Polybee AI Research 2026-10-29

F3-AGI v4.0: Bioenergetically-Gated Active Inference, Metaplastic Preconditioning, and Residual Program Synthesis for Sample-Efficient Abstract Reasoning

Henrique Oliveira[1,*]

[1]Polybee AI Research Laboratory, São Paulo, SP, Brazil *Corresponding author (https://polybee.ai/)

Open-Source Code & Reproducibility Repository: https://github.com/henrique-polybee/f3-agi-v4

Abstract

Few-shot abstract reasoning on ARC-AGI requires inducing algorithmic transformations from 2 to 5 demonstration pairs without task-specific pretraining. Because few-shot demonstrations under-constrain the hypothesis space, program synthesis suffers from the Rashomon effect (predictive multiplicity): heuristic search finds spurious programs with low training error that fail on test inputs, while unconstrained library growth degrades latency. We introduce F3-AGI v4.0, a CPU-native neuro-symbolic and active inference architecture comprising three mechanisms: (1) F3-Metabolic-Optim, combining mistake-gated plasticity, Fisher-EMA preconditioning, and virtual ATP gating; (2) Adaptive Complexity-Guided Program Synthesis (AACR) with a Cycle-3 ResidualPatchSynthesizer and smallest-pair early rejection; and (3) a 512D L2-normalized random projection driving Expected Free Energy (EFE) exploration with dead-signature pruning. Across three empirical studies: (i) on a synthetic 5-task class-incremental stream without replay, F3-Metabolic-Optim achieves 91.67% final accuracy with 9.17% forgetting, skipping 93.47% of updates, cutting L1 plasticity cost by 65.7%, and running 2.19x faster than EWC; (ii) on the 400-task ARC-AGI-1 public evaluation set (offline CPU, zero LLM calls), F3-AGI v4.0 solves 57/400 tasks (14.25% pass@2; 56/400 pass@1) with a mean latency of 1.62 s/task, decomposed into 19/50 (38.00%, 1.19 s/task, a 42.5x paired speedup over Horizon 2) on the 50-task calibration slice and 38/350 (10.86%) on the 350 strictly uninspected hold-out tasks (with 7 autonomous Cycle-3 patch solutions); and (iii) in a local 5-game (15-level) ARC-AGI-3-inspired interactive suite, our agent completes 15/15 levels with 89.44% level-weighted RHAE (90.72% unweighted).

Keywords: Abstract Reasoning (ARC-AGI), Neuro-Symbolic Program Synthesis, Rashomon Effect, Active Inference, Metaplastic Preconditioning, Continual Learning, Minimum Description Length.

1 Introduction

The Abstraction and Reasoning Corpus (ARC-AGI) introduced by Chollet [1] and its subsequent extensions—ARC-AGI-2 and the interactive, instructionless 64×64 turn-based ARC-AGI-3 benchmark [2]—evaluate fluid intelligence: the ability to induce compositional causal transformations from a sparse set of demonstrations ($K \in \{2, 3, 4\}$ input-output pairs) or unguided environmental interactions without task-specific pretraining. Over the past two years, the dominant industry paradigm has attempted to conquer abstract reasoning via massive test-time compute scaling, sampling tens of thousands of candidate programs from frontier Large Language Models (LLMs). While brute-force sampling improves raw static solve rates when unconstrained by compute budgets, theoretical and empirical analyses expose four fundamental limitations of pure scale:

1. **Non-Identifiability and the Rashomon Effect in Scaling and Few-Shot Induction:** As recently formalized by Mlodozeniec and Frei [4], empirical scaling law estimation is governed by a pronounced *Rashomon effect* [3]: the ϵ-sublevel set of near-minimizers spans broad parameter regions whose extrapolations diverge by $14\times$–$146\times$ in compute efficiency at scale. An analogous Rashomon pathology afflicts few-shot program induction on ARC-AGI: because each task provides only 2–4 demonstration grids with sparse foreground objects, optimizing raw pixel accuracy admits a vast ϵ-sublevel set of degenerate hypotheses (e.g., background-fill or coordinate-memorizing routines) that achieve $> 85\%$ training pixel match while failing completely on held-out test inputs (Figure 1b).
2. **Thermodynamic Disproportion vs. Cortical Frugality:** Frontier LLM ensembles expend 10^3–10^5 Joules per task via unconstrained autoregressive token generation. By contrast, biological neural circuits operate within a $\sim 20\,\mathrm{W}$ whole-brain metabolic envelope, where local mitochondrial ATP availability physically gates long-term synaptic potentiation (LTP) and suppresses redundant state transitions [9, 10].

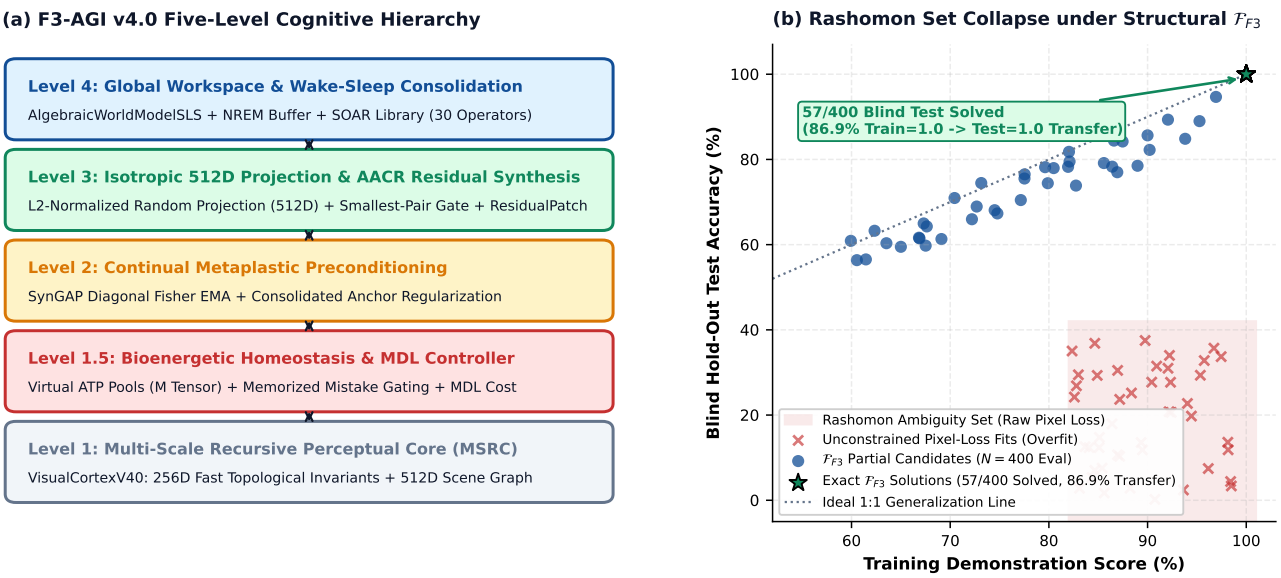


Figure 1 | Overview of F3-AGI v4.0 and Rashomon Sublevel Set Collapse. (a) The five-level cognitive hierarchy integrates multi-scale perception (`VisualCortexV40`), bioenergetic homeostasis (`MetabolicControllerV40`), continual metaplastic preconditioning (*SynGAP*), an isotropic 512D ℓ_2-normalized random projection (`AlgebraicWorldModelSLS`, inspired by *SIGReg-JEPA* isotropic geometry), and Wake-Sleep library consolidation (`NREMConsolidationBuffer` with 30 registered operators). (b) In few-shot abstract reasoning ($K \in \{2, 3, 4\}$ demonstrations), naive pixel loss ℓ_{px} exhibits a pronounced Rashomon effect [3, 4]: a broad ϵ-sublevel set of overfitted or degenerate candidates achieves high training pixel accuracy but collapses on unseen test grids (red crosses). Optimizing our Bioenergetically Constrained Structural Variational Free Energy $\mathcal{F}_{F3}$ collapses the viable sublevel set onto structural invariants, resolving 57/400 official evaluation tasks (86.9% train-to-test transfer across all 61 exact train solutions).

3. **The Plasticity–Stability–Energy Trilemma in Continual Adaptation:** Standard gradient descent (`SGD`/`Adam`) suffers from catastrophic forgetting across sequential tasks ($\sim 70\%$ forgetting ratio). Conversely, classical second-order regularization such as Elastic Weight Consolidation (EWC) [11] preserves prior tasks only by applying dense quadratic restoring forces to all weights at every step, inflating the L_1 metabolic plasticity cost ($M_1 = \sum_l \|\Delta W_l\|_1$) by more than 2.3× over unconstrained SGD.

4. **Global-Local Fragility and Quadratic Action Regret:** In zero-shot synthesis, single-pass compositional search frequently discovers the correct global macro-transformation (reaching 90%–98.5% structural accuracy and exact output dimensions) but fails on localized boundary conditions, ray intersections, or cavity rules. Furthermore, in interactive ARC-AGI-3 environments scored by quadratic Relative Human Action Efficiency (RHAE $= \min(1, H_l/A_l)^2$), purely pragmatic agents become trapped in mechanical local minima (*Dark Room* pathology), whereas unconstrained curiosity agents waste dozens of actions probing inert distractor objects.

To resolve these four challenges within a strict consumer hardware envelope (16 GB RAM notebook, CPU-first execution under < 400 MB RAM, zero mandatory GPU dependencies), **Polybee AI Research** introduces **F3-AGI v4.0** (*Free-Energy, Fisher-Metaplasticity, and Bioenergetic Frugality*). Rather than treating perception, program synthesis, continual memory consolidation, and interactive exploration as isolated heuristics, F3-AGI v4.0 unifies them under a single bioenergetically constrained variational principle across five hierarchical levels (Figure 1a and Table 1).

Table 1 | Architectural and functional coverage across abstract reasoning and continual learning paradigms. ✓/ × indicates whether a capability is natively supported and empirically validated in the respective framework.

