538 broken optima at $R = 0.5$ m outscore their own starting layout, so a system following this objective would prefer them.

Table 23: Scenes whose optimum breaks a declared placement, by objective and by the bound on translation. Intervals are percentile bootstrap over scenes, 10,000 replicates, seed 20260924. The covalent column is the corresponding row of the ligand audit at its own scale, shown for shape, not for numerical comparison. The gated arm is post-hoc.

Objective	R=0.05 m	R=0.10 m	R=0.25 m	R=0.50 m
Relation-blind	31.5 [28.5, 34.7]	72.0 [69.1, 75.0]	91.4 [89.4, 93.3]	95.7 [94.3, 97.0]
Relation-aware repair	31.3 [28.2, 34.5]	61.2 [57.9, 64.5]	67.4 [64.2, 70.6]	70.0 [66.9, 73.1]
Repair with a placement gate*	0.0	0.0	0.0	0.0

The control does not separate at the smallest bound. The frozen protocol required the relation-aware objective to break placements materially less often, and said that otherwise the result does not support relation blindness as the cause. At $R = 0.05$ m it does not: 31.5% against 31.3%, with overlapping intervals and 414 against 411 objects. Separation appears only as the bound grows, reaching 3.9× at the object level at $R = 0.5$ m, and even there the repaired objective still breaks a placement in 70.0% of scenes. The covalent audit's middle arm fell to 30.5%; this one does not.

The asymmetry has a plain reading. Excluding the declared pair from the penalty removes the incentive to break the placement but supplies no incentive to keep it. Unlike a ligand held in its pocket, nothing else holds a small object to its parent once that pair is excluded. The remedy is the one the covalent audit already identifies: multiplying by an explicit gate on the declared relation leaves no broken optimum at any bound, in either domain. We therefore read this panel as portability of the *failure* and of the *remedy*, and as a boundary on the repair. A corrected measurement is not by itself a compliance guarantee, and how much of the gap it closes is domain-dependent.

Choosing among layouts of one scene. The optimizer arm moves one layout; this arm chooses among several. Protocol frozen before any candidate existed (`scene_selection_audit_20260924/PROTOCOL.md`, SHA-256 `d42d8231...07bdf1b`, 2026-09-24T15:05:29Z; runner hashed at 15:06:09Z). For each of the 847 cohort scenes the pool is the original layout plus sixteen perturbations, four each of jitter (± 0.05 m on every movable object), slide (± 0.30 m horizontally on one), drop (0–0.20 m downward on one) and shove (± 0.50 m on one), seeded 20260924. Four types with different failure modes are used deliberately, since a pool distorted alike cannot produce a selection difference. Candidates are a stand-in for alternative layouts: they are not outputs of a scene generator, and they are not produced by optimizing either score, which would make the comparison circular.

On the 844 pools that contain both a layout meeting the declared placements and one breaking them, the relation-blind count selects a breaking layout on 44.2% of scenes ($[40.9, 47.6]$); the relation-aware count on 20.5% ($[17.8, 23.3]$); paired difference 23.7 points ($[19.9, 27.6]$, interval excluding zero, so the control separates here as the frozen rule required). Unlike the optimizer arm, the repair recovers more than half the gap.

Two limits are structural rather than incidental. The original layout is always in the pool and always meets the requirement, so a correct option exists by construction and these rates are not a difficulty estimate. And the best-of-n curve is not a clean monotone result: candidates enter in generation order, so small pools contain only the mildest perturbation type and the relation-blind count looks better there (100% satisfied at n=2 against 79.5% for the relation-aware count at n=16). We report the curve in `results.json` and draw no conclusion from its shape.

Scope. A rigid-body local optimizer is not a scene generator. This measures reachability of relation-breaking optima by local optimization of the score, not the behaviour of any published scene-synthesis system, and no claim is made that any such system reported an incorrect number. ProcTHOR houses are procedurally generated rather than scanned, and reconstructed AABBs over-report intersection for rotated and concave objects. The gated arm was added after the two frozen arms were read and is labelled post-hoc throughout.

A.14 A DECODER THAT OPTIMIZES THE SCORE, OUTSIDE GEOMETRY

The covalent and scene panels both put a relation-blind score inside a loop that optimizes it, and both are geometric. This section asks whether the same thing happens where the carrier is text. Minimum Bayes-risk (MBR) decoding selects the candidate maximizing mean similarity to the other samples, so the metric is the decoding objective and, because the other samples serve as pseudo-references, it needs no reference at decode time. That is what the code panel could not supply: chrF against a canonical solution cannot enter a generator's loop, since the canonical solution is what is being generated.

The protocol was frozen before any candidate was generated (`term_mt_audit_20260924/PROTOCOL.md`, SHA-256 `0a0ed37b...1e807eb8`, 2026-09-24T14:56:52Z) and fixed the decision rule below. One deviation was recorded before generation and is logged with both file hashes: the sampling call had to set `num_beams=1` explicitly, because both checkpoints ship `num_beams=4` in their generation config and `transformers` rejects 16 return sequences against it. Ancestral sampling is what the protocol specifies, so this implements it rather than changing it.

Setup. WMT21 terminology task, English–French, TICO-19 domain, files pinned by SHA-256. Each source segment carries inline `<term src tgt>` annotations declaring which source term must be rendered as which target term, with alternatives separated by `|`; 498 segments carry at least one. Relations are read only from those annotations. The requirement is met when one declared alternative for every declared term occurs in the output, matched case-insensitively with accents preserved — a check that consumes neither the metric nor the reference. Candidates are 17 outputs per model (one beam-4 search and 16 ancestral samples, temperature 1.0, top-p 0.95, seed 20260924) from `Helsinki-NLP/opus-mt-en-fr` and `facebook/nllb-200-distilled-600M`, 34 per segment.

The decoder follows the metric away from the requirement. Table 24 gives the requirement-satisfaction rate of each decoder. MBR with a chrF utility satisfies the declared terminology on 87.6% of segments; the same MBR with the utility gated on the requirement reaches 94.8%. The paired difference is $+7.23$ points (95% CI $[5.02, 9.64]$, 10,000 segment bootstrap replicates, seed 20260924), which excludes zero and meets the frozen rule. Beam search on the smaller model is the weakest arm at 81.1%; because it optimizes model log-probability rather than the terminology checker, this comparison is descriptive and does not identify why a term is dropped.

Available alternatives and avoidable selection loss. A post-hoc replay separates candidate availability from the choice made by MBR. Of 498 pools, 472 contain at least one candidate that satisfies every declared term. Ungated MBR selects a compliant output on 436, so 36/472 (7.6%) of the pools with an available compliant alternative are lost through selection. The other 26 pools contain no compliant alternative. Gating recovers the 36 by construction; its gain quantifies avoidable selection loss on this frozen pool, not independent evidence that the checker improves translation quality. An exact-term mismatch can reflect inflection or a valid synonym rather than a semantic error. We have not obtained a blinded human quality assessment, so the aggregate endpoint is declared-term compliance only; the common-cold case below illustrates one directly interpretable semantic error.

Table 24: Requirement satisfaction by decoder on 498 WMT21 terminology segments. Intervals are percentile bootstrap over segments, 10,000 replicates, seed 20260924. The gated arm multiplies the MBR utility by the requirement indicator; it is the analogue of the covalent bond gate, not a new metric.

Decoder	Objective	Declared terminology met
Beam search, opus-mt	model log-probability	81.1 [77.7, 84.5]
Beam search, NLLB-600M	model log-probability	87.8 [84.7, 90.6]
MBR, chrF utility	relation-blind, reference-free	87.6 [84.5, 90.4]
MBR, gated utility	relation-aware	**94.8 [92.8, 96.6]**

Reference response and terminology checking. Replacing a segment's reference with another segment's changes which candidate chrF selects on 94.1% of 442 cohort segments, establishing reference sensitivity rather than an isolated response to the requirement. Replacing instead the declared target term *inside* the reference with another term from the corpus breaks the requirement on 340 of 380 segments while costing a median of only 0.085 chrF, and changes the selection on 21.3%. These are interventions on reference text while the original declared terminology stays fixed. They do not by themselves establish semantic reading; the output-side term checks supply the complementary verification evidence.

Does the same objective penalise a compliant human translation? The covalent panel contains a measurement-level result this one otherwise lacks: the published rule rejects structures that satisfy the requirement by construction. We ask the same of the decoding objective already in use rather than of a separate acceptance rule (`PROTOCOL_REFERENCE_ARM.md`, SHA-256 `3ab2bc58...81c938f2`, frozen 2026-09-25T03:51:42Z, before the endpoint was computed; post-hoc with respect to the main protocol). Because MBR scores a candidate by its mean chrF against the *other* candidates, the human reference can be placed into the pool as one more candidate and ranked by exactly the same objective. There is no circularity: the objective never consults it as a reference.

On the 478 segments whose reference meets the declared terminology, the chrF-MBR winner is a model candidate that *violates* the terminology, while the reference satisfies it, on 9.2% of segments ($[6.7, 11.9]$). The reference itself ranks 15.5th of 35 on average and falls outside the top five on 86.6% of segments.

Two limits belong with those numbers. The gated control is uninformative here: multiplying the utility by the requirement indicator makes a violating winner impossible, so its 0.0% is arithmetic rather than evidence, and the frozen decision rule for this arm is therefore not usable. And MBR consensus favours machine-typical text, while a single reference is not the space of acceptable translations, so a model candidate outranking it is not by itself an error — only the 9.2% that also break the declared terminology are. This endpoint is not numerically comparable with the covalent reference-rejection rate: it is a selection event with a different denominator and score.

One segment, read directly. The rates above are easier to trust with a case in front of them. Segment 1005 of the English–French set discusses the 1967 Tristan da Cunha *common cold* data; the task declares *common cold* as *rhume commun*, *rhumes communs* or *rhume*, and *infectious* as *infectieux*. Eight of the 34 candidates satisfy both, and so does the human reference. The beam output of `opus-mt-en-fr` reads "données communes sur le froid": it has attached *common* to *data* and taken *cold* in its temperature sense, so the disease is gone. MBR ranks that candidate first, at utility 62.8, ahead of the `nllb-600M` beam output at 59.9, which keeps *rhume commun*. Gating the same utility selects the latter. Both candidates are unmodified beam outputs; the mixed-model MBR pipeline is constructed for this experiment, not documented as a production default. The segment can be replayed from the frozen `candidates.json`.

Held-out language pair. The protocol's fourth endpoint repeats endpoints 2 and 3 on a second language pair. WMT21 released inline term annotations only for English–French, so we rebuilt the English–Chinese set the way the task itself built one: the task's own published Chinese terminology list (589 entries) applied to TICO-19 English–Chinese. The matching rule

was fixed in `PROTOCOL_HELDOUT.md` (SHA-256 `c77777aa...caf2b442`, frozen 2026-09-24T19:02:33Z) before any Chinese output existed. The declared relation still comes from a published artefact and never from the reference or the model. The organisers' own segment selection is not reproduced, however, so this is a reconstruction of the same construction rather than the released file. Of 3,071 segments, 952 declare a term whose reference can meet it; 1,116 declare none and 1,003 have a reference that does not meet all declared targets, both dropped by the frozen rule. The estimated effect therefore conditions on reference-compatible terminology. This is a target-language transfer within the same TICO-19 source corpus, not an independently collected content holdout. Exact overlap of the retained English-source subsets was not measured in this revision, so no source-disjoint generalization claim is made.

