SUCCESS IS NOT EVIDENCE: VERIFIABLE AUTORESEARCH FOR CRYO-ELECTRON TOMOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous scientific agents are graded by execution success, but execution success is not scientific validity. In cryo-electron tomography (cryo-ET), a picker emits coordinates whether or not particles lie at them, a segmentation returns a mask that can be exactly empty, and a refinement reports a resolution its own half-maps may not support—all with exit code zero. We introduce CRUCIBLE, an autonomous cryo-ET research loop organized around *executable evidence*: measurements an independent verifier recomputes from stored artifacts, never from tool self-report. A masked admissibility policy acts on the artifact store, and a gated reward scores each trajectory over four evidence tiers with an explicit availability flag, so "not checkable" is never conflated with "bad". We prove that with sufficiently separated tier weights this reward orders trajectories exactly as a fixed scientific priority does (validity, non-vacuity, localization, resolution), while an ungated scalar can rank the worse policy higher. To our knowledge CRUCIBLE is the first autoresearch agent for structural cryo-ET: the whole multi-stage subtomogram-averaging reconstruction is the policy's action space, and the policy is trained on the verifier's evidence. Experiments carry a published twelve-stage protocol to completion, confirm resolution claims by independent recomputation, and validate picking at particle level (precision 0.57, recall 0.71). The same pipeline generates the controls that test our claims; four weaken or vanish under them.

1 INTRODUCTION

An autonomous agent operating scientific software inherits a property that tool-use benchmarks never test: execution can succeed computationally while producing nothing scientifically valid. In cryo-ET, a template matcher emits coordinates whether or not particles lie at them, a segmentation network returns a mask whose volume fraction can be exactly zero, and a refinement reports an FSC resolution that recomputation from its own half-maps may not reproduce. All three exit 0. The dominant agent pattern—plan, call tools, read stdout—optimizes a proxy that diverges from its target precisely where scientific cost is highest.

Cryo-ET is where that divergence is most consequential: the gain from automating it and the cost of getting it wrong are both large. Annotation is the field's rate-limiting step, and community benchmarks exist chiefly to measure how well it can be automated (Peck et al., 2025; Ermel et al., 2026). Method choice is dataset-specific, and a bad choice still produces artifacts that look finished—an average of noise renders as a blob, a refinement prints a resolution either way. What is needed is an evidence interface, not a better prompt.

CRUCIBLE supplies that interface by organizing the loop around *executable evidence*: a measurement a verifier recomputes from stored artifacts—half-maps, coordinate files, masks, logs—independently of the tool that claimed it. Three design commitments follow. *Admissibility before score*: actions are masked by a predicate computed from the artifact store, and outputs failing format or plausibility checks receive zero reward rather than a small one, so validity cannot be traded for task score. *A hierarchy, not a scalar*: localization and resolution occupy distinct tiers with distinct availability indicators, and a missing resolution is reported as missing—never as zero, never

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.

substituted by a cheaper metric. *Controls as first-class outputs*: the pipeline that produces a claim produces the paired, sign-tested, size-matched control that could refute it.

Each commitment is forced by a result rather than a preference. We prove that a gated trajectory reward orders two trajectories exactly as a fixed scientific priority does—validity, then non-vacuity, then localization accuracy, then resolution—provided the tier weights are separated widely enough (Thm. 1). Two negative results motivate the gates: when an outcome cannot be measured for some runs and measurability depends on the specimen rather than on chance, a scalar reward conflates the two and can rank the worse policy higher (Prop. 2); and keeping only the top-scoring particles raises measured purity even when the score carries no information, so a purification claim needs a random control of the same size (Prop. 3).

Position. Autonomous experimental agents have closed propose–execute–learn loops over chemical synthesis, ML engineering and wet-lab robotics; within cryo-EM, reinforcement learning has optimized data collection, CNNs screen specimens, and agentic workflows now reach single-particle processing and model building (Sec. 2). None puts a *structural cryo-ET* reconstruction—a 3D density built by subtomogram averaging—inside such a loop, and there the cheapest observation available, a tool's exit code, is the least trustworthy. To our knowledge CRUCIBLE is the first autoresearch agent that does. The claim is one of system class, not of outcome: Sec. 4.3 and 4.6 measure how far the present loop still falls short of unattended discovery.

Contributions.

1. **A formal account of when a scalar reward may stand in for a scientific priority.** Evidence tiers over a checksummed artifact store, each carrying an explicit availability flag (Def. 1); evidence sufficiency (Def. 3); a proof that no policy improvement can trade validity for task score (Prop. 1); and the condition on tier weights under which one scalar reproduces the priority order exactly (Thm. 1)—together with a statement that the weights we ship do not meet that condition.
2. **The first autoresearch agent for structural cryo-ET, built on that account.** A masked admissibility policy acting on artifacts; a planner/orchestrator split that lets the contract-following half be a small locally trained policy; a power-aware critic with two evidence-triggered escape hatches; instruments admitted only through a four-stage gate; an independent verifier recomputing gold-standard FSC from stored half-maps; and a multi-turn group-relative policy update (Eq. 7).
3. **An evaluation discipline in which every claim ships with the control that could refute it.** Four claims weaken or vanish under controls our own pipeline generated, and all four stay in the main text: a denoising benefit whose seven paired cases cannot reach below $p = 0.016$; held-out RL superiority indistinguishable from its supervised initialization; a membrane-proximity purification effect below its size-matched control; and autonomous instrument creation, where the gate rejects every candidate.

2 Related Work

Language-model agents that interleave reasoning with tool calls (Yao et al., 2023; Qin et al., 2024; Shinn et al., 2023) extend to scientific workflows—closed-loop laboratory control (Boiko et al., 2023; Bran et al., 2024; Lu et al., 2024), ML engineering benchmarks (Huang et al., 2024), and self-evolving wet-lab protocols (ProtoPilot developers, 2026)—but treat tool stdout as observation, so their success metrics are computed by the tool that produced them. Closest to our setting, EM-Seek (Chen et al., 2025) routes one controller over specialist workers for electron microscopy, unifies returns into one record, keeps a provenance tree, and repairs tool arguments before execution; CRUCIBLE inherits that return contract and adds an admissibility mask computed from an artifact store, an evidence-tiered reward with an availability flag, and a trained local policy for the tool-call step. On the training side, group-relative objectives (Shao et al., 2024; Ahmadian et al., 2024; Liu et al., 2025) learn from scalar trajectory rewards, and multi-turn rollout middleware records generations at token level so interleaved trajectories can be stitched into training examples (Xu et al., 2026; Zhang et al., 2026). Our trainer adds two domain choices: tier-0 violations contribute zero rather than negative gradient, and the optimized scalar is gated (Prop. 2).

Modern subtomogram-averaging (STA) chains compose motion correction and tilt-series processing (Tegunov & Cramer, 2019; Zheng et al., 2022), denoising and missing-wedge correction (IsoNet developers, 2022), picking (Bepler et al., 2019; Wagner & Raunser, 2020; Moebel et al., 2020; Liu et al., 2024; Rice et al., 2023), segmentation (Isensee et al., 2021; Lamm et al., 2022; Cryo-SAM developers, 2024), and refinement with gold-standard FSC (Scheres, 2012; Zivanov et al., 2018; Punjani et al., 2017; Rosenthal & Henderson, 2003; TODO: verify authors, 2019), over benchmark infrastructure converging on a common data interface (Ermel et al., 2026; Peck et al., 2025). These systems automate a fixed recipe; CRUCIBLE makes recipe choice the policy's action space and the pipeline's artifacts its evidence. Autonomy in cryo-EM has so far lived upstream of reconstruction—RL over microscope navigation (CryoRL developers, 2024), specimen screening (SmartScope developers, 2022), parallelized acquisition pipelines (SPARTA developers, 2024; NextPyp developers, 2023)—and agentic workflows have reached single-particle processing (CryoAgent developers, 2026) and model building (StructAgent developers, 2026; CryoForge-MB developers, 2026); none closes a loop over cryo-ET subtomogram averaging, and none trains its policy on verifier-recomputed evidence.

3 METHOD

Three failure modes motivate the design (incident register in Appendix B). **F1, inadmissible calls:** in one RL round 71% of dispatched actions were denoise calls on already-denoised volumes. **F2, silent vacuity:** an all-zero segmentation returned exit code 0 and propagated into picking; a draft figure quoted a resolution copied from a reference column with no supporting artifact. **F3, selection illusions:** top-k "purification" raised purity by up to $1.81\times$ under subsampling alone. The mask φ answers F1, the evidence hierarchy with its availability indicator answers F2, and the gated reward plus size-matched controls answers F3.

3.1 SYSTEM OVERVIEW

The loop is a state machine over the artifact store with six nodes (Fig. 1). **Observe** re-reads the store and re-computes which tools are admissible, so no component hand-maintains a description of the world. **Plan** is a hosted model emitting a three-line contract (PLAN/EXPECT/BLOCKED_BY) in words, deliberately *not* the action record: method choice needs world knowledge, whereas rendering a choice as a valid tool call is contract-following that a small local model does well and can be trained on. The split is not cosmetic: a planner seeing only the latest observation re-proposes at iteration 7 what failed at iteration 2. **Orchestrate** is that local policy, emitting one structured action under the mask of Eq. 1. **Execute** runs it and scores it with the same tiered reward the trainer optimizes. **Critique** returns one of `continue`, `need_data`, `need_tool`, `stop`; anything else is escalated rather than allowed to end the run quietly. **Route** dispatches back to observe, to the data scout, to the instrument gate, or to the end.

Four executor rules apply, each adopted after a specific error. (i) An action naming an inadmissible tool scores a hard zero: renormalizing over "no tiers available" once gave such an action 1.0, above a picker that actually ran. (ii) Exit code 0 is not success—the declared artifact must exist at the declared path. (iii) An unmeasurable tier is not measured, never 0.0. (iv) A step that produced no measurable outcome is capped at 0.20; without the cap the untrained base was the best policy on that metric.

The critic is power-aware: it is told that an exact two-sided test on n paired cases cannot go below $p = 2/2^n$, so "not significant" is a property of sample size, not of the effect. Two escape hatches fire only on evidence: `need_data` dispatches a data scout ranking portal candidates by expected distribution shift; `need_tool` opens an instrument gate where the model writes no code but its declarative specification must survive four admission stages (structural validity, flag reality, execution, product check). In the runs we report the gate fired and rejected every candidate (Appendix G).

3.2 CONTROL INTERFACE: MASKED POLICIES OVER AN ARTIFACT STORE

An execution is a finite-horizon sequence over states $s_t = (c_t, B_t)$, where c_t is the linguistic context and B_t an artifact store: a content-addressed, checksummed set of files (tomograms, masks, coordi-

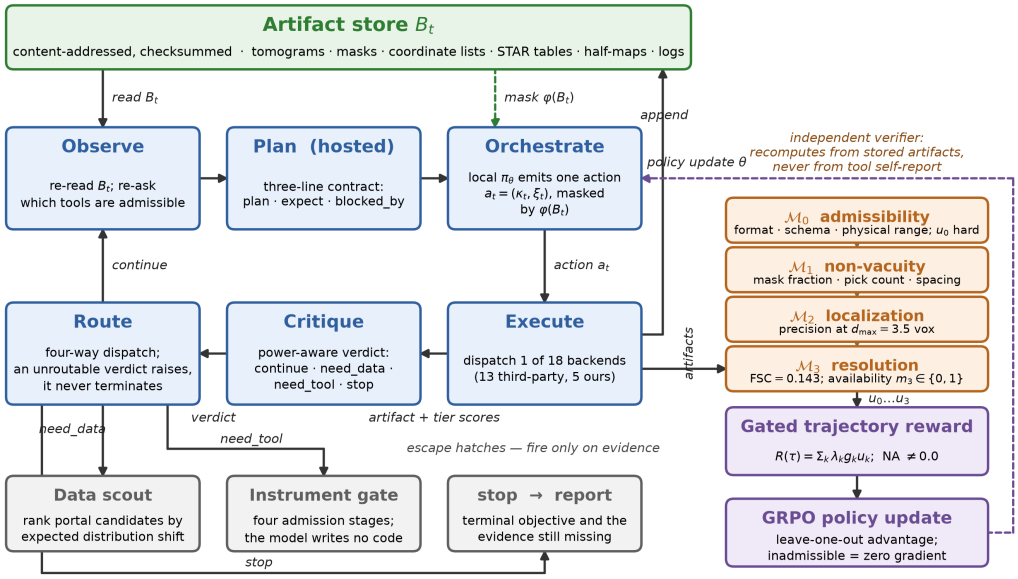


Figure 1: **The CRUCIBLE loop, its evidence hierarchy, and two escape hatches.** The artifact store B_t is the only shared state: every node re-reads it and every execution appends to it, checksummed. The admissibility mask $\varphi(B_t)$ restricts the local policy *before* sampling; the critic's four-way verdict routes back into the loop or out through a hatch. On the right, the four evidence tiers score each executed action independently of tool self-report, and the gated reward with its GRPO update returns trained weights to the orchestrator (dashed).

nate lists, STAR tables, half-maps, logs). Actions $a_t = (\kappa_t, \xi_t)$ pair an action type κ_t from a fixed vocabulary $\mathcal{K}$ with a structured parameter record ξ_t from a shared schema.

Not every syntactically valid action is scientifically admissible: denoising an already-denoised tomogram wastes an expensive call and corrupts caches. We encode admissibility as a predicate $\varphi(s, a) \in \{0, 1\}$ computed from B_t by provenance and checksum checks, and control with a masked policy

$$\pi_\theta^\varphi(a \mid s) \;=\; \frac{\pi_\theta(a \mid s)\,\varphi(s, a)}{\sum_{a'} \pi_\theta(a' \mid s)\,\varphi(s, a')}, \tag{1}$$

so inadmissible actions have zero probability *before* sampling, not merely negative reward after. Tool execution is a deterministic operator $T : (s, a) \mapsto s'$ appending artifacts to B; a trajectory is $\tau = (s_0, a_0, \ldots, s_H)$.

φ is an episode property, not a permanent restriction: a tool is inadmissible only because an upstream artifact is missing, so once the store acquires it the same tool becomes admissible with no code change (Sec. 4.3). The precondition must also be salient: on a 20-state routing benchmark, declaring availability per episode raised accuracy from 60% to 95% (Appendix F).

3.3 THE CONTROLLED PROCESS

The action vocabulary covers a twelve-stage STA reconstruction executed by 18 dispatchable backends (13 wrapping third-party engines, 5 self-built instruments). Appendix Table 13 lists each stage with its engine, pixel size, artifact appended to B_t, and the content gate that makes the next stage admissible. Provenance: the twelve stages on EMPIAR-13566 were run by two shipped idempotent drivers invoking the same backends the loop dispatches, with human intervention for environment faults; what the system contributes there is *measurement*—the verifier's independent recomputation of tier 3—not scheduling. The loop's own traces are the two runs in Sec. 4.3.

3.4 Evidence Hierarchy and Gated Reward

Definition 1 (Evidence tiers). *For a trajectory prefix ending in state s, the verifier computes four tier measurements:*

$$\mathcal{M}_0(s):\ \textit{admissibility and format} \qquad \mathbf{1}[\textit{parse ok}]\cdot\mathbf{1}[\textit{schema ok}]\cdot\mathbf{1}[\textit{physical ranges ok}], \tag{2}$$

$$\mathcal{M}_1(s):\ \textit{non-vacuity and physics} \qquad \big(\textit{mask fraction},\ \textit{pick count},\ \textit{spacing bounds},\ \textit{CTF fits}\big), \tag{3}$$

$$\mathcal{M}_2(s):\ \textit{localization precision} \qquad (P, R, F_\beta)\ \textit{vs. curated annotations } G \textit{ at tolerance } d_{\max}, \tag{4}$$

$$\mathcal{M}_3(s):\ \textit{reconstruction measurement} \qquad \big(m_3(s),\ \rho(s)\big), \quad m_3 \in \{0,1\}, \tag{5}$$

where m_3 is an explicit availability indicator (gold-standard half-maps exist and FSC is computable) and $\rho(s) = \min\{s^\star : \mathrm{FSC}(s^\star) \le 0.143\}$, recomputed by the verifier from stored half-maps (Rosenthal & Henderson, 2003), never taken from tool self-report. A missing measurement is recorded as $m_3 = 0,\ \rho =$ not measured, never imputed as $0.0\,\text{Å}$.

Definition 2 (Gated trajectory reward). *With tier scores $u_k \in [0,1]$ and gates $g_k(\tau) = \prod_{j<k} \mathbf{1}[u_j(\tau) \ge \vartheta_j]$, the trajectory reward is*

$$R(\tau) = \sum_{k=0}^{3} \lambda_k\, g_k(\tau)\, u_k(\tau), \qquad \lambda_0 > 0,\ \lambda_k \ge 0, \tag{6}$$

with $u_3(\tau) = m_3(\tau)\cdot\widetilde{\rho}(\tau)$ and $u_0 \in \{0,1\}$ a hard indicator. A tier-0 failure zeroes all downstream gates, so the whole reward is 0 rather than merely small.