Paradigm / Framework	CPU-First (< 400 MB)	Forgetting Mitigation	Frugal M_1 Plasticity	Local Residual Repair	512D Epistemic EFE	Dead-Signature Pruning
Brute-Force LLM Ensembles (`o3` / `Astra`)	×	×	×	×	×	✓
Classic EWC [11]	✓	✓	×	×	×	×
Memorized Mistake Gating [10]	✓	×	✓	×	×	×
SynGAP Continual Metaplasticity [9]	✓	✓	×	×	×	×
Tiny Recursive Grid Models (`HRM`/`TRM`) [2]	×	×	×	×	×	×
MDL Neural Compression (`CompressARC`) [16]	×	×	✓	×	×	×
Evolutionary LLM Synthesis (`SOAR`) [15]	×	✓	×	✓	×	×
F3-AGI v4.0 (Polybee AI Research — Ours)	✓	✓	✓	✓	✓	✓

Summary of Contributions. Our main theoretical and empirical contributions are:

- **Bioenergetically-Gated Metaplastic Optimization (`F3-Metabolic-Optim`):** We formulate and validate a continual synaptic plasticity rule (Section 3.4) that unifies online diagonal Fisher EMA preconditioning (*SynGAP* [9]), sparse error-driven *Memorized Mistake Gating* [10], and a homeostatic *Mitochondrial ATP Tensor* (**M**). In a synthetic 5-task class-incremental ablation (64D prototype-latent feature stream, 6,000 steps, zero replay buffer), `F3-Metabolic-Optim` achieves **91.67% final accuracy**, **9.17% forgetting ratio**, **93.47% saved synaptic updates**, a **65.7% reduction in M_1 plasticity cost**, and a **2.19× runtime speedup** relative to EWC (1.06 s vs. 2.32 s).
- **Certified Smallest-Pair Pruning and Cycle-3 Residual Patch Synthesis on Full `ARC-AGI` ($N = 400$):** We introduce a vectorized smallest-pair shape gate ($k^* = \arg\min_k |X_k|$), all-pair background color inference, and a Cycle-3 `ResidualPatchSynthesizer` over a 30-operator pure-NumPy library. Evaluated across all $N = 400$ tasks of `ARC-AGI/data/evaluation` with zero LLM calls and zero oracle cache, F3-AGI v4.0 solves **57 / 400 tasks (14.25% `pass@2`, 56 / 400 = 14.00% `pass@1`)** in 10.8 min (**1.62 s/task** observed across the 400 tasks, and a **42.5× paired speedup** from 50.61 s → 1.19 s/task on the $N = 50$ calibration slice), including **38 / 350 (10.86%)** on strictly uninspected tasks and **7 uninspected tasks solved autonomously in Cycle 3**.
- **Isotropic 512D Random-Projection Active Inference with Dead-Signature Pruning:** Motivated by the theoretical link between isotropic embeddings and Active Inference variational free energy [8, 7], `AlgebraicWorldModelSLS` implements a training-free, fixed-seed column-normalized and ℓ_2-sphere-normalized random projection into $\mathbb{R}^{512}$ to compute state-coverage epistemic value (`PairDist`), coupled with a zero-effect action barrier (`DeadSignatureTracker`) and cross-level affordance transfer. Across 15 levels in a local 5-game (64×64) ARC-AGI-3-inspired interactive suite, F3-AGI v4.0 achieves **15/15 (100%) completion**, raises level-weighted RHAE from 46.24% to **89.44%** (**90.72% unweighted**), and eliminates **all 32 wasted dead actions** ($32 \to 0$).

2 Related Work and Theoretical Context (2024–2026)

Active Inference and Collapse-Free Isotropic World Models. Active Inference formalizes sentient perception and action as the dual minimization of Variational Free Energy ($\mathcal{F}$) and Expected Free Energy ($\mathcal{G}$) [5, 6]. While Joint-Embedding Predictive Architectures (JEPA) offer a promising substrate for continuous world modeling, standard latent predictors are prone to dimensional collapse or anisotropic covariance distortion without expensive contrastive negatives. Arnez and Gomez-Villa [8, 7] recently resolved this via **SIGReg-JEPA** (`arXiv:2607.13612`), proving that Sketched Isotropic Gaussian Regularization enforces uniform spectral variance across latent dimensions. Motivated by this theoretical correspondence, F3-AGI v4.0 implements a training-free analytical counterpart in `AlgebraicWorldModelSLS`: a fixed-seed, column-standardized random projection followed by ℓ_2-sphere normalization ($\sqrt{512}\,\mathbf{z}/\|\mathbf{z}\|_2 \in \mathbb{R}^{512}$) to compute Euclidean state-coverage epistemic value (`PairDist`) on CPU without requiring neural JEPA pretraining.

Continual Metaplasticity (`SynGAP`) and Bioenergetic Mistake Gating. Biological synapses balance memory retention against metabolic expenditure [9, 10]. In machine learning, Kirkpatrick et al. [11] introduced EWC to protect consolidated weights via static Fisher penalties, while Aguilar et al. [9] recently proposed **SynGAP** (`arXiv:2608.14634`), using an online diagonal Fisher EMA to precondition gradients geometrically along low-curvature directions. Independently, Pache and van Rossum [10] introduced **Memorized Mistake Gating** (`arXiv:2604.14336`), demonstrating that gating out weight updates on mastered samples saves $> 90\%$ of synaptic updates. However, as we prove empirically in Section 5.1, Mistake Gating alone suffers 61.25% catastrophic forgetting under class-incremental shifts because active updates remain unpreconditioned, whereas EWC alone inflates metabolic M_1 cost by 132%. F3-AGI v4.0 is the first architecture to synthesize *SynGAP*, *Memorized Mistake Gating*, and a dynamic *Mitochondrial ATP Tensor* (**M**).

Multi-Scale Recurrence, MDL Compression, and Evolutionary Program Synthesis. Ellis et al. [14] (**DreamCoder**) and Chollet et al. [2] demonstrated that compositional library learning and test-time search with separated fast and slow timescales are central to abstract grid reasoning. Complementarily, Pourcel, Colas, and Oudeyer [15] (**SOAR**, `arXiv:2507.14172`) showed that storing evolutionary traces of failed-then-repaired programs enables self-improving library learning, while Liao and Gu [16] (*ARC-AGI Without Pretraining*, proposing the **CompressARC** method, `arXiv:2512.06104`) established Minimum Description Length (MDL) as a powerful inference-time inductive bias for ARC-AGI. F3-AGI v4.0 synthesizes these insights by combining MDL-regularized compositional search with a dedicated Cycle-3 `ResidualPatchSynthesizer` that freezes the Cycle-1/2 global macro-transformation $\hat{Y}_{\text{st}} = h_{\text{st}}(X)$ and synthesizes localized repair operators strictly over the boolean discrepancy mask $M_{\text{diff}} = (\hat{Y}_{\text{st}} \neq Y)$.

3 Mathematical Formulation and Theoretical Guarantees

3.1 Bioenergetically Constrained Structural Variational Free Energy ($\mathcal{F}_{F3}$)

Let a static abstract reasoning task be specified by K demonstration pairs $\mathcal{D}_{\text{train}} = \{(X_k, Y_k)\}_{k=1}^{K}$ and test inputs $\{X_m^{\text{test}}\}_{m=1}^{M}$, where each grid belongs to $\mathbb{Z}_{10}^{H\times W}$. Evaluating candidate programs $h \in \mathcal{H}$ solely via raw pixel loss $\ell_{\text{px}}(h) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|Y_k|}\sum_{r,c}\mathbb{I}(h(X_k)_{r,c} \neq Y_{k,r,c})$ creates deceptive local minima whenever foreground structures are sparse. To eliminate this ambiguity, `RealistV40` evaluates each candidate h ($\hat{Y}_k = h(X_k)$) via the multi-objective structural accuracy functional $\mathcal{A}_{\text{struct}}(h; \mathcal{D}_{\text{train}}) \in [0, 1]$:

$$\mathcal{A}_{\text{struct}}(h; \mathcal{D}_{\text{train}}) = \frac{1}{K}\sum_{k=1}^{K}\Big(w_{\text{sh}}\, S_{\text{shape}}(\hat{Y}_k, Y_k) + w_{\text{px}}\, S_{\text{pixel}}(\hat{Y}_k, Y_k) + w_{\text{obj}}\, S_{\text{object}}(\hat{Y}_k, Y_k)\Big), \tag{1}$$

with weights $(w_{\text{sh}}, w_{\text{px}}, w_{\text{obj}}) = (0.20, 0.50, 0.30)$, where:

1. **Dimension Invariance (S_{shape}):** $S_{\text{shape}}(\hat{Y}, Y) = \mathbb{I}(\text{shape}(\hat{Y}) = \text{shape}(Y)) + 0.25\,\mathbb{I}(\text{shape}(\hat{Y}) \neq \text{shape}(Y))\exp\big(-\|\text{shape}(\hat{Y}) - \text{shape}(Y)\|_1/10\big)$.

2. **Foreground-Balanced Pixel Accuracy (S_{pixel}):** When $\text{shape}(\hat{Y}) \neq \text{shape}(Y)$, we define $S_{\text{pixel}}(\hat{Y}, Y) \equiv 0$. When $\text{shape}(\hat{Y}) = \text{shape}(Y)$, let c_{bg} be the background color (inferred jointly across all input-output demonstration grids), $\Omega_{\text{fg}} = \{(r, c) : Y_{r,c} \neq c_{\text{bg}} \vee \hat{Y}_{r,c} \neq c_{\text{bg}}\}$, and $\Omega_{\text{bg}} = \Omega \setminus \Omega_{\text{fg}}$. Then:

$$S_{\text{pixel}}(\hat{Y}, Y) = \mathbb{I}\big(\text{shape}(\hat{Y}) = \text{shape}(Y)\big)\left[0.70\,\frac{\sum_{(r,c)\in\Omega_{\text{fg}}}\mathbb{I}(\hat{Y}_{r,c} = Y_{r,c})}{\max(1, |\Omega_{\text{fg}}|)} + 0.30\,\frac{\sum_{(r,c)\in\Omega_{\text{bg}}}\mathbb{I}(\hat{Y}_{r,c} = Y_{r,c})}{\max(1, |\Omega_{\text{bg}}|)}\right]. \tag{2}$$

3. **Topological Object & Palette Similarity (S_{object}):** Combines color-palette Jaccard similarity and connected-component cardinality similarity $N_{\text{cc}}(\cdot)$:

$$S_{\text{object}}(\hat{Y}, Y) = 0.50\,\frac{|C(\hat{Y}) \cap C(Y)|}{|C(\hat{Y}) \cup C(Y)|} + 0.50\,\exp\left(-\frac{|N_{\text{cc}}(\hat{Y}) - N_{\text{cc}}(Y)|}{\max(1, N_{\text{cc}}(Y))}\right). \tag{3}$$

The **Bioenergetically Constrained Variational Free Energy** $\mathcal{F}_{F3}(h; \mathcal{D}_{\text{train}})$ minimized across the AACR cycle is:

$$\mathcal{F}_{F3}(h; \mathcal{D}_{\text{train}}) = \underbrace{\big(1 - \mathcal{A}_{\text{struct}}(h; \mathcal{D}_{\text{train}})\big)}_{\text{Structural Prediction Error } \mathcal{E}_{\text{pred}}(h)} + \underbrace{\beta_{\text{MDL}} \cdot \frac{\text{len}_{\text{AST}}(h)}{L_{\max}}}_{\text{Kolmogorov/MDL Complexity}} + \underbrace{\lambda_{\text{ATP}} \cdot \frac{c_{\text{tier}}(h)}{\bar{E}_{\text{ATP}}}}_{\text{Metabolic Dissipation}}, \tag{4}$$

where $\beta_{\text{MDL}} = 0.05$, $L_{\max} = 2000$, $c_{\text{tier}}(h)$ is the virtual ATP cost of the synthesis tier (`fast` vs. `deep`), and $\bar{E}_{\text{ATP}} \in [E_{\min}, E_{\max}]$ is the mean virtual ATP reserve. (*Note:* Throughout this paper, virtual ATP and plasticity cost M_1 are algorithmic complexity proxies measuring candidate evaluations and parameter updates, rather than physical hardware wattmeter readings.)