`analyze.py` ran unmodified, byte-identical to the English–French run. The direction holds and the effect is larger. MBR with the chrF utility keeps the prescribed term on 75.6% of the 952 segments ($[72.9, 78.4]$); gating the same utility raises it to 94.1% ($[92.5, 95.6]$). The paired difference is $+18.5$ points ($[16.1, 20.9]$), against $+7.2$ points ($[5.0, 9.6]$) on English–French: same sign, interval excluding zero, so the frozen transfer rule is met. Beam search on NLLB keeps it on 70.9% ($[68.1, 73.7]$). The other endpoints replicate too: the chrF-argmax candidate fails the requirement on 15.4% of the 876-segment cohort ($[13.0, 17.8]$) against 8.8% on English–French; a wrong declaration changes the selection on 94.3% of segments against 94.1%; and breaking the declared term inside the reference changes it on 27.2% against 21.3%.

Two departures are recorded. The frozen `analyze.py` looks up the Marian beam arm by the literal model name of the English–French run, so the `opus-mt-en-zh` beam outputs were generated but not tabulated. We report that arm as not computed rather than patch the analyzer, and no endpoint's decision rule depends on it. The generator also reads the main protocol file to record its hash, which had to be supplied alongside the addendum; the hash it recorded is that of the frozen main protocol.

What is weak here, and what was not run. The selection endpoint is small: on the 442 segments whose pool contains both a satisfying and a violating candidate, the chrF-argmax candidate fails the requirement on 8.8% ($[6.1, 11.5]$). These term-check failures, code archive non-pass labels and geometric violations are different endpoints and should not be compared as a common error rate. The native-optimization gap, while it meets the frozen rule, is modest at 7.2 points here, although the held-out English–Chinese replication above reaches $+18.5$. Two checkpoints are not a survey of translation systems, MBR here reranks a fixed sample pool rather than modifying beam search itself, and terminology is one declared relation among many.

A.15 When does relation-blind scoring change a ranking?

The panels in this appendix include both reversals and nulls. This section states what separates them. It is a post-hoc analysis over frozen result files; no experiment was rerun and no quantity was refitted.

The condition is exact. Write p_i for a candidate's projected score, a_i for its relation-aware score, and $d_i = a_i - p_i$ for the distortion the projection introduces. For $p_i \neq p_j$ (ties are handled separately), $a_i - a_j = (p_i - p_j) + (d_i - d_j)$, so candidates i and j swap places between the two evaluators exactly when

$$\frac{d_i - d_j}{p_i - p_j} < -1. \tag{3}$$

The magnitude of d does not appear. What matters is how d varies across candidates relative to how p does. If every candidate is distorted by the same amount, $d_i - d_j = 0$ and no ranking can change however large that amount is.

One number summarises a panel. Let $\beta = \mathrm{cov}(a, p)/\,\mathrm{var}(p)$ be the slope of the relation-aware score regressed on the projected score. Then $\beta < 1$ is equivalent to $\mathrm{corr}(p, d) < 0$: the projection spreads candidates further apart than the relation-aware evaluator endorses, and $\beta = 1$ says only that the distortion is uncorrelated with the projected score, not that it is constant. Individual pairs can reverse at any value of β, because (3) is a statement about one pair and β is an average over all of them. On the full ImageNet ReaL panel, $\beta = 1.19$ and three of 561 point estimates reverse; none

survives the interval test. A separate relation-bearing subset has nine point reversals and is reported below.

β summarises reversals; it does not predict them. β is not independent evidence about reversal. Summing (3) over pairs gives the identity

$$\beta \;=\; \frac{\sum_{i<j}(a_i - a_j)(p_i - p_j)}{\sum_{i<j}(p_i - p_j)^2},$$

so β is exactly the $(p_i - p_j)^2$-weighted mean of the per-pair ratio $(a_i - a_j)/(p_i - p_j)$, verified to six decimals on every panel. A panel with $\beta < 0$ is one whose weighted average pair already reverses. We therefore report β as a compact description of a panel, and as a screen computable from per-candidate means alone when per-example data for a paired bootstrap are unavailable — not as a discovered law that predicts reversal.

The substantive observation is the one the decomposition isolates: the *mean size* of the distortion does not by itself say whether a ranking changes. ImageNet ReaL understates 34 public checkpoints by 26.0 points on average and reverses nothing robustly, while the knowledge-graph panel distorts by 8.9 points and reverses 25 of 36 pairs. Distortion does not order the panels by reversal: the three largest-distortion panels that are not the covalent gate — GoEmotions (30.7 points), SQuAD (27.7) and ImageNet ReaL (26.0) — all reverse nothing, while the knowledge-graph panel at 8.9 points reverses most of its pairs. Rank correlation between mean distortion and reversal fraction over the fourteen panels is $+0.08$; the Pearson value is $+0.43$, carried entirely by the single high-leverage covalent row. With only two panels reversing at all, neither number is strong evidence, and we rest the point on the direct comparison. Table 25 reports every panel with paired projected and relation-aware scores, and Figure 4 plots them.

Table 25: Every panel with paired projected and relation-aware scores, ordered by β. Intervals are percentile bootstrap over candidates, 10,000 replicates, seed 20260924. Distortion is the mean of $|d_i|$ in the panel's own metric points, so its magnitude is descriptive within a panel rather than a common cross-domain scale. †sacreBLEU is already on a 0–100 scale. Reversals are counted by each panel's own registered rule; β is scale-free.

Panel	n	β	95% CI	$\mathrm{corr}(p, d)$	distortion	point rev.	robust rev.
Covalent validity gate	6	-1.02	$[-4.75, -0.96]$	-1.00	74.2	11/15	8/15 run pairs
Knowledge graphs, cross-architecture	9	-0.61	$[-0.81, -0.40]$	-0.98	8.9	27/36	25
DistMult seeds only	3	1.00	[0.83, 1.17]	0.02	9.3	0/3	0
GoEmotions, full test	5	1.08	[0.87, 1.23]	0.90	4.7	0/10	0
ImageNet ReaL	34	1.19	[1.10, 1.27]	0.82	26.0	3/561	0
SQuAD, full validation	5	1.20	[1.10, 1.25]	1.00	12.2	0/10	0
TriviaQA aliases	5	1.21	[0.65, 2.76]	0.61	9.4	1/10	0
TriviaQA aliases, held-out	5	1.25	[0.92, 1.58]	0.80	10.0	0/10	0
WebNLG multi-reference†	9	1.25	[1.01, 1.38]	0.84	11.0	4/36	0
SQuAD, relation-bearing	5	1.65	[1.19, 1.83]	1.00	27.7	0/10	0
GoEmotions, relation-bearing	5	1.66	[0.27, 2.16]	0.78	30.7	1/10	0
TransE seeds only	3	2.27	[1.21, 2.79]	0.95	7.0	0/3	0
BIRD text-to-SQL	13	2.53	[2.09, 3.06]	0.92	22.2	6/78	0
ComplEx seeds only	3	2.54	[1.32, 7.26]	0.78	10.5	0/3	0

What this does and does not establish. Both equations here are identities, so nothing in this section is a fitted model, and correspondingly none of it is a prediction: β re-expresses the reversals rather than forecasting them. The claim we do make is the negative one — that mean distortion magnitude alone does not determine whether a ranking changes in these panels — and it rests on fourteen panels of which four (n=5) are individually underpowered. Because the panels use different metrics, their distortion magnitudes are not measurements on a shared scale. The panels are also not independent: several share a benchmark, and the covalent rows share tasks. Whether a new domain has small or large β is an empirical question about its candidate pool, not a property of the task, and nothing here bounds how often relation-blind scoring misranks in general.

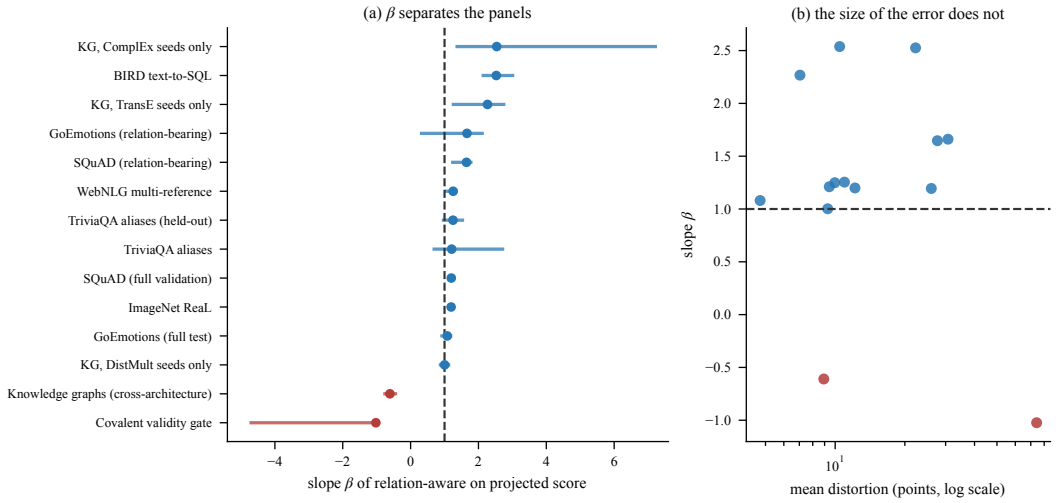

Figure 4: Across these panels the slope separates reversals from nulls; the mean size of the measurement error does not. Red marks panels whose rankings reverse. Left: β with its bootstrap interval, ordered; the line at $\beta = 1$ is where the distortion is uncorrelated with the projected score, not where it vanishes. Right: the same β against mean distortion, which does not order the panels.

A.15.1 Boundary tests, evaluator screens and null results

These tests bound the claims in Sections 6 and 8. They include both observed counterexamples and experiments that did not meet their prespecified criteria. They are not pooled with the four main-domain audits.

These experiments are auxiliary scope or boundary probes. They do not contribute to the main empirical claim and are not a prevalence estimate. Positive, weak, and null outcomes are retained to delimit transfer rather than to enlarge the evidence base.

Text-to-SQL boundary result (frozen null). Before inspecting predictions, we mechanically selected all 13 complete SQLite system files at a pinned BIRD Mini-Dev commit and evaluated the common 500 SELECT-only questions. The projected score is lowercase lexical SQL-token F1; the relation-aware score is official set-valued execution equivalence on the released read-only databases. Token F1 ranges from 0.541 to 0.703 and execution accuracy from 0.214 to 0.630. Six of 78 system pairs reverse in point estimate, but none has 95% paired-question intervals excluding zero under both scores. The predeclared requirement of three robust reversals involving three base models therefore fails. This panel shows per-query construct disagreement but is not a third hard witness and is not pooled with code generation.

