Skepthical review: *Regional Brain Morphometry and Adaptive Foraging Reveal Age-Related Cognitive Flexibility and Resilience Trends in Egyptian Fruit Bats*

Summary

This manuscript investigates cognitive aging and putative neural correlates of cognitive resilience in Egyptian fruit bats using (i) a multi-phase spatial foraging task with derived metrics for learning efficiency, perseveration (short/long-term), shifting, and learning consolidation (Sec. 2.2), and (ii) atlas-based regional morphometry from $b = 0$ (b0) images of a DTI acquisition, extracting ROI volumes and mean b0 signal intensities across 24 regions (Sec. 2.3). DNA methylation age (DNAm age) is used as the primary aging axis, with sex and colony as covariates, and FDR correction applied to families of tests (Sec. 2.4). The most robust reported effects are (a) a negative association between DNAm age and short-term perseverative errors (older bats making fewer short-term perseverative errors; Sec. 3.2.1) and (b) a positive association between DNAm age and mean b0 intensity in ROI 14 (Sec. 3.3.1). A DNAm age × ROI 19 intensity interaction on Phase-1 learning consolidation (Correct_Box_Preference) is presented as suggestive of resilience but does not survive FDR correction (Sec. 3.4.2). The study is timely and potentially impactful due to its non-traditional, long-lived model and thoughtful behavioral decomposition, but key reporting inconsistencies (cohort N/demographics), insufficient operationalization of "resilience," unclear anatomical identification of ROIs, limited validation/normalization of b0 intensity as a quantitative marker, and gaps in methodological/statistical transparency currently prevent confident interpretation and reproducibility. Tightening cohort accounting, clarifying the moderation/multiplicity framework, improving MRI intensity handling, naming/localizing ROIs, and aligning claims with FDR-supported evidence would substantially strengthen the paper.

Strengths

- Important and timely question—cognitive aging and candidate resilience mechanisms—in a long-lived, ecologically relevant and underused mammalian model (Egyptian fruit bat) (Sec. 1).
- Behavioral task design is multi-phase and the derived metrics go beyond overall accuracy to probe learning efficiency, perseveration (short/long-term), shifting, and consolidation (Sec. 2.2.1–2.2.2).
- Creative use of available imaging (b0 volumes from DTI) to derive atlas-based regional volumes and intensities, with substantial visual QC documentation (Sec. 2.3; Figs. 2–17).

- Use of DNAm age as a biologically grounded aging marker, and inclusion of key covariates (sex, colony) with stated regression and moderation models (Sec. 2.1; Sec. 2.4).
- The age association with reduced short-term perseverative errors is robust (FDR-significant) and counter-intuitive, offering a concrete behavioral phenomenon for follow-up mechanistic work (Sec. 3.2.1).
- Clear ambition to separate confirmatory (main effects) from exploratory (brain–behavior interactions) analyses, and to control multiplicity via FDR (Sec. 2.4.2–2.4.3; Sec. 3.4).
- Figure 1 serves as an effective conceptual overview linking behavior, brain metrics, and interaction logic (though labeling/claims need tightening).

Major issues

1. **Cohort description is internally inconsistent across Abstract, Sec. 2.1/Table 1, Sec. 3.1, Sec. 4.2, and Fig. 1D (e.g., $N = 31$ vs $N = 33$; differing sex/colony counts; DNAm age ranges 6.6–15.1 vs 6.62–13.84/13.8). Sec. 3.1 also contains apparent template/table artifacts (e.g., irrelevant fields such as "New Surgeon," "Single (Award/Alter.)," and implausible age ranges). These inconsistencies make it unclear which animals contribute to which analyses and undermine all reported inferential statistics, especially interaction models that are highly sensitive to missingness and sample size.**

   Recommendation: Provide a single, definitive cohort accounting and propagate it consistently everywhere (Abstract; Sec. 2.1; Table 1; Sec. 3.1; Sec. 4.2; Fig. 1D). Include: starting N, exclusion criteria (behavioral incompletion, imaging QC failures, missing DNAm age), and final N per analysis type (behavior-only; brain-only; moderation). Add a concise CONSORT-style flow diagram or table listing per-metric N (important for Sec. 3.2–3.4). Remove template artifacts in Sec. 3.1 and replace with a clean demographic summary (mean±SD, min–max for DNAm age; sex/colony counts). If different Ns are unavoidable (e.g., some metrics undefined for some bats), report them explicitly for each regression in Sec. 3 and/or in Tables 2–4.