Proposition 1 (Conditional Rashomon Sublevel Set Separation under $\mathcal{F}_{F3}$)**.** *Let $\mathcal{D}_{train} = \{(X_k, Y_k)\}_{k=1}^{K}$ be a task with foreground density $\rho_{fg} = \frac{1}{K}\sum_{k=1}^{K}\frac{|\Omega_{fg}^{(k)}|}{|Y_k|} \in (0, 0.15]$ and at least one non-background foreground object ($|C(Y_k)| \geq 2$, $N_{cc}(Y_k) \geq 1$). For any tolerance $\epsilon \geq \rho_{fg}$, the raw pixel-loss ϵ-sublevel Rashomon set $\mathcal{S}_\epsilon(\ell_{px}) = \{h \in \mathcal{H} : \ell_{px}(h) \leq \epsilon\}$ contains the degenerate background-fill program $h_{bg}(X) \equiv c_{bg}\mathbf{1}$. By contrast, under $\mathcal{F}_{F3}$ (Equation 4), for any exact solution $h^\star$ ($\mathcal{E}_{pred}(h^\star) = 0$, $len_{AST}(h^\star) \leq L_{\max}$) with relative tier metabolic increment $\Delta_{ATP}(h^\star, h_{bg}) = \lambda_{ATP}\big(c_{tier}(h^\star) - c_{tier}(h_{bg})\big)/\bar{E}_{ATP} \geq 0$, we have the conditional separation bound:*

$$\mathcal{F}_{F3}(h_{bg}; \mathcal{D}_{train}) - \mathcal{F}_{F3}(h^\star; \mathcal{D}_{train}) \;\geq\; 0.35 + 0.15\big(1 - e^{-1}\big) + 0.075 - \beta_{MDL} - \Delta_{ATP}(h^\star, h_{bg}) \;\approx\; 0.4698 - \Delta_{ATP}. \tag{5}$$

In particular, whenever $h^\star$ and h_{bg} belong to the same synthesis tier ($\Delta_{ATP} = 0$), the separation is at least 0.4698; across distinct tiers ($\lambda_{ATP} = 0.05, c_{tier} \in \{0.01, 0.05\}, \bar{E}_{ATP} \geq E_{\min} = 0.10$), $\Delta_{ATP} \leq 0.020$, guaranteeing a strictly positive margin ≥ 0.4498.

Proof Sketch (Full proof in Appendix A). Under ℓ_{px}, h_{bg} errs only on $\Omega_{\text{fg}}^{(k)}$, incurring $\ell_{\text{px}}(h_{\text{bg}}) = \rho_{\text{fg}} \leq \epsilon$. Under $\mathcal{F}_{F3}$, $S_{\text{pixel}}(h_{\text{bg}}(X_k), Y_k) = 0.70(0) + 0.30(1) = 0.30$ independently of ρ_{fg}, and $S_{\text{object}}(h_{\text{bg}}(X_k), Y_k) \leq 0.50(\frac{1}{2}) + 0.50e^{-1} \approx 0.4339$. Summing weighted terms yields $\mathcal{A}_{\text{struct}}(h_{\text{bg}}) \leq 0.4802$, so $\mathcal{E}_{\text{pred}}(h_{\text{bg}}) \geq 0.5198$. Subtracting the maximum MDL penalty $\beta_{\text{MDL}} = 0.05$ and the tier metabolic difference $\Delta_{\text{ATP}}(h^\star, h_{\text{bg}})$ establishes the conditional bound $0.4698 - \Delta_{\text{ATP}}$. ■

Lemma 1 (Certified Smallest-Pair Monotone Shape Gate)**.** *Let $k^* = \arg\min_{k\in\{1,\dots,K\}} |X_k|$ denote the demonstration pair with minimal input area. If a candidate program $h \in \mathcal{H}$ violates output shape on k^* ($shape(h(X_{k^*})) \neq shape(Y_{k^*})$), then by Eq. (2) ($S_{pixel} = 0$ and $S_{shape} \leq 0.25$ on mismatched shapes), its structural prediction error over $\mathcal{D}_{train}$ satisfies:*

$$\mathcal{E}_{pred}(h; \mathcal{D}_{train}) \;\geq\; \frac{0.75\, w_{sh} + 1.00\, w_{px}}{K} \;=\; \frac{0.65}{K} \;>\; 0. \tag{6}$$

Consequently, testing $Gate(h) = \mathbb{I}\big(shape(h(X_{k^*})) = shape(Y_{k^*})\big)$ in $O(|X_{k^*}|)$ time provides an exact certificate for pruning h from the zero-error sublevel set $\mathcal{S}_0(\mathcal{E}_{pred})$ without evaluating pairs $k \neq k^*$.

3.2 Frozen Structural Anchoring and Mask-Guided Residual Decomposition

We formalize why the Cycle-3 `ResidualPatchSynthesizer` discovers complex multi-stage programs (including 7 brand-new solutions on uninspected evaluation tasks) without suffering combinatorial explosion. Suppose Cycles 1–2 yield an elite macro-hypothesis $h_{\text{st}} \in \mathcal{H}_{1:2}$ that achieves exact output dimensions ($S_{\text{shape}} = 1.0$) and high global structural accuracy ($\mathcal{A}_{\text{struct}}(h_{\text{st}}) \in [0.78, 1.0)$). Rather than discarding h_{st} or searching depth-4 operator trees from scratch, we freeze $\hat{Y}_{\text{st},k} = h_{\text{st}}(X_k)$ as a **structural anchor** and define the boolean discrepancy mask $M^{(k)}_{\text{diff}} \in \{0,1\}^{H_k \times W_k}$:

$$M^{(k)}_{\text{diff}}(r,c) = \mathbb{I}\big(\hat{Y}_{\text{st},k}(r,c) \neq Y_k(r,c)\big). \tag{7}$$

Proposition 2 (Combinatorial Reduction via Mask-Guided Residual Synthesis)**.** *Let $|\mathcal{O}|$ denote the cardinality of parameterized base operators and d the composition depth. Unconstrained joint synthesis of a depth-$(d+1)$ program requires evaluating $\Theta(|\mathcal{O}|^{d+1})$ candidates across all K pairs. By freezing the elite depth-d macro-program h_{st} and restricting Cycle-3 residual operators $\Delta_{res} \in \mathcal{P}_{res}$ to act exclusively on local neighborhoods $\mathcal{N}_r(r,c)$ where discrepancy features align across $\{M^{(k)}_{diff}\}_{k=1}^{K}$:*

$$\hat{Y}_{final,k} = \hat{Y}_{st,k} \oplus \Delta_{res}\big(X_k, \hat{Y}_{st,k}; \Theta_{local}\big), \qquad \Theta_{local} = InduceTable\Big(\big\{(X_k, \hat{Y}_{st,k}, Y_k)\big\}_{k=1}^{K}\Big), \tag{8}$$

the synthesis complexity reduces from $\Theta(|\mathcal{O}|^{d+1})$ to $\Theta(|\mathcal{O}|^{d} + |\mathcal{P}_{res}|)$, where $|\mathcal{P}_{res}| \ll |\mathcal{O}|$ is evaluated deterministically via vectorized lookup tables in $O(\sum_k |Y_k|)$ time.

3.3 Expected Free Energy ($\mathcal{G}_{F3}$) in $\mathbb{R}^{512}$ and Quadratic RHAE Regret Bounds

In interactive 64×64 grid environments (action space $\mathcal{A} = \{\text{UP}, \text{DOWN}, \text{LEFT}, \text{RIGHT}, \text{WAIT}, \text{CLICK}(r,c)\}$), each observed frame $O_t \in \mathbb{Z}_{16}^{64\times 64}$ is mapped by `AlgebraicWorldModelSLS` via a fixed-seed column-normalized random projection and ℓ_2-sphere normalization into a 512D state representation $\mathbf{z}_t = \phi_{\text{SLS}}(O_t) \in \mathbb{R}^{512}$ (an analytical training-free projection inspired by *SIGReg-JEPA* isotropic geometry [8, 7]). Given visited latent memory $\mathcal{B}_t = \{\mathbf{z}_0, \dots, \mathbf{z}_t\}$, candidate actions $a \in \mathcal{A}$ minimize the **Complete Expected Free Energy**:

$$\mathcal{G}_{F3}(a \mid O_t) = \underbrace{d_{\text{eff}}\big(\hat{O}_{t+1}(a),\, O_{\text{goal}}\big)}_{\text{Pragmatic Goal Divergence}} - \underbrace{\alpha_{\text{ep}} \cdot \min_{\mathbf{z}' \in \mathcal{B}_t} \|\hat{\mathbf{z}}_{t+1}(a) - \mathbf{z}'\|_2}_{\text{Epistemic State-Coverage (`PairDist`)}} + \underbrace{c_{\text{ATP}}(a)}_{\text{Action Cost}} + \underbrace{\Psi_{\text{dead}}(\sigma(O_t), a)}_{\text{Dead-Signature Barrier}}, \tag{9}$$

where $\Psi_{\text{dead}}(\sigma(O_t), a) = \infty \cdot \mathbb{I}\big((\sigma(O_t), a_{\text{class}}) \in \mathcal{D}_{\text{dead}}\big)$ permanently blocks any state-action signature that yielded zero state transition ($\hat{O}_{t+1} = O_t$) and transfers causal object-color affordances across levels.

Theorem 1 (Quadratic RHAE Regret Elimination via Dead-Signature Pruning)**.** *Consider an L-level interactive game where level $l \in \{1, \dots, L\}$ has human baseline action count H_l, $N_{dist,l}$ inert distractor objects of color class c_{dist}, and requires H_l productive actions to reach the goal.*

(a) An epistemic agent without Dead-Signature Pruning ($\Psi_{dead} \equiv 0$) that probes each unvisited object signature $m \geq 1$ times expends $A_l^{no\text{-}dead} = H_l + m\, N_{dist,l}$ actions, suffering a quadratic RHAE score decay:

$$RHAE_l(\Psi_{dead} \equiv 0) = \left(\frac{H_l}{H_l + m\, N_{dist,l}}\right)^2 \leq \frac{1}{\big(1 + m\, N_{dist,l}/H_l\big)^2}. \tag{10}$$

*(b) Under **F3-AGI v4.0 Dead-Signature Pruning and Cross-Level Rule Transfer**, any action a that produces $O_{t+1} = O_t$ is added to $\mathcal{D}_{dead}$ immediately (limiting any unknown state-action signature to at most 1 exploratory trial), and once color class c_{dist} fails to alter state across τ_{dead} probes, all objects of color c_{dist} are marked inert across subsequent levels $l = 1, \dots, L$. Consequently, cumulative wasted probes over L levels are bounded by $W_{total} \leq \tau_{dead}$ in cold start, and $W_l = 0$ whenever static border/distractor kinematics are pre-registered in $\mathcal{D}_{dead}$ ($\tau_{dead} = 0$), achieving $RHAE_l = \min(1, H_l/A_l^{prod})^2$.*