The reference-oracle deficit. `oracle_deficit.py` computes the reference-oracle deficit on both benchmarks. The covalent branch reads only `corrected_all.csv`: the untyped extremum is relation-covered in 98.4% of the 4,587 structures, and the published gate rejects 98.7% of these specification-faithful references. The knowledge-graph branch reads only the answer-multiplicity distribution and, under exchangeability among known correct alternatives, returns expected raw MRR $\mathbb{E}_k[H_k/k] = 0.3680$, against 0.3683 for the separately measured SHA-ordered oracle. The residual reflects the fixed ordering rather than the exchangeable expectation. Neither branch touches model output. These quantities diagnose an evaluator's treatment of correct references; they do not upper-bound arbitrary predictors, which may score higher by exploiting the erased relation.

The relation is not expressible in the released interface. The screen above is a statement about definitions. For the rule whose consequence we measured, the same point can be checked in the released implementation, and it is narrower and more concrete than a definitional claim. PoseBusters exposes an exclusion list for its intermolecular distance test — hydrogens, the protein, organic cofactors, inorganic cofactors, waters — and defaults to excluding hydrogens alone, with

the clash cutoff at 0.75 that we audit. Its bond information is read from the ligand molecule, so an inter-molecular bond has no representation. A user who knows that a particular protein–ligand pair is covalently bonded therefore has no way to say so, and the released documentation does not name the case. We inspected `posebusters/modules/intermolecular_distance.py` and `distance_geometry.py` at the repository's main branch.

This is the sense in which the false penalty is reachable without contrivance: the tool already carves out relation types whose short contacts are expected, and a declared covalent bond is not among them. We are not claiming the authors erred; the rule is documented for non-bonded pairs. We are recording that the case is silent rather than refused, so a covalent benchmark run through it returns a near-total rejection with no warning attached.

Screening published evaluator definitions. `evaluator_screen.py` reimplements ten scoring rules at the abstract signature and probes each. Scene rules are reduced to axis-aligned boxes rather than meshes, which is faithful for the property under test because erasure depends on which pairs are eligible, not on how finely intersection is computed. A row is a statement about a definition. It does not assert that the authors erred, that a system misreports on its own data, or that any published number is wrong—only that the rule cannot condition on a relation the task declares. Where we measured the consequence, the table says so.

Table 26: Erasure screen over published scoring rules. "Admissible τ" is the interval of thresholds satisfying R1 and R2 on the witness pair, empty exactly under erasure. The four non-erasing rows consume the relation explicitly.

Scoring rule	Source	Erasing	Measured consequence
Minimum-distance ratio	Buttenschoen et al. (2024)	Yes	98.7% of 4,587 crystals rejected
Clash count	Harris et al. (2023)	Yes	4.86 $\rightarrow$ 1.16 per complex
Probe-style primitive	Chen et al. (2010)	Yes	3.16 $\rightarrow$ 1.10 per 1,000 atoms
Raw entity ranking	Bordes et al. (2013)	Yes	oracle MRR 0.368
Single-designated-label top-1	Beyer et al. (2020)	Yes	3/561 model pairs reverse
All-pairs scene collision	Tam et al. (2025)	Yes	96.7% of 1,000 scenes rejected
Filtered entity ranking	Bordes et al. (2013)	No	oracle MRR 1.000
Multi-label (ReaL) accuracy	Beyer et al. (2020)	No	6/34 per-image ranks change
Relation-aware scene collision	this work	No	86.4% rejected; 10.3% recovered
Graph-aware sterics	this work	No	95.8% of the same 4,587 accepted

Prospectively frozen public-API holdout. To separate the RIA prediction from case selection, we froze the evaluator configurations, dataset URL, intervention, graph IDs, four mechanism thresholds, and one stronger corpus prediction before downloading MUTAG or executing either API. The dated protocol is project-local rather than third-party registered; its deposited SHA-256 is `3cbf99c17d645fad986c47b8290e6d7ac51fdecbabc56ad19d0f8fbe03424209`. MUTAG contains 188 typed molecular graphs (Morris et al., 2020). We preserve each topology and edge-label histogram while rotating label assignments, then call the unmodified NetworkX WL-hash and graph-edit-distance APIs (Hagberg et al., 2008; Shervashidze et al., 2011). Default configurations respond on 0/188 WL cases and 0/32 digest-selected GED cases; passing the edge attribute or categorical matcher raises both to 100%. Thus all four frozen mechanism predictions pass. The stronger prediction does not: edge-aware hashing was predicted to reduce cross-class hash collisions, but both configurations already have zero. We retain this failed endpoint and conclude only that the holdout corroborates information sensitivity, not a MUTAG classification gain.

Scene-collision measurement. `scene_collision_audit.py` processes all 1,000 houses in the pinned ProcTHOR-10K validation split (Deitke et al., 2022): 68,838 objects, 3,335,440 within-scene pairs, and 41,660 declared parent–child placement edges. It reconstructs world AABBs from the pinned asset dimensions and stored poses, then uses the same fixed rule in every arm: intersection with more than 1 mm penetration on all three axes. Relations come only from the stored `children` hierarchy and are never inferred from geometry.

The untyped rule flags 21,670 pairs and rejects 967/1,000 scenes. Of those pairs, 11,929 (55.0%) are direct declared edges. Excluding direct edges (r1) leaves 9,741 flagged pairs and rejects 864

scenes; hence 103 scenes (10.3%) fail only because of relation-covered contact. Excluding graph-distance-two pairs as well (r2) leaves 8,003 flagged pairs and rejects 806 scenes. As the mandatory sensitivity control, each scene receives one rigidly translated overlap between objects more than two relation edges apart. Both r1 and r2 catch all 1,000 injected faults. The audit retains and reports 139 non-related pairs with AABB IoU above 0.10 as possible missing annotations rather than excluding them.

ProcTHOR is procedurally generated rather than scanned, and reconstructed AABBs can over-report intersection for rotated or concave meshes. The parent–child hierarchy also combines support and attachment. These limitations affect the magnitude, so the result establishes a corpus-level consequence for the rule, not an error in any system's published results. The original five-object scene remains an executable unit test: all-pairs collision rejects its faultless declared-containment case, while relation-aware collision accepts it and still detects an unrelated injected overlap.

Image-classification measurement. The primary project-local protocol and SHA-256 companion were frozen before computing ReaL outcomes. It includes every configuration with released 50,000 by 5 prediction and probability arrays at pinned `cgnorthcutt/benchmarking-keras-pytorch` commit `74df938db0f266947abedb2ec9f741af65e2d8ed` (Northcutt, 2019): 13 Keras and 21 PyTorch configurations. The 68 source arrays are individually checksum-pinned. We reconstruct the source `ImageFolder` order from official filename–WNID assignments and recover the repository's 34 reported original-label accuracies within 0.02 percentage points. The sole non-rounding match is PyTorch ResNet-152: its released arrays give 77.606% while the README prints 77.62%; this tolerance was fixed after the sanity check stopped and before any ReaL score was computed.

Both scores then use the same 46,837 images with nonempty ReaL label sets. Among all 561 model pairs, three strictly reverse order, six model ranks change by one place, and Kendall $\tau = 0.9893$. The largest reversal is Keras ResNet-50 versus PyTorch VGG13-BN: the designated-label gap is $+0.0811$ points, whereas the ReaL gap is -0.4825 points. The scoring change favors the former on 1,357 images and the latter on 1,532, with a net difference of 264 images. Prespecified paired bootstrap intervals (10,000 image resamples, seed 20260909) are $[-0.278, 0.442]$ and $[-0.852, -0.132]$ percentage points, respectively. Thus the ReaL difference excludes zero, but the original difference does not. The other two reversals are also not significant on both sides; consequently zero of three meet the frozen robust-inversion criterion. The defensible conclusion is that relation-aware scoring changes concrete choices on actual public model outputs, not that such reversals are frequent or statistically robust across a model population. Related architectures and frameworks are not treated as independent samples.

After observing the outcome, we added one explicitly exploratory interval for the attribution contrast already required by the protocol: (ReaL gap) minus (designated-label gap). For the largest reversal it is $[-0.843, -0.293]$ percentage points, excluding zero. Thus the scoring change affects the two configurations differently, even though it does not make the pair a robust inversion under the frozen two-interval rule. The post-outcome status and unchanged decision rule are recorded in a separate dated note.

As a separate scale audit, every unique model–image-size configuration in both public timm result files at pinned commit `a23518b6625b98a2b711e82061f2f2224f057668` (Wightman, 2019) is included: 1,556 configurations and no unmatched rows. Among 1,209,338 pairs with unequal released aggregate top-1 values in both evaluations, 43,003 reverse order (3.56%); 1,522 ranks move, by up to 238 places ($\tau_b = 0.9287$). These files contain rounded aggregate outputs rather than paired predictions, so this larger audit has no image-level uncertainty analysis and is corroborative rather than a substitute for the 34-configuration experiment above.

`vision_multilabel_audit.py` reads two pinned label files and no images: the reassessed ILSVRC-2012 label sets (Beyer et al., 2020) and the designated validation labels. Of 50,000 images, 3,163 carry no declared valid label and are excluded by the annotation's own instruction, leaving 46,837. Two distinct effects reduce the credit an erasing rule gives a specification-faithful predictor, and they factor exactly: on 4,673 images (10.0%) the designated label is not declared valid at all, which is label noise; on the remaining 42,164 the designated label is declared valid but 15.3% of those images admit more than one, which is erasure. Only the second is reported as the 8.5% deficit. The two scoring rules carry no geometry, so the vision rows also show that the probe of Section A.12

tests a signature rather than a spatial construct. This label-only calculation is separate from the downstream ranking audit; neither analysis claims that either published score was miscomputed.

Second KG dataset (prospective boundary result). Before outcome inspection, we checksum-froze the standard WN18RR (Dettmers et al., 2018) train/validation/test splits and transferred the complete FB15k-237 recipe without dataset-specific tuning: DistMult, TransE, and ComplEx at seeds 11–13, with filtered validation MRR selecting each checkpoint. Filtering changes 49.8–53.6% of the 6,268 test queries. DistMult's raw MRR is 0.222–0.241 and filtered MRR 0.356–0.358; ComplEx moves from 0.216–0.242 to 0.356–0.360; TransE moves from 0.153–0.155 to 0.193–0.195. However, none of the nine matched-seed architecture comparisons reverses, so 0/9 are robust reversals and the predeclared promotion gate fails. WN18RR thus reproduces large query-level reference deficits but not the FB15k-237 model-order consequence. This boundary result is not pooled with FB15k-237.

Relation-bearing vision subset (post-hoc null). After the full-dataset null, we froze a model-independent subset rule before reanalysis: retain the 6,448 images with at least two ReaL labels for which the original designated label remains acceptable. Among the same 34 public models and 561 pairs, nine point estimates reverse between designated-label and acceptable-set top-1 accuracy. None has paired bootstrap intervals excluding zero under both evaluators, and none survives the additional two-sided paired exact-test Benjamini–Hochberg sensitivity gate. Thus even the relation-bearing subset remains a null for robust model-order reversal.

Controlled multi-label objective test (null). We also froze a GoEmotions (Demszky et al., 2020) test before fitting: word/bigram TF–IDF plus the same linear logistic family, trained either on the first stored label or on the complete 28-label multi-hot target, across five seeds. On all 837 multi-label test examples, first-label training is favored by the projected evaluator by 2.7–3.8 points, with every interval excluding zero. The complete-label model is favored by the relation-aware evaluator by 0.5–1.6 points, but all five intervals include zero; hence 0/5 seeds pass the predeclared robust-reversal gate. A separate five-public-checkpoint pilot has one point reversal among ten pairs on the 837 relation-bearing examples, but zero paired-bootstrap-robust reversals; the full 5,427-example test has none. Because public checkpoints use heterogeneous training pipelines and two expose only generic label IDs, this pilot is retained only as a bounded null, not as a model comparison. These retained nulls prevent the molecular/KG evidence from being generalized into an unsupported three-domain prevalence claim.