2. **The operational definition, statistical implementation, and strength of claims regarding "cognitive resilience" are currently misaligned with the evidence (Sec. 1; Sec. 2.4.3; Sec. 3.4.1–3.4.2; Sec. 4.3–4.4). The moderation equation is inconsistently written (interaction term not clearly shown in places), the moderation search space is ambiguous ("promising trends" vs "all pairs"), and the key ROI 19 interaction is uncorrected/non-FDR-significant but described with strong language ("compelling/strong/buffering/mitigated"), risking overinterpretation in a modest-N, cross-sectional design.**

Recommendation: Define cognitive resilience explicitly in Sec. 1 with citations (e.g., preserved function relative to age-related burden), and state clearly that here it is operationalized as a DNAmAge × brain-metric interaction (cross-sectional proxy, not longitudinal/clinical resilience). Correct and standardize the moderation model equation in Sec. 1 and Sec. 2.4.3 to include DNAmAge, BrainMetric, and DNAmAge× BrainMetric. In Sec. 2.4.3 and Sec. 3.4.1, specify exactly: (i) which behavior outcomes entered moderation, (ii) which brain metrics (24 ROIs × volume/intensity), (iii) whether any pre-screening occurred and if it was pre-registered or post hoc, (iv) the total number of interaction tests, and (v) the exact FDR family and procedure (e.g., BH at $q = 0.05$). Throughout (Abstract; Fig. 1C; Sec. 3.4.2; Sec. 4.3–4.4), clearly label ROI 19 as exploratory and report both raw and FDR-adjusted p-values; replace causal/buffering wording with correlational language and emphasize hypothesis-generation.

3. **Mean b0 signal intensity is treated as a region-specific biological marker without sufficient acquisition/preprocessing detail or controls for non-biological intensity scaling (Sec. 2.3.1–2.3.2; Sec. 3.3.1; Sec. 4.3). b0 intensity is not inherently quantitative and can vary with coil sensitivity/bias field, receiver gain, session/scanner differences, subject positioning, EPI distortions (if DTI-EPI), motion, and partial voluming. Without explicit intensity normalization and distortion/bias handling, ROI intensity effects (including ROI 14 and ROI 19) could reflect acquisition/processing artifacts or global scaling rather than aging biology.**

   Recommendation: Expand Sec. 2.3.1 to report key MRI acquisition parameters: scanner model and field strength, sequence type (EPI DTI?), TR/TE, voxel size, number of directions/b-values, number of b0 volumes, and whether parameters were identical across bats/sessions. In Sec. 2.3.2, document preprocessing affecting intensity comparability: brain extraction, bias-field correction, denoising, motion/eddy correction, EPI distortion correction (topup/fieldmap) or rationale if absent, and any intensity normalization (e.g., divide ROI mean by whole-brain mean/median; z-score within subject; histogram matching). Add robustness checks: (i) include global b0 intensity (whole-brain mean/median) as a covariate in ROI-intensity models or analyze relative intensity (ROI/global), and (ii) report whether the ROI 14 age effect remains after such adjustment. Temper mechanistic interpretations in Sec. 4.3 (gliosis/iron/water etc.) as speculative unless supported by quantitative MRI or histology; frame as hypotheses for follow-up.

4. **ROI labels are largely numeric (ROI 14/19 central to conclusions) without anatomical names, laterality, or localization (Sec. 2.3.2; Sec. 3.3.1; Sec. 3.4.2; Sec. 4.3–4.4). This prevents readers from evaluating biological plausibility, comparing to prior literature (e.g., hippocampal/striatal/cortical systems for navigation/flexibility), or interpreting why specific ROIs might relate to perseveration or consolidation.**

Recommendation: Add an atlas/ROI mapping table (preferably in Sec. 2.3.2 or Supplement with clear in-text pointers) listing ROI 1–24 with: anatomical name, laterality (if applicable), broad class (cortical/subcortical/white matter), and brief description of atlas provenance. Provide a figure showing the atlas in template space with ROI 14 and ROI 19 highlighted. In Sec. 3.3–3.4 and Discussion, refer to ROIs by anatomical names in addition to indices (e.g., "StructureName (ROI 14)"). If the atlas does not support reliable anatomical correspondence (e.g., composite parcels), state this limitation explicitly and correspondingly soften functional claims.