Proposition 1 (Non-compensability of validity). *Under Eq. (6), any trajectory with $u_0 < \vartheta_0$ receives $R = 0$, while any admissible trajectory receives $R \ge \lambda_0\vartheta_0 > 0$. No policy improvement can trade format or physical validity for task score.*

Definition 3 (Evidence sufficiency). *A subset $S \subseteq B_H$ of the final artifact store is* sufficient *for claim c, written $S \models c$, if the verifier's measurement of c is a function of S alone. The indicator m_3 of Def. 1 signals whether B_H contains a sufficient set for the resolution claim; $m_3 = 0$ means* not checkable from this trajectory*, not false.*

Theorem 1 (Separated gated rewards reproduce the priority order). *Let $u, u' \in [0,1]^4$ be tier vectors with identical gate vectors $g = g'$, and let weights be δ-separated: $\lambda_k > \frac{1}{\delta}\sum_{j>k}\lambda_j$ for all k, some $\delta \in (0,1]$. Let k be the first coordinate with $|u_k - u'_k| \ge \delta$. Then (a) if $g_k = 1$, $\mathrm{sign}(R_\lambda - R'_\lambda) = \mathrm{sign}(u_k - u'_k)$; (b) if $g_k = 0$, $R_\lambda = R'_\lambda$. Consequently, on the fully admissible set the gated scalar is order-isomorphic to the lexicographic order on (u_0, u_1, u_2, u_3) at granularity δ.*

Remark 1 (Deployed weights do not satisfy separation). *The weights we ship, $\lambda = (0.10, 0.10, 0.30, 0.50)$, do* not *satisfy that separation. Non-compensability still holds in the deployed reward, but from the parse-failure hard zero and the 0.20 cap rather than from order-isomorphism (details in Appendix A).*

Proposition 2 (Scalar rewards invert rankings under MNAR outcomes). *Let $m \in \{0,1\}$ be measurement availability and $u \in [0,1]$ quality. A scalar $R' = m\,u$ satisfies $\mathbb{E}_\pi[R'] = \Pr_\pi(m = 1)\mathbb{E}_\pi[u \mid m = 1]$. For policies A, B with $\mathbb{E}_A[u \mid m = 1] > \mathbb{E}_B[u \mid m = 1]$, ranking reverses whenever $\Pr_A(m = 1)/\Pr_B(m = 1) < \mathbb{E}_B[u \mid m = 1]/\mathbb{E}_A[u \mid m = 1]$. Since availability correlates with specimen difficulty, m is not missing at random and the reversal condition is met on real workloads.*

Proposition 3 (Top-k purity inflation under a null score). *Let n particles contain K true positives and let score q be independent of labels. Purity of top-k is $\hat{p}_k = X/k$ with $X \sim \mathrm{Hypergeometric}(n, K, k)$: $\mathbb{E}[\hat{p}_k] = K/n$ and $\mathrm{Var}[\hat{p}_k] = \frac{p(1-p)}{k}\frac{n-k}{n-1}$. At small k the null band is wide ($p = 0.5, n = 200, k = 20$: sd ≈ 0.106), so apparent purification is what the null predicts. A purification claim must beat a size-matched random control.*

Proofs in Appendix D.

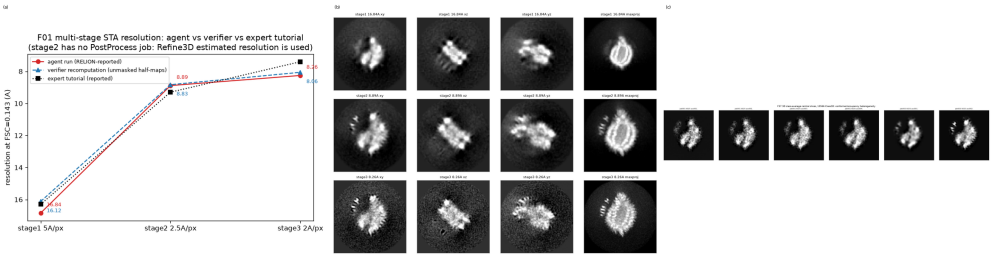


Figure 2: **End-to-end execution, independently verified (EMPIAR-13566).** (a) Per-stage resolution at FSC=0.143 from three sources: engine report, verifier recomputation from stored unfiltered half-maps, and expert tutorial. (b) Central slices and maximum-intensity projections per stage. (c) Central slices of final 3D class averages.

3.5 Policy Update

With small groups G, a shared baseline is biased; and trajectories violating tier 0 carry no usable signal, so they contribute no gradient. Advantages use leave-one-out baselines with no standard-deviation normalization (Shao et al., 2024; Ahmadian et al., 2024; Liu et al., 2025):

$$A_i = R(\tau_i) - \frac{1}{G-1}\sum_{j\neq i} R(\tau_j), \quad J(\theta) = \mathbb{E}\Big[\sum_i z_i \sum_t \min\big(r_{i,t}A_i,\ \mathrm{clip}(r_{i,t}, 1\pm\varepsilon)A_i\big)\Big] - \beta_{\mathrm{KL}} D_{\mathrm{KL}}, \tag{7}$$

with ratio $r_{i,t} = \pi^{\varphi}_{\theta}(a_{i,t} \mid s_{i,t}) / \pi^{\varphi}_{\theta_{\mathrm{old}}}(a_{i,t} \mid s_{i,t})$ and zero-gradient mask $z_i = \mathbf{1}[u_0(\tau_i) \geq \vartheta_0]$. Multi-turn trajectories are reassembled from token-level logs at the inference proxy (Xu et al., 2026; Zhang et al., 2026).

4 Experiments

Setup. Policies are local 4B/9B models (base, supervised-finetuned, group-relative RL); planning uses a hosted model while all tool control and verification run locally. The primary case is EMPIAR-13566 (nucleosome tutorial); secondary cases span EMPIAR-10440/CZII (seven tomograms with curated point annotations (Peck et al., 2025; Ermel et al., 2026)), EMPIAR-10445 (held out for cross-domain picker validation), and datasets 10491, 10005, 10302, 10510, 11899, and the 11845 vault negative control (Iudin et al., 2023). Matching tolerance is $d_{\max} = 3.5$ voxels; resolution is at FSC $= 0.143$ (Rosenthal & Henderson, 2003), verifier-recomputed from stored half-maps. Every volume, mask, coordinate list and log is retained under a sandbox tree from which every figure is generated by a shipped script (Appendix J).

4.1 End-to-End Verification

A twelve-stage STA protocol was executed on EMPIAR-13566 by the shipped drivers of Sec. 3.3. What is under test is the *measurement* half of the system: given artifacts an engine wrote, can an independent verifier reproduce the engine's resolution claim? Across three stages (5, 2.5, 2 Å/px) the engine reports 16.84/8.89/8.26 Å against the tutorial's 16.3/9.3/7.4 Å, and our verifier recomputes FSC from stored unfiltered half-maps to 16.12/8.83/8.06 Å (Fig. 2). Every residual is smaller than one Fourier shell step, so both place the 0.143 crossing in the same shell.

The chain matches or exceeds the tutorial at stage 2 (8.89 vs 9.3 Å) and trails at the final stage (8.26 vs 7.4 Å, a 0.86 Å gap). The gap is not explained by particle number (**123,384** exported subtomograms from all 14 tilt series against the tutorial's 120,420) nor by protocol deviation at stages 2–3. Two differences a single run cannot separate remain: class selection (visual in the tutorial, programmatic in ours) and replication. We report the gap as unattributed (Appendix G.4).

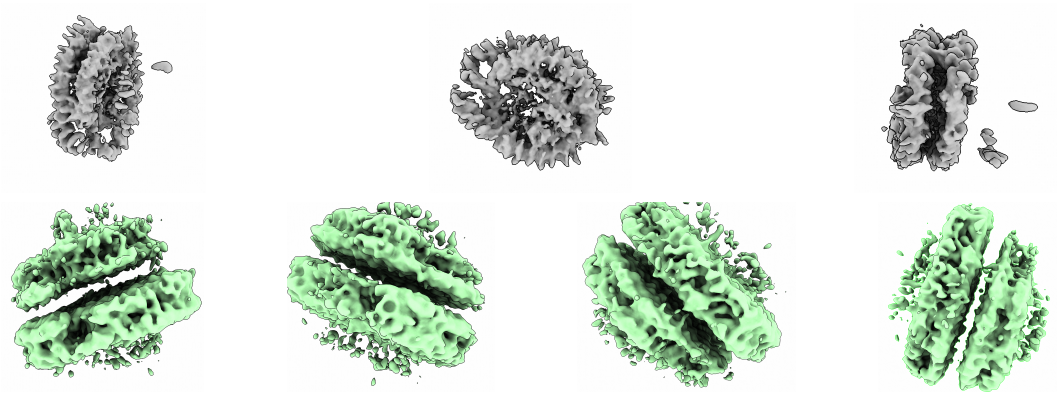


Figure 3: **Qualitative comparison against the human-guided reference tutorial (EMPIAR-13566).** Isosurface renderings of the final reconstructed nucleosome map. Top row (a–c): produced by a human following the EasyMode reference tutorial. Bottom row (d–g): produced autonomously by CRUCIBLE. Both recover the nucleosome disk morphology with resolved surface features; the comparison is qualitative and complements the per-stage FSC analysis of Sec. 4.1.

4.2 COMPARISON WITH THE HUMAN-GUIDED REFERENCE TUTORIAL

To check that the autonomous loop's end product is structurally meaningful and not merely self-consistent, we render the final reconstructed map from the autonomous run against the map produced by a human following the EasyMode reference tutorial on the same EMPIAR-13566 tilt series (Fig. 3), with isosurfaces drawn at a common threshold. Both pipelines recover the nucleosome disk with resolved surface features; the autonomous reconstruction shows the same overall morphology and comparable surface detail. This is a qualitative agreement check only: it complements, and does not replace, the per-stage FSC comparison of Sec. 4.1, in which the autonomous chain trails the tutorial by 0.86 Å at the final stage, and no superiority claim is drawn from renderings.

4.3 CLOSED-LOOP ROUTING AND FAILURE RECOVERY

Two complete closed-loop traces are read end to end (Table 4; narration in Appendix G). Two traces are an instrumented case study, not an experiment: they show that a behaviour occurs but cannot say how often.

Run A (3 iterations, 0 errors) tunes a template matcher three times. The admissible action space grows from **6 to 8 tools between iterations 1 and 2 with no code change**, because the first pick artifact makes consumer tools admissible via Eq. 1. Tier-3 resolution stays not measured throughout, recorded as not-measured rather than 0.0 Å. Because tier 2 is a measurement and not a log line, the loop sees that the picker returned 1022 picks for a requested `top_n` of 700, and that two parameter settings gave byte-identical scores: what the planner spent two iterations tuning was never the binding constraint. At iteration 2 the critic fires `need_tool`; the instrument gate rejects all candidates.

Run B (4 iterations, 0 errors) isolates repair and refusal. Validate-and-fix rewrites an out-of-bound `top_n` ($10 \rightarrow 100$) before launch; the physics tier reports that `metric_threshold` is silently ignored—a defect no exit code exposes, since the run returns 0 and writes 179 picks. All four iterations resolved to the same cached artifact, so the tuned parameter never changed what was measured: a cache-key defect visible only because artifact paths are recorded per iteration. At iteration 4 the critic detects plan/action incoherence and stops. Both runs are evidence-complete for tiers 0–2 and not checkable at tier 3.

Four silent failures were localized and recovered without human intervention (Fig. 5): an all-zero segmentation traced to missing alignment files, a mask/map box mismatch, pick-count recovery, and a STAR column-name collision. In every case tier 1, not the exit code, raised the alarm (Appendix G.7).

4.4 Transfer and Paired Comparisons

One vocabulary and schema transfer without retuning across four registers: raw and preprocessed slices on all seven CZII tomograms, segmentation for two targets from different agent-selected models, and per-class Dice on held-out sets (Fig. 6). All eleven reconstructed chains are tabulated with provenance and failure category (Appendix Table 3); no row is dropped. Three are supported by recomputation to within one Fourier shell step; three show engine/verifier disagreement up to 4.87 Å; two have poor measurements; three lack sufficient artifacts.

Applied to denoising over the seven CZII tomograms—same picker (TomoTwin (Rice et al., 2023), $d_{\max} = 5.0$), ground truth and voxels, only the preprocessing input differing—the denoised input wins 6 of 7 pairs (median gain 0.0332 in $F_{\beta=4}$). A two-sided sign test (Conover, 1999) with $n = 7$ cannot go below $p = 0.016$, so we report this as directionally consistent but unproven (Fig. 7b, Appendix Table 8).

4.5 Particle-Level Validation

Against curated CZII annotations (136 particles, six species), the union of trained-picker outputs yields 169 picks of which 96 lie within $d_{\max}$ (precision 0.57, recall 0.71). Matched sites coincide with ring-like particle density while unmatched sites sit on background, consistently across depths (Fig. 9a,b). Per-species metrics show mask-guided picking dominating template matching on the two abundant species—ribosome $F_{\beta=4}$ 0.642 vs 0.102, ferritin 0.790 vs 0.395—and transferring across four rarer species at 0.762–0.850 (Appendix Table 5).

The guided picker is a nnU-Net (Isensee et al., 2021) trained on the seven CZII tomograms and admitted through the instrument gate. Its cross-domain acceptance test on EMPIAR-10445 passes on four of six species; the rejected five-fold ensemble is documented in Appendix G.5.

4.6 Training Effects

On a 12-prompt retrospective evaluation the RL-trained 4B policy scores 0.0637 mean reward against 0.0354 for its base—a factor of 1.80 (Fig. 8a,b)—and the action mix shifts: template-matching picks fall from 37.2% to 3.5% while STA rises to 48.3% (c). Two caveats: one seed, and prompts from the training corpus, so this licenses changed behaviour, not generalization. Three trained 9B variants are *indistinguishable* on a held-out diagnostic (total reward 0.7245, identical tool choices; Appendix Table 10). We claim learned tool selection, not held-out RL superiority.

4.7 Null Controls

The particle funnel (Fig. 10a) tracks attrition from raw picks through refinement; the runtime profile (b) shows denoising dominating wall-clock, motivating the mask's refusal to re-denoise. Measured denoise action share per checkpoint: 13.9–32.8%.

Our logs measured purity inflation of 1.40–1.81× from subsampling alone. Testing a concrete claim—that membrane proximity filters false-positive picks (EMPIAR-10301)—the result is *negative*: purity near membrane (0.0412) and far (0.0701) both fall below the size-matched random control (mean 0.0773 over five draws). Purity rises merely by keeping fewer particles, exactly as Prop. 3 predicts, and only the control exposes it (Appendix G.7).

4.8 Ablations

Nine quantitative ablations (full details in Appendix F): (i) removing mask guidance collapses ribosome $F_{\beta=4}$ from 0.642 to 0.102; (ii) base→RL ×1.80 on training prompts, 9B GRPO adds nothing over SFT on held-out; (iii) paired denoising 6/7, not significant; (iv) denoise action share 13.9–32.8%; (v) a parsing defect left tier 3 inactive for rounds 1–8, fix shifts policy toward resolution-bearing actions; (vi) removing size-matched control turns inflation into an apparent effect; (vii) per-episode availability declaration raises routing from 60% to 95%; (viii) five-fold ensembling neutral in-domain, negative cross-domain at 5× cost; (ix) neutral-score cap restores correct ordering ($0.20 < 0.40 < 0.656$).

5 CONCLUSION

CRUCIBLE reframes autonomous scientific execution as a measurement problem: the agent's outputs are artifacts, and its progress is whatever an independent verifier can recompute from them. Evidence tiers, admissibility masks, gated rewards and availability flags keep validity from being traded for score, and keep a missing outcome from masquerading as a zero one. Experiments carry a published protocol through all twelve stages with every resolution claim independently recomputed, transfer across datasets, validate at particle level, and—through paired and size-matched controls—demote several attractive claims to directional or null. Held-out RL superiority and tutorial-level end-to-end resolution remain open, now as objectives whose criteria are declared before the runs that will test them. Because every measurement is recomputable from checksummed artifacts, a future run can re-score any past trajectory under a revised reward without re-executing it; the artifact store thus turns a sequence of experiments into a cumulative, re-auditable record rather than a set of snapshots.

Limitations. The present loop is a measurement instrument, not an unattended discoverer. The two closed-loop traces run three and four iterations and stop before subtomogram averaging, so no resolution claim rests on them; the loop's longer-horizon behaviour is evidenced instead by the reproduction and recovery experiments. The instrument gate has so far only fired and *refused*, never produced: the five instruments in the execution path were written by us and admitted through the same gate. End-to-end resolution trails the expert tutorial by 0.86 Å at the final stage, a gap we could not attribute to particle number or protocol deviation. Training uses a single seed with prompts drawn from the training corpus, so it licenses changed behaviour, not generalization. Each of these is reported in the main text together with the control or measurement that exposes it, and the next training comparison is pre-declared—same prompts, three seeds, deployed versus δ-separated weights as a declared arm—in Appendix K.