Multi-reference QA and generation screens (frozen nulls). We next froze three acceptable-reference-set audits before model outputs or scores. On SQuAD (Rajpurkar et al., 2016), five public extractive-QA checkpoints were evaluated on the 4,655 examples with at least two distinct normalized answers. Accepting all declared answers raises individual exact match by 10.1–32.7 points, but produces 0/10 point or robust pairwise ranking reversals. On a deterministic 2,000-question TriviaQA (Joshi et al., 2017) alias panel, five public checkpoints produce one point reversal and zero robust reversals. A prospectively frozen confirmation then excluded those 2,000 questions and evaluated the same five checkpoints on all 2,176 remaining eligible questions in the pinned shard. It produces 0/10 point and 0/10 robust reversals, so the initial point reversal does not replicate. Finally, all nine complete official WebNLG 2017 systems (Gardent et al., 2017) were scored by the same case-sensitive SacreBLEU configuration against reference 0 or all three official references. Four of 36 system orders reverse in point estimate, but none has 95% paired bootstrap intervals excluding zero under both evaluators. These experiments show large per-output credit deficits and possible selection changes, but fail the predeclared third-domain gate; none enters the main empirical claim.

Domain-agnostic relation-intervention audit. `relation_probe.py` implements the reading test (RIA). It also calculates the error forced by information the evaluator cannot see: outputs that look identical to it may require different verdicts. `fibre_error_bound` groups weighted binary labels by the representation visible to the evaluator and computes Eq. 4 exactly. `relation_intervention_audit` accepts paired covered and uncovered relation states, then reports response rate, correct-direction rate, signed margin, global threshold feasibility, and the balanced erasing-class error floor. For every screened witness that floor is 50%: the difference is that the

six erasing rules attain zero response and have no feasible threshold, whereas all four relation-aware rules respond correctly and separate the arms.

The lower-level function `is_relation_erasing` evaluates the same carrier under four relation variants—as declared, with E emptied, with one edge removed, and with one edge retyped—and reports whether the score is invariant on this finite probe. A changed score disproves global erasure; an unchanged score is a local witness, not by itself a proof about every possible input. `erasure_witness` constructs a pair requiring opposite verdicts from the risk argument in Appendix A.16 and returns the threshold interval satisfying R1 and R2, which is empty exactly when the evaluator is erasing. `run_probe.py` runs both on a protein–ligand complex and on a rigid assembly. The verdicts agree across domains: the untyped evaluator is erasing and admits no threshold; the graph-aware evaluator is not erasing and admits $\tau \in (0.517, 1.317]$ for the complex and $\tau \in (0.450, 1.844]$ for the assembly, both containing the published operating point.

`corpus_intervention_audit.py` then applies the same summary to all 4,587 deposited carriers using the paired canonical score columns. The untyped rule has response and correct-direction rates 0; the graph-aware rule has both rates 0.9839, mean signed margin 0.3599 and zero wrong-direction cases. At the published threshold it changes 4,336 decisions from rejection to acceptance. The global interval is empty because residual short contacts make a single steric threshold insufficient for every carrier. RIA exposes this rather than calling the repair perfect, and R5 assigns excluded relation geometry to the separate centre test.

Partial-erasure calibration. `partial_erasure_calibration.py` treats the response indicator at a randomly sampled carrier as Bernoulli with population mean ρ. For each of six response rates and three query budgets, it draws 2,000 independent audit replicates under seed 20260903. Table 27 reports the hardest nonzero rates plus the zero-response control. The distribution-free interval is conservative, as expected. At $\rho = 0.10$, 512 probes reduce mean absolute error to 0.0105 and give a strictly positive 95% lower bound in every replicate; the zero-response control produces none. The closed-form sufficient budgets for absolute error 0.10, 0.05, and 0.025 at 95% confidence are 185, 738, and 2,952 probes.

Table 27: RIA calibration under partial relation response. "Detected" is the fraction of 2,000 runs whose Hoeffding lower endpoint is above zero.

True ρ	Probes	Mean estimate	MAE	Detected (%)
0.00	512	0.000	0.0000	0.0
0.10	32	0.101	0.0418	1.2
0.10	128	0.100	0.0210	20.6
0.10	512	0.100	0.0105	100.0
0.25	32	0.247	0.0607	55.4
0.25	128	0.249	0.0300	100.0
0.25	512	0.250	0.0156	100.0

The tolerance ϵ defines the estimand ρ_ϵ rather than being hidden in the test. If each score evaluation has bounded numerical error η, choosing $\epsilon > 2\eta$ prevents jitter alone from creating a response; stochastic evaluators should instead be queried repeatedly before forming the paired difference.

A.16 Loss-aware and partial-relation derivations

Two separate arguments live in this section, and they answer different questions.

- **Proposition 1** (the molecular result): once the bonded pairs are dropped, the clash score's *maximum* sits at a state with no bond. This is what makes the failure reachable by anything that maximizes the score.
- **Theorem 1** (the general result): an evaluator cannot recover a distinction that never reached its inputs, and the size of the unavoidable error is exactly the amount of opposite-verdict mixing inside what it can see.

The first is proved immediately below; the second occupies the rest of the section.

For a finite declaration-dependent mask $\mathcal{A}(E)$, define the clipped non-bonded score $\bar{s}_E(X) = \min\{1, \min_{(i,j)\in\mathcal{A}(E)} \|x_i - p_j\|/a_{ij}\}$ with finite positive a_{ij}. This is the object used by the repaired molecular score s_2: it drops the pairs created by the declared bond but adds no positive check that the bond formed.

Proof of Proposition 1. Starting from $X_0 \in \mathcal{D}_E$, translate the entire ligand by tv along the ray guaranteed by the domain assumption. The reverse triangle inequality gives $\|x_i + tv - p_j\| \geq t - \|x_i - p_j\|$. Finiteness implies that all eligible normalized distances eventually exceed one, so $\bar{s}_E = 1$. Meanwhile the declared cross-component bond length can be made arbitrarily large. Hence an invalid state attains the score's global maximum. □

Proof of Theorem 1. The idea is the one stated in the main text: group outputs by what the evaluator can see, and note that it must give every member of a group the same answer. A group is called a *fibre*. If one fibre holds outputs that require opposite verdicts, the best the evaluator can do is pick whichever answer is cheaper there and be wrong about the rest. Equation 4 does exactly that—least costly decision per group, summed—and the refinement inequality that follows it says that handing the evaluator more information can only shrink the total.

A typed output is $X = (Z, E, \sigma)$: a carrier Z, task-defining typed relations E, and a joint provenance tag σ. An evaluator observes $T = (Z, W)$ with $W = r(Z, E)$ its relation message, which may be all of E, part of it, a noisy summary, or none. Here *carrier* means the output without its declared relations, and *message* means the relation information actually supplied to the evaluator.

Definition 1 (Relation-type erasure)**.** *An evaluator is* relation-erasing *if its action factors through $q(X) = Z$: it is constant on every carrier fibre and cannot condition on E. It is* partially erasing *when it factors through a non-injective message $W = r(Z, E)$.*

This is the object the main text calls a score that does not read the relation.

For finite target Y, finite action set $\mathcal{A}$, loss L, and a finite-support observed message T with $\mu_t(y) = P(T = t, Y = y)$, the minimum risk of every deterministic or privately randomized evaluator observing only T is

$$R^\star_T = \sum_t \min_{a\in\mathcal{A}} \sum_y \mu_t(y) L(a, y). \tag{4}$$

The inner minimum chooses the best action for one visible group; the outer sum adds its error to that of the other groups. Randomizing the action cannot improve this linear conditional risk.

Let $T_2 = (Z, W)$ refine $T_1 = g(T_2)$. For a coarse fibre t, define the unnormalized conditional loss of action a in refined fibre u as $\ell_u(a) = \sum_y P(T_2 = u, Y = y) L(a, y)$. Its contribution to the coarse Bayes risk is $\min_a \sum_{u:g(u)=t} \ell_u(a)$, whereas the refined contribution contributes $\sum_{u:g(u)=t} \min_a \ell_u(a)$. The latter cannot be larger. Equality holds iff one action lies in every positive-mass $\arg\min_a \ell_u(a)$: equality of the sum requires every nonnegative optimality gap to vanish. Summing over t proves Theorem 1. The same risk formula covers a stochastic relation channel by taking T_2 to include its observed message. Private evaluator randomness cannot lower risk because conditional loss is linear over the action simplex; its minimum occurs at a deterministic extreme point.

For budgeted recovery, each subset S with $|S| \leq k$ defines a refinement $T_S = (Z, E_S)$. Applying the preceding inequality whenever a relation is added gives $R^\star_{\text{full}} \leq R^\star_{k+1} \leq R^\star_k \leq R^\star_Z$ after minimizing over admissible subsets. Equality with the full-information risk holds exactly when at least one admissible S merges only refined fibres that share a Bayes action. This is a finite combinatorial characterization, not a claim that selecting the best subset is computationally easy.

For binary actions under false-accept and false-reject costs, the two constant actions in a fibre incur $c_{\text{FA}}\mu_t(0)$ and $c_{\text{FR}}\mu_t(1)$, giving the stated minimum. For K-class 0–1 loss, choosing the largest mass gives $1 - \sum_t \max_y \mu_t(y)$. For one relevant item with posterior probabilities $p_i(t)$, a ranking π has conditional reciprocal-rank utility $\sum_i p_i(t)/\pi(i)$. The rearrangement inequality maximizes it by sorting $p_i(t)$ decreasingly, yielding $\sum_j p_{(j)}(t)/j$; Hits@k analogously sums the k largest masses. With multiple relevant items, dependencies enter the relevance-set loss and the general finite-action formula applies directly.

Conditional Fano gives $H(Y \mid Z, W) \le h_2(\epsilon) + \epsilon \log_2(K-1)$ for $\epsilon \le 1 - 1/K$. Subtracting this from $H(Y \mid Z)$ yields the necessary bound $I(Y; W \mid Z) \ge [H(Y \mid Z) - h_2(\epsilon) - \epsilon \log_2(K-1)]_+$. If W is generated from (Z, E) plus independent noise, conditional data processing also gives $I(E; W \mid Z) \ge I(Y; W \mid Z)$. This is a necessary bit budget, not a sufficient rule for choosing relations.