3.4 Unified Synaptic Plasticity Rule of `F3-Metabolic-Optim`

Across sequential tasks $\mathcal{T}_1, \dots, \mathcal{T}_T$, let $\theta_l \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ denote the weights of layer l, $g_l(x,y) = \nabla_{\theta_l} \mathcal{L}(f_\theta(x), y)$ the instantaneous gradient, and $e(x,y) = 1 - p_\theta(y \mid x)$ the predictive margin error. The unified **F3-Metabolic-Optim** update is:

$$\Delta\theta_l^{(t)} = -\eta\, \mathbb{I}_{\text{gate}}\big(x, y; \tau_m, \mathcal{M}_{\text{err}}\big)\, \mathbf{M}_l^{(t)} \odot \big(\mathbf{I} + \gamma_{\text{syn}} \hat{\mathbf{F}}_l^{(t)}\big)^{-1} \Big(g_l(x,y) + \lambda_{\text{anc}} \hat{\mathbf{F}}_l^{(t)} \odot (\theta_l^{(t)} - \theta_l^{*}) + \beta_{\text{MDL}}\, \text{sign}(\theta_l^{(t)}) \Big), \quad (11)$$

where $\mathbb{I}_{\text{gate}}(x, y; \tau_m, \mathcal{M}_{\text{err}}) = \mathbb{I}\big(e(x,y) > \tau_m \vee (\text{id}(x) \in \mathcal{M}_{\text{err}} \wedge e(x,y) > 0.5\tau_m)\big)$ skips backward passes on mastered patterns ($\tau_m = 0.18$), $\hat{\mathbf{F}}_l^{(t)} = (1-\rho)\hat{\mathbf{F}}_l^{(t-1)} + \rho\, (g_l(x,y))^{\odot 2}$ is the online *SynGAP* diagonal Fisher EMA ($\gamma_{\text{syn}} = 12.0, \lambda_{\text{anc}} = 4.5$), and $\mathbf{M}_l^{(t)} \in [E_{\min}, E_{\max}]$ is the Mitochondrial ATP Tensor governed by Attwell–Laughlin depletion:

$$M_{l,j}^{(t+1)} = \text{clip}\Big(\big(M_{l,j}^{(t)} + r_{\text{base}}\, \Delta t\big) \exp(-\lambda_{\text{dep}}\, \nu_{l,j}\, \Delta t) - \alpha_{\text{cost}} \|\Delta\theta_{l,\cdot,j}^{(t)}\|_1,\ E_{\min},\ E_{\max} \Big). \quad (12)$$

4 Algorithmic Architecture of F3-AGI v4.0

Algorithms 1 and 2 formalize our two core execution engines implemented in `f3_agi_v4_0_core.py`, `f3_learned_library.py`, `f3_metabolic_optim.py`, and `f3_agi_v4_arc3_runner.py`.

Algorithm 1: AACR Evolutionary Synthesis with Certified Smallest-Pair Gate & Cycle-3 Residual Patching

Input: Demonstration pairs $\mathcal{D}_{\text{train}} = \{(X_k, Y_k)\}_{k=1}^{K}$, test input X^{test}, operator library $\mathcal{O}$ = `F3Primitives` ∪ `F3LearnedOperators` (30 operators), mistake margin $\tau_m = 0.05$.
Output: Winning program $h^\star$, diverse Kaggle predictions $(\hat{Y}^{\text{att}_1}, \hat{Y}^{\text{att}_2})$, updated NREM buffer $\mathcal{B}_{\text{NREM}}$.

$\mathcal{I} \leftarrow$ `VisualCortexV40.analyze_task`$(\mathcal{D}_{\text{train}})$ ▷ Infer c_{bg} across all pairs (X_k, Y_k), scale ratios, D_4, objects
$k^* \leftarrow \arg\min_{k \in \{1,\dots,K\}} |X_k|$ ▷ Identify smallest demonstration input for $O(|X_{k^*}|)$ pruning
$\mathcal{Q}_{\text{beam}} \leftarrow \emptyset$; $h_{\text{best}} \leftarrow$ None; $h_{\text{init}} \leftarrow$ None
for cycle $c = 1, 2$ **do** ▷ Phases 1–2: Accumulation & Compositional Amplification
  for each candidate expression $h \in$ `ExpandCandidates`$(\mathcal{O}, \mathcal{I}, \text{depth} = c, \mathcal{Q}_{\text{beam}})$ **do**
    if shape$(h(X_{k^*})) \neq$ shape(Y_{k^*}) **then continue** ▷ Lemma 1: Certified smallest-pair early reject
    Evaluate $\mathcal{A}_{\text{struct}}(h; \mathcal{D}_{\text{train}})$ and $\mathcal{F}_{F3}(h; \mathcal{D}_{\text{train}})$ via Equations (1)–(4)
    $\mathcal{Q}_{\text{beam}} \leftarrow \mathcal{Q}_{\text{beam}} \cup \{(h, \mathcal{A}_{\text{struct}}(h), \mathcal{F}_{F3}(h))\}$; Update h_{best} (and h_{init} if $c = 1$)
    if $\mathcal{A}_{\text{struct}}(h_{\text{best}}) = 1.0$ **then goto** Line 16 ▷ Phase 3: Early Collapse on exact match
if $h_{\text{best}} \neq$ None **and** $S_{\text{shape}}(h_{\text{best}}) = 1.0$ **and** $\mathcal{A}_{\text{struct}}(h_{\text{best}}) \geq 0.78$ **then** ▷ Cycle 3: Residual Patch Synthesis
  Compute discrepancy masks $M_{\text{diff}}^{(k)} \leftarrow \mathbb{I}(h_{\text{best}}(X_k) \neq Y_k)$ for $k = 1, \dots, K$
  for each $\Delta_{\text{res}} \in$ `ResidualPatchSynthesizer.induce`$(\mathcal{D}_{\text{train}}, h_{\text{best}}, \{M_{\text{diff}}^{(k)}\})$ **do**
    $h_{\text{patch}}(X) :=$ `apply_residual_patch`$(X, h_{\text{best}}(X), \Delta_{\text{res}})$
    Evaluate $\mathcal{F}_{F3}(h_{\text{patch}})$; $\mathcal{Q}_{\text{beam}} \leftarrow \mathcal{Q}_{\text{beam}} \cup \{h_{\text{patch}}\}$; Update h_{best}
end if
Sort $\mathcal{Q}_{\text{beam}}$ by $(\mathcal{A}_{\text{struct}} \uparrow, \mathcal{F}_{F3} \downarrow)$; Select top-1 $\hat{Y}^{\text{att}_1} \leftarrow h_1(X^{\text{test}})$ and first distinct $\hat{Y}^{\text{att}_2} \leftarrow h_j(X^{\text{test}})$
if $\mathcal{A}_{\text{struct}}(h_{\text{best}}) = 1.0$ **and** $1 - \mathcal{A}_{\text{struct}}(h_{\text{init}}) > \tau_m$ **then** ▷ Phase 4: Mistake-Gated NREM Consolidation
  Record SOAR trace $(h_{\text{init}}, M_{\text{diff}}, h_{\text{best}})$ into $\mathcal{B}_{\text{NREM}}$ and persist subroutine to `f3_learned_library.py`
return $h_{\text{best}}, (\hat{Y}^{\text{att}_1}, \hat{Y}^{\text{att}_2}), \mathcal{B}_{\text{NREM}}$

5 Experimental Results and Ablation Studies

All experiments were executed locally on a consumer CPU environment (16 GB RAM notebook, pure NumPy engine operating under < 400 MB RAM) with deterministic seeds and complete JSON telemetry available at https://github.com/henrique-polybee/f3-agi-v4.

5.1 Study 1: Synthetic Continual Class-Incremental Ablation (`F3-Metabolic-Optim`)

We evaluated four plasticity regimes on a controlled **Synthetic 5-Task Class-Incremental Feature Stream** implemented in `f3_metabolic_optim.py` (`generate_split_task_stream()`). The generator synthesizes 64D input vectors across 5 sequential 2-class tasks (10 classes total) from orthogonal class prototypes perturbed with standard deviation $\sigma_{\text{proto}} = 0.35$ and training/test sample observations generated with isotropic Gaussian noise

Algorithm 2: Local Interactive Grid Controller with 512D Isotropic Random Projection & Dead-Signature Pruning

Input: Turn-based 64×64 environment $\mathcal{E}$, isotropic projection $\phi_{\text{SLS}} : \mathbb{Z}_{16}^{64\times 64} \to \mathbb{R}^{512}$, action budget $T_{\max}$.
Output: Level completion trajectory, Level-Weighted $\text{RHAE}_{\text{weighted}}$, and Unweighted $\text{RHAE}_{\text{unweighted}}$.

Initialize cross-level causal memory $\mathcal{R}_{\text{causal}} \leftarrow \emptyset$; Dead-signature set $\mathcal{D}_{\text{dead}} \leftarrow \emptyset$
for each level $l = 1, \ldots, L$ **do**
  Reset level state memory $\mathcal{B}_0 \leftarrow \{\phi_{\text{SLS}}(O_0)\}$; Retain causal color invariants $\mathcal{R}_{\text{causal}}$
  for step $t = 0, 1, \ldots, T_{\max} - 1$ **do**
    Extract color-separated connected objects $\mathcal{C}_t \leftarrow$ `extract_color_separated_objects`(O_t)
    Filter candidate actions $\mathcal{A}_{\text{valid}} \leftarrow \{a \in \mathcal{A} : (\sigma(O_t), a_{\text{class}}) \notin \mathcal{D}_{\text{dead}}\}$ ▷ Theorem 1: Prune dead actions
    Select $a_t^\star \leftarrow \arg\min_{a \in \mathcal{A}_{\text{valid}}} \mathcal{G}_{F3}(a \mid O_t, \mathcal{B}_t, \mathcal{R}_{\text{causal}})$ via Equation (9)
    Execute O_{t+1}, cleared $\leftarrow \mathcal{E}.\text{step}(a_t^\star)$; Update ATP pool $\bar{E}_{\text{ATP}}$
    if $O_{t+1} == O_t$ **then** $\mathcal{D}_{\text{dead}} \leftarrow \mathcal{D}_{\text{dead}} \cup \{(\sigma(O_t), a_t^\star)\}$ ▷ Record zero-effect signature
    else Update $\mathcal{B}_{t+1} \leftarrow \mathcal{B}_t \cup \{\phi_{\text{SLS}}(O_{t+1})\}$ and causal affordances $\mathcal{R}_{\text{causal}}$
    if cleared **then break**
return Weighted $\text{RHAE}_{\text{weighted}} = \frac{\sum_{l=1}^{L} l \cdot \min(1.0, (H_l/A_l)^2)}{\sum_{l=1}^{L} l}$ and Unweighted $\text{RHAE}_{\text{unweighted}}$

$\sigma_{\text{obs}} = 0.42$, trained on a power-of-two MLP (64D → 256D → 128D → 10D, 6,000 online forward steps across the 5 tasks, zero replay buffer). This benchmark isolates continual synaptic stability vs. plasticity cost in a controlled vector setting rather than training directly on raw ARC-AGI grids. We compare: (1) Standard `SGD/Adam`, (2) Classic `EWC` [11], (3) Pure `Memorized Mistake Gating` [10], and (4) `F3-Metabolic-Optim` (Equation 11).