REPRODUCIBILITY STATEMENT

Every figure is generated from artifacts by scripts shipped with the source bundle; every numeric claim carries an artifact reference in Appendix J. Metric definitions, reward versions, data partitions and tool interfaces are in Appendices A–I. Missing outcomes are reported as not measured and never as zero.

ETHICS STATEMENT

This study analyses existing cryo-ET data computationally and reports no new human or animal experiments. Negative and non-significant results appear in the main text. Human review remains necessary for structural interpretation, dataset permissions and resource allocation.

REFERENCES

Arslan Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, et al. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12248–12267, 2024. doi: 10.18653/v1/2024.acl-long.662.

Tim Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J. Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 16:1153–1160, 2019. doi: 10.1038/s41592-019-0575-8.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

Andres M. Bran, Sam Cox, Oliver Schilter, et al. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Guangyao Chen, Wenhao Yuan, and Fengqi You. Bridging electron microscopy and materials analysis with an autonomous agentic platform. EMSeek, CUAISci, Cornell University, 2025. Citation block transcribed verbatim from the bundled reference repository (`EMSeek/README.md`).

William J. Conover. *Practical Nonparametric Statistics*. Wiley, 3rd edition, 1999.

Cryo-SAM developers. Cryo-SAM: segment anything in cryo-electron tomograms with a foundational vision model. Software release used in this work, 2024. No peer-reviewed record for this exact release was located in Crossref or Europe PMC on 2026-09-25; cited as software rather than with an unverified venue.

CryoAgent developers. CryoAgent: agentic single-particle cryo-em processing. Concurrent system referenced in this work, 2026. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

CryoForge-MB developers. CryoForge-MB: agentic model building for single-particle cryo-em. Concurrent system referenced in this work, 2026. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

CryoRL developers. CryoRL: reinforcement learning over microscope navigation. System referenced in this work, 2024. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

Utz Heinrich Ermel, Jonathan Schwartz, Zhuowen Zhao, Daniel Ji, Ariana Peck, Yue Yu, Mohammadreza Paraan, Bridget Carragher, Achilleas S. Frangakis, and Kyle I. S. Harrington. copick: An open dataset interface and toolkit for collaborative annotation and analysis of cryo-electron tomography data. *Protein Science*, 35:e70578, 2026. doi: 10.1002/pro.70578.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *International Conference on Machine Learning (ICML)*, 2024.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021. doi: 10.1038/s41592-020-01008-z.

IsoNet developers. IsoNet: isotropic-resolution reconstruction and denoising for cryo-electron tomography. Software release used in this work (denoising and missing-wedge correction stages), 2022. No peer-reviewed record matching this release was located in Crossref or Europe PMC on 2026-09-25; cited as software rather than with an unverified venue.

Andrii Iudin, Ardan Patwardhan, Jose Salavert-Torres, et al. EMPIAR: the electron microscopy public image archive. https://www.ebi.ac.uk/empiar/, 2023. Data source for EMPIAR-10005, 10301, 10302, 10440, 10445, 10491, 10510, 11845, 11899, 13566.

Lorenz Lamm, Ricardo D. Righetto, Wojciech Wietrzynski, Matthias Pöge, Antonio Martinez-Sanchez, Tingying Peng, and Benjamin D. Engel. MemBrain: A deep learning-aided pipeline for detection of membrane proteins in cryo-electron tomograms. *Computer Methods and Programs in Biomedicine*, 224:106990, 2022. doi: 10.1016/j.cmpb.2022.106990.

Gaoyang Liu, Tongkun Niu, Mengxuan Qiu, Yao Zhu, Fei Sun, and Zhuo Yang. DeepETPicker: Fast and accurate 3d particle picking for cryo-electron tomography using weakly supervised deep learning. *Nature Communications*, 15:2090, 2024. doi: 10.1038/s41467-024-46041-0.

Zichen Liu, Changyu Xia, Aditya Bajpai, Jie Zhang, Haiwen Hu, Ziqi Liu, Yanzhe Chen, Ziyi Wang, Tito Ji, Yifan Zhang, et al. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D. Righetto, Wojciech Wietrzynski, Simon Albert, Dorit Larivière, Emile Fourmentin, Serge Pfeffer, Julio Ortiz, Wolfgang Baumeister, Tingying Peng, Benjamin D. Engel, and Charles Kervrann. Deep learning improves macromolecule identification in 3d cellular cryo-electron tomograms. bioRxiv preprint, 2020.

NextPyp developers. NextPyp: a unified platform for cryo-et and subtomogram averaging. Platform referenced in this work, 2023. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

Ariana Peck, Yue Yu, Jonathan Schwartz, Anchi Cheng, Utz Heinrich Ermel, Joshua Hutchings, Saugat Kandel, Dari Kimanius, Elizabeth A. Montabana, Daniel Serwas, Hannah Siems, Feng Wang, Zhuowen Zhao, Shawn Zheng, Matthias Haury, David A. Agard, Clinton S. Potter, Bridget Carragher, Kyle Harrington, and Mohammadreza Paraan. A realistic phantom dataset for benchmarking cryo-ET data annotation. *Nature Methods*, 22:1819–1823, 2025. doi: 10.1038/s41592-025-02800-5.

ProtoPilot developers. ProtoPilot: self-evolving wet-lab protocols. System referenced in this work, 2026. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

Ali Punjani, John L. Rubinstein, David J. Fleet, and Marcus A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14:290–296, 2017. doi: 10.1038/nmeth.4169.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Xin Lan, Yan Qian, Yaxi Zhao, Ning Zhao, Jinyuan Lan, Xin Ma, Lanxin Hu, Wen Chen, Zhihao He, Dahua Wong, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations (ICLR)*, 2024.

Gabriel Rice, Thorsten Wagner, Moritz Stabrin, Oleg Sitsel, Daniel Prumbaum, and Stefan Raunser. TomoTwin: generalized 3d localization of macromolecules in cryo-electron tomograms with structural dissimilarity metrics. *Nature Methods*, 20:871–880, 2023. doi: 10.1038/s41592-023-01878-z.

Peter B. Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, 333(4):721–745, 2003.

Sjors H. W. Scheres. RELION: Implementation of a bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, 180(3):519–530, 2012.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Dayxia Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

SmartScope developers. SmartScope: autonomous specimen screening for cryo-em. System referenced in this work, 2022. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

SPARTA developers. SPARTA: machine-learning-enabled automation of cryo-em experiments. System referenced in this work, 2024. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

StructAgent developers. StructAgent: agentic structural modeling. Concurrent system referenced in this work, 2026. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

Dimitry Tegunov and Patrick Cramer. Real-time cryo-electron microscopy data preprocessing with Warp. *Nature Methods*, 16:1146–1152, 2019. doi: 10.1038/s41592-019-0580-y.

TODO: verify authors. Subtomogram averaging with gold-standard refinement, 2019. UNVERIFIED PLACEHOLDER: no registry record located on 2026-09-26; replace with the verified record before submission.

Thorsten Wagner and Stefan Raunser. The evolution of SPHIRE-crYOLO particle picking and its application in automated cryo-EM processing workflows. *Communications Biology*, 3:61, 2020. doi: 10.1038/s42003-020-0790-y.

Binfeng Xu, Hao Zhang, Shaokun Zhang, Songyang Han, Mingjie Liu, Jian Hu, Shizhe Diao, Zhenghui Jin, Yunheng Zou, Michael Demoret, Jan Kautz, and Yi Dong. Polar: Agentic RL on any harness at scale. arXiv preprint arXiv:2605.24220, 2026. Citation block transcribed verbatim from the bundled reference repository (`ProRL-Agent-Server-stable/README.md`).

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Hao Zhang, Mingjie Liu, Shaokun Zhang, Songyang Han, Jian Hu, Zhenghui Jin, Yuchi Zhang, Shizhe Diao, Ximing Lu, Binfeng Xu, et al. ProRL agent: Rollout-as-a-service for RL training of multi-turn LLM agents. arXiv preprint arXiv:2603.18815, 2026. Citation block transcribed verbatim from the bundled reference repository (`ProRL-Agent-Server-stable/README.md`).

Shawn Zheng, Grant Wolff, Gene Greenan, Zhuowen Chen, et al. AreTomo: An integrated software package for automated marker-free, motion-corrected cryo-electron tomography. bioRxiv preprint, 2022.

Jasenko Zivanov, Takanori Nakane, Björn O. Forsberg, Dari Kimanius, Wim J. H. Hagen, Erik Lindahl, and Sjors H. W. Scheres. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife*, 7:e42166, 2018. doi: 10.7554/elife.42166.

A METRIC DEFINITIONS AND REWARD VERSIONS

Precision/recall/F_β use tolerance $d_{\max} = 3.5$ voxels on CZII (10.012 Å/px) with greedy one-to-one matching; $\beta = 4$ follows the challenge's recall-biased metric, while the training reward uses tier-2 precision plus tier-3 resolution (project decision 2026-08-20). Resolution is the smallest shell frequency at which the verifier-recomputed FSC ≤ 0.143. Mask fraction is the voxel share above 0.5 in the binarized segmentation. Reward versions are tagged in logs (v2: precision-only after a tier-parsing defect; v3: tiered with resolution gate active); the v2 defect is retained as a natural ablation (rounds 1–8 equivalent to precision-only) and is documented in the research log. The deployed configuration, transcribed from `Crucible/training/reward_resolution.py::compute_reward` rather than paraphrased, is $\lambda = (w_{\text{format}}, w_{\text{physics}}, w_{\text{task}}, w_{\text{resolution}}) = (0.10, 0.10, 0.30, 0.50)$; the weighted sum is divided by the total weight of the tiers *present* for that step, so an episode that never reached averaging is scored on what it did rather than penalized for a missing tier 3. Tier-2 credit is $\text{clip}(F_\beta/0.2, 0, 1)$—$F_\beta = 0.2$ earns full credit, a calibration chosen when the best measured F_β in the corpus was 0.0853. Tier-3 credit is $m_3 \cdot \widetilde{\rho}$ and returns not measured rather than 0.0 whenever resolution is absent or implausible. The renormalization has one hole that was measured rather than anticipated: a step with no measurable tier at all (denoising, segmentation—neither produces a pick set) is scored on format and physics alone and reached 1.0, above the 0.456 that the best pick in the same 17-case sweep earned. Hence the cap at 0.20, applied only when both measurable tiers are absent; the post-fix ordering is denoise $0.20 <$ picks-nothing $0.40 <$ good pick 0.656. Remark 1 records that this weight vector does not satisfy the separation hypothesis of Theorem 1, and the cap is one of the two mechanisms that supply non-compensability without it.

The gates of Eq. (6) are not instantiated in the shipped reward. This is the largest gap between the formal object and the code, so we state it against source rather than let the transcription imply agreement. Eq. (6) gates tier k on every lower tier crossing a threshold, $g_k = \prod_{j<k} \mathbf{1}[u_j \geq \vartheta_j]$; `Crucible/training/reward_resolution.py` contains no ϑ_j at all. What it does is *availability* gating: a term enters the weighted sum only if its measurement exists—tier 2 iff `task_fbeta is not None`, tier 3 iff `r is not None` (lines 289–293)—and the sum is then divided by the total weight of the terms that entered (lines 294–295), which is the renormalization above. Tier 0 is graded, not binary: 1.0 for a valid parse naming a registered tool, 0.25 for a valid parse naming an unregistered one, 0.0 for unparseable JSON, a non-object, or a missing/non-string `tool` key (lines 187–199); on total parse failure the reward short-circuits to $w_{\text{format}} \cdot f$ with physics forced to 0 (lines 278–279), which is where the hard zero of Def. 2 actually comes from. Tier 1 is a genuine three-point indicator $\{0, 0.5, 1.0\}$ (lines 202–232). `UNMEASURABLE_CAP = 0.20` is line 124, applied at lines 296–299 only when both measurable tiers are absent. Weights are the default arguments of `compute_reward` (lines 259–260), not a named constant, and the GRPO trainer separately weights three reward *functions* at $[0.1, 0.1, 0.8]$ (`grpo_orchestrator.py:309`), which is a different axis from the four tiers and should not be confused with λ. Consequently Prop. 1 is a statement about Eq. (6), and what survives into the shipped reward is its conclusion—non-compensability—supplied by the short-circuit and the cap rather than by the product gates. A run that instantiates ϑ_j and re-measures is future work, not a claim.

FSC estimator convention (relocated from Definition 1). The convention is part of the definition, so we fix it once rather than per experiment: half-maps are used *unmasked* and *without* phase randomization, the crossing is obtained by linear interpolation between adjacent shells, and $\min\{\cdot\}$ in $\rho(s) = \min\{s^\star : \text{FSC}(s^\star) \leq 0.143\}$ is a *first*-crossing rule. That last choice is load-bearing rather than cosmetic—recomputed curves are not monotone in practice (the stage-3 curve of Sec. 4.1 rises from 0.295 to 0.466 between shells 19 and 25 before its final descent), and a monotone-ized or last-crossing variant would return a different number. Two rows of Table 3 (EMPIAR-10491 r1, EMPIAR-10005 r2) were recomputed earlier by a *masked* variant of the same estimator, so we do not compare their gaps to the unmasked residuals at sub-shell level. Because the convention also differs from the engines' own (masked, phase-randomized, discrete shell grid), verifier-versus-engine agreement is an estimator-to-estimator comparison that bounds disagreement between conventions; it is not an independent confirmation of the engine's value. Implementation: `scripts/direct_fsc.py`.

B INCIDENT REGISTER

Table 1: Observed failures that motivated each mechanism (F1–F3, Sec. 3).

Date	Incident (artifact)	Mechanism motivated
2026-08-22	71% of dispatched actions were denoise-on-denoised no-ops until timeout (RL v3 dispatch logs)	admissibility mask φ
2026-09-08	all-zero segmentation from a defective fallback model; exit code 0; zero picks (EMPIAR-11845)	tier-1 content checks
2026-09-08	STAR column collisions crashed WarpTools export with four distinct duplicate-key errors	schema verification, `star_clean`
2026-09-08	5 Å/px mask (box 64) fed to a 2.5 Å/px map (box 96) aborted PostProcess	per-stage mask gate
2026-08-14/16	subsampling alone inflated purity 1.40–1.81×; 5/5 classification products negative vs size-matched control	Prop. 3, controls
2026-09-22	draft figure hardcoded a resolution copied from the tutorial column, no artifact	verifier recomputation

C METHOD–EXPERIMENT MAPPING

D PROOFS

Proposition 1. If $u_0 < \vartheta_0$ then $g_k = 0$ for all $k \geq 1$ and u_0's own gate $g_0 \equiv 1$ multiplies $u_0 < \vartheta_0$; by Definition 2 the tier-0 term is $\lambda_0 u_0$ but tiers 1–3 vanish; we set the tier-0 score itself to zero on parse/schema failure, giving $R = 0$. Any admissible trajectory has $R \geq \lambda_0 \vartheta_0 > 0$. For $R' = u_3$, take a vacuous output with $m_3 = 1$ and degenerate but nonzero $\widetilde{\rho}$; it exceeds an admissible trajectory whose refinement failed ($m_3 = 0 \Rightarrow R' = 0$). □

Proposition 2. $\mathbb{E}[mu] = \mathbb{E}[\mathbb{E}[mu \mid m]] = \Pr(m = 1)\mathbb{E}[u \mid m = 1]$ since $m \in \{0, 1\}$. The ranking reversal condition follows by direct comparison of the two products. MNAR: refinement failure probability increases with specimen difficulty, so $m \not\perp u$. □

Proposition 3. Top-k selection under score-independence is a simple random sample of size k without replacement from n items with K successes, hence hypergeometric with the stated mean and variance. The numeric example substitutes $p = 0.5, n = 200, k = 20$. □

Theorem 1. Write $R_\lambda - R'_\lambda = \sum_{j=0}^{3} \lambda_j g_j (u_j - u'_j)$. By hypothesis $u_j = u'_j$ for $j < k$, so those terms vanish and $R_\lambda - R'_\lambda = \lambda_k g_k (u_k - u'_k) + \sum_{j>k} \lambda_j g_j (u_j - u'_j)$. Since $g_j u_j \in [0, 1]$, each tail summand lies in $[-\lambda_j, \lambda_j]$, hence the tail is bounded in magnitude by $\sum_{j>k} \lambda_j < \delta \lambda_k$. Case (a): $g_k = 1$ gives $|\lambda_k g_k (u_k - u'_k)| = \lambda_k |u_k - u'_k| \geq \lambda_k \delta > \sum_{j>k} \lambda_j \geq |\text{tail}|$, so the sign of the sum equals the sign of $u_k - u'_k$. Case (b): $g_k = 0$ means some coordinate $i < k$ falls below ϑ_i; but then for every $j > k$ the gate $g_j = \prod_{i<j} \mathbf{1}[u_i \geq \vartheta_i]$ also contains that factor and vanishes, and since $g = g'$ the same holds for the primed trajectory, so both sums equal the (identical) prefix and $R_\lambda = R'_\lambda$. On $\{g \equiv \mathbf{1}\}$ case (a) applies at every first-differing coordinate with margin $\geq \delta$, which is exactly order-isomorphism to the lexicographic order at resolution δ; letting the separation ratio diverge makes $\delta \to 0^+$ admissible for any fixed finite set of trajectories, giving convergence of the argmax to the lexicographic argmax. □

Table 2: Each mechanism, the failure that motivated it, the experiment that tests it, and the verdict.