Finally, if a detectable response occurs with probability at least ρ_0, independence gives $P(\text{no response}) \le (1-\rho_0)^N$, whose inversion is the RIA power bound. For paired differences $D_i \in [-1, 1]$, Hoeffding gives $P_0(\bar{D} \ge c) \le e^{-Nc^2/2}$ under $\mathbb{E}D \le 0$, and $P_\Delta(\bar{D} < c) \le e^{-N(\Delta - c)^2/2}$ under $\mathbb{E}D \ge \Delta > c$. Setting $c = \sqrt{2\log(1/\alpha)/N}$ proves the level and power statements in the main text. For a lower bound, restrict D to $\{-1, 1\}$ and compare $P_0 = \mathrm{Ber}(1/2)$ with $P_\Delta = \mathrm{Ber}((1+\Delta)/2)$. The optimal sum of errors is $1 - \mathrm{TV}(P_0^N, P_\Delta^N)$, and the two-point inequality gives $\alpha + \beta \ge \frac{1}{2}\exp[-N\mathrm{KL}(P_0 \| P_\Delta)]$. Since this one-sample divergence is $\Theta(\Delta^2)$ near zero, every uniformly valid test needs $N = \Omega(\Delta^{-2})$ for fixed nontrivial errors, matching the Hoeffding upper rate up to constants. These calculations quantify a declared alternative; they cannot certify global erasure from a finite null response.

For the minimax-regret statement, fix any Z-only rule f. For every $P \in \mathcal{P}$, optimality of the fibre rule gives $R_P(f) - R^\star_{(Z,E)}(P) \ge R^\star_Z(P) - R^\star_{(Z,E)}(P) = \Delta(P, L)$. Taking the supremum over P, then the infimum over f, gives

$$\inf_f \sup_{P \in \mathcal{P}} [R_P(f(Z)) - R^\star_{(Z,E)}(P)] \ge \sup_{P \in \mathcal{P}} \Delta(P, L). \tag{5}$$

The comparator is a distribution-specific relation-aware Bayes oracle; replacing it by a composite minimax optimum would require additional assumptions. In particular, if no coarse fibre contains task-opposed states, Δ is zero and no universal positive gap exists.

Proposition 3 (Selection regret under a missing relation)**.** *For candidate utilities $u_j \in [0, 1]$, let a score policy have regret $\mathcal{R}(s) = \mathbb{E}[\max_j u_j - \sum_j \pi_s(j) u_j]$. If two equiprobable relation states share the evaluator observation and their unique optimal candidates swap with utility gap γ, every relation-blind policy incurs regret at least $\gamma/2$, while a relation-aware policy can incur zero. Conversely, if $\|s - u\|_\infty \le \epsilon$, deterministic score maximization has regret at most 2ϵ.*

In words: when two different requirements look identical to the score but call for different winners, no relation-blind rule can be right about both, and it gives up at least half the utility gap between those winners; a rule that reads the relation can give up nothing. The second half is the ordinary bound in the other direction — a score everywhere within ϵ of true utility loses at most 2ϵ — so the first half is about missing information, not about approximation error.

The first claim follows by summing the two statewise errors of the same selection distribution; the second is the standard two-approximation argument. It establishes possible selection regret; it does not by itself predict amplification or an empirical curve in n.

For Proposition 3, let a Z-only policy select the first candidate with probability p. In the first relation state its regret is $\gamma(1-p)$ and in the second it is γp; averaging equiprobable states gives $\gamma/2$. A relation-aware policy selects the statewise optimum. For the upper bound, let a maximize s and b maximize u. Then $u_b - u_a \le (u_b - s_b) + (s_a - u_a) \le 2\epsilon$.

For routing, let A be acceptance and C the event that the decision meets the registered contract. If conditional loss on $A \cap C$ is at most R_E, abstention costs at most λ, and loss on $A \cap C^c$ is at most $L_{\max}$, then the disjoint events give $R_{\text{route}} \le P(A \cap C) R_E + P(A^c)\lambda + P(A \cap C^c) L_{\max}$. An external bound $P(A \cap C^c) \le \eta$ gives the corresponding last term $\eta L_{\max}$. Passing a finite audit does not itself establish η or R_E for future inputs; this conditional accounting identity is not a deployment-risk guarantee inferred from the probes.

`assembly_demo.py` builds 200 assemblies in each of four populations under a fixed seed: correct, a distal violation that drives an unrelated part into the bracket, an over- or under-inserted joint, and a badly angled joint. Radii, the ratio form and $\tau = 0.75$ are carried over unchanged.

Discrimination between correct assemblies and each corruption behaves as it does on crystal structures: for the distal violation, AUROC rises from 0.648 untyped to 1.000 graph-aware; for the badly angled joint it rises from 0.470—below chance, the same inversion the covalent bond produces—to 1.000. For the over-inserted joint neither steric reading discriminates (0.520 either way), which is precisely the coverage gap R5 exists to close and which the separate joint test fills.

Table 28: Acceptance (%) on the rigid-assembly instance, 200 assemblies per population, seed 20260901. A valid evaluator accepts the first row and rejects the other three. The pattern matches the crystallographic result exactly, including the inverted response to an over-long declared relation.

Population	Untyped	Graph-aware	Joint test	Composite
Correct assembly	0.0	100.0	100.0	**100.0**
Distal violation	0.0	0.0	100.0	**0.0**
Over/under-inserted joint	47.5	100.0	0.0	**0.0**
Badly angled joint	0.0	100.0	0.0	**0.0**

Regression tests in `analysis/tests/test_erasure.py` pin this down. The probe labels both evaluators correctly in both domains. The witness admits no threshold under erasure, and admits the published one after the correction. Untyped sterics reject correct assemblies, and the exclusion restores them without amnestying the distal violation. An impossible joint passes the corrected steric test and is caught only by the composite. The new regression cases also verify the exact fibre-wise bound, finite-sample guarantee, partial-response calibration, the frozen holdout, and invalid inputs.

The rigid-assembly example is retained as a synthetic unit test, not as evidence about a published assembly system. The knowledge-graph and ProcTHOR experiments supply measured cross-domain precedents; together they show that the failure follows from the evaluator's information loss, making the fibre-risk argument in Section 2 a statement about evaluators rather than about molecules.

A.17 Comparison across steric primitives

To distinguish a relation-typing problem from a package-specific bug, we apply three steric primitives derived from published definitions to the same 247 loadable drug-like crystal complexes. They are the PoseBusters normalized-distance gate, the PoseCheck-style absolute-overlap count, and the Probe-style overlap count underlying the MolProbity clashscore (Chen et al., 2010). For each reference we create one 2.60 Å over-long target bond and one distal 1.0 Å corruption. The "untyped" versions examine every protein–ligand pair; the "relation-aware" versions apply the identical primitive after removing cross-link 1–2 and 1–3 pairs. This is a formula-level comparison using common coordinates and radii, not execution of the complete MolProbity or Phenix pipelines.

On the unmodified crystals, the declared bond is the worst untyped pair in 247/247 cases. The untyped PoseBusters gate accepts 2.4%, versus 92.3% after relation-aware exclusion. Mean PoseCheck-style counts fall from 4.86 to 1.16 per complex, and mean Probe-style counts fall from 3.16 to 1.10 per 1,000 atoms. Table 38 shows that none of the three steric primitives replaces the reaction-centre test: graph-aware sterics recover distal-clash sensitivity, whereas the separate centre score detects an over-long bond. For comparison, the separate reaction-centre score has AUROC 0.967 on the over-long target bond and 0.509 on the distal clash, confirming that the two components measure complementary constructs.

B Selection results and preregistered boundary test

This appendix supports Section 4. Figure 5 places the mixed-model and within-generator results side by side; Table 29 states the two registered hypotheses and their outcomes. The same tasks are used, but candidate sources and pool-size grids differ. Detailed residual errors and endpoint checks are in Section C.2.

Preregistration record for the within-generator test. Table 29 restates the two timestamped hypotheses next to what was observed, so that the preregistration can be checked without reading the figure. The analysis was run once, on 292 tasks and 74 clusters, after the hypotheses, the task-inclusion rule, and the analysis procedure were publicly timestamped and before any new candidate was generated.

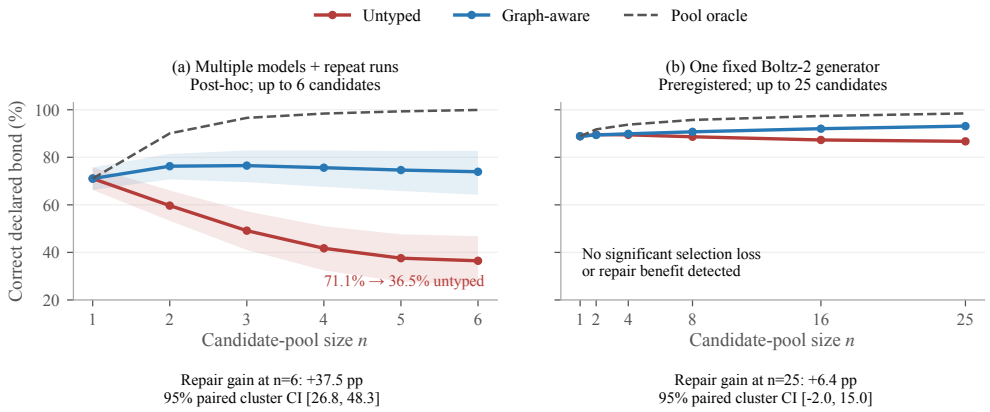

Figure 5: Selection cost and its boundary, on the same 292 tasks and 74 clusters. (a) Exact post-hoc subset analysis of six cross-model/repeat-run candidates. (b) Preregistered analysis of 25 samples from one fixed generator. Both panels use the same vertical scale; their candidate sources and size grids differ. Left shading shows pointwise 95% cluster-bootstrap intervals; the right curves are point estimates. The paired repair contrasts below each panel, not overlap of curve intervals, quantify uncertainty in the gain. The prospective gain interval includes zero, as does the selection-loss interval $[-5.1, 9.6]$ pp. Neither contrast establishes a universal effect of increasing sample count.

Table 29: The two preregistered predictions and their outcomes. Neither is supported; both are reported as registered rather than reinterpreted.

	Registered prediction	Observed	Verdict
H1	Original-score selection declines as samples rise from 1 to 25	88.9% → 86.7% ($[-5.1, 9.6]$ pp)	Not supported
H2	Repaired score improves selection within the same generator	93.1% at $n = 25$ (interval spans zero)	Not established

B.1 Temporal and link-type sensitivity

Splitting at 1 January 2023 gives 3,621 earlier and 966 later structures. The covalent-aware pass rate is 95.6% [94.9,96.3] before the split and 96.4% [95.0,97.4] after it. A reference distribution built only from pre-2023 drug-like complexes places 92.2% of later structures inside its nominal 5th–95th percentile band. Entries with only a `covale` edge pass at 96.6%; entries also containing metal coordination pass at 93.7%.

C Selection records, reproducibility and exploratory extensions

The main selection and displacement results are documented in Sections C.2 and C.3. Selector training and the three-task steering pilot are separate exploratory extensions (Sections C.4 and C.5); their outcomes do not replace the main cohorts.

The anonymous artifact records checksum-pinned best-of-n inputs, frozen per-query ranks for the five-public-checkpoint KG panel, and all nine controlled training checkpoints with configurations and traces. It also records the nine-run WN18RR replication, the independent TriviaQA holdout, before-outcome protocols, domain-neutral RIA code, audit outputs, and regression tests. One command verifies the best-of-n input hashes, recomputes its headline table and cluster bootstrap, and replays both KG analyses and tests. Third-party public checkpoints and raw molecular coordinates are not redistributed; the packaged ranks are sufficient for their reported decision results. Deposited-reference and corruption construction is documented in Appendix A.2 but is not reconstructed from raw PDB coordinates by the one-command replay; the HumanEval+ panel is retained as frozen

per-task scores rather than third-party generation files. No archival DOI is claimed in this working draft.