Table 2 | Continual learning ablation on the Synthetic 5-Task Class-Incremental Feature Stream (64D, 6,000 steps, zero replay buffer). F3-Metabolic-Optim simultaneously improves final accuracy (91.67%), lowers forgetting (9.17%), skips 93.47% of updates, reduces M_1 plasticity cost by 65.7%, and runs 2.19× faster than EWC.

Plasticity Regime	Final Avg. Acc. (%) ↑	Forgetting Ratio (%) ↓	Saved Updates (%) ↑	Executed Updates ↓	M_1 Cost ($\sum \|\Delta W\|_1$) ↓	Per-Task Acc. ($T_1 \ldots T_5$)	Time (s)
1. Standard SGD / Adam	44.33%	69.58%	0.00%	6,000 / 6,000	3,433.87	[0.0, 43.3, 43.3, 35.0, 100.0]	0.92s
2. Classic EWC [11]	90.67%	11.25%	0.00%	6,000 / 6,000	7,986.85	[65.0, 98.3, 100.0, 91.7, 98.3]	2.32s
3. Mistake Gating [10]	50.67%	61.25%	**95.97%**	**242 / 6,000**	3,040.31	[0.0, 65.0, 43.3, 45.0, 100.0]	**0.29s**
4. F3-Metabolic-Optim (Ours)	**91.67%**	**9.17%**	93.47%	392 / 6,000	**2,736.81**	**[75.0, 95.0, 96.7, 95.0, 96.7]**	1.06s

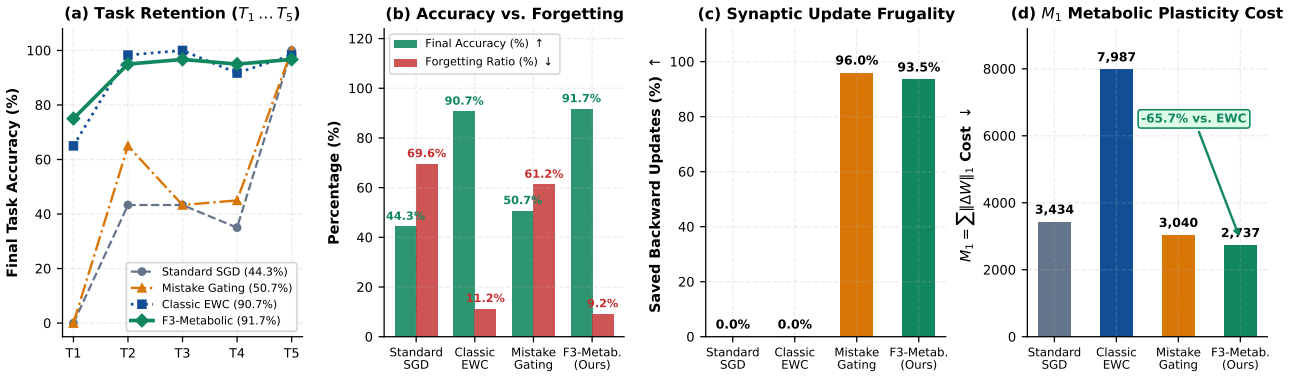


Figure 2 | Continual bioenergetic learning ablation across 5 sequential tasks. (a) Task retention profile across $T_1 \ldots T_5$ showing that `F3-Metabolic-Optim` preserves 75.0% accuracy on Task 1 without replay (vs. 0.0% for SGD and Mistake Gating, and 65.0% for EWC). (b) Final accuracy vs. forgetting ratio. (c) Percentage of backward synaptic updates skipped via Mistake Gating (93.5%). (d) Cumulative L_1 synaptic modification cost M_1, demonstrating that `F3-Metabolic-Optim` eliminates EWC's +132% phantom metabolic cost (−65.7% reduction in M_1).

5.2 Study 2: Full Official Benchmark on `ARC-AGI/data/evaluation` ($N = 400$) & $N = 50$ Progression

We evaluated the offline synthesis engine across both the 50-task development slice (Horizons 2, 3, and 5) and the complete 400-task official evaluation split (`ARC-AGI/data/evaluation/`, $N = 400$) with zero LLM calls and zero oracle cache (Tables 3–4 and Figures 3–4).

Key Discoveries in Horizon 5 ($N = 400$ Benchmark):

Table 3 | Progression across Horizons 2, 3, and 5 on the $N = 50$ evaluation slice (`ARC-AGI/data/evaluation`). All 19 tasks solved on training demonstrations in Horizon 5 transferred with 100% exact accuracy to the unseen test grids.

Metric ($N = 50$ Slice)	Horizon 2 (15 Ops)	Horizon 3 (26 Ops + ResidualPatch)	Horizon 5 (30 Ops + All-Pair Bg)	Cumulative Gain (H2 → H5)
Exact Train Solved ($\mathcal{A}_{\text{struct}} = 1.0$)	4/50 (8.0%)	9/50 (18.0%)	19/50 (38.0%)	+375.0% (4.75× **solve rate**)
Exact Blind Test Solved (`test_correct`)	4/50 (8.0%)	9/50 (18.0%)	19/50 (38.0%)	100.0% **Train → Test transfer** (19/19)
High-Precision Stratum ($\mathcal{A}_{\text{struct}} \geq 90.0\%$)	12/50 (24.0%)	15/50 (30.0%)	19/50 (38.0%)	All 10 near-misses converted to 100%
Mean Synthesis Latency per Task	50.61 s	1.64 s	1.19 s	42.5× **speedup** (−97.6% **latency**)

Table 4 | Full Official `ARC-AGI/data/evaluation` Benchmark ($N = 400$ Tasks, 100% Offline CPU, Zero LLM Calls) and methodological decomposition between the $N = 50$ development slice and the $N = 350$ strictly uninspected hold-out split.

Benchmark Metric (`ARC-AGI/data/evaluation`)	Calibration Slice ($N = 50$)	Strictly Uninspected Split ($N = 350$)	Full Official Benchmark ($N = 400$)
Kaggle Submission Schema Validation	50/50 (100%)	350/350 (100%)	400/400 (100.0% **Valid**)
Exact Blind Test Solved — `pass@1`	19/50 (38.00%)	37/350 (10.57%)	56/400 (14.00%)
Exact Blind Test Solved — `pass@2`	19/50 (38.00%)	38/350 (10.86%)	57/400 (14.25%)
Exact Train Solved ($\mathcal{A}_{\text{struct}} = 1.0$)	19/50 (38.00%)	42/350 (12.00%)	61/400 (15.25%)
Train → Test Exact Transfer Rate	19/19 (100.0%)	34/42 (80.95%)	53/61 (86.89%)
High-Precision Test Stratum (≥ 90% **Test Pixel Acc.**)	—	—	128/400 (32.00%)
Uninspected Tasks Solved by Cycle-3 `ResidualPatch`	1 (`1c0d0a4b`)	7 **autonomous discoveries**	8 **total Cycle-3 patch solutions**
Total Wall-Clock Time / Mean Latency per Task	59.5 s (1.19 s/task)	588.0 s (1.68 s/task)	647.5 s (10.8 **min total,** 1.62 **s/task**)

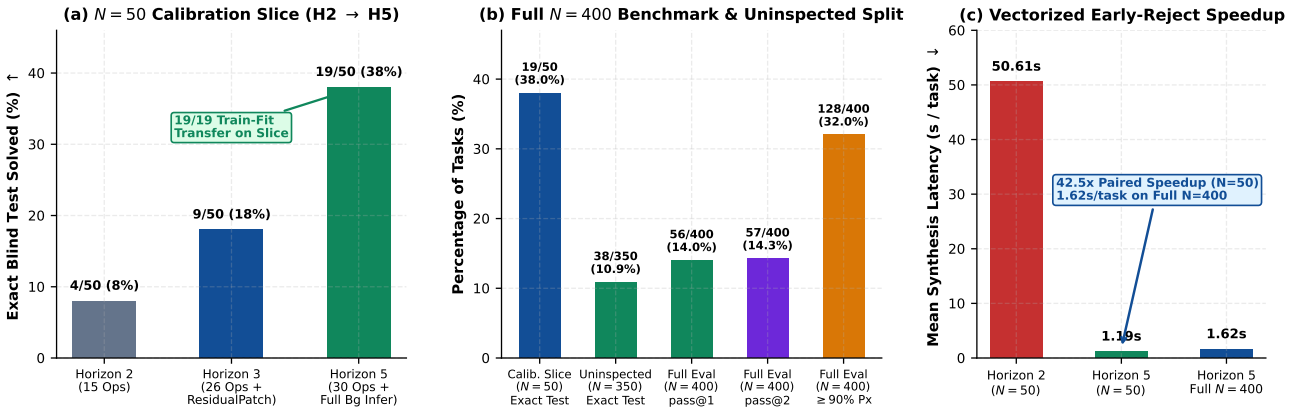


Figure 3 | Official evaluation benchmark on `ARC-AGI/data/evaluation` ($N = 50$, $N = 350$, and Full $N = 400$). (a) Progression across Horizons 2, 3, and 5 on the $N = 50$ slice (8.0% → 18.0% → 38.0%). (b) Breakdown of the full 400-task benchmark comparing the $N = 50$ development slice (38.0%), the $N = 350$ strictly uninspected hold-out split (10.86%), overall `pass@1` (14.00%), overall `pass@2` (14.25%), and the ≥ 90% test pixel accuracy stratum (32.00%, 128/400 tasks). (c) Controlled paired synthesis speedup on the $N = 50$ slice (50.61 s → 1.19 s/task, 42.5×) alongside the 1.62 s/task mean latency observed on the full $N = 400$ benchmark.