Mechanism	Motivating failure	Experiment	Verdict
Masked admissibility φ	F1 denoise waste	RQ8, Abl.(iv)	waste quantified (13.9–32.8%); causal gain not claimed
Evidence tiers + m_3	F2 vacuity, uncheckable	RQ1, RQ4, RQ7	reported vs recomputed agree within 0.20–0.72 Å; not measured reported
Independent verifier	F2 copied scalar	RQ1, RQ5	FSC recomputed from stored half-maps
Gated reward, Prop. 1	hacking by vacuity	RQ3, RQ6	vacuous runs score 0; behaviour shifts
Prop. 2	ranking confounded by availability	RQ6, RQ4	held-out superiority not established
Prop. 3 + controls	F3 selection illusion	RQ9	5/5 negative vs size-matched control
GRPO/LOO/z_i	small-G bias, failed-traj noise	RQ6, Abl.(ii)	+80% retrospective; held-out $\equiv$ SFT
Episode-level φ + salience	one-answer floor on 20 routing cases	RQ2, Abl.(vii)	action space grows 6$\rightarrow$8 in-run; the 25% violation figure retracted as instrument-created
Validate-and-fix repair layer	out-of-bound and silently ignored parameters	RQ2	repair written into the observation and the tier score; a cache-key defect found by reading traces
Four-stage instrument gate	three invented command lines shipped	RQ2, RQ7	gate rejected every autonomous candidate; self-built instruments admitted with stored gate evidence
Neutral-score cap (0.20)	inaction outscored execution (base 0.1912)	Abl.(ix)	ranking inverts after the cap (0.0787 vs 0.1414); no-product step never scores 1.0 again
Power-aware critic	"not significant" read as "no effect"	RQ5, RQ9	test floor $2/2^n$ stated in the prompt; 6/7 directional at $p_{\min} = 0.016$
Artifact-keyed caching	cache key missing an input, faked equivalence	RQ2	defect exposed by per-iteration artifact paths; rule now requires the key to cover every output-affecting input

Table 3: **Every reconstructed case, with measurement provenance, failure category, and no row dropped.** This is the single table behind Sec. 4.4: "Reported" is the value written by the reconstruction engine; "recomputed" is our verifier's independent gold-standard FSC from the stored unfiltered half-maps; the measurement path names the exact job directory, engine field or verifier script the number was read from, and, where a published protocol exists, its reference value. Unmasked recomputations use `scripts/direct_fsc.py` (first 0.143 crossing, linearly interpolated); the EMPIAR-10491 r1 and EMPIAR-10005 r2 rows were recomputed earlier by a *masked* variant and are therefore not comparable to the primary-case residuals at sub-shell level (Definition 1). not measured means *not measured* and is never recorded as 0.0 Å. The m_3 column is the availability indicator of Definition 1: 1 where our verifier could recompute an FSC from retained half-maps, 0 where it could not. It is 1 on 6 of the 11 rows, which is the count Prop. 2 quotes; carrying it as a column is what makes that number checkable rather than asserted. $m_3 = 0$ does not by itself force category (i), and two rows show why the distinction matters. EMPIAR-10302 template-matching is (iii) because its own `PostProcess` record declares `low_resol_join_halves`$= 12$, certifying that the reported 11.71 Å lies below the resolution at which its halves were joined—a verdict the sibling M-segment row independently supports at 18.16 Å. EMPIAR-10302 `ribosome2d` is (ii) because its own record shows a mis-selected class at 34.29 Å. In both the chains artifacts certify the verdict while our independent recomputation is unavailable, and we mark that difference in m_3 rather than let the category column conceal it. The last column assigns each row to the taxonomy of Definition 3: (0) supported by recomputation to within one Fourier shell step, (i) no sufficient artifact set, (ii) a sufficient set with a poor measurement, (iii) a value the set does not support. Aborted, degraded and vacuous runs are listed, not dropped.

Dataset / chain	Rep. (Å)	Recomp. (Å)	Measurement path and status	m_3	Cat.
EMPIAR-13566 nucleosome, stage 1 (5 Å/px)	16.84	16.12	`PostProcess/job005` _rlnFinalResolution; tutorial 16.3	1	(0)
EMPIAR-13566 nucleosome, stage 2 (2.5 Å/px)	8.89	8.83	`Refine3D/job002` _rlnEstimatedResolution (no `PostProcess` job); tutorial 9.3	1	(0)
EMPIAR-13566 nucleosome, stage 3 (2 Å/px)	8.26	8.06	`PostProcess/job002` _rlnFinalResolution; tutorial 7.4	1	(0)
EMPIAR-10491 apoferritin, fresh-species r1	10.52	6.79	engine self-report vs direct masked half-map FSC (`pytom/fsc_apo15b_r1.json`)	1	(iii)
EMPIAR-10491 apoferritin, tutorial 25-step chain	11.34	not measured	log-recorded refinement; half-maps not retained	0	(i)
EMPIAR-10005 tubulin, combo3 r2 (3 tomograms)	19.26	14.39	engine self-report vs direct masked half-map FSC (`mt10005_m/combo3/fsc_r2.json`)	1	(iii)
EMPIAR-10302 ribosome, template-matching chain	11.71	not measured	`tm10302_25/PostProcess/job020`; optimistic (`low_resol_join_halves`$= 12$)	1	(iii)
EMPIAR-10302 ribosome, M segment (honest)	18.16	—	M-estimator half-map FSC; worse than its input, retained as a negative result	1	(ii)
EMPIAR-10302 `ribosome2d`, mis-selected class	34.29	not measured	`ribosome2d/PostProcess/job003`; degraded chain, retained	1	(ii)
EMPIAR-11845 vault (negative control)	not measured	not measured	0 particles: no reconstruction exists to measure	0	(i)
EMPIAR-11899 lamella	not measured	not measured	chain incomplete (11–19 picks per tomogram); never refined	0	(i)

Table 4: Iteration-level records of the two closed-loop runs of Sec. 4.3, transcribed from `sandbox_runs/autoresearch_trace.json` (Run A, checkpoint `verl_v6_kl100_step25_hf`) and `sandbox_runs/graph_e2e_step90_vf.json` (Run B, checkpoint `verl_v6_kl100_s25_step90_hf`). Both runs terminate on a critic `stop` with zero recorded errors. ρ (tier 3) is not measured in all seven attempts and is never written as 0.0 Å. The tier annotation the loop itself saw is reproduced here: in Run A the physics tier flags an out-of-bound parameter that the executor runs anyway; in Run B the repair layer rewrites it and the physics tier reports a parameter the target tool silently ignores. Run B's four iterations share one artifact hash (`cache_template_match_d803ba0c7ac4`, 0.1 s each), which is the cache-key defect discussed in Sec. 4.3.

Step	Planner contract (PLAN, abbreviated)	Action (tool, key params)	Reward	F_β	Picks	Critic
A1	template matching at a moderate threshold; expect $F_\beta > 0.02$, <5000 picks	`template_match` thr 0.7, `top_n` 10	0.2069	0.0023	301	`continue`
A2	raise `top_n` to 500–1000, lower threshold to 0.6	`template_match` thr 0.6, `top_n` 700	0.4123	0.0041	1022	`need_tool`→gate rejects
A3	`top_n` 1000, threshold 0.55, so candidates can feed STA	`template_match` thr 0.55, `top_n` 1000	0.4123	0.0041	1022	`stop`
B1	denoise first, "clean tomograms are foundational"	`template_match` `metric_threshold` 1, `top_n` 10→100 (repaired)	0.2000	0.0	179	`continue`
B2	raise `top_n` to 100, within valid bounds	`template_match` `metric_threshold` 1, `top_n` 100	0.3000	0.0	179	`continue`
B3	`top_n` 1000 and drop the ignored parameter	`template_match` `metric_threshold` 1, `top_n` 1000	0.3000	0.0	179	`need_tool`→gate rejects
B4	switch to a learning-based picker	`template_match` `top_n` 10→100 (repaired)	0.2000	0.0	179	`stop`

Table 5: Picking ablation on CZII TS_5_4 against curated annotations ($d_{\max} = 3.5$ vox). Mask-guided (nnUNet) picking dominates template-based (easymode) picking on the two abundant species; the per-species rows below the rule show the same guided picker transferring across the four rarer species. This is the table behind the head-to-head numbers quoted in Sec. 4.5.

Species	Method	n_{pred}	Precision	Recall	$F_{\beta=4}$
ribosome	easymode (template)	3	1.000	0.097	0.102
ribosome	nnUNet-guided	60	0.350	0.677	0.642
ferritin	easymode (template)	39	0.462	0.391	0.395
ferritin	nnUNet-guided	39	0.923	0.783	0.790
β-galactosidase	nnUNet-guided	14	0.571	0.800	0.782
PP7 VLP	nnUNet-guided	11	0.818	0.818	0.818
thyroglobulin	nnUNet-guided	33	0.697	0.767	0.762
β-amylase	nnUNet-guided	12	0.583	0.875	0.850

Table 6: Cross-domain acceptance test of the self-trained segmentation-as-localization picker (`nnunet_particle_pick`, five-fold nnU-Net trained on the seven EMPIAR-10440 tomograms, fold 0 deployed), evaluated on the five unseen EMPIAR-10445 tomograms with ground truth rescaled by $4.990/10.012$. The declared criterion was $F_1 \geq 0.46$ on four of six species: met, with two species reported as failures rather than dropped. "Declared" rather than "pre-registered" because the threshold is first recorded in the log entry that reports the outcome (Appendix G.5, limit (h)). Artifact `sandbox_runs/phantom_pick_test.json`.

Species	n_{GT}	n_{pred}	TP	Precision	Recall	F_1
ribosome	285	373	247	0.662	0.867	0.751
apo-ferritin	471	350	285	0.814	0.605	0.694
virus-like particle	8	18	8	0.444	1.000	0.615
β-galactosidase	35	21	13	0.619	0.371	0.464
thyroglobulin	69	65	19	0.292	0.275	0.284
β-amylase	28	3	1	0.333	0.036	0.065

Table 7: Single fold versus five-fold majority-vote ensemble for the same picker (ablation (viii)). In-domain rows are from EMPIAR-10440 TS_99_9 (tolerance 5.0 px); the cross-domain row is EMPIAR-10445 TS_100_3 (tolerance 9.0 px, eight ground-truth points). The ensemble never helps where it matters and costs $5\times$ inference per rollout step, so the registered instrument deploys fold 0. Artifacts `ensemble_pick_compare.log`, `ensemble_pick_compare_thyro.log`, `ensemble_pick_compare.json`; the cross-domain ribosome comparison is research-log-recorded only, its per-run JSON having been overwritten (flagged in Sec. 4.5).

Domain	Tomogram	Species	Picks	TP	F_1 single	F_1 ensemble
in-domain	TS_99_9	ribosome	63→63	61→61	0.9531	0.9531
in-domain	TS_99_9	thyroglobulin	47→46	45→46	0.9677	1.0000
cross-domain	TS_100_3	thyroglobulin	11→6	2→1	0.2105	0.1429
cross-domain	TS_100_3	ribosome†	—	—	0.7878	0.8140

Table 8: Paired preprocessing control, CZII: same run, same voxels, same ground truth, same picker, same tier—only the input volume differs (denoised vs filtered). Measurement uses the TomoTwin picker at the loose tolerance $d_{\max} = 5.0$ voxels on *unfiltered* raw picks, so pick counts are $O(10^3)$ and precision is $O(10^{-3})$ by construction; this table is therefore *not* comparable to Table 5 (nnUNet-guided, $d_{\max} = 3.5$ voxels). Δ = filtered − denoised in $F_{\beta=4}$; 6/7 negative (denoised better), median -0.0332; two-sided sign test $n = 7$, minimum attainable $p = 2/2^7 = 0.016$ (not significant).

TS	$F_{\beta=4}$ den.	$F_{\beta=4}$ filt.	Δ	TP den.	TP filt.	n_{GT}	n_{pred} den.	n_{pred} filt.
5_4	0.0716	0.0458	−0.0258	21	19	31	4493	6555
6_4	0.1546	0.0770	−0.0776	53	36	74	4643	6759
6_6	0.0395	0.0302	−0.0093	11	12	23	4372	6387
69_2	0.0731	0.0782	+0.0051	24	30	37	4988	5930
73_6	0.0914	0.0485	−0.0429	29	20	46	4658	6271
86_3	0.1215	0.0795	−0.0420	35	31	55	4019	5747
99_9	0.1284	0.0952	−0.0332	48	40	65	5314	6106

Table 9: Sampled action mix (%) per training checkpoint, 12 prompts each, aggregated over all sampled actions in the rollout (artifact: `phaseB3_chain_checkpoint-*_hist.json`). n is the number of sampled actions the percentages are computed over; it varies because rollouts differ in length, so the ckpt24 row ($n = 56$) is the noisiest and should not be read as a trend. `other` collects `run_tool`/`hold`; `memb.` collects membrane segmentation tools (CryoSAM, MemBrain-seg).

Checkpoint	denoise	segment	tomotwin	STA	template	evaluate	memb.	other	n
ckpt8 (sft_v9)	13.9	26.3	37.2	15.3	5.1	0.7	0.0	1.5	137
ckpt24 (sft_v9)	37.5	35.7	3.6	17.9	0.0	1.8	1.8	1.8	56
ckpt30 (sft_v14)	31.8	26.0	7.3	22.4	1.0	9.4	2.1	0.0	192
ckpt45 (probe_v17)	32.8	27.1	0.5	26.6	0.0	7.8	5.2	0.0	192
ckpt57 (probe_v18)	22.1	17.4	3.5	48.3	3.5	5.2	0.0	0.0	172

Table 10: Held-out 12-prompt diagnostic: trained 9B variants indistinguishable (ablation of the RL stage).

Model	format	physics	task	total
SFT-LoRA 9B v2	1.0	1.0	0.1736	0.7245
GRPO 9B full r3	1.0	1.0	0.1736	0.7245
GRPO 9B r5	1.0	1.0	0.1736	0.7245

E Additional tables

F Ablation detail: salience of the admissibility declaration, and reward shaping

Table 11: The 20-state routing benchmark behind ablation (vii), `scripts/eval_routing.py`. Only the prompt differs between conditions; model, states and scoring are held fixed. The comparison point is a single *best degenerate baseline*—the accuracy of answering one tool for all twenty states—computed from the acceptable-set arithmetic and shared by all three conditions: always `log_blob` 60.0% (12/20), always `tomotwin` 55.0% (11/20), always `template_match` 35.0% (7/20). Conditions 1 and 2 *are* degenerate policies: each answers one tool twenty times and scores its own constant's baseline, so their accuracy carries no routing information. Condition 3 is the first in which the answer varies with the state, and 19 of the 20 states are answered acceptably.

Prompt condition	Answer distribution	Acc. (%)	Δ vs 60%	Failing state
1. tool text with cost only	`log_blob` ×20	60.0	+0.000	8 states, all constant
2. + measured quality	`tomotwin` ×20	55.0	−0.050	9 states, all constant
3. + per-episode availability declaration	`tomotwin` 10, `template_match` 7, `log_blob` 3	95.0	+0.350	R16 only

The single failure under condition 3 is state R16, a budget of three GPU-minutes: the policy chose `template_match` where our acceptable set contained only `log_blob`. We record our own doubt about that key rather than claim 19/20—an FFT scan over a full convolution is not obviously cheaper than a blob detector at that budget, so the acceptable set may be too strict and the true accuracy may be 100%. Two caveats are part of this result. First, all three conditions use the same model on the same twenty states, so the benchmark measures *compliance* with a stated constraint and not discovery of an unstated one: a policy that inferred preconditions from artifact provenance without being told them would score high here for the wrong reason, and this instrument cannot tell the difference. Testing discovery would require hiding the constraints back in the state, which is exactly conditions 1 and 2—so the two questions need two benchmarks and we conflated them into one. Second, under condition 3 the untrained base model and the RL checkpoint answer item by item identically—same 19/20, same distribution, same failing state—so the benchmark cannot detect a

training effect; it is a prompt-sensitivity probe, not an RL evaluation. Condition 2 is the informative failure: adding better information (measured quality per tool) *lowered* accuracy, because these states do not share a best answer and a constant answer becomes a wrong constant once the prompt makes quality salient. That is the mechanism claim of Sec. 3.2 in negative form—salience without per-episode declarations does not help and can hurt, and the reason is that the model responds to the most prominent sentence in the prompt rather than reading preconditions cell by cell.