C.1 RIA-BASED EVALUATOR ROUTING

The routing protocol was checksum-frozen before computing outcomes. This experiment screens declaration coverage: it selects an evaluator only if its stored RIA screen reports relation response, the required direction, and a feasible witness threshold. This is not a pass of the satisfaction audit; the selected graph-masked score can still prefer a broken bond. If the declared relation is missing or cannot be resolved, the router returns ABSTAIN; it never falls back to an erasing evaluator. This operates on evaluator contracts and relation-schema status, not ground-truth test labels.

For FB15k-237, raw ranking fails the screen and filtered ranking passes. Raw MRR selects TransE; routing selects filtered MRR and therefore ComplEx+RP, matching the declared relation-aware target. This is one model-choice decision, not an estimated error rate. For the four primary molecular runs, the untyped 0.75 threshold disagrees with the strict graph-aware verdict on 902/1,168 resolved model–task cases. The router uses the strict evaluator on all 1,168 and abstains on the eight unresolved cases (two per run), with zero unsafe fallbacks. Because the target is contract-defined, zero routed disagreement is a consistency property rather than independent evidence of physical correctness. Protocol, input hashes, per-run counts, and tests are stored in `RIA_ROUTING_PROTOCOL.md` and `ria_routing_results.json`.

C.2 SELECTION ERRORS AND INDEPENDENT GEOMETRIC ENDPOINTS

This section expands Sections 4.1 and 4.4. Table 30 compares bond, pose and angle outcomes; Table 31 attributes the remaining selection errors to invalid maxima, ties or absent valid candidates. All use the same frozen candidate pool and tie rule as the main analysis.

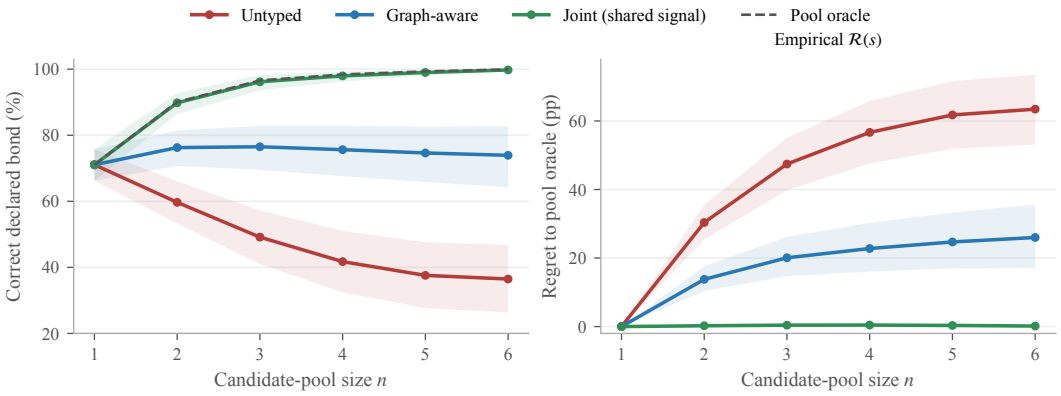

Figure 6: Detailed post-hoc six-panel sweep, including empirical selection regret $\mathcal{R}(s)$ (right), not a fitted theory curve. Shading gives pointwise 95% cluster-bootstrap intervals. Graph exclusion does not use the bond-distance endpoint; joint is a shared-signal upper bound, not independent validation. Figure 5 contrasts this heterogeneous-pool result with the preregistered within-model test.

The reaction-centre band $[73.9°, 153.3°]$ had to accept 95% of deposited structures before any prediction was examined; two earlier specifications failed that bar and are recorded in the deviation log.

The post-hoc diagnostic keeps the exact score, tie rule, candidate pool, and cluster weights used in Section 4.1. Table 31 decomposes all residual graph-only error. A "strict maximum" means every candidate attaining the maximum graph-excluded score has an incorrect declared bond even though another candidate in the pool is correct. Mixed ties average uniformly, as in the primary analysis.

The within-pool oracle is 99.92%, so 26.00 of the 26.08 residual points are avoidable selection regret rather than missing candidate support. All tasks in this panel are drug-like and all six endpoint parses

Table 30: Selection under each rule on three endpoints at pool size six, as referenced from the main text. The reaction-centre row uses only the angles at the two declared link atoms and is independent of the repaired score, which never reads a bond distance. Intervals are 95% paired cluster bootstraps.

Endpoint at $n = 6$	Original s_0	Repaired s_2	Difference
Correct declared bond	36.5	73.9	$+37.5$ $[26.8, 48.3]$
Pose within 2 Å†	29.4	33.4	$+4.0$ $[-1.1, 9.3]$
Reaction-centre geometry	44.8	72.5	$+27.7$ $[17.4, 38.0]$

†Pose uses the 266 tasks and 68 clusters with complete graph-symmetry correspondence; the other rows use all 292 tasks and 74 clusters. The accepted reaction-centre band is $[73.9^{\circ}, 153.3^{\circ}]$, calibrated on deposited structures alone and frozen before any candidate was labelled.

Table 31: Graph-only best-of-six failure decomposition. Contributions use equal weight per homology cluster; task counts are unweighted.

Category	Tasks	Cluster mass (%)	Failure contribution (pp)
All maximizers correct	246	73.88	0.00
Incorrect strict maximum	44	25.96	25.96
Mixed-correctness exact tie	1	0.08	0.04
No correct candidate	1	0.08	0.08

resolve. By anchor residue, cluster-weighted correct-bond selection is 96.7% for SER (79 tasks), 83.3% for CYS (161), and 65.1% for LYS (39); ASP has 48.7% (6), while ASN, GLU, and HIS have zero in only 1, 3, and 3 tasks, respectively. These small strata are diagnostic rather than inferential. The dominant failure is objective incompleteness: after bonded neighbors are removed from a nonbonded steric score, a geometrically clean but unformed declared bond can outrank a candidate that actually forms it. Increasing the exclusion radius cannot by itself add that missing positive bond evidence. The released task table records candidate availability, number of maximizers, anchor class, and selection probability.

Relation to the regret bound. Proposition 3 establishes that missing relation information can cause selection regret; it does not predict its growth with pool size. Each target here provides one declared relation state, so the experiment does not identify a distribution over conflicting relation states. The measured gap to the pool oracle is therefore reported directly, without fitting a theoretical amplification curve.

C.3 Bond-axis displacement audit

Section 3 reports the prespecified 0–2 Å grid. Table 32 also gives the separately fixed 0–8 Å grid. Both start from the same 272 initially bonded predictions; these are interventions on frozen outputs, not fresh generations.

The original RIA intervention changes the declared relation while fixing all coordinates. It asks whether the evaluator reads the relation. We add the dual intervention: keep the declaration fixed and change only whether the coordinates satisfy it. This distinction matters because graph exclusion can pass the first intervention while remaining indifferent to formation of the excluded bond.

All 292 frozen Boltz-2 outputs resolve under their recovered runtime YAML and strict atom identities. The main denominator comprises the 272 whose declared bond is already in the prespecified 1.2–2.2 Å window at $t = 0$. We translate the complete ligand rigidly along the oriented declared-bond axis, away from the protein endpoint, at $t = 0, 0.25, \ldots, 8$ Å. This preserves all intraligand distances and increases the declared distance by exactly t. Scores are clipped to $[0, 1]$ and exact maximum ties are averaged. The 0–2 and 0–8 Å grids were both fixed before outcome inspection.

All 65 graph-excluded failures on the 0–2 Å orbit are strict preferences, not artifacts of tie handling. Three invalid points already attain the clipped score ceiling of one and are therefore global maximizers of the score. The remaining 62 establish strict preference on the registered finite orbit; the

Table 32: Verifying intervention on 272 initially bond-valid predictions. "Valid at max" averages bond validity across all score-maximizing grid points; "strict invalid win" counts tasks where the best invalid score exceeds the best valid score. No valid–invalid top-score ties occur.

	0–2 Å		0–8 Å	
Score	Valid at max	Strict invalid win	Valid at max	Strict invalid win
Untyped sterics	9.9%	245/272	9.6%	246/272
Graph-excluded sterics	76.1%	65/272	71.3%	78/272
Graph × bond gate	100.0%	0/272	100.0%	0/272

separate rigid-translation proposition establishes the existence of a global invalid maximizer along any admitted unbounded ray.

Case study: one structure, two bond states. The largest positive invalid-minus-valid margin on the registered 0–2 Å orbit occurs in the 9X2I Boltz-2 prediction. Row 1 of Figure 3 shows its actual protein pocket and ligand coordinates at $t = 0$, and the same protein with the complete ligand rigidly translated by 1.75 Å at the registered maximum. The two molecular views share a camera; the dashed segment indicates the declared SG(Cys184)–C17(ligand) endpoint pair, not an inferred bond from the coordinate file. The views were ray-traced in PyMOL from the frozen coordinates, with its automatic cross-chain bond inference disabled. This is a controlled counterfactual and not an independently generated pose or a molecular-dynamics trajectory.

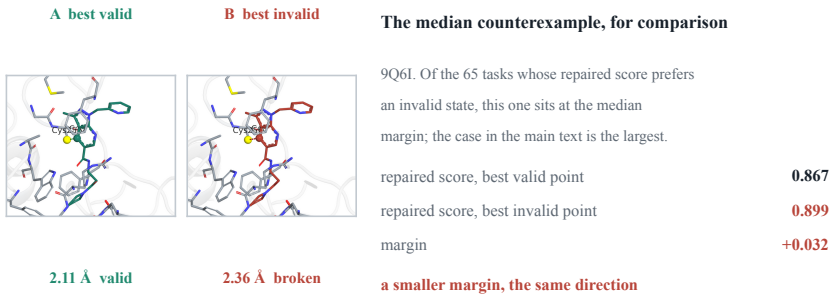

Figure 7: **The median counterexample.** The main text shows the largest invalid-minus-valid margin among the 65 tasks; this is the median one, rendered the same way from the same kind of registered displacement. The margin is smaller, $+0.032$ rather than $+0.147$, and points the same way.

Figure 7 shows the median case so that the main-text example is not mistaken for the typical one.

Case study: independent predictions, opposite choices. The frozen six-candidate panel also contains a concrete selection example at 9HU8. Its untyped steric score selects the Boltz-1 prediction, whose declared Cys19 SG–ligand C25 distance is 3.45 Å; graph-excluded sterics selects the independent Boltz-2 prediction at 1.96 Å. The earlier pose-only labels for this case are not used: 9HU8 is excluded from the corrected 266-task graph-correspondence cohort. Figure 3(**c**) shows the actual model coordinates. We rigidly aligned 178 protein Cα atoms (0.42 Å RMSD) only to obtain comparable camera views; this alignment does not change either internal bond distance or score. This post-hoc case illustrates the measured panel reversal, not its population frequency.

The last arm shares the evaluated bond window and is therefore a constructive control, not independent evidence. The translations are task-defined counterfactuals rather than generated-pose samples. The result establishes the objective-completeness failure predicted by Proposition 1; it does not estimate how often a generator or optimizer reaches these states. Per-task traces, frozen source hashes, and invariance checks are included in the artifact.