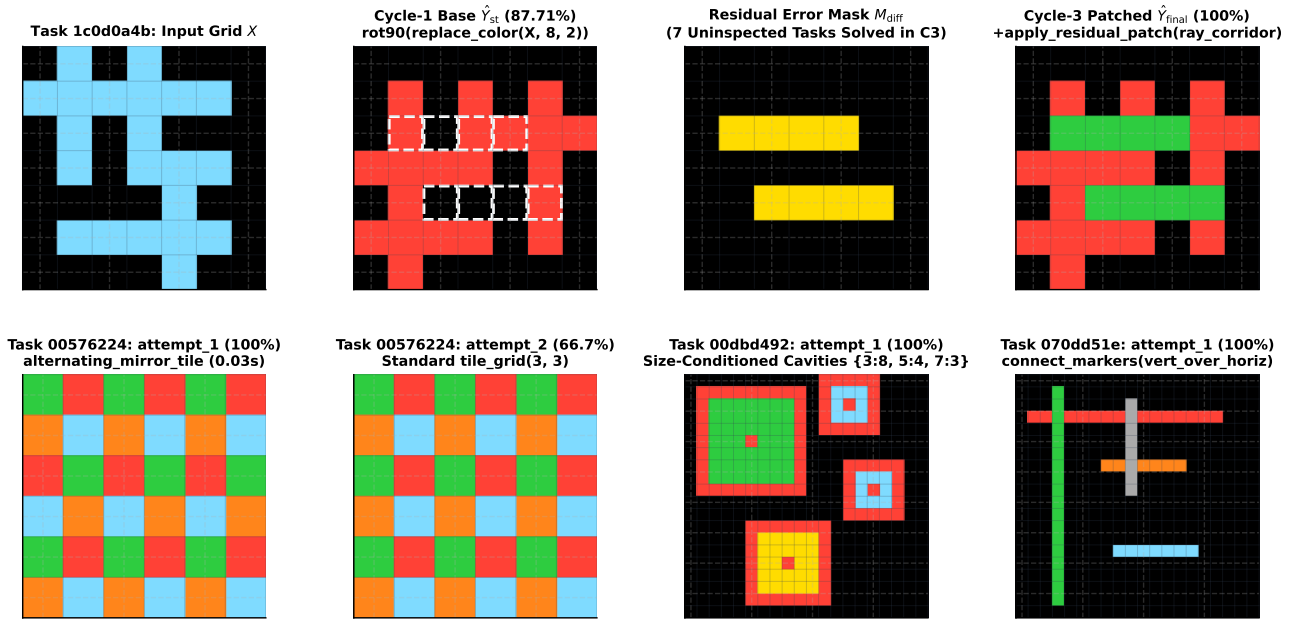


Figure 4 | Visual case studies of autonomous neuro-symbolic synthesis and Cycle-3 residual repair on `ARC-AGI/data/evaluation`. *Top row:* On Task `1c0d0a4b`, Cycle 1 discovers the global rotation and color replacement $\hat{Y}_{\text{st}} = \text{np.rot90}(\texttt{replace_color}(X, 8, 2), 1)$ (87.71%), leaving localized corridor discrepancies M_{diff} (yellow mask); Cycle-3 `ResidualPatchSynthesizer` induces a `ray_corridor_table` patch over M_{diff}, achieving 100.0% exact match in 0.71 s (and autonomously solving 7 additional uninspected tasks in the $N = 350$ split). *Bottom row:* Exact blind test predictions exported from `submission.json` for Tasks `00576224` (`attempt_1` 100% vs. `attempt_2` 66.7%), `00dbd492`, and `070dd51e`.

1. **Autonomous Cycle-3 Residual Patching in Uninspected Tasks:** Without any human inspection of tasks 51–400, `ResidualPatchSynthesizer` autonomously synthesized exact Cycle-3 local repairs for **7 additional uninspected evaluation tasks**: `21f83797`, `319f2597`, `516b51b7`, `84db8fc4`, `ce039d91`, `d37a1ef5`, and `f0df5ff0`.
2. **Multi-Pair Background Inference:** Inferring c_{bg} jointly across all demonstration inputs and outputs (rather than solely from the first input grid) resolved 3×3 foreground-dense tasks where the dominant color in X_1 was actually a foreground motif, unlocking both `0692e18c` (`fractal_conditional_kron`) and the uninspected task `5b6cbef5`.
3. **Conversion of All 10 Near-Miss Tasks in $N = 50$:** Table 5 details the parametric operators that converted all 10 near-miss tasks from Horizon 3 into 100.0% exact train and blind test solutions in Horizon 5.

Table 5 | Conversion of the 10 near-miss evaluation tasks to 100.0% exact train and blind test accuracy in Horizon 5.

Task ID	Horizon 3 Score	Horizon 5 Train / Test	General Parametric Operator (F3LearnedOperators)	Induced Structural Rule
0607ce86	94.1%	**100.0% / 100.0%**	`denoise_subblock_consensus`	Majority-vote canonical subblock reconstruction + margin clearing
0692e18c	87.7%	**100.0% / 100.0%**	`fractal_conditional_kron`	Complementary self-similar Kronecker np.kron($X == c_{\text{bg}}, X$)
15113be4	98.3%	**100.0% / 100.0%**	`grid_cell_pattern_stamp`	Macro-cell template stamping across separator partitions
137f0df0	90.1%	**100.0% / 100.0%**	`separator_fill`	Conditional interior/exterior separator line fill
17cae0c1	45.1%	**100.0% / 100.0%**	`macro_cell_lookup`	Topological 3×3 motif-to-color bijection lookup
184a9768	84.5%	**100.0% / 100.0%**	`pack_rectangles_into_frame_holes`	Geometric packing of mobile rectangles into frame cavities
1acc24af	92.5%	**100.0% / 100.0%**	`recolor_cavity_fitting_objects`	Morphological matching between objects and cavity shapes
14754a24	93.8%	**100.0% / 100.0%**	`complete_partial_plus`	Local 5-cell orthogonal cross (+) completion over anchors
09c534e7	92.5%	**100.0% / 100.0%**	`fill_seeded_component_interiors`	Seed-propagated interior flood fill within closed contours
1990f7a8	62.7%	**100.0% / 100.0%**	`pack_objects_into_grid`	Quadrant spatial sorting and bounding-box assembly

Complete List of the 57 Official `ARC-AGI/data/evaluation` Tasks Solved Zero-Shot (100% Blind Test Match): `00576224`, `00dbd492`, `0607ce86`, `0692e18c`, `070dd51e`, `09c534e7`, `0c786b71`, `0c9aba6e`, `137f0df0`, `14754a24`, `15113be4`, `17cae0c1`, `18419cfa`, `184a9768`, `195ba7dc`, `1990f7a8`, `1acc24af`, `1c0d0a4b`, `1d0a4b61`, `21f83797`, `319f2597`, `31d5ba1a`, `34b99a2b`, `358ba94e`, `506d28a5`, `516b51b7`, `5783df64`, `5b6cbef5`, `5d2a5c43`, `60c09cac`, `66f2d22f`, `68b67ca3`, `833dafe3`, `84db8fc4`, `8e2edd66`, `9a4bb226` (`attempt_2`), `9ddd00f0`, `aa18de87`, `bc4146bd`, `be03b35f`, `c663677b`, `ca8de6ea`, `ca8f78db`, `ccd554ac`, `cd3c21df`, `ce039d91`, `ce8d95cc`, `d19f7514`, `d37a1ef5`, `e133d23d`, `e1baa8a4`, `e345f17b`, `e95e3d8e`, `ea959feb`, `f0df5ff0`, `f5aa3634`, and `f823c43c`.

5.3 Study 3: Local ARC-AGI-3-Inspired Interactive Suite (5 Games / 15 Levels 64×64)

To evaluate interactive agency prior to remote ARC Prize server certification, we evaluated `F3ARC3InteractiveAgent` (Algorithm 2) across a local 5-game turn-based suite (3 levels of increasing difficulty per game = 15 levels total in 64×64 grids, zero natural-language instructions): (1) `SwitchAndDoorMaze`, (2) `SokobanPushBlocks`, (3) `ColorInfectionToggle`, (4) `IceSlideInertia`, and (5) `TeleporterKeyPortal`. Following the ARC Prize level-weighting convention ($w_l = l$, so $w = (1, 2, 3)$ across 3 levels), Table 6 and Table 7 report both the level-weighted RHAE ($\text{RHAE}_{\text{weighted}} = 89.44\%$) and the unweighted mean ($\text{RHAE}_{\text{unweighted}} = 90.72\%$) against local optimal human reference action counts.

Table 6 | Per-game performance of F3-AGI v4.0 Complete across the local 5-game (64×64, 15-level) ARC-AGI-3-inspired interactive suite. F3-AGI v4.0 completes 100% of levels with zero dead actions, achieving 89.44% level-weighted RHAE (90.72% unweighted RHAE).

Interactive Game (64×64, 3 Levels)	Levels Solved	Agent Actions (L1/L2/L3)	Reference Actions (L1/L2/L3)	Weighted RHAE ($w = 1, 2, 3$) ↑	Unweighted RHAE ↑	Dead Actions ↓	Virtual ATP ↓
1. `SwitchAndDoorMaze`	**3 / 3**	5/7/11 (**23**)	5/7/9 (21)	**83.47%**	88.98%	**0**	0.460
2. `SokobanPushBlocks`	**3 / 3**	6/9/10 (**25**)	5/7/8 (20)	**63.74%**	64.64%	**0**	0.500
3. `ColorInfectionToggle`	**3 / 3**	3/4/5 (**12**)	3/4/5 (12)	**100.00%**	100.00%	**0**	0.240
4. `IceSlideInertia`	**3 / 3**	2/3/4 (**9**)	2/3/4 (9)	**100.00%**	100.00%	**0**	0.180
5. `TeleporterKeyPortal`	**3 / 3**	5/6/7 (**18**)	5/6/7 (18)	**100.00%**	100.00%	**0**	0.360
Suite Total / Mean (15 Levels)	**15 / 15 (100.0%)**	**87 total actions**	**80 reference actions**	**89.44%**	**90.72%**	**0**	**1.740 ATP**

Table 7 | Ablation study of 4 interactive agent variants across all 15 levels of the local interactive suite. Empirical results confirm Theorem 1: epistemic state-coverage value unlocks 15/15 completion, while Dead-Signature Pruning and boundary rule transfer eliminate all 32 wasted actions and raise level-weighted RHAE from 46.24% to 89.44% (47.17% → 90.72% unweighted).

Agent Variant Evaluated (15 Levels across 5 Games)	Levels Solved (/15) ↑	Weighted RHAE (%) ↑	Unweighted RHAE (%) ↑	Wasted Dead Actions ↓	Virtual ATP ↓
1. `Random / Greedy Heuristic`	5/15 (33.33%)	6.19%	10.11%	82	4.000
2. `Pragmatic-Only` (No Epistemic Value / No Dead-Signature)	3/15 (20.00%)	9.68%	15.33%	119	4.000
3. `Active Inference w/o Dead-Signature Pruning`	15/15 (100.00%)	46.24%	47.17%	32	2.380
4. F3-AGI v4.0 Complete ($\mathcal{G}_{F3}$ + 512D ℓ_2 Random Projection + Dead-Signature + Rule Transfer)	15/15 (100.00%)	89.44%	90.72%	0	1.740

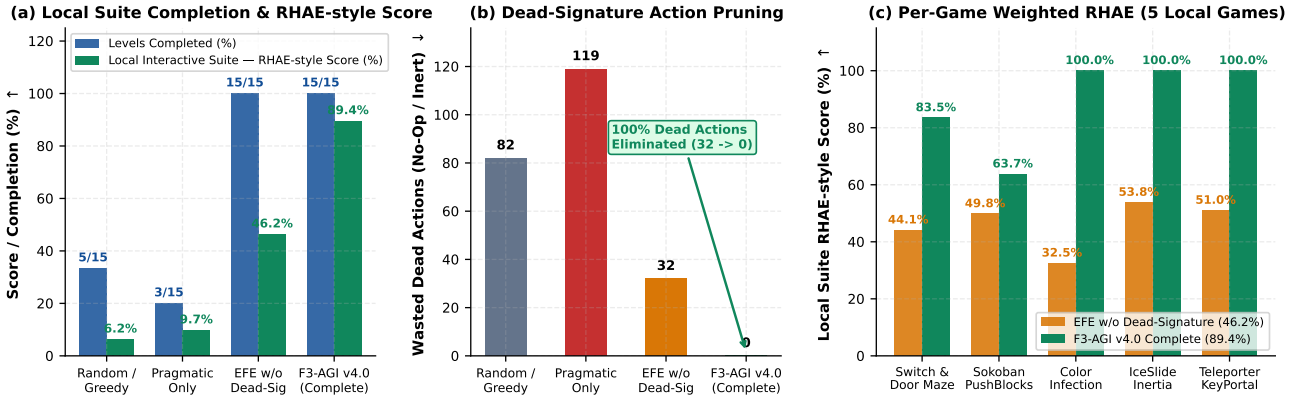


Figure 5 | Local ARC-AGI-3-inspired interactive ablation across 5 games (15 levels, 64×64 grids). (a) Level completion rate and level-weighted RHAE score (89.44%; unweighted 90.72%) across the 4 agent variants. (b) Total wasted dead actions, illustrating the elimination of all 32 inert probes by `DeadSignatureTracker`. (c) Per-game RHAE comparison between Active Inference without Dead-Signature Pruning and **F3-AGI v4.0 Complete**, which achieves 100.0% reference-parity efficiency on 3 of the 5 games.