A separate magnitude claim is retracted here and in ablation (vii). The first quantification of this mechanism in production traces reported that 25% of episodes opened with a tool the dispatcher would refuse. That figure was created by the instrument that measured it: the check tested whether a picking action preceded a consuming action in the recorded order, without asking whether the profile already held a usable pick set from an earlier stage. Split by that condition, genuine violations were zero in both corpora. The mechanism claim survives—declaring availability per episode changes the measured first-action distribution, with a prompt-driven segmenter falling from 25% to 1% of first actions, an uncalibrated template matcher from 10% to essentially zero, and denoising becoming the most common legitimate first step at 53%—but the magnitude claim does not, and the ordering check now runs as a gate item (`scripts/check_trajectory_order.py`) that defaults to the highest-numbered corpus rather than a hardcoded filename.

Table 12: Measurement regime of the 17-case reward sweep behind ablation (ix) (`sandbox_runs/multicase17/aggregate.json`). The rightmost column is the reason the sweep matters more than any single number in it: at $n = 5$ cases no exact test can reach conventional significance, at $n = 15$ it can.

Quantity	Value	Note
cases completed	17 / 17	no case failed to produce an artifact
F_β measurable	15 / 17	two produced no pick set
median F_β	0.0102	best case 0.0853 (CZII TS_5_4, `template_match`)
F_β exactly 0.0	5	pick calibration is dataset-scoped
resolution measurable	0 / 17	$m_3 = 0$ throughout; no case reached STA in one round
reward distribution	0.20–0.75	median 0.2402, mean 0.3051, sd 0.1782
exact two-sided p floor at $n = 15$	6×10^{-5}	versus 0.0625 at $n = 5$
best-F_β case, pre-cap	0.456	renormalization over present tiers only
best denoise step, pre-cap	1.000	*above* the pick that measured something
same three steps, post-cap	$0.20 < 0.40 < 0.656$	denoise < picks-nothing < good pick

Qualitative ablations (relocated from Sec. 4.8). (i) *Negative control* (Fig. 5a): without tier-$\mathcal{M}_1$ content checks the all-zero segmentation on EMPIAR-11845 would have propagated into picking as a silent success, since its exit code was 0 and its mask fraction was 0.00%. (ii) *Crop-level evidence* (Fig. 9b): aggregate precision 0.57 decomposes into visually distinct matched sites, which coincide with ring-like particle density, and unmatched sites, which sit on background—the decomposition the scalar hides and the reason tier 2 is reported with its per-species breakdown. (iii) *Behavioural drift* (Fig. 8c): the stacked action mix per checkpoint is the qualitative signature that RL changed behaviour even where held-out scores did not separate (template matching 37.2%→3.5%, subtomogram averaging rising to 48.3%, over 137 and 172 sampled actions). (iv) *Mask-boundary zoom* (Fig. 4c): segmentation boundaries follow chromatin density, which is the precondition for mask-guided picking rather than a cosmetic result. (v) *Refinement-stage progression* (Fig. 21): holding specimen, picker and mask fixed, density is qualitatively ablated by the number of refinement stages—stage-1 class averages are orientation-streaked blobs, stage-2 isosurfaces resolve the nucleosome disc, stage-3 adds the detached density lobes—tracking the quantitative 16.84→8.89→8.26 Å trajectory of Fig. 2a. (vi) *Preprocessing on/off and cross-case transfer* (Fig. 22): one tomogram with denoising disabled versus enabled shows the contrast gain the paired control of Sec. 4.4 measures and fails to make significant, while pick overlays and refined maps on EMPIAR-10302 and EMPIAR-10005 show the transferred toolchain producing specimen-appropriate structure without retuning.

G EXTENDED EXPERIMENT DETAIL

This appendix carries the material the main body points at rather than repeats: the measured instances of the two negative propositions, the full narration of the two closed-loop traces, both escape hatches rule by rule, the attribution analysis behind the primary-case resolution gap, the acceptance test and rejected ensemble of the self-trained picker, the walk-through of Table 3, and the tool provenance inventory. The three secondary experiments that the page limit moved out of the main body—RQ3 diagnosis and recovery, RQ4 cross-dataset transfer, RQ8 yield and runtime—are reproduced in full in Appendix G.7, and the six qualitative ablations in Appendix F. Nothing here is a claim the main text makes without it.

G.1 MEASURED INSTANCES OF PROPOSITIONS 2 AND 3

Availability is decreasing in difficulty (Prop. 2). The correlation the proposition needs is not hypothetical in our corpus. Across the eleven chains of Table 3, a resolution value is engine-reported for nine but verifier-recomputable for six ($m_3 = 6/11$); across the seventeen-case reward sweep of ablation (ix) it is recomputable for none ($m_3 = 0/17$). The two chains with no measurement at all are the vault negative control (0 particles, so no reconstruction exists to measure) and the lamella run (11–19 picks per tomogram, never refined)—the two hardest specimens in the corpus. Availability is thus decreasing in difficulty, exactly the MNAR structure, and a scalar $R' = m_3 u_3$ would place both at 0.0 alongside chains that reconstructed successfully but whose half-maps the verifier will not accept: the EMPIAR-10302 template-matching refinement reports 11.71 Å with an empty verifier column (its `low_resol_join_halves`$= 12$ makes the estimate optimistic), and under R' it is indistinguishable from a run that produced nothing.

The null band at the configuration actually run (Prop. 3). At the configuration of Sec. 4.7 ($n = 1940$ picks, $K = 107$ ground-truth points matched one-to-one, $\hat{p} = 0.0552$, $k = 970$) the null sd is $\sqrt{\frac{p(1-p)}{k}\frac{n-k}{n-1}} \approx 0.0052$: a large n makes the band narrow in absolute terms, and the two strata sit at -2.7 and $+2.9$ sd from $\hat{p}$ (near 0.0412, far 0.0701), so the stratification by membrane distance is real. That is not the claim under test. The claim is purification *at fixed size*, and the size-matched random subset scores above both strata (mean 0.0773, max 0.0835 over five draws). Reading "far from membrane is purer" as $0.0701 > 0.0552$ is exactly the subsampling inflation the proposition names: it compares a k-subset against the full set rather than against another k-subset, and the finite-population factor $\frac{n-k}{n-1}$ is what makes the naive comparison look decisive when it is not. The one-to-one matcher in `scripts/phaseC_membrane_context.py` claims each ground-truth point at most once *within a stratum*, which is why the strata are disjoint in picks but not additive in true positives ($40 + 68 = 108$ against 107 overall): one ground-truth point lies inside both distance bands and is claimed once in each.

G.2 THE TWO CLOSED-LOOP TRACES OF SEC. 4.3, IN FULL

Run A (`autoresearch_trace.json`; 3 iterations, 3 attempts, 0 errors). Iteration 1: the planner proposes template matching, the orchestrator emits `top_n`$= 10$, the physics tier flags it as outside the declared bound $[100, 40000]$, the executor runs it anyway and measures reward 0.2069, F_β 0.0023, 301 picks, no resolution, in 258 s; the critic returns `continue` with the bottleneck "threshold/top_n too restrictive". Iteration 2 lowers the threshold and raises `top_n` to 700 (reward 0.4123, F_β 0.0041, 1022 picks); the critic returns `need_tool`—"flooded picks with very low F_β despite tuning"—the instrument hatch fires, and it answers *"candidates found but none passed the validation gate"*. Iteration 3 retunes again (threshold 0.55, `top_n` 1000, 207 s of real compute) and reproduces the *identical* measurement—1022 picks, F_β 0.0041—and the critic returns `stop`. Two facts here are visible *only* because admissibility is recomputed from the artifact store each iteration: the orchestrator's action space grows from 6 to 8 tools between iterations 1 and 2 with no code change, because the first pick artifact makes consumer tools admissible; and `resolution_A` is `null` in all three attempts, recorded as not-measured rather than as 0.0 Å. A third fact is visible only because tier 2 is a measurement and not a log line: the picker returned 1022 picks for a requested `top_n` of 700, and two different parameter settings gave byte-identical scores, so the parameter the planner spent two iterations tuning was never the binding constraint.

Run B (`graph_e2e_step90_vf.json`; 4 iterations, 0 errors) isolates repair and refusal—and one defect of our own. The executor's validate-and-fix stage rewrites an out-of-bound `top_n`: $10 \rightarrow 100$ before launch. The physics tier then states why in the observation: *"`template_match` does not accept [`metric_threshold`] (silently ignored, so the action would not do what its rationale says)"*—a defect no exit code exposes, since the run returns 0 and writes 179 picks. The reward for that iteration is exactly 0.20, the value executor rule (iv) assigns to a step with no measurable product, and it rises to 0.30 once the bound is satisfied: the difference comes from the physics tier, not from a better pick set, because F_β is 0.0 throughout. Reading the traces also exposes an instrument defect we did not know about: all four iterations resolved to the *same* cached artifact (identical content hash, 0.1 s each), so the parameter the loop was tuning is not part of the cache key and never changed what was measured. That is a direct violation of our own rule that a key must cover every input affecting the output, it is visible only because the artifact path is recorded per iteration, and it invalidates any claim that this run "searched" over `top_n`. At iteration 3 the critic routes `need_tool`, and the gate answers *"candidates found but none passed the validation gate"*: the hatch fired and no autonomously-authored tool entered the execution path. At iteration 4 the critic detects plan/action incoherence—an incoherence present from iteration 1, where the plan asks for denoising and the action calls template matching—and stops ("still using `template_match` despite plan to switch to a learning-based picker").

Read in the categories of Definition 3, both runs are evidence-complete for tiers 0–2 and *not checkable* at tier 3 ($m_3 = 0$ in every attempt, because this configuration stops before subtomogram averaging), so no resolution claim is made from them; the run in which tier 3 becomes available is RQ1. The loop's decisions are reproducible from the traces, and its most consequential decision in Run B was a refusal.

G.3 The two escape hatches: data scout and instrument gate

Data scout (`need_data`), relocated from Sec. 3.1. The scout ranks portal candidates by expected distribution shift—organism novelty first, then sample type, then voxel spacing—because the question under test is whether the policy generalizes, which a seventh run of an already-covered organism cannot answer. Downloads are byte-verified against the portal manifest before they are registered into the artifact store, since 17 of 69 files in one early download batch reported complete while truncated; an unverified file in B_t is worse than a missing one, because every downstream gate that reads provenance would then be reading a checksum of the wrong bytes.

Instrument gate (`need_tool`). The gate is the only node that could put model-influenced behaviour into the execution path, so its output is treated as untrusted until four checks pass in order. *Stage 1, structural:* every declared field is present, the argument template contains no shell metacharacters, and the executable resolves on the sandbox `PATH`. *Stage 2, flag reality:* every flag the rendered command emits must appear in the tool's own `-help` output. This stage exists because three invented command lines reached this project's scripts before anyone ran the tool they named; a specification can be fluent and still describe a program that does not exist. *Stage 3, execution:* the command runs inside the sandbox and exits 0 within a timeout. *Stage 4, product:* an artifact of the declared type exists at the declared path afterwards, checked by suffix and by content (a volume must have a non-trivial header, a coordinate file must parse), because exit 0 has meant "nothing written" here more than once. Only then is the specification registered, and the registration stores the evidence each stage saw—`-help` text, exit code, artifact path and checksum—so admission is auditable rather than a leap of faith. The model writes no code at any point: it emits a declarative specification (executable, argument template, declared product path and suffix) that the dispatcher interprets, and no generated Python is ever executed. In both reported runs the gate fired once and rejected every candidate it was given, so no autonomously-authored tool entered the execution path; the five self-built instruments in use (Sec. 3.3) were authored by us and admitted through the same four stages.

G.4 Attribution of the primary-case resolution gap

The 0.86 Å gap between the agent's final stage (8.26 Å engine-reported, 8.06 Å verifier-recomputed) and the expert tutorial's 7.4 Å is reported in Sec. 4.1 as *unattributed*. This appendix records what

was measured in the attempt to attribute it, because the negative results are what make the label honest.

Excluded: particle number. The chain carries 123,384 exported subtomograms from all 14 tilt series against the tutorial's 120,420 picks; 41,777 class-selected particles against its ∼35,000; and 33,266 particles entering the final refinement. Every count is at or above the tutorial's. The particles that do enter are also well behaved: class occupancy is 29.2% for the selected class, Euler-angle coverage shows no orientation collapse, and shift distributions are centred (Figs. 12, 13).

Excluded: protocol deviation at stages 2–3. The re-exports (box 96 at 2.5 Å/px, box 128 at 2.0 Å/px), the alignment-free $K{=}4$ and $K{=}2$ classifications with `skip_align`, and the final refinement ran the tutorial's commands with the tutorial's parameters; the command lines are in the artifact tree and reproduced in the research log.

Not excluded: class selection. The tutorial selects its final class visually, from the $K{=}2$ classification, at 29.2% occupancy. Our chain selects programmatically—the largest class distribution subject to a rotation-accuracy bound below 900°—and the two rules can pick different classes on the same run. A single run cannot separate "a different class" from "a worse refinement of the same class".

Not excluded: replication. 0.86 Å is one realization of a stochastic refinement, not a gap with a variance attached. No seed replication was run for the primary chain, so the gap carries no interval and cannot be compared to any published per-run spread.

A contrast that shows the difference matters. An earlier variant of the same chain (`nucleosome_3v2`) terminated at 17.07 Å and *was* particle-limited: 1,561 particles after class selection and 846 into refinement, i.e. 26× and 41× below the tutorial's counts. That variant's gap is attributable and large; the primary chain's is not attributable and small. Keeping both in the artifact tree, and refusing to transfer the first explanation to the second, is the discipline Sec. 4.1 claims.

G.5 The self-trained picker: acceptance test and rejected ensemble

The guided picker of Table 5 is a five-fold nnU-Net we trained on the seven EMPIAR-10440 tomograms (datasets 501–503), with fold 0 deployed and registered as a declarative instrument; segmentation becomes localization by taking each connected component's centroid as a coordinate. Its acceptance test is *cross-domain*: EMPIAR-10445, five tomograms never seen in training, with ground truth rescaled by the pixel-size ratio $4.990/10.012$. The criterion was $F_1 \geq 0.46$ on four of six species. We call it *declared*, not pre-registered: the research log announces this acceptance test two entries before the run but writes the numeric threshold only in the entry that reports the outcome, and no script encodes it as a pass/fail test, so the honest claim is a criterion stated in the record alongside its result rather than one committed to before the result existed. It passes on four (ribosome 0.75, apoferritin 0.69, virus-like particle 0.62, β-galactosidase 0.46) and fails on two (thyroglobulin 0.28, β-amylase 0.07—the smallest particles, with the fewest training instances); per-species precision, recall and counts are in Table 6. We report the failures because a criterion met by six of six would have been evidence that the held-out set was in-domain rather than that the picker generalizes.

A second declared decision went the other way. Ensembling all five folds by majority vote is *exactly* neutral in-domain on the strong class—ribosome on TS_99_9: F_1 0.9531 for the single fold and 0.9531 for the ensemble, with identical 63-pick sets and 61 true positives (artifact `ensemble_pick_compare.log`)—and gains $+0.032$ in-domain on the weak one (thyroglobulin on TS_99_9: $0.9677 \to 1.000$, 47 picks/45 TP → 46/46), yet *loses* 0.068 on the same species once the tomogram is cross-domain (thyroglobulin on TS_100_3: $0.2105 \to 0.1429$, 11 picks/2 TP → 6 picks/1 TP over eight ground-truth points; artifact `ensemble_pick_compare.json`)—all at five times the inference cost, a cost paid on every rollout step during training. A third comparison, ribosome on the same hard tomogram, gained $+0.026$ ($0.788 \to 0.814$ under that script's per-species radius tolerance; the 0.75 quoted above uses a fixed tolerance, so the two are not directly comparable); we flag it as log-recorded with its per-run JSON overwritten by the subsequent thyroglobulin

run, rather than quietly re-running it to restore the artifact. Ensembling is a model-quality lever, not a domain-shift lever, so the deployed instrument keeps a single fold (ablation (viii), Table 7). Both decisions are recorded with their numbers rather than only their outcomes, which is the same discipline the reward hierarchy imposes on the agent—and neither is a pre-registration in the timestamped sense, a distinction Sec. G.5 states rather than papers over.