5. **Outcome distributions and modeling choices for behavioral metrics (counts, times, proportions) are not convincingly justified for linear regression assumptions (Sec. 2.2.2; Sec. 2.4.1–2.4.2; Sec. 3.2). Perseveration outcomes are small-range counts with likely zero inflation; latency measures are typically skewed/censored; Correct_Box_Preference is a proportion with variable denominators ("after first correct") and an unclear log transform that may be invalid if zeros occur. These issues can bias estimates/p-values and complicate interpretation of effect sizes.**

   Recommendation: In Sec. 2.2.2, provide explicit formulas for each metric (including Correct_Box_Preference) and specify handling of edge cases (e.g., no post-discovery entries; failure to find correct box). In Sec. 2.4, justify the modeling family per metric and report transformations precisely (e.g., $\log_{10}(x+1)$, logit, arcsine-sqrt). Strongly consider re-fitting key outcomes with appropriate models: negative binomial/Poisson (or zero-inflated) GLMs for count perseveration; survival/AFT or appropriately transformed time models for latencies; beta regression or binomial (successes/total) models for proportions like Correct_Box_Preference. At minimum, add sensitivity analyses showing that the main inferences (especially Perseverative_Errors_STM age effect and any ROI 19 interaction trend) are robust across reasonable alternative model families and/or transformations.

6. **Multiplicity handling and the "universe" of tests are not transparently defined, especially for moderation (Sec. 2.4.2–2.4.3; Sec. 3.4.1). The manuscript alternates between moderation on a subset of "promising trends" and an "extensive series across all brain–behavior pairs." Without an explicit count of tested hypotheses per family, readers cannot interpret FDR-adjusted results, nor judge the evidential strength of uncorrected p-values (e.g., $p = 0.004$ for ROI 19 interaction).**

   Recommendation: In Sec. 2.4, define separate hypothesis families and test counts (e.g., age→behavior; age→brain volumes; age→brain intensities; brain×age moderation per behavior metric). Report the exact number of tests included in each FDR correction and the procedure (e.g., BH). In Sec. 3.2–3.4, consistently label raw vs FDR-adjusted p-values and provide effect sizes with 95% CIs. If moderation tests were pre-screened, describe the screening rule, whether it used the same data (risking circularity), and

consider presenting both (i) a fully exploratory all-pairs analysis with stringent correction and (ii) a smaller, pre-specified hypothesis set (if defensible) analyzed confirmatorily.

7. **Regional volume analyses appear not to control for total brain size/intracranial volume, limiting interpretation of null or localized volumetric findings (Sec. 2.3.2; Sec. 3.3; Sec. 4.2). Without adjusting for total brain volume (or equivalent), regional volume associations may reflect global size differences (including sex/allometry) rather than region-specific effects.**

   Recommendation: Compute total brain volume (or intracranial volume, if feasible from the same atlas/mask) and either (i) include it as a covariate in ROI volume regressions, or (ii) analyze normalized volumes (ROI/total brain) and justify the choice. Report both absolute and size-adjusted results (at least in Supplement) and state clearly in Sec. 2.4.2 and Sec. 3.3 which approach is primary. Also clarify whether voxel dimensions/resampling affect volume computation (native vs template space) (Sec. 2.3.2).

8. **Core reproducibility details are missing or fragmented across behavioral, imaging, and statistical pipelines (Sec. 2.2–2.4), and key result tables are referenced but not adequately presented (Tables 2–4). The current description is insufficient for replication and for evaluating robustness (diagnostics, influential points, missingness patterns).**

   Recommendation: Strengthen Methods: (i) Sec. 2.2: arena geometry and box layout, phase durations and criteria, reward details, habituation/training, event coding, and explicit rules for aborted/incomplete phases; (ii) Sec. 2.3: atlas provenance and validation, registration direction and parameters (rigid/affine/nonlinear; cost function; interpolation), masking/erosion, and QC criteria; (iii) Sec. 2.4: software and package versions, covariate coding, missing-data handling (complete-case per model vs other), and routine diagnostics (residuals, heteroskedasticity, Cook's distance). Provide complete model outputs in Tables 2–4 (or Supplement): N, $\beta \pm$SE, 95% CI, (adjusted) R^2 or pseudo-R^2, and both raw and FDR p-values. Include at least leave-one-out or influence sensitivity for the headline effects (Perseverative_Errors_STM; ROI 14 intensity). State whether code/data (or a de-identified derivative) will be shared.