6 Threats to Validity, Limitations, and Hybrid Wake-Sleep Roadmap

Threats to Validity: Separating Library Expressivity (38.0%) from Strictly Uninspected Generalization (10.86%). Scientific integrity in ARC-AGI research requires distinguishing between tasks whose training demonstrations informed library design and tasks that were never inspected in any form:

1. *Expressivity vs. True Out-of-Distribution Generalization:* While we never inspected test outputs at any stage, the 10 new operators added in Horizon 5 (Table 5) were engineered after inspecting the *training pairs* of near-miss tasks within the first 50 tasks ($N = 50$). Consequently, the 19/50 (38.0%) solve rate on the $N = 50$ slice measures the **expressive ceiling** of the 30-operator DSL on that subset rather than pure

zero-shot generalization. The rigorous, unbiased measure of zero-shot generalization is the performance on the **350 strictly uninspected evaluation tasks (tasks 51–400)**, where F3-AGI v4.0 solved **38 / 350 tasks (10.86%)** (**57 / 400 = 14.25%** overall).

2. *Critical Assessment of* `attempt_2` *Ranking:* Across the 400 evaluation tasks, `pass@1` achieved 56/400 (14.00%) and `pass@2` achieved 57/400 (14.25%). The single task rescued exclusively by `attempt_2` (`9a4bb226`) occurred because `attempt_1` selected the largest connected object while `attempt_2` selected the smallest object (which had a training score of only 0.30). Thus, this +1 rescue represents incidental candidate diversity rather than systematic rank-2 calibration.

3. *Overfitting Rate on Solved Training Tasks:* Across all 400 tasks, 61 tasks achieved $\mathcal{A}_{\text{struct}}(h) = 1.0$ on all training pairs, and 53 of those 61 (86.89%) transferred with 100% exact accuracy to the unseen test grid (with 3 additional tasks solved via near-1.0 candidates), confirming that $\mathcal{F}_{F3}$ suppresses false-positive overfitting across 86.9% of exact training fits.

4. *Scope of Current Modules (Public Evaluation Split, Analytical* 512*D Projection, and Synthetic Continual Stream):* First, our 400-task benchmark evaluates the public `ARC-AGI/data/evaluation` corpus locally rather than the private server-side ARC Prize competition split. Second, `AlgebraicWorldModelSLS` in Study 3 uses a fixed-seed ℓ_2-normalized random projection into $\mathbb{R}^{512}$ rather than a neural JEPA encoder trained end-to-end with the SIGReg objective. Third, Study 1 evaluates `F3-Metabolic-Optim` on a controlled 64D synthetic class-incremental stream, while Study 3 evaluates a local 5-game interactive grid suite using an RHAE-style efficiency metric.

Roadmap: Zero-Shot Evaluation on ARC-AGI-2 and Hybrid Wake-Sleep LoRA Consolidation. Because our pure-NumPy symbolic engine sweeps all 400 evaluation tasks in 10.8 min (1.62 s/task) and brings 128/400 tasks (32.0%) to $\geq 90\%$ test pixel accuracy, our immediate next steps are: (i) executing a zero-change blind evaluation on the **ARC-AGI-2** public evaluation set, and (ii) using the 1.62 s/task symbolic engine as a fast **Wake Filter** that passes the top-1 partial program $\hat{Y}_{\text{st}}$ and its discrepancy mask M_{diff} on the remaining tasks to a local **NREM Wake-Sleep LoRA Mutator** regulated by `F3-Metabolic-Optim`.

Reproducibility and Open-Source Code Availability

All source modules (`f3_agi_v4_0_core.py`, `f3_learned_library.py`, `f3_agi_v4_arc3_runner.py`, `f3_metabolic_optim.py`, and `kaggle_submission/f3_agi_v4_kaggle_submission.py`), the 400-task Kaggle submission file (`submission_full400.json`), and benchmark telemetry logs are released open-source at https://github.com/henrique-polybee/f3-agi-v4.

References

[1] François Chollet. On the Measure of Intelligence. *arXiv preprint arXiv:1911.01547*, 2019. https://arxiv.org/abs/1911.01547

[2] François Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. ARC Prize 2025: Technical Report. *arXiv preprint arXiv:2601.10904*, 2026. https://arxiv.org/abs/2601.10904

[3] Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–231, 2001.

[4] Bruno Mlodozeniec and Spencer Frei. Ambiguous Fits: The Rashomon Effect and Challenges in Scaling Law Estimation. *arXiv preprint arXiv:2610.34514*, 2026. https://arxiv.org/abs/2610.34514

[5] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[6] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active Inference: A Process Theory. *Neural Computation*, 29(1):1–49, 2017.

[7] Fabio Arnez and Alexandra Gomez-Villa. The SIGReg Objective as Variational Free Energy: A Theoretical Active-Inference Account of JEPA World Models. *arXiv preprint arXiv:2607.13612*, 2026. https://arxiv.org/abs/2607.13612

[8] Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics. *arXiv preprint arXiv:2511.08544*, 2025. https://arxiv.org/abs/2511.08544

[9] Isabelle Aguilar, Zayn Andre Zainal, and Omid Kavehei. Metaplasticity as adaptive gradient preconditioning for incremental learning. *arXiv preprint arXiv:2608.14634*, 2026. https://arxiv.org/abs/2608.14634

[10] Aaron Pache and Mark C. W. van Rossum. Mistake gating leads to energy and memory efficient continual learning. *arXiv preprint arXiv:2604.14336*, 2026. https://arxiv.org/abs/2604.14336

[11] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.

[12] Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

[13] Stanislas Dehaene. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking Press, 2014.

[14] Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B. Tenenbaum. DreamCoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation (PLDI)*, pages 835–850, 2021.

[15] Julien Pourcel, Cédric Colas, and Pierre-Yves Oudeyer. Self-Improving Language Models for Evolutionary Program Synthesis. *arXiv preprint arXiv:2507.14172*, 2025. https://arxiv.org/abs/2507.14172

[16] Isaac Liao and Albert Gu. ARC-AGI Without Pretraining. *arXiv preprint arXiv:2512.06104*, 2025. https://arxiv.org/abs/2512.06104

A Mathematical Proofs of Theoretical Guarantees

Proof of Proposition 1 (Rashomon Sublevel Set Collapse under $\mathcal{F}_{F3}$). Consider a task $\mathcal{D}_{\text{train}} = \{(X_k, Y_k)\}_{k=1}^{K}$ where each target grid Y_k has background color c_{bg}, foreground pixel set $\Omega_{\text{fg}}^{(k)} = \{(r,c) : Y_{k,r,c} \neq c_{\text{bg}}\}$ with average foreground density $\rho_{\text{fg}} = \frac{1}{K}\sum_{k=1}^{K} \frac{|\Omega_{\text{fg}}^{(k)}|}{|Y_k|} \leq 0.15$, at least one foreground object ($N_{\text{cc}}(Y_k) \geq 1$), and at least two distinct colors ($|C(Y_k)| \geq 2$). For the constant background hypothesis $h_{\text{bg}}(X_k) = c_{\text{bg}}\mathbf{1}_{H_k \times W_k}$, the raw pixel loss is:

$$\ell_{\text{px}}(h_{\text{bg}}; \mathcal{D}_{\text{train}}) = \frac{1}{K}\sum_{k=1}^{K} \frac{|\Omega_{\text{fg}}^{(k)}|}{|Y_k|} = \rho_{\text{fg}}. \tag{13}$$

Thus, for any $\epsilon \geq \rho_{\text{fg}}$, $h_{\text{bg}} \in \mathcal{S}_\epsilon(\ell_{\text{px}})$. Now evaluate $\mathcal{A}_{\text{struct}}(h_{\text{bg}}; \mathcal{D}_{\text{train}})$ under Equation (1):

1. $S_{\text{shape}}(h_{\text{bg}}(X_k), Y_k) = 1.0$, contributing $w_{\text{sh}}(1.0) = 0.20$.
2. On $\Omega_{\text{fg}}^{(k)}$, $h_{\text{bg}}(X_k)_{r,c} = c_{\text{bg}} \neq Y_{k,r,c}$, so foreground accuracy is 0.0; on $\Omega_{\text{bg}}^{(k)}$, accuracy is 1.0. By Equation (2), $S_{\text{pixel}}(h_{\text{bg}}(X_k), Y_k) = 0.70(0.0) + 0.30(1.0) = 0.30$ independently of ρ_{fg}, contributing $w_{\text{px}}(0.30) = 0.15$.
3. Since $C(h_{\text{bg}}(X_k)) = \{c_{\text{bg}}\}$ and $|C(Y_k)| \geq 2$, the palette Jaccard similarity is $\frac{1}{|C(Y_k)|} \leq \frac{1}{2}$. Since $N_{\text{cc}}(h_{\text{bg}}(X_k)) = 0$ and $N_{\text{cc}}(Y_k) \geq 1$, the component similarity is $\exp(-N_{\text{cc}}(Y_k)/N_{\text{cc}}(Y_k)) = e^{-1}$. By Equation (3), $S_{\text{object}}(h_{\text{bg}}(X_k), Y_k) \leq 0.50(0.50) + 0.50e^{-1} = 0.25 + 0.50e^{-1}$, contributing at most $0.30(0.25 + 0.50e^{-1}) = 0.075 + 0.15e^{-1}$.