G.6 Walk-through of Table 3

Eleven chains, four datasets beyond the primary case, and three distinct failure categories plus one supported category—this is what "transfer" means here, and it is a weaker and more useful claim than uniform success. Three rows are category (0), the primary case at all three stages, where reported and recomputed values fall in the same Fourier shell. Three are category (iii), where the engine's own scalar disagrees with a recomputation from the artifacts the engine wrote: by 4.87 Å on EMPIAR-10005 tubulin (19.26 reported vs 14.39 recomputed), by 3.73 Å on EMPIAR-10491 apoferritin (10.52 vs 6.79), and by an unquantified margin on the EMPIAR-10302 template-matching chain, whose `low_resol_join_halves`$=12$ makes its 11.71 Å optimistic by construction and whose half-maps the verifier therefore does not accept. Two are category (ii), where a sufficient set exists and the measurement is genuinely poor: the M-estimator segment returns 18.16 Å, *worse than its own input*, and the mis-selected-class chain returns 34.29 Å; both are retained, because a table of only the successful chains is exactly the selection effect Proposition 3 warns about. Three are category (i), where no measurement exists to report: the vault negative control yields zero particles (and the tier-1 content check is what caught it, Sec. 4.3), the lamella chain stops at 11–19 picks per tomogram, and one apoferritin chain did not retain its half-maps. A scalar-reward agent would have scored all three category-(i) rows as 0.0 Å—indistinguishable from the worst possible reconstruction—and would have had no mechanism to notice that two of its own reported resolutions were unsupported by its own artifacts. Transfer is therefore evidenced not by the good rows but by the fact that the taxonomy, the availability indicator and the verifier classify all eleven without exception.

G.7 Secondary experiments in full

The experiments summarized in Sec. 4.3–4.7 of the main body are stated here in full, with the same motivations and the same numbers.

Autonomous diagnosis and recovery in long-horizon execution (Sec. 4.3). *Motivation (F2, tier 1 and the diagnose/rollback actions):* long-horizon execution fails mostly by silent vacuity with zero exit codes, so content-level checks must localize and recover it. Four representative failures were localized to root cause and recovered without human intervention (Fig. 5): (a) an all-zero segmentation on EMPIAR-11845 (mask fraction 0.00%, zero picks) traced to missing alignment sidefiles that forced a degraded re-alignment; (b) a mask/map box mismatch (64 vs 128) that aborted post-processing, fixed by regenerating the mask at the map box; (c) pick-count recovery after the (a) fix; (d) a STAR column-name collision that silently mis-reads coordinates, caught by schema verification. In each case the verifier's content-level check (tier $\mathcal{M}_1$), not the exit code, raised the alarm—exit codes were zero in (a) and (d).

Cross-dataset and cross-target transfer (Sec. 4.4). *Motivation (control interface):* one vocabulary and one schema are justified only if they transfer without retuning. The evidence is of four kinds: raw tomogram slices across datasets, paired filtered/denoised slices for all seven CZII tomograms, segmentation boundaries for two distinct targets produced by different agent-selected models, and per-class Dice of the trained segmentation model on two held-out evaluation sets (Fig. 6). Radial power spectra of the two preprocessing variants (Fig. 16) show the transfer is not achieved by suppressing high-frequency information, which is the trivial way a denoiser can make slices look cleaner. Resolution values travel along different measurement paths and each is labelled in Table 3: 8.26 Å engine-reported for the primary case, 6.79 and 14.39 Å verifier-recomputed for apoferritin and tubulin, 11.71 Å engine-reported only, its half-maps not accepted by the verifier.

Anisotropy and the missing wedge (Sec. 4.4). Orthogonal slices and Fourier planes of the seven CZII tomograms expose the missing wedge imposed by the $\pm 45^\circ$ tilt geometry (Fig. 7c,d), which

bounds how any z-direction resolution claim may be read; the engine's own Fourier completeness for the three refined maps of the primary case is $0.947/0.954/0.972$, the scalar counterpart of the same defect. No isotropic-resolution claim is made anywhere in this paper. The 2D power spectra (Fig. 19) show CTF rings present in the raw tilt and attenuated by the filtered and denoised variants, which is the frequency-domain evidence that the preprocessing pair of Table 8 differs in the way it is supposed to.

What reinforcement training changes (Sec. 4.6). Two ratios in this project round to 1.80 and measure different things, so we separate them explicitly. The RL-trained 4B policy scores 0.0637 mean reward against 0.0354 for its base on a 12-prompt retrospective evaluation, a factor of 1.80 (Fig. 8a,b). The shaping fix of ablation (ix) moves the *9B* task reward from 0.0787 to 0.1414, also 1.80, on a different model, a different reward component and a different prompt set. We flag the coincidence because two ratios that round identically invite conflation; neither corroborates the other. The first is a single seed, so it is a point estimate with no variance, and its prompts are drawn from the corpus the policy trained on, so it licenses a claim about changed behaviour and not about generalization. Behaviourally the sampled action mix shifts: template-matching picks fall from 37.2% to 3.5% while subtomogram averaging rises to 48.3% (Fig. 8c; full mix and sample sizes—137 and 172 actions at the two endpoints—in Table 9), and rollout loop statistics are stable (d). Against that, on a held-out 12-prompt diagnostic set three trained 9B variants (SFT-LoRA and two GRPO runs) are *indistinguishable*: total reward 0.7245, task reward 0.1736, identical tool choices (Table 10); and on the 20-state routing benchmark of ablation (vii) the untrained base and the RL checkpoint answer item by item identically—same 19/20, same distribution, same failing state. The claim we therefore make is learned tool-selection behaviour, not held-out RL superiority.

Yield and runtime orchestration (Sec. 4.7). *Motivation (F1, admissibility mask):* the mask is aimed at measurable waste, so this experiment quantifies the attrition and runtime profile it targets. The particle funnel (Fig. 10a) quantifies attrition from raw picks through cleaning, classification and refinement; the runtime profile (b) shows learning-based denoising dominating wall-clock, which motivates the availability-layer mask φ that rejects denoising already-denoised inputs (Eq. 1). Per-tomogram mask fraction tracks pick count (Fig. 17), cross-validating the segmentation and picking modules against each other.

Selection claims against their null (Sec. 4.7). The claim under test is concrete: that membrane proximity filters false-positive picks (EMPIAR-10301; MemBrain-seg membrane, TomoTwin picks, curated ribosome annotations, tolerance 5 voxels, keep-fraction 0.5). The result is negative. Purity near the membrane is 0.0412 and far from it 0.0701; both fall below the size-matched random control (mean 0.0773, max 0.0835 over five draws). Overall purity is 0.0552 (107 of 1940 picks matched one-to-one) and each stratum holds $k = 970$ picks, so the structured verdict `beats_size_matched_control` is false for both. All five classification products carry the fields `before/after/size_matched_control/verdict` and all five report negative; the acceptance script recomputes the verdict direction from the numbers and rejects prose that contradicts them. One detail we report rather than smooth over: the strata are disjoint in picks but not additive in true positives ($40 + 68 = 108$ against 107 overall), because the one-to-one matcher in `scripts/phaseC_membrane_context.py` claims each ground-truth point at most once *within* a stratum and one point lies inside both distance bands. It does not move the verdict direction, but the stratified purities are not a partition of the overall one. The subsampling-only inflation our logs measured—1.40/1.58/1.65/1.81× at keep-fractions 0.5/0.25/0.1/0.05 (`purity_confound.json`)—is the effect the control exists to remove.

H Supplementary figures

Figure 4 renders the artifact types the loop reads and writes; Figures 5, 6, 7, 8, 9 and 10 support RQ3, RQ4, RQ5, RQ6, RQ7 and RQ8 of the main text and are placed here to keep the main body within the page limit; each is referenced from the section it supports. Figures 11–22 are supplementary throughout.

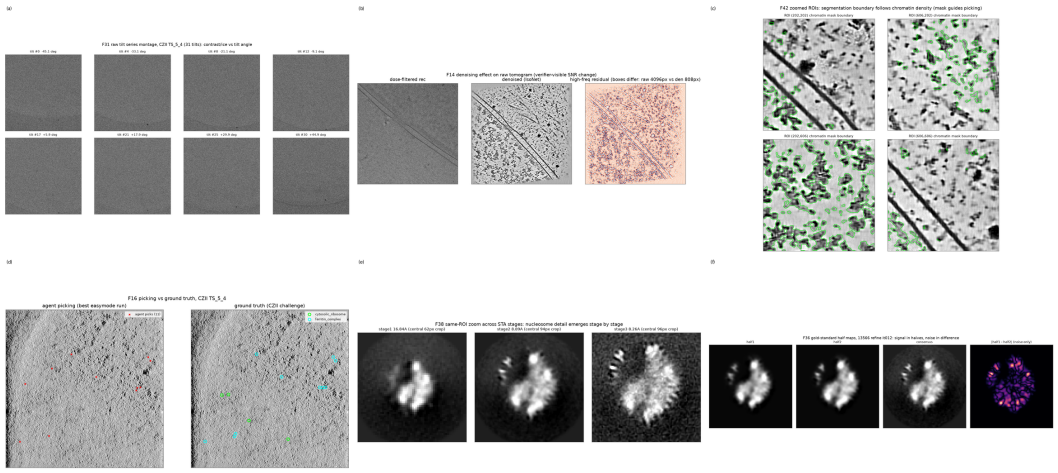


Figure 4: **The agent's observable state space, as images.** Each panel is a verifier-readable artifact of one stage: (a) raw tilts (8 of 31, $-45.1°$ to $+44.9°$); (b) filtered/denoised/difference triplet on one CZII tomogram; (c) segmentation boundary zoom (chromatin, EMPIAR-13566); (d) picks versus curated ground truth; (e) one ROI across subtomogram-averaging stages; (f) half-maps, consensus, and their difference. The agent acts on and is scored over exactly these objects, not over log text.

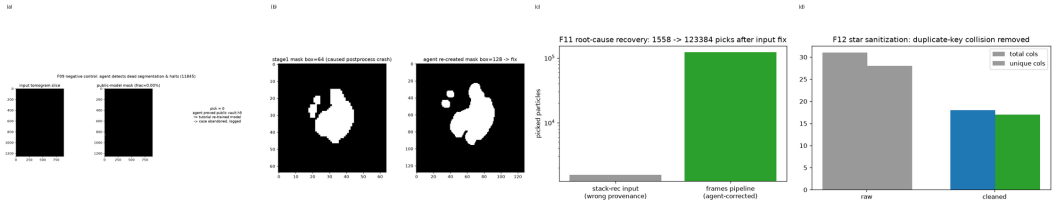


Figure 5: **Autonomous diagnosis and recovery (RQ3).** (a) All-zero segmentation negative control; (b) mask box mismatch and regenerated mask; (c) pick counts before/after root-cause fix; (d) STAR column collision before/after schema cleaning. In (a) and (d) the exit codes were zero and only the tier-1 content check raised the alarm.

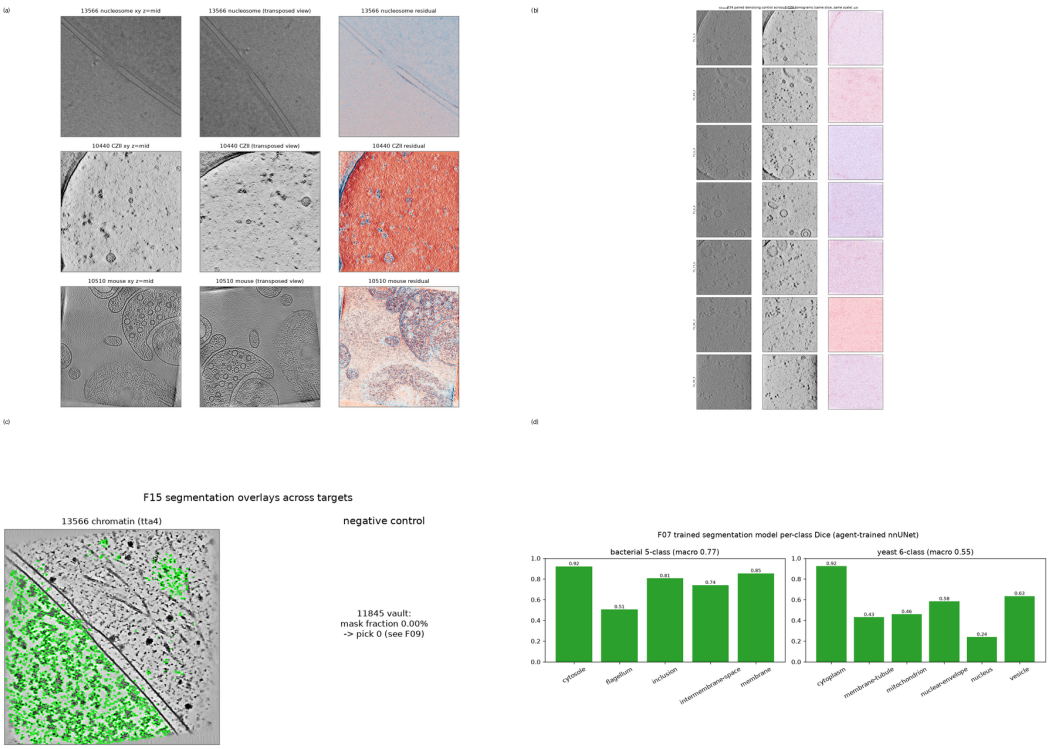


Figure 6: **Generalization (RQ4).** (a) Raw tomogram slices across datasets, identical contrast; (b) paired filtered/denoised/difference for all seven CZII tomograms; (c) segmentation contours for chromatin (13566) and membrane (10510), produced by different agent-selected models; (d) per-class Dice on held-out evaluation sets.

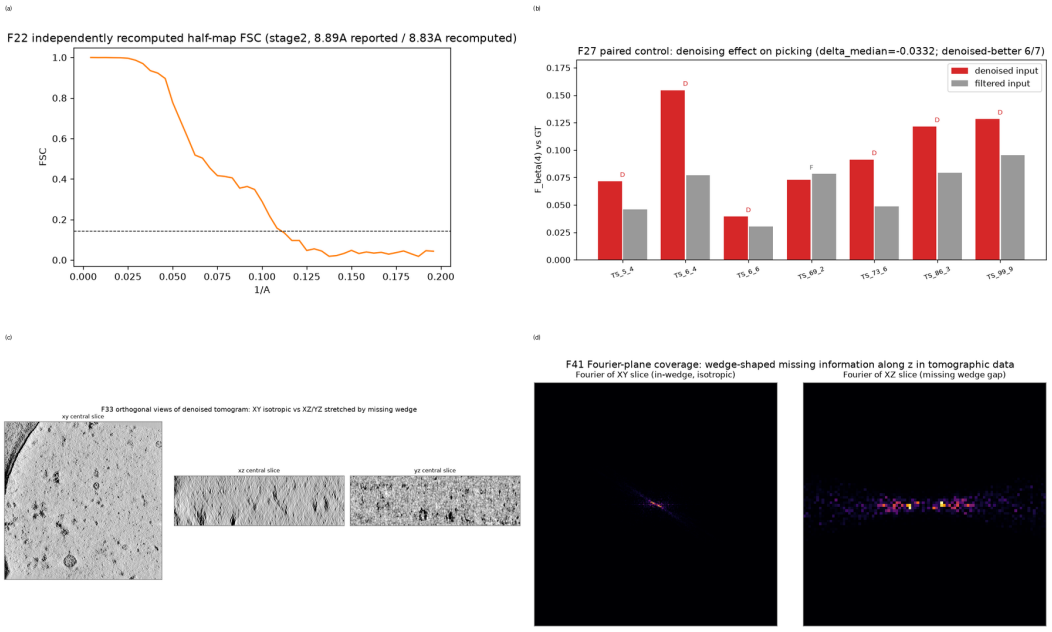


Figure 7: **Verifier-grounded QC (RQ5).** (a) FSC recomputed from half-maps vs reported; (b) paired denoised-vs-filtered picking control, 7 tomograms; (c) orthogonal slices; (d) Fourier planes showing the missing wedge.

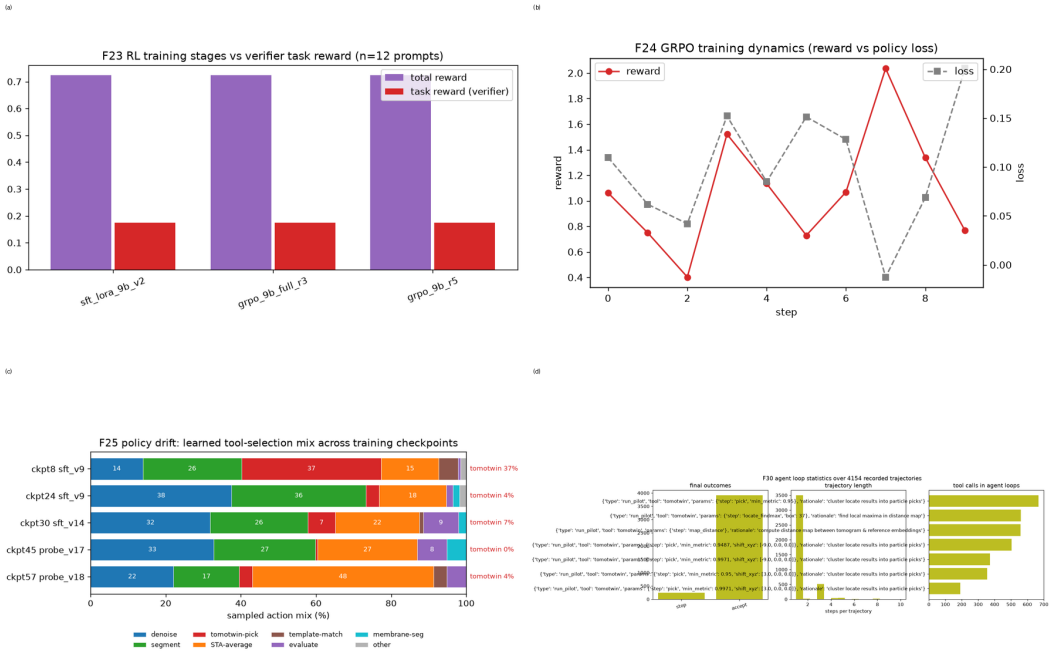


Figure 8: **Policy training effects (RQ6).** (a) Evaluation reward by training stage; (b) reward and loss trajectories; (c) sampled action mix per checkpoint (137 and 172 sampled actions at the two endpoints, Table 9); (d) rollout loop statistics. Panels (a) and (c) are the qualitative and quantitative halves of the same finding: sampled behaviour shifts while held-out scores do not separate.