Minor issues

1. Figures and captions sometimes promote exploratory interpretations without clearly flagging multiplicity and FDR outcomes (notably Fig. 1C; also Figs. 19–21) and contain inconsistencies (Fig. 1D demographics; mismatched caption/panel counts in Fig. 19; potential line-direction/caption mismatch in Fig. 21).

   Recommendation: Revise Fig. 1C caption/labeling to explicitly state "exploratory (uncorrected)" and include both raw and FDR-adjusted p-values for the interaction; align Fig. 1D with corrected cohort demographics. For Figs. 19–21, reconcile captions with

plotted content, add 95% CI ribbons, show raw data points, annotate N per panel, specify whether variables are transformed and whether plots reflect covariate-adjusted partial effects, and use a colorblind-safe palette.

2. Behavioral metric definitions leave important ambiguities (Sec. 2.2.2): what exactly counts as an "entry," how "first five entries" is anchored (phase start vs first entry), whether entries include correct+incorrect, and how missing/censored cases are treated (e.g., failure to find the correct box within five entries or within phase duration).

   Recommendation: Add explicit operational definitions and edge-case handling in Sec. 2.2.2 (and/or a small table). Report per-metric missingness in Sec. 3.1/3.2 and assess whether missingness correlates with DNAm age, sex, or colony.

3. Interpretation of the robust behavioral finding (older bats show fewer short-term perseverative errors) risks being overly narrow (Sec. 4.1–4.2). Reduced "perseveration" could also reflect altered exploration strategy, motivation, or activity levels with age.

   Recommendation: Add analyses or descriptive controls related to overall activity/exploration (e.g., total entries, movement/flight counts if available, time-in-arena) and discuss alternative interpretations. Where possible, show whether age effects persist after controlling for overall entry rate/engagement.

4. Potential batch/session confounds are not addressed for DNAm age and/or MRI (Sec. 2.1; Sec. 2.3; Sec. 2.4). With modest N, even mild batch effects can generate apparent age or ROI effects.

   Recommendation: Clarify whether DNAm assays were performed in one batch or multiple (and whether batch was modeled). For MRI, report whether scanning occurred across multiple sessions/days and whether acquisition parameters changed; consider adding session/batch covariates if applicable or at least provide descriptive plots showing DNAm age distribution by sex/colony/session.

5. QC presentation is heavy in the main text (Figs. 2–17) and may distract from primary results; these figures are largely repetitive.

   Recommendation: Move most registration QC panels to Supplement, keeping 1–2 representative examples in the main text plus a quantitative summary (e.g., overlap/Dice metrics, registration cost values, or failure rates) to document pipeline performance more efficiently.

6. Ethics and animal welfare procedures are not clearly stated despite behavioral testing and MRI (Sec. 2.1–2.3).

   Recommendation: Add an explicit ethics statement (end of Sec. 2.1): committee approvals/permit numbers, anesthesia/monitoring for MRI, recovery procedures, and steps taken to minimize stress.

7. Keywords under the Abstract appear to be template artifacts unrelated to the topic (e.g., “Astronomy data reduction”).

   Recommendation: Replace with domain-appropriate keywords (e.g., cognitive aging; cognitive flexibility; cognitive resilience; Egyptian fruit bat; DNA methylation age; spatial foraging; DTI; brain morphometry; atlas-based ROIs).

Very minor issues

1. Notation is inconsistent across the manuscript (e.g., DNAm age vs DNAmAge vs DNAm_age; “b0” vs “b=0”; inconsistent “Phase 1/phase 1”), and some equations appear incompletely typeset (especially moderation equations) (Sec. 1; Sec. 2.4).

   Recommendation: Standardize variable naming and terminology throughout, and ensure all equations compile and explicitly include intercept, main effects, and interaction term where relevant (Sec. 1; Sec. 2.4).

2. Multiple typographical/formatting artifacts appear (line-break word splits; duplicated “Figure X” in captions; remnants like “Project Summary/Total Cohort Summary”) (various sections; figure captions).

   Recommendation: Proofread from the compiled manuscript; remove template artifacts; standardize caption format; and ensure section heading formatting is consistent with journal style.

3. Standardization/transformations are not always reflected in interpretability of coefficients (Sec. 2.4.1; Sec. 3.2–3.4): it is unclear whether reported β are per 1 SD of DNAm age or per year, and whether outcomes were transformed/standardized.