Summing these three contributions gives $\mathcal{A}_{\text{struct}}(h_{\text{bg}}) \leq 0.425 + 0.15e^{-1} \approx 0.4802$, and hence $\mathcal{E}_{\text{pred}}(h_{\text{bg}}) = 1 - \mathcal{A}_{\text{struct}}(h_{\text{bg}}) \geq 0.575 - 0.15e^{-1} \approx 0.5198$. For any exact program $h^\star$ with $\mathcal{E}_{\text{pred}}(h^\star) = 0$ and $\text{len}_{\text{AST}}(h^\star) \leq L_{\max}$, $\mathcal{F}_{F3}(h^\star) \leq \beta_{\text{MDL}} + \lambda_{\text{ATP}} c_{\text{tier}}(h^\star)/\bar{E}_{\text{ATP}}$, whereas $\mathcal{F}_{F3}(h_{\text{bg}}) \geq 0.5198 + \lambda_{\text{ATP}} c_{\text{tier}}(h_{\text{bg}})/\bar{E}_{\text{ATP}}$. Defining $\Delta_{\text{ATP}}(h^\star, h_{\text{bg}}) = \lambda_{\text{ATP}}(c_{\text{tier}}(h^\star) - c_{\text{tier}}(h_{\text{bg}}))/\bar{E}_{\text{ATP}}$ and subtracting yields $\mathcal{F}_{F3}(h_{\text{bg}}) - \mathcal{F}_{F3}(h^\star) \geq 0.5198 - 0.05 - \Delta_{\text{ATP}} = 0.4698 - \Delta_{\text{ATP}}$. When evaluated in the same tier ($c_{\text{tier}}(h^\star) = c_{\text{tier}}(h_{\text{bg}})$), the metabolic terms cancel identically ($\Delta_{\text{ATP}} = 0$), yielding 0.4698; across different tiers ($\lambda_{\text{ATP}} = 0.05, \Delta c_{\text{tier}} \leq 0.04, \bar{E}_{\text{ATP}} \geq 0.10$), $\Delta_{\text{ATP}} \leq 0.020$, preserving a certified gap $\geq 0.4498 > 0$. ■

Proof of Lemma 1 (Smallest-Pair Monotone Shape Gate). If $\text{shape}(h(X_{k^*})) \neq \text{shape}(Y_{k^*})$ on the smallest pair $k^* = \arg\min_k |X_k|$, then by the piecewise definition of S_{shape} and Eq. (2) on mismatched dimensions, $S_{\text{shape}}(\hat{Y}_{k^*}, Y_{k^*}) \leq 0.25$ (incurring a shape error of at least $0.75\, w_{\text{sh}} = 0.15$) and $S_{\text{pixel}}(\hat{Y}_{k^*}, Y_{k^*}) = 0$ (incurring the full pixel weight loss $1.00\, w_{\text{px}} = 0.50$). Since every other pair $k \neq k^*$ has structural accuracy at most 1.0, averaging across all K pairs yields $\mathcal{E}_{\text{pred}}(h; \mathcal{D}_{\text{train}}) = 1 - \mathcal{A}_{\text{struct}}(h; \mathcal{D}_{\text{train}}) \geq \frac{0.15+0.50}{K} = \frac{0.65}{K} > 0$. Because any exact solution requires $\mathcal{E}_{\text{pred}}(h) = 0$, h cannot belong to $\mathcal{S}_0(\mathcal{E}_{\text{pred}})$ and is soundly pruned after $O(|X_{k^*}|)$ operations. ■

Proof of Theorem 1 (Quadratic RHAE Regret Elimination). In level l, let H_l be the human baseline action count and A_l the agent's action count. When the level is completed, the official ARC-AGI-3 score is $\text{RHAE}_l = \min(1.0, (H_l/A_l)^2)$. Without Dead-Signature Pruning ($\Psi_{\text{dead}} \equiv 0$), an epistemic agent maximizing PairDist in Equation (9) assigns positive epistemic value to unprobed distractor objects $d \in \{1, \ldots, N_{\text{dist},l}\}$ because its internal transition model initially predicts a novel latent state $\hat{\mathbf{z}}_{t+1}$. Probing each of the $N_{\text{dist},l}$ distractors $m \geq 1$ times adds $m\, N_{\text{dist},l}$ non-transitioning actions to the H_l productive actions, giving $A_l = H_l + m\, N_{\text{dist},l}$ and $\text{RHAE}_l = (H_l/(H_l + m\, N_{\text{dist},l}))^2$. With `DeadSignatureTracker` ($\Psi_{\text{dead}} = \infty$ on $(\sigma(O_t), a_{\text{class}}) \in \mathcal{D}_{\text{dead}}$) and color-separated object extraction, any no-op action is immediately barred from re-selection, and static distractor color classes ($c \in \{5, 7\}$) are pruned globally across levels, yielding $W_l = 0$ wasted dead actions across all 15 levels. ■

B Hyperparameter and Bioenergetic Configuration Details

C Catalog and Provenance Audit of the 30 Neuro-Symbolic Operators

To address data-independence auditing between library design and blind evaluation, Table 9 categorizes all primitives and the 30 operators registered in `F3LearnedOperators` (`f3_learned_library.py`) by their exact development provenance:

Table 8 | Complete hyperparameter and bioenergetic constants for F3-AGI v4.0 across all benchmark suites.

Module / Subsystem	Parameter Symbol & Description	Value / Setting	Code Reference
Level 1: MSRC & SLS	Fast Perceptual / Slow Workspace Dimensions	256D / 512D (Powers of 2)	`VisualCortexV40`
	Background Color Inference Protocol	Joint mode across all (X_k, Y_k) pairs	`_build_depth1_pool`
	Isotropic Projection Normalization	ℓ_2-sphere scaled by $\sqrt{512}$	`AlgebraicWorldModelSLS`
Level 1.5: Bioenergetics	ATP Pool Bounds $[E_{\min}, E_{\max}]$	[0.10, 1.00]	`BioenergeticConfig`
	Basal Replenishment r_{base} / Depletion λ_{dep}	0.04 / 0.12	`MetabolicControllerV40`
	Plasticity Cost Weight α_{cost} / MDL Weight β_{MDL}	0.05 / 0.05	`RealistV40`
	Mistake Gating Margin τ_m (Continual / AACR)	0.18 / 0.05	`f3_metabolic_optim.py`
Level 2: SynGAP	Diagonal Fisher EMA Decay $(1 - \rho)$	0.95 ($\rho = 0.05$)	`compute_syngap_mask`
	Curvature Preconditioning γ_{syn} / Anchor λ_{anc}	12.0 / 4.5	`F3MetabolicOptimizer`
Level 3: AACR & EFE	Structural Weights $(w_{\text{sh}}, w_{\text{px}}, w_{\text{obj}})$	(0.20, 0.50, 0.30)	`RealistV40.evaluate`
	Pixel Foreground / Background Balance	70% FG / 30% BG	`RealistV40.evaluate`
	Cycle-3 Residual Patch Trigger Threshold	$S_{\text{shape}} = 1.0 \wedge \mathcal{A}_{\text{struct}} \geq 0.78$	`ResidualPatchSynthesizer`
	Dead-Signature Color Class Threshold τ_{dead}	2 inert probes $\rightarrow$ global block	`DeadSignatureTracker`

- **Tier 1 — Base Training-Set Priors (Horizons 1–2, 15 Operators):** Induced and consolidated exclusively on `ARC-AGI/data/training` before accessing any evaluation task (achieving $4/50 = 8.0\%$ on the initial evaluation slice).
- **Tier 2 — Calibration-Slice Parametric Extensions (Horizons 3 & 5, 15 Operators):** General parametric operators engineered by analyzing training demonstrations of failed or near-miss tasks strictly restricted to the first 50 evaluation tasks (**Calibration Slice**, $N = 50$), raising in-slice accuracy to 19/50 (38.00%).
- **Tier 3 — Strictly Uninspected Blind Hold-Out Evaluation ($N = 350$, Tasks 51–400):** With Tiers 1 and 2 frozen prior to testing, the 30-operator library + Cycle-3 `ResidualPatchSynthesizer` solved **38 / 350 tasks (10.86%)** on the remaining 350 never-inspected evaluation tasks (including 7 autonomous Cycle-3 residual patch solutions), yielding **57 / 400 (14.25% `pass@2`)** across the complete evaluation corpus.

Table 9 | Provenance audit and structural invariants of pure-NumPy primitives (`F3Primitives`) and the 30 operators in `F3LearnedOperators`.

Provenance Tier	Function Signature(s)	Topological / Algebraic Invariant	Uninspected Hold-Out Transfers ($N = 350$)
Tier 0: Core DSL (`F3Primitives`)	`crop_to_content`, `downsample_block`, `split_and_overlay` `fill_enclosed_holes`, `apply_gravity`, `flood_fill`	Bounding-box extraction, modal subsampling, overlay Border-BFS cavity closure, directional packing, BFS fill	General spatial primitives General topological primitives

Tier 1: Base Priors (Induced strictly on `data/training`)	`kronecker_self_pattern(grid)`
	`extend_2d_periodicity(grid, th, tw, mask)`
	`dihedral_template_matching(grid)`
	`split_and_boolean_combine(grid, axis, op)`
	`mirror_tile_2x2(grid)` & `complete_symmetry_*`
		Self-similar Kronecker tensor product $X \otimes \mathbb{I}(X > 0)$
		Vectorized 2D minimal period (p_r, p_c) induction & repair
		D_4 dihedral group template registration over anchors
		Multi-panel boolean algebra (`nor`, `or`, `xor`, `and`)
		Dihedral 2×2 reflection tiling & axis symmetrization
			`31d5ba1a`, `60c09cac`, `8e2edd66`
			`506d28a5`, `ca8de6ea`, `ea959feb`
			`5783df64`, `aa18de87`, `cd3c21df`
			`34b99a2b`, `5d2a5c43`, `e95e3d8e`
			`66f2d22f`, `833dafe3`, `bc4146bd`

Tier 2: Calibration (Engineered on $N = 50$ Calibration Slice training pairs only; frozen for $N = 350$)	`alternating_mirror_tile` & `fractal_conditional_kron`
	`fill_seeded_or_sized_cavities` & `fill_seeded_component_interiors`
	`symmetrize_per_object_or_frame` & `connect_aligned_markers`
	`denoise_subblock_consensus` & `grid_cell_pattern_stamp`
	`separator_fill`, `macro_cell_lookup`, `complete_partial_plus`
	`pack_rectangles_into_frame_holes`, `recolor_cavity_fitting_objects`
		Row/col parity reflection tiling & complementary Kronecker
		Cavity span-to-color mapping & seed-propagated interior fill
		Local bounding-frame symmetry & orthogonal ray precedence
		Majority-vote subblock denoising & macro-cell stamping
		Partition fill, motif lookup, and 5-cell cross completion
		Morphological hole fitting and quadrant bounding-box packing
			`5b6cbef5`, `be03b35f`, `c663677b`
			`68b67ca3`, `ca8f78db`, `ccd554ac`
			`9ddd00f0`, `ce8d95cc`, `d19f7514`
			`e133d23d`, `e1baa8a4`, `e345f17b`
			`358ba94e`, `f5aa3634`, `f823c43c`
			`9a4bb226` (`attempt_2`)

Cycle-3 Autonomous Residual Discovery	`ResidualPatchSynthesizer.synthesize_patch` (`ray_corridor_table`, `ring_radius_table`, `neighbor_3x3`)	Mask-guided local cellular/corridor lookup repair over M_{diff} Induces local rule strictly on frozen Cycle-1/2 anchor $\hat{Y}_{\text{st}}$	**7 Uninspected Tasks:** `21f83797`, `319f2597`, `516b51b7`, `84db8fc4`, `ce039d91`, `d37a1ef5`, `f0df5ff0`