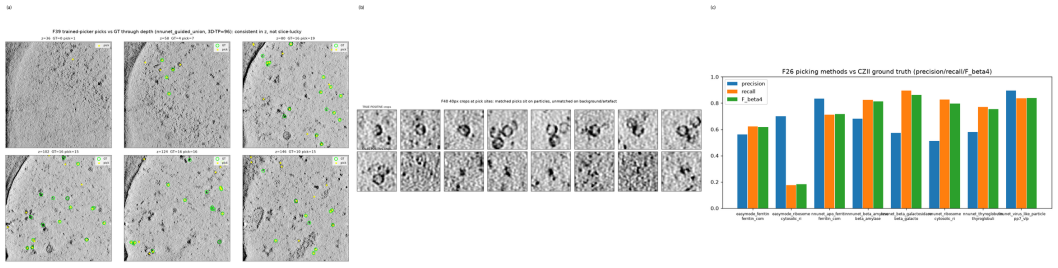


Figure 9: **Particle-level picking validation (RQ7).** (a) Picks vs ground truth at six depths; (b) 40-voxel crops at matched (top) and unmatched (bottom) sites—matched sites coincide with ring-like particle density, unmatched sites sit on background; (c) precision/recall/$F_{\beta=4}$ per method and species.

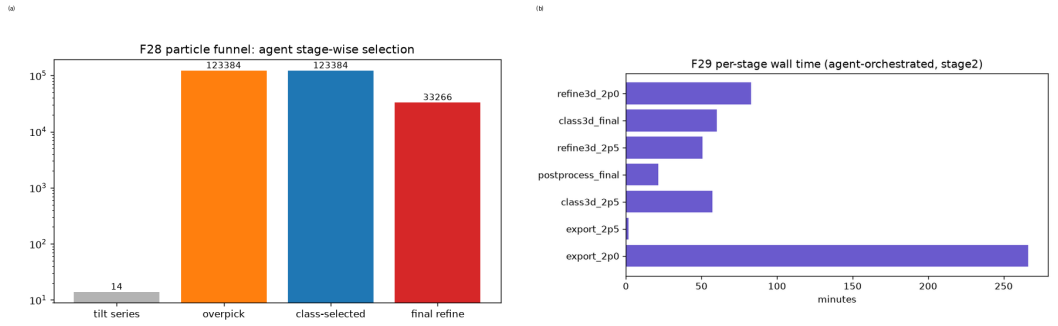


Figure 10: **Orchestration (RQ8).** (a) Particle-count funnel from raw picks through cleaning, classification and refinement; (b) per-stage wall-clock share, in which learning-based denoising dominates and thereby motivates the availability-layer mask of Eq. 1.

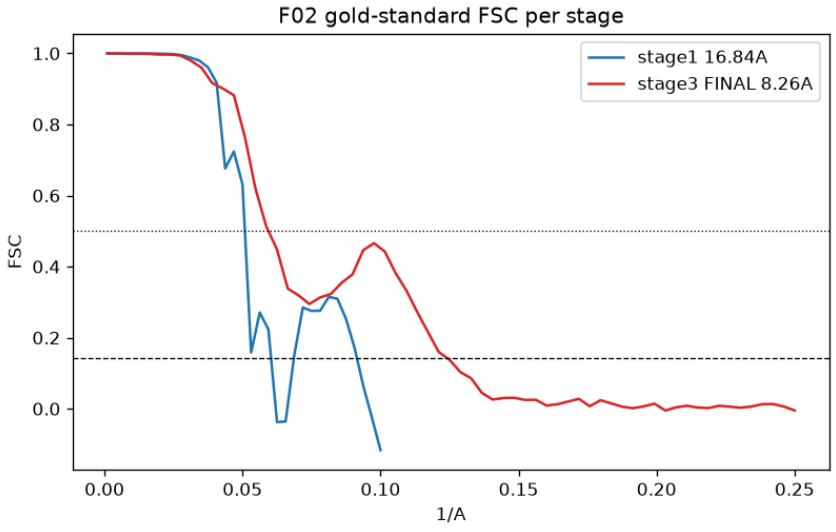


Figure 11: Per-stage gold-standard FSC curves with 0.143/0.5 thresholds (supports Fig. 2a).

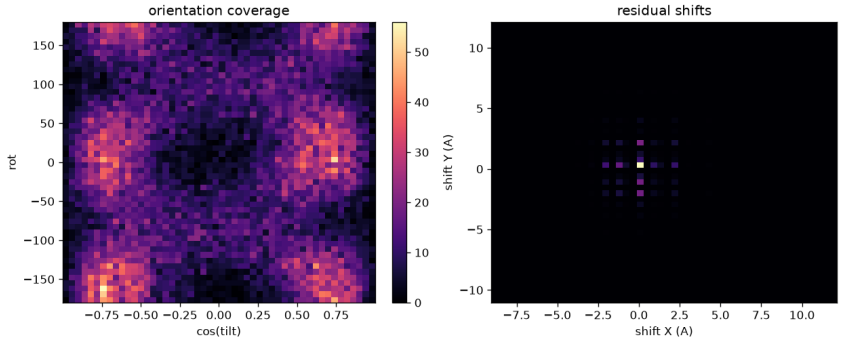


Figure 12: Euler-angle and shift distributions of the final refinement; no orientation collapse (supports Fig. 2).

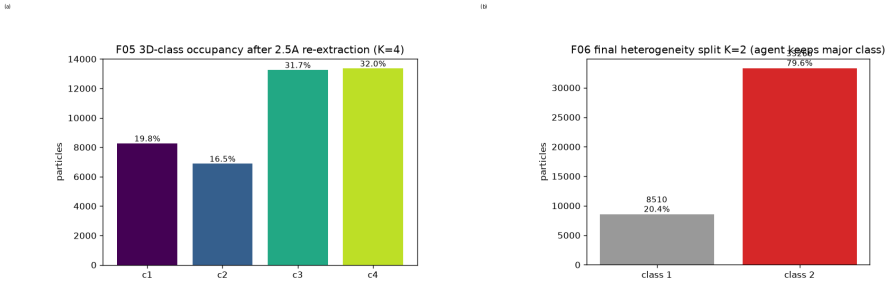


Figure 13: Class occupancy for K=4 and the final K=2 split with the selected class (supports Fig. 2c).

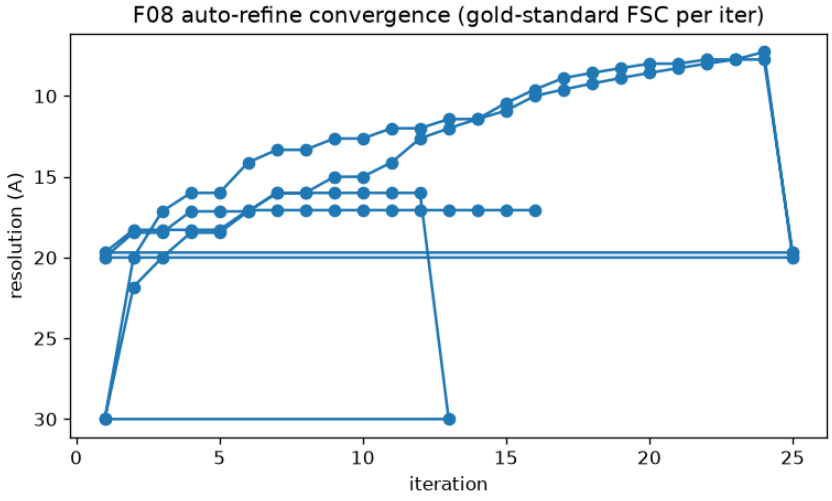


Figure 14: Resolution-versus-iteration convergence parsed from the refinement log (supports Fig. 2a).

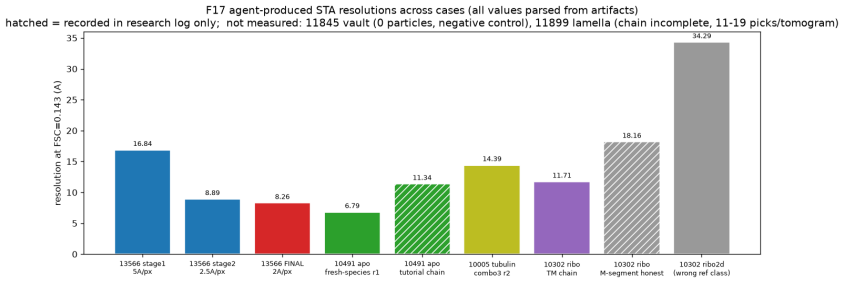


Figure 15: Final resolution per reproduced case, exclusions listed (supports Fig. 6).

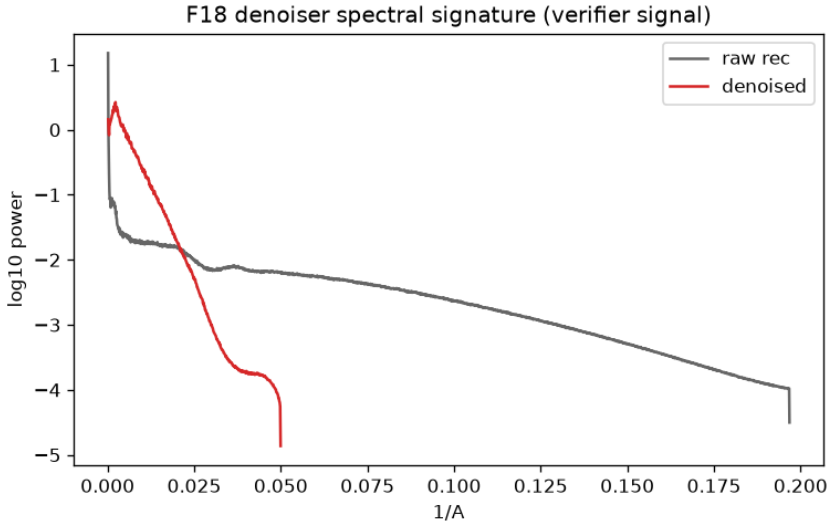


Figure 16: Radial power spectra, filtered vs denoised: high-frequency noise suppression (supports Figs. 6b and 7).

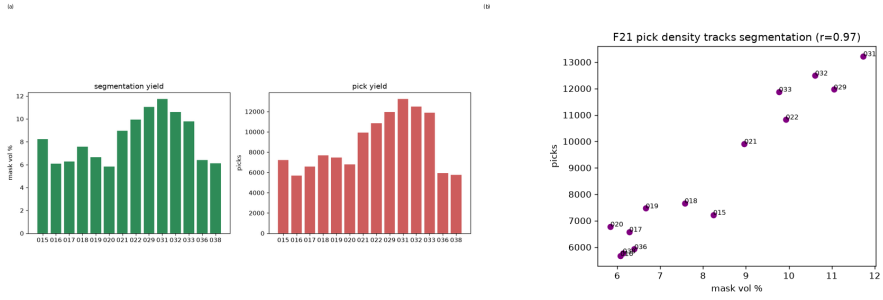


Figure 17: Per-tomogram mask fraction and pick count, and their cross-module correlation (supports Fig. 10).

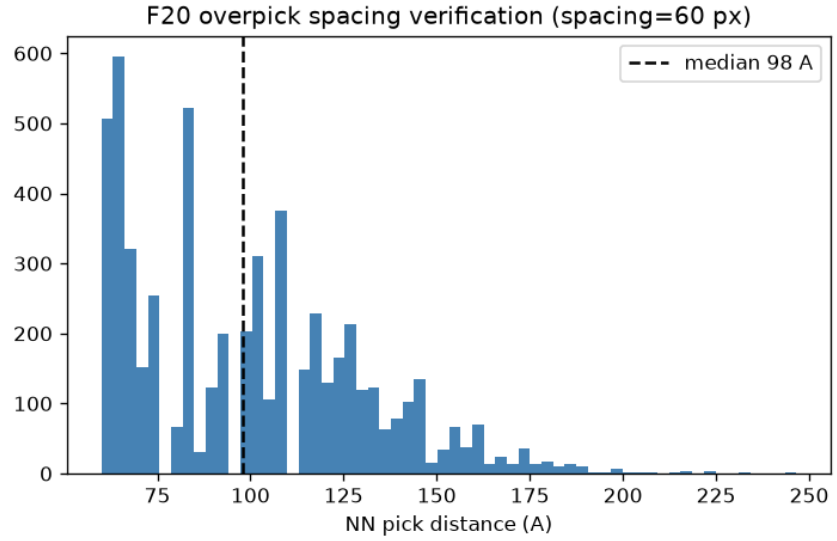


Figure 18: Nearest-neighbour pick spacing against the particle-diameter lower bound (supports Fig. 9).

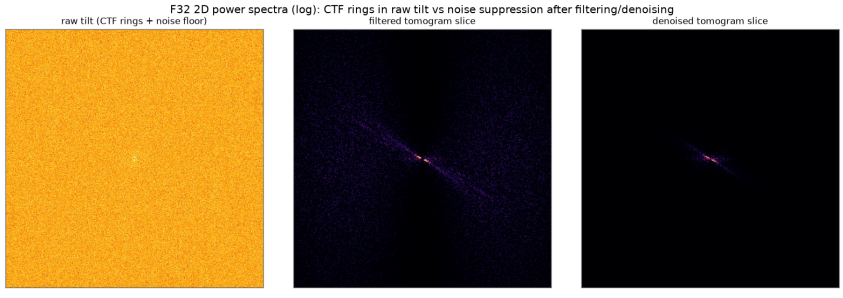


Figure 19: 2D power spectra (log): CTF rings in a raw tilt vs filtered and denoised slices (supports Figs. 4b and 7d).

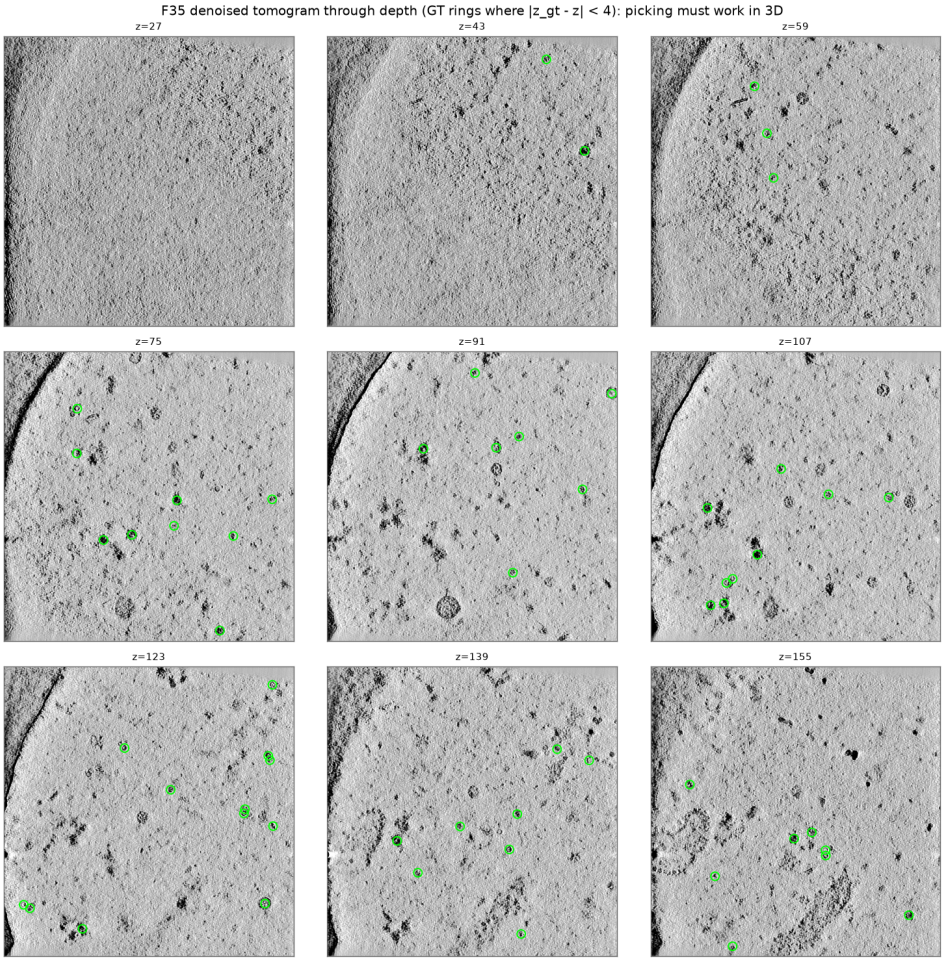


Figure 20: Denoised slices at nine depths with ground-truth rings: picking must be evaluated in 3D (supports Fig. 9a).