A non-borderline frozen-panel case: 7I90. For illustration, a post-hoc screen of the 266 pose-resolvable tasks in the frozen six-model panel retains tasks whose repaired-score maximizers all

have a declared distance at least 2.7 Å, and whose pool contains an alternative passing both the bond and corrected pose checks. Seven tasks meet this rule; 7I90 is the lexicographically first. This outcome-based screen selects an example, not a new prevalence estimate. Table 33 reads the original score and distance records and the graph-symmetry-corrected pose labels; it is not a new coordinate regeneration or a deployment trace.

Table 33: Frozen predictions for 7I90. The declared protein atom is Lys11 NZ; the ligand atom is C29 for Boltz and C6 for Chai/Protenix. Bold marks the unique contact-score maximizer. Pose is the independent corrected 2 Å endpoint; the gate below uses only the registered bond window.

Candidate	s_0	s_2	Bond (Å)	Bond pass	Pose pass
Boltz-2	0.5550	0.7815	1.8315	Yes	No
Chai-1	0.4629	0.6478	1.5275	Yes	No
Boltz-1	**0.8665**	**0.8665**	5.9193	No	No
Protenix	0.3545	0.7813	1.1698	No	No
Boltz-1 repeat	0.5466	0.5466	3.3863	No	No
Chai-1 repeat	0.4470	0.8090	1.4750	Yes	Yes

Both contact scores select Boltz-1; a bond gate followed by the repaired score selects the Chai-1 repeat. Its pose pass is not used by the gate, though the post-hoc example screen did require such an alternative to exist. The case therefore illustrates the selection consequence without serving as an unbiased estimate of pose benefit. It concerns our audited ranking, not model-native confidence. The earlier independently replayed 10AK case is retained below.

Executable benchmark replay: 10AK. As a post-hoc verification added during manuscript revision, we choose the lexicographically first task among the 44 strict-invalid-win tasks in Table 31. This fixes the illustrative case as 10AK; it is neither an external holdout nor a new estimate of failure prevalence. The replay takes all six archived model outputs, reconstructs the original and graph-local scores using the frozen scoring function, selects their maxima, and separately measures the named atom distance. All twelve scores match the frozen panel within 10^{-10}, and all six independent distance measurements match the archived distances within 10^{-9} Å.

Table 34: Selection replay on six unmodified predictions for 10AK. The declared protein endpoint is Cys212 SG; the ligand endpoint is C49 in the Boltz output naming and C21 in Chai/Protenix. Scores are clipped to one. Bold marks each score's unique maximizer. The checker uses the fixed 1.2–2.2 Å window.

Candidate	Original s_0	Repaired s_2	Bond (Å)	Checker
Boltz-2	0.6638	**0.8975**	2.3232	Reject
Chai-1	0.4892	0.5743	1.7123	Pass
Boltz-1	0.7011	0.7011	13.4979	Reject
Protenix	0.4666	0.5984	1.6332	Pass
Boltz-1 repeat	**0.7565**	0.7565	14.9825	Reject
Chai-1 repeat	0.5042	0.8223	1.7647	Pass

The original winner even passes the original steric gate ($s_0 \geq 0.75$), although its required bond is absent. Graph-local repair changes the winner but still selects a bond beyond the accepted window. Filtering by the independent bond check before ranking by either score selects the Chai-1 repeat; its repaired steric score also exceeds 0.75. This is a consistency control using the same bond endpoint, not an independent chemistry or pose improvement. All six candidates fail the corrected pose-only endpoint, so this case demonstrates bond verification and does not demonstrate pose rescue. The accompanying replay includes source hashes, the six original coordinate files, task declarations, a separate endpoint checker, and the full decision trace. Original native confidence scores are not present in this archive, so the result concerns the paper's benchmark reranker, not a model's native selection rule. Original generation provenance qualifications are retained.

C.4 Exploratory selector training and model-order diagnostics

Post-hoc model-order audit. Because Table 37 suggested a possible order change, we resampled the 74 frozen homology clusters 10,000 times and retained all six pairwise comparisons. Only Boltz-2 versus Protenix reverses in point estimate: named-pair pass favors Boltz-2 by 2.05 points (95% CI $[-6.41, 13.83]$), whereas graph sterics favors Protenix by 1.71 points (Boltz-2-minus-Protenix CI $[-5.14, 2.56]$). Neither side excludes zero, and none of six pairs is a robust reversal. This analysis was specified after the aggregate counts were known. We therefore treat it as a boundary result, not a third molecular ranking claim; the hard molecular evidence remains the candidate-selection reversal in Section 4.

This corrective extension updates model parameters using the 292 targets with resolved graph scores for all four frozen candidate predictions. It does not generate coordinates. The protocol was recorded before the corrective rerun, after endpoint audit counts were known; it is exploratory, not an independent preregistration. The original 75 homology-cluster IDs were sorted by a seeded SHA-256 key and assigned round-robin to five folds. Subsetting leaves 74 clusters and preserves their original fold assignments and feature vocabulary.

Ligand adjacency is reconstructed from the saved SMILES and protein adjacency from the input sequence. The saved SMILES atom index defines the ligand endpoint. Complete heavy-atom names and elements must match the output, with ordered matching additionally required for Boltz and Protenix. Missing or ambiguous identities fail closed. Reconstructing output identity does not establish which constraints a provider used during generation.

A shared MLP maps a candidate's model identity and the task's declared anchor class to a scalar logit (16 tanh hidden units); a four-way softmax yields the selection policy. Anchor classes come from the original task manifest. No reference coordinates, RMSD, target ID, or validity score is given as a network input. Three reward arms share architecture, initialization, folds, and optimizer budget: the published continuous distance ratio clipped to $[0, 1]$, the similarly clipped graph-exclusion ratio, and the latter multiplied by a declared-pair distance gate (1.2–2.2 Å). This gate is not a full reaction-centre geometry test. The bond endpoint reuses the gate definition, so it is not independent chemical validation; pose-only outcomes are also retained.

We maximize exact expected reward over the four actions minus $0.02\,\mathrm{KL}(\pi\|\mathrm{Uniform})$ using 500 full-batch Adam steps (learning rate 0.01) for each of five initializations and five folds, giving 75 training runs. Targets are weighted equally within clusters and clusters equally within the training fold. Held-out probabilities, learned checkpoints, and training curves are saved. Analytic gradients are checked against finite differences. The bond endpoint measures the two declared atoms, not the old `link_formed` or nearest-contact `bond_valid` flags. A later correspondence audit withdrew the historical element-only Hungarian RMSD: it could select a mapping inconsistent with molecular topology. Pose is now computed only on the 266 tasks (68 clusters) for which all four candidates have a complete heavy-atom graph isomorphism preserving elements, bond orders, aromaticity, and the declared ligand anchor. The protein alignment is unchanged; the other 26 tasks do not enter the pose denominator. This correction changes 59 of 1,596 candidate-level 2 Å labels but leaves every bond endpoint and every trained policy fixed.

Table 35: Out-of-fold expected success under corrected pose correspondence, averaged over seeds and then equally over 68 retained clusters. Percentages use a different weighting from the target-weighted main model table. Higher is better.

Policy / training reward	Pose and pair distance	Pair distance	Pose only
Uniform policy	24.4	75.6	31.0
Untyped reward	14.8	35.8	27.7
Graph-exclusion reward	27.6	81.5	31.7
Graph exclusion plus distance gate	28.2	87.2	32.4
Joint steric–bond–angle reward	30.1	87.1	32.7
Fixed Boltz-2	32.6	87.2	35.1

Under the corrected endpoint, graph-exclusion training improves combined pose–bond success over untyped training by 12.7 percentage points, with paired cluster-resampling interval $[5.0, 21.0]$

(10,000 resamples). The distance-gated arm improves it by 13.4 points $[4.8, 22.4]$. After a separate frozen protocol, the joint steric–bond–angle arm reaches 30.1%, $+2.5$ points over graph exclusion $[-0.2, 5.8]$ and -2.5 points versus fixed Boltz-2 $[-5.2, 0.3]$. It does not establish superiority over the fixed model; pose-only is 2.4 points lower $[-4.8, -0.4]$ while bond success differs by -0.1 points $[-3.5, 3.2]$. These intervals resample seed-averaged out-of-fold cluster outcomes under the fixed fitting procedure; they omit full training-set resampling uncertainty. The experiment supports objective-induced learning, not superior selection performance or generative fine-tuning. All original and corrected records are retained rather than silently overwritten.

Angle-calibration boundary. Declared-graph angle measurement enumerates every heavy-atom neighbour on each side of the link; a nearest-atom rule can choose a non-neighbour or omit other bonded neighbours. We therefore discarded the earlier nearest-neighbour calibration, reprocessed deposited structures with the same declared graphs, and froze the $[73.935°, 153.278°]$ band before prediction outcomes were labelled. On 313 measurable structures from a separate 500-complex gate set untouched by earlier attempts, 307 (98.1%) pass the required structure-level rule; 317 unresolved links fail closed and are reported outside the acceptance denominator. Two earlier bands that failed the 95% false-rejection gate remain in the deviation log. The selector's broad angle factor was trained before this calibrated endpoint and is not retroactively relabelled as centre validity.

C.5 Separate three-task physical-steering pilot

This diagnostic changes the stock pair mask on three tasks. It differs from the 100-task isolated-objective experiment in Section 5, whose four arms are specified in Section E.2. Their task counts and effects are not pooled.

The pinned Boltz-2 physical-guidance mask removes every atom pair between two chains joined by a declared covalent connection. On the three fixed protein–ligand inputs this leaves zero cross-chain VDW pairs. The process-local adapter changes only pair eligibility: it excludes the declared bond and its graph-distance-two neighbourhood (5, 4, and 4 pairs) while retaining 15,931, 65,036, and 142,446 distal pairs. The installed package and checkpoint are not edited.

Fresh and archived parser paths produce identical compared potential features. Before inference, actual full-featurizer tensors pass CPU and CUDA checks for atom names, elements, order, padding, chain and endpoint identity, together with analytical/autograd gradient agreement. We then freeze three tasks, seeds 20260914 and 20260915, and four samples per task–seed–arm. FK steering is off in all arms. The off arm disables physical gradient guidance; original and local enable it with identical schedules and differ only in VDW support. All 18 cells and 72 CIFs complete; hashes written by each process match the downloaded outputs, and all endpoints resolve.

Table 36: Three-task physical-steering diagnostic. Each cell has eight candidates. Ratio and overlap entries are candidate means; the parent task, not a candidate, is the independent unit.

Task	Endpoint	Off	Original	Graph-local
9KA6	minimum eligible ratio	.8330	.8347	.8461
	overlaps below .75	.125	.125	.000
	bond / pose success	8/8 / 0/8	8/8 / 0/8	8/8 / 0/8
9RAJ	minimum eligible ratio	.8146	.8065	.8095
	overlaps below .75	.000	.000	.000
	bond / pose success	8/8 / 1/8	8/8 / 1/8	8/8 / 1/8
9XZZ	minimum eligible ratio	.8521	.8521	.8521
	overlaps below .75	.000	.000	.000
	bond / pose success	8/8 / 8/8	8/8 / 8/8	8/8 / 8/8

Local versus original support improves the mean minimum ratio by .0114 and .0030 on 9KA6 and 9RAJ and ties on 9XZZ. It removes the sole threshold failure (23/24 to 24/24) without changing discrete bond or pose outcomes. This verifies an active, local inference repair but does not establish a better generator: the panel has three size-selected parents, only one candidate crosses the steric threshold, and pose success is unchanged. The archived predictions used elsewhere in the paper had physical and FK steering disabled, so their errors are not attributed to this implementation.