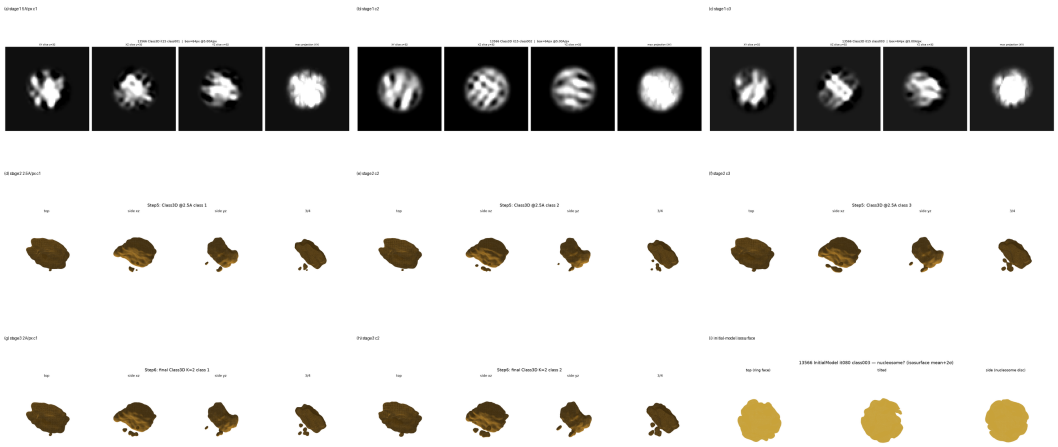


Figure 21: **Qualitative ablation of refinement depth (EMPIAR-13566).** Rows hold specimen, picker and mask fixed and vary only the number of refinement stages: (a–c) stage-1 (5 Å/px) 3D class averages as orthogonal slices and projections; (d–f) stage-2 (2.5 Å/px) class isosurfaces; (g–h) stage-3 (2 Å/px) final class isosurfaces; (i) the ab-initio initial-model isosurface for comparison. Density sharpens monotonically with stage count, the visual counterpart of the 16.84→8.89→8.26 Å trajectory (supports Figs. 2a and **??**). Sources: `Class3D_it015_class00*.png`, `vis_step5_class3d_2p5A_class00*.png`, `vis_step6_class3d_final_class00*.png`, `InitialModel_it080_class003_isosurface.png`.

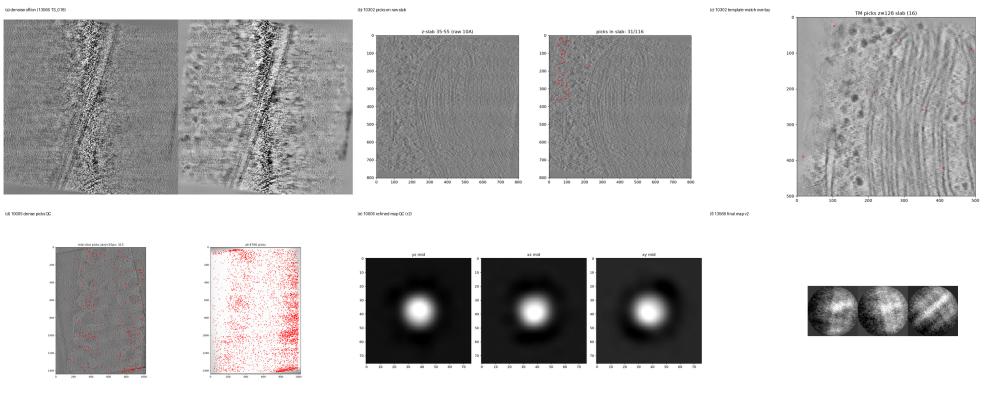


Figure 22: **Qualitative ablation of the preprocessing input and cross-case transfer.** (a) The same EMPIAR-13566 TS_018 tomogram slice with denoising disabled (left) and enabled (right): the contrast gain is what the paired sign test of RQ5 quantifies and fails to make significant. (b,c) EMPIAR-10302 raw slab with agent picks and the template-match overlay. (d) EMPIAR-10005 dense-pick QC. (e) EMPIAR-10005 refined map central slices (ring-like density). (f) EMPIAR-13566 final map. Together these show the transferred toolchain yields specimen-appropriate structure on four datasets without retuning (supports Figs. 6 and 7b). Sources: `ts018_denoise_cmp.png`, `pick_overlay.png`, `tm_overlay.png`, `qc_picks_dense.png`, `qc_map_r2.png`, `final_v2_map.png`.

I TOOL INTERFACES

Eighteen dispatchable actions share one JSON schema (action type, parameter record, tier tag); the dispatcher enforces the admissibility predicate φ from artifact provenance before launch and returns $(\text{OK}, \text{STATUS}, \text{TEXT}, \text{META})$. Third-party engines (WarpTools, RELION5, nnUNet, TomoTwin, MemBrain-seg, CryoSAM, easymode) retain their original licenses and versions pinned in the environment manifest.

Table 13: **The twelve-stage reconstruction process, as executed for EMPIAR-13566.** Stage, executing engine (TP third-party, SB self-built), pixel size, artifact appended to the store, and the verifier gate that makes the next stage admissible.

Stage	Engine	Å/px	Artifact	Gate
1 align+reconstruct	WarpTools (TP)	10.0	`*_10.00Apx.mrc`	$\geq$12 tomograms
2 denoise	IsoNet/easymode (TP)	10.0	`denoised/*.mrc`	$\geq$12, non-vacuous
3 segment chromatin	Ais/easymode (TP)	10.0	`segmented/*.mrc`	mask fraction > 0
4 pick	Ais/easymode (TP)	10.0	`coordinates/*`	picks > 0, spacing bound
5 STA export	WarpTools (TP)	5.0	`particles.star`	schema clean (SB)
6 initial model/class	RELION5 (TP)	5.0	`InitialModel/`, `Class3D/`	classes occupied
7 refine	RELION5 (TP)	5.0	`Refine3D/`, `run_class001.mrc`	half-maps exist
8 mask+postprocess	RELION5 (TP)	5.0	`postprocess.mrc`, `.star`	box match; FSC (SB)
9 re-export	WarpTools (TP)	2.5	`particles.star`	schema clean (SB)
10 class+refine	RELION5 (TP)	2.5	`run_model.star`	half-maps exist
11 select+re-export	RELION5+Warp (TP)	2.0	`particles.star` (846)	class occupancy
12 refine+postprocess	RELION5 (TP)	2.0	`postprocess.mrc`, `.star`	FSC recomputed (SB)

Provenance inventory (relocated from Sec. 3.3). Of the **18 dispatchable backends**, **13** wrap third-party engines—WarpTools, RELION5, easymode/IsoNet, Ais, MemBrain-seg, CryoSAM, TomoTwin, DeepETPicker, ProPicker, nnU-Net, and IMOD/AreTomo where alignment side-files are absent, with easymode and Ais each contributing more than one backend—and **5** are instruments we built because no adequate public tool fills the control or measurement role: a Laplacian-of-Gaussian blob baseline picker, an FFT template matcher with EMDB template retrieval, a mask-to-coordinate converter, an end-to-end segment-then-pick chain, and the STA/FSC probe behind `sta_average`. A sixth self-built instrument, the coordinate-frame offset probe, implements the `diagnose` action type rather than a tool name. "Third-party" here qualifies the engine, not the weights: `nnunet_particle_pick` runs nnU-Net (Isensee et al., 2021) weights we trained ourselves (datasets 501–503) and validated on a cross-domain held-out set (Table 6), so a backend can be somebody else's code and still be our measurement. Beyond tools, the framework contributes the dispatcher with pre-condition checks, a declarative tool registry through which new tools enter as JSON specifications rather than generated code, the sandbox executor, the four-tier verifier including independent FSC recomputation, the reward parser, and the multi-turn GRPO trainer.

J ARTIFACT INDEX

Every main-text number traces to: `postprocess.star` / `run_model.star` (_rlnFinalResolution, _rlnEstimatedResolution) and `postprocess_fsc.dat` (Figs. 2a, 11); verifier recomputations `sandbox_runs/paper_figs/fsc_recompute/13566_stage*_unmasked.json`, `datasets/10491_apoferritin/pytom/fsc_apo15b_r1.json`, `sandbox_runs/mt10005_m/combo3/fsc_r2.json` (Table 3); particle counts of Sec. 4.1

from datasets/13566_nucleosome/relion/nucleosome/particles.star (123,384 rows), nucleosome/nucleosomes.star (41,777), nucleosome_3/particles.star (33,266), with engine values in nucleosome/PostProcess/job005/postprocess.star (16.842105), nucleosome_2/Refine3D/job002/run_model.star (8.888889) and nucleosome_3/PostProcess/job002/postprocess.star (8.258065); Class3D/run_it*_class00*.mrc (Figs. 2c, 13); segmentation/pick/mask artifacts of EMPIAR-11845 and MaskCreate jobs (Fig. 5); per-dataset denoised/filtered MRCs (Figs. 6, 7b–d); step2_denoise_paired/summary_loose.json (Table 8); reward_compare_r5_n12.json, grpo_r5.log, phaseB3_chain_checkpoint-*_hist.json (Fig. 8, Tables 9 and 10); indep_pick_czii/eval_report.json and anno_gt/*.ndjson (Fig. 9, Table 5); STAR coordinate counts and stage logs (Fig. 10); RELION/initial-model rendering PNGs under sandbox_runs/easymode_tuts/nucleosome13566/final/ (Class3D_it015_*.png, vis_step5_*.png, vis_step6_*.png, InitialModel_it080_class003_isosurface.png, paperfigs/r13566_015_a_pipeline4panel.png) and the cross-case QC renderings nucleosome/ts018_denoise_cmp.png, workshop10302/pick_overlay.png, workshop10302/tm_overlay.png, mt10005_chain/0106/pytom/qc_picks_dense.png, mt10005_m/0106/qc_map_r2.png, nucleosome/final_v2_map.png (Figs. **??**, 21, 22). selection-control artifacts sandbox_runs/phaseC/purity_confound.json and sandbox_runs/phaseC/membrane_context_10301.json (Sec. 4.7); closed-loop traces sandbox_runs/autoresearch_trace.json and sandbox_runs/graph_e2e_step90_vf.json with their run log autoresearch_run.log (Sec. 4.3, Table 4); picker acceptance and ensemble artifacts phantom_pick_test.json, ensemble_pick_compare.log, ensemble_pick_compare_thyro.log, ensemble_pick_compare.json (Tables 6 and 7); the 17-case sweep sandbox_runs/multicase17/ (ablation (ix)); availability and ordering instruments sandbox_runs/availability_probe/, scripts/record_teacher_trajectories.py::tool_availability, scripts/check_trajectory_order.py, and the 20-state routing benchmark scripts/eval_routing.py with its three prompt conditions (ablation (vii)); membrane-control script scripts/phaseC_membrane_context.py, whose one-to-one matching routine produces the non-additive stratum counts disclosed in Sec. 4.7; figure-generation scripts: scripts/paper_figs_30.py, scripts/paper_figs_img2.py, scripts/make_paper_composites.py, scripts/make_paper_qual.py.

Artifacts we cannot ship. Four cited file names are absent from the current tree and we list them rather than let the index imply otherwise: the training-era reward comparisons reward_compare_r5_n12.json and reward_compare_9b_task_n12.json, the run log grpo_r5.log, and the per-checkpoint action histograms phaseB3_chain_checkpoint-*_hist.json; they were produced on the training cluster and the derived objects—Fig. 8, Tables 9, 10, and the ablation (ix) numbers—are shipped in their place, with the research-log entries that recorded them at the time. The cross-domain ribosome ensemble comparison (last row of Table 7) likewise survives only as a log entry, its JSON overwritten by the later thyroglobulin run. Numbers whose artifacts are present are the ones we ask a reader to verify first; these four are the ones we would have to re-run.

Citation provenance. No bibliography entry in this paper was written from memory. Each was transcribed from a Crossref, Europe PMC or DOI-registry response by scripts/xref_verify.py, xref_verify2.py, xref_verify3.py, xref_verify4.py and xref_doi.py, which are shipped so the audit is repeatable; the header of references.bib records the same. Fuzzy bibliographic queries returned plausible but wrong records for several items (a preprint for a journal article, an unrelated book chapter for a method paper), so verification switched to exact-title queries and direct DOI lookups, and every field that survived is registry-backed. Two third-party tools used in the pipeline (Cryo-SAM, IsoNet) have no peer-reviewed record we could locate as of the submission date; they are cited as explicitly labelled software entries stating that fact rather than as articles with invented metadata. The same discipline applies to the two orchestration references we build on: the EMSeek platform

and the Polar/ProRL agent server are cited from the bundled repositories, with the note that a public citation should replace them if one exists.

K LIMITATIONS AND PROSPECTIVE PROTOCOL

Held-out RL superiority is not established (Table 10); the denoising benefit is directional only (Table 8); final-stage resolution trails the expert tutorial by 0.86 Å and that gap is *unattributed*—particle number and protocol deviation were both excluded by measurement (123,384 exported subtomograms, 41,777 class-selected, 33,266 refined, against the tutorial's 120,420 and ∼35,000; stage-2/3 commands run verbatim), leaving visual-versus-programmatic class selection and the absence of replication as candidates we cannot separate from one run. Eight further limits, (a)–(h), are specific to what this revision reports rather than to the framework. (a) The two closed-loop traces of Sec. 4.3 are short—three and four iterations—and stop before subtomogram averaging, so tier 3 is unavailable in both and no resolution claim rests on them; the loop's long-horizon behaviour is evidenced by the reproduction and recovery experiments instead. (b) In Run B the cache key omitted the parameter being tuned, so that run did not in fact search over `top_n`; we report the defect and the artifact path that exposes it rather than reinterpreting the four iterations as a sweep. (c) No autonomously authored instrument has yet passed the four-stage gate, so the escape hatch is demonstrated as *firing and refusing*, not as producing; the five instruments in the execution path were written by us and admitted through the same gate. (d) The self-trained picker fails on two of six species cross-domain (F_1 0.28 and 0.07, Table 6), and five-fold ensembling does not repair them. (e) One magnitude claim was withdrawn during the work: the 25% ordering-violation rate that appeared to justify episode-level availability was created by the measuring instrument (ablation (vii)), which is why the mechanism claim and the magnitude claim are stated separately. (f) The reward actually deployed, $\lambda = (0.10, 0.10, 0.30, 0.50)$ with renormalization over present tiers, does not satisfy the δ-separation hypothesis of Theorem 1 for any $\delta \leq 1$ (Remark 1); the theorem therefore characterizes an alternative regime, and the implementation's non-compensability comes from a parse-failure hard zero and the 0.20 unmeasurable cap rather than from order-isomorphism; the product gates g_k of Eq. (6) are not instantiated in the shipped code at all, which gates on tier availability instead (Appendix A). A certified vector exists $(0.70, 0.20, 0.07, 0.03$ at $\delta \approx 0.51)$ but puts almost no mass on resolution, and we have not measured which of the two produces better science. (g) The verifier's tier-3 estimator is unmasked, is not phase-randomized, and takes the first crossing of a curve that is non-monotone in practice; all resolution numbers in this paper are therefore directional under a declared convention, no directional-FSC or anisotropy analysis is performed, and a single scalar is quoted for a reconstruction whose Fourier completeness is 0.947–0.972. The transfer claim of Sec. 4.4 rests on eleven chains over five datasets beyond the primary case, of which three yield no measurement at all. (h) Neither acceptance criterion in Appendix G.5 is a pre-registration in the timestamped sense: the log announces each test before the run but records the numeric threshold in the entry that reports the outcome, and no script encodes either threshold as a pass/fail gate. We report both as *declared* criteria and flag the weaker status wherever they are cited. The next training comparison is therefore fixed here, in this paragraph, before the run exists—same 12 prompts, three seeds, base vs SFT vs GRPO, tiered reward v3 with resolution-tier parsing verified, and the weight vector entering as a declared arm (deployed λ versus a δ-separated λ) rather than as a tuned constant; the success criterion is a non-overlapping seed interval on tier-2 precision, not on total reward. That sentence is the record; we make no stronger claim for it.