C.6 RELATIONBENCH DEVELOPMENT PROTOTYPE

The runnable prototype contains twelve rule families, 24 projected/aware configurations, and 640 constructed carriers across graph matching, vision, question answering, KG ranking, and molecular contact evaluation (3,072 configuration–case evaluations). Graph adapters use WL hashing/isomorphism; vision and QA compare single-reference projection with declared acceptable sets; KG adapters cover MRR, Hits@1, Hits@10, and mean rank; molecular adapters cover minimum contact and clash count. Except for NetworkX calls, these are local definition adapters rather than independent production packages, and configuration counts are not prevalence estimates.

The diagnostic records response, direction, margins, and repeated/irrelevant- relation controls. All twelve projected configurations are invariant and all twelve aware configurations respond in the required direction on these development witnesses (thresholded Hits rules respond only when the rank crosses their threshold). This is a functionality check, not held-out detection accuracy or a source-reviewed metric registry; we make no larger-scale claim.

D SUPPORTING MOLECULAR MEASUREMENTS

D.1 THE CORRECTED MEASUREMENT

Repairing the rule changes what the comparison says, so we report the corrected numbers rather than only the defect. On the 292 targets where the declared endpoints and output graphs resolve, we measure whether each model actually formed the bond the task declares, using declared atom identities rather than nearest contacts. We also record whether the structure passes the repaired steric test.

Table 37: Strict declared-identity audit on the common 292-target subset.

Model	Named pair 1.2–2.2 Å	Graph steric pass
Boltz-1	149 (51.0%)	200 (68.5%)
Boltz-2	272 (93.2%)	251 (86.0%)
Chai-1	266 (91.1%)	249 (85.3%)
Protenix	266 (91.1%)	256 (87.7%)

Boltz-2, Chai-1, and Protenix form the declared bond in nine cases out of ten, whereas Boltz-1 does so in half; the steric column separates that question from non-bonded plausibility. These are the quantities a covalent benchmark should report, and the published rule cannot report either of them: it rejects essentially every structure in the table, including the ones that are correct.

Component boundary. Graph exclusion restores a non-bonded steric test but cannot validate the excluded bond. We therefore report the declared-bond distance separately and calibrate a reaction-centre angle endpoint only on deposited structures (Section 4.4). Exclusion relocates this construct; it does not certify it.

Inference-side locality check. The opposite boundary appears in optional physical steering: the pinned Boltz-2 implementation excludes the complete connected chain pair, leaving zero protein–ligand VDW pairs on three admitted covalent inputs (Passaro et al., 2025). Replacing only that support by the same graph-local exclusion retains 15,931–142,446 distal pairs. In a frozen 72-candidate pilot (three size-selected tasks, two seeds, three arms), local versus original steering raises the task-mean minimum eligible ratio on two tasks and ties on one. Steric passage changes from 23/24 to 24/24. Declared-bond success remains 24/24 and pose success 9/24 under both arms. This is a mechanism check, not an efficacy or prevalence claim; full controls are in Appendix C.5.

The defect is not primitive-specific. On the same 247 loadable drug-like crystals, we implement three published steric primitives with and without identical relation-aware pair eligibility. Across a 2.60 Å over-long target bond and a distal 1.0 Å clash, the repair restores target-bond discrimination without sacrificing distal-clash sensitivity (Table 38). A separate reaction-centre score detects the

over-long bond with AUROC 0.967; steric exclusion alone is not used to certify it. Full definitions and count-scale results are in Appendix A.17.

Table 38: AUROC on 247 clean references and matched corruptions. "Aware" changes only eligible atom pairs; values below 0.5 are directionally inverted.

	Over-long target bond		Distal clash	
Primitive	Untyped	Aware	Untyped	Aware
PoseBusters ratio	0.075	0.817	0.506	0.788
PoseCheck count	0.517	0.800	0.754	0.782
Probe-style count	0.532	0.746	0.756	0.780

D.2 Expanded evaluator-routing check

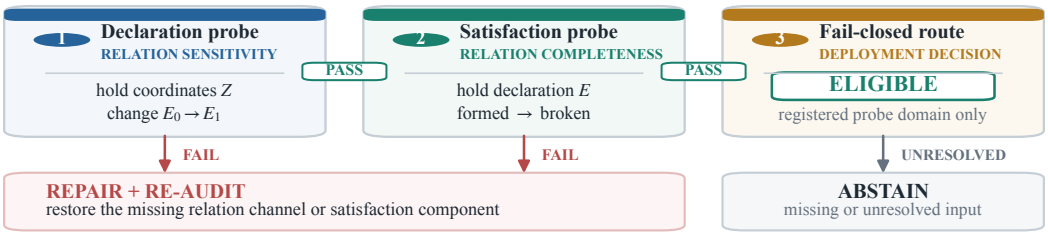

Figure 8: **Audit outcomes and follow-up actions.** Both probes report on recorded inputs and candidate sets. A failed required response or an invalid winner gives a counterexample to the tested condition. Passing all recorded checks means no counterexample was found; it does not certify untested outputs. Unresolved checks require abstention. The routing experiment below evaluates declaration coverage only.

A before-outcome routing experiment uses the declaration-side RIA screen: nonzero response, correct direction, and a feasible witness threshold. It does not implement a full two-sided pass. Missing or unresolvable relations cause abstention, never fallback to the erasing rule. On FB15k-237 it routes raw to filtered ranking. Across all 36 pairwise decisions among the nine controlled checkpoints and all 10 among the public panel, raw ranking disagrees with the relation-aware contract in 34/46 decisions; routing reduces this to zero without fallback. Across 1,176 molecular model–task cases, it routes all 1,168 resolvable cases to graph-aware sterics, prevents 902 projected disagreements, and abstains on eight with zero unsafe fallbacks. These counts measure agreement with the declared routing contract and absence of fallback to the erasing rule. They do not establish physical validity or satisfaction completeness of the selected score.

E Reachability audits: protocols and full results

Both audits in Section 5 were frozen before any structure was produced: protocol, cohort, and code hashes are in the artifact, neither analysis was run more than once, and no strength, schedule, or weight was chosen after seeing an outcome. We state the evidential difference plainly. The prospective best-of-n experiment of Section 4.2 carries a public timestamp created before its first candidate existed; these two audits do not. Their manifest was written after they ran and attests to content integrity only, so their pre-specification rests on the frozen protocol files and the working record rather than on a third-party anchor.

E.1 Local rigid-body optimizer

This is the search experiment in Section 5, not a generation experiment. Table 39 extends its reported 0.5 Å and 5 Å bounds to the full grid.

Cohort: the 272 archived predictions whose declared bond is already valid. Objective: one score at a time, maximized by a deterministic Hooke–Jeeves pattern search over six rigid-body degrees of freedom, with rotations bounded at 30° and translations bounded by $R \in \{0.5, 1, 2, 5\}$ Å. The ligand's internal geometry is preserved on every accepted step (maximum intra-ligand drift below 10^{-8} Å) and the protein is fixed; 0 failures. Table 39 reports the fraction of tasks whose optimum is bond-invalid. "Invalid and higher-scoring" coincided with "invalid" in every cell: every broken optimum also outscored its start. The median final bond under the repair stays near 2 Å (1.99 Å at $R = 0.5$, 2.03 Å at $R = 5$) because the failures are a split, some bonds stretched past 2.2 Å and others compressed below 1.2 Å, whichever raises the smallest eligible contact ratio.

Table 39: Fraction of the 272 tasks whose rigid-body optimum is bond-invalid.

Objective	$R = 0.5$ Å	$R = 1$ Å	$R = 2$ Å	$R = 5$ Å
untyped s_0	93.0%	95.2%	96.0%	96.0%
graph-local s_2	30.5%	34.2%	34.9%	34.9%
s_2 with the bond gate	0.0%	0.0%	0.0%	0.0%

A rigid-body optimizer is not a generator. This measures reachability of relation-invalid optima by local optimization of the score, not the behaviour of any sampler, and it does not estimate how often a generator visits such states.

Figure 3(**b**) shows one optimum in full. The search is deterministic, so the pose was regenerated from the frozen record rather than stored: on 9P1B it moves the ligand 0.43 Å and rotates it 9.9°, stretching the declared Lys190 NZ–B10 bond from 1.84 to 2.33 Å while the repaired score rises from 0.815 to its clipped ceiling of one. The protein is fixed and the ligand's internal geometry is preserved to within 10^{-8} Å.

E.2 Four-arm guidance experiment

This protocol produces Table 3 in Section 5. Arms 1 and 2 isolate pair eligibility; arm 3 adds the declared-bond term; arm 4 retains the complete stock objective.

Cohort: 100 tasks drawn in SHA-256 order, seed 20260916, from the same 272; two replicates per task with seeds fixed in advance; four arms per replicate, so 800 generations. Every arm reuses the task, the frozen MSA, and the replicate's noise seed, so arms are compared within a replicate.

The generator's stock steric term excludes every atom pair between chains joined by a declared bond, which is neither the transferred untyped rule nor our graph-local repair. Swapping only its mask would confound two changes: removing the bond penalty and adding distal repulsion the stock stack never applied. Arms 1 and 2 therefore both use a potential whose pair set we define, and differ in exactly one thing. The stock stack also contains a positive term on the declared bond; keeping it silently would mean arm 2 was not mask-only, so it enters only in arm 3, at its stock weight. Arm 4 runs the unmodified official stack as a system reference; arms 1–3 isolate the steric term and omit the other stock potentials, which is declared rather than hidden.

Primary endpoint: the declared bond length at the final structure, its validity window $[1.2, 2.2]$ Å, and the bond length as a function of diffusion step, recorded during generation by a hook that reads coordinates and returns exactly what the wrapped potential returns. Registered hypotheses: H-steer1, arm 2 exceeds arm 1; H-steer2, arm 3 exceeds arm 2, tested by paired comparison within replicate. The interpretation of a null H-steer2 was fixed in advance and is quoted in Section 5.

The artifact retains per-call bond-distance traces for all four arms; Table 3 reports their final bond-validity outcomes.

All 800 runs produced a parseable structure. Arm-1 traces leave the valid window by the second guidance call and settle at a median 2.73 Å; arms 2–4 remain between 1.5 and 1.67 Å throughout. The audit ran across two GPUs by a parity rule that never splits a task's eight runs across machines. 265 runs were produced on both machines by an overlap in scheduling: they agree on bond validity in 265 of 265 cases, with median absolute difference in declared bond length 0.000 Å and maximum

0.030 Å, so the hardware split does not carry the result; that agreement also bounds the run-to-run nondeterminism of the pipeline on this endpoint.

The released artifact contains the frozen protocol and cohort, the potential definitions and runner, the per-structure declared-bond measurements, the per-call traces, and the analysis that turns them into Table 3. The generation itself is reproducible from the frozen cohort seeds against the vendored commit.