   Recommendation: State explicitly which predictors were z-scored, whether outcomes were transformed and/or standardized, and interpret key β accordingly (e.g., per 1 SD DNAm age). Consider adding a compact transformation table in Methods or Supplement.

4. Voxel-volume and ROI-volume computation details are potentially confusing (Sec. 2.3.2): it is unclear whether volumes are computed in native space with native voxel dimensions or after resampling in template space, which affects mm^3 calculations.

   Recommendation: Clarify the space in which voxel counting occurs (native vs resampled), the voxel dimensions used for mm^3 conversion, and confirm consistency across subjects after any resampling/registration.

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains light mathematics focused on defining behavioral metrics, computing ROI volumes/intensities from imaging data, and specifying linear regression and moderation (interaction) models with covariates and FDR correction. There are no multi-step analytical derivations; the key audit points are definition consistency, dimensional consistency, and consistency of statistical-model descriptions across sections.

Checked items

1. ✔ **Moderation (resilience) regression model specification** (Introduction (end of p.2) and Sec. 2.4.3, p.5)
    - **Claim:** Behavioral_Metric is modeled as a linear function of DNAm age, a brain metric, their interaction, plus covariates (sex, origin colony).
    - **Checks:** symbol/definition consistency, algebra/structure sanity-check
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Linear model with an interaction term DNAmAge× BrainMetric, Covariates are additive main effects
    - **Notes:** Model form Behavioral $\sim \beta_0 + \beta_1 \cdot \mathbf{Age} + \beta_2 \cdot \mathbf{Brain} + \beta_3(\mathbf{Age} \times \mathbf{Brain})$ + covariates is internally coherent. Notation varies slightly (DNAmAge vs DNAm age; Brain Metric vs Regional_Brain_Metric) but does not change the algebraic meaning.
2. ✔ **Main-effect regression models for age effects** (Sec. 2.4.2, p.5)
    - **Claim:** Separate multiple linear regressions are run for each behavioral metric and each brain metric with predictors DNAmAge + Sex + Origin_Colony.
    - **Checks:** model structure sanity-check, symbol/definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Sex and origin colony are encoded as covariates, Each metric is modeled separately (univariate outcomes)
    - **Notes:** The stated model structures are consistent and suitable for the described screening of age effects; no algebraic issues.
3. ⚠ **Predictor standardization (z-scoring) vs interpretation of coefficients** (Sec. 2.4.1, p.5; Results Sec. 3.2.1 (p.6) and Sec. 3.3.1 (p.7))
    - **Claim:** All continuous predictor variables (including DNAm age and brain metrics) are standardized before regression; reported β coefficients summarize effects.
    - **Checks:** definition consistency, units/scale interpretability
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** z-scoring applies to predictors only (as stated), Outcomes may be transformed but not necessarily standardized

- **Notes:** If predictors are z-scored, β corresponds to change in outcome per 1 SD increase in predictor, not per 1 year of age. The paper interprets effects qualitatively (older vs younger), which is compatible, but does not clarify whether reported β are from standardized-predictor fits, whether DNAm age was z-scored in the final reported models, and whether outcomes were standardized. This prevents a fully consistent interpretation audit.

4. ✔ **Voxel volume calculation and ROI volume dimensional consistency** (Sec. 2.3.2, p.4)
    - **Claim:** ROI volume = (number of ROI voxels)$\times(0.5\,\mathrm{mm} \times 0.5\,\mathrm{mm} \times 1.0\,\mathrm{mm} = 0.25\,\mathrm{mm}^3)$ yielding mm^3.
    - **Checks:** dimensional/units, arithmetic sanity-check
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Voxel dimensions are 0.5 mm, 0.5 mm, 1.0 mm, Counting voxels is done in a space with those voxel dimensions
    - **Notes:** $0.5 \times 0.5 \times 1.0 = 0.25$ and units multiply to mm^3, so the volume computation is dimensionally correct.

5. ⚠ **b0 averaging operation** (Sec. 2.3.1, p.4)
    - **Claim:** The first three volumes are $b = 0$ images and are voxel-wise averaged to create a single 3D b0 image.
    - **Checks:** definition completeness, algebra/operation sanity-check
    - **Verdict:** UNCERTAIN; confidence: low; impact: minor
    - **Assumptions/inputs:** First three 4D volumes are indeed $b = 0$, Voxel-wise average is computed as $(I_1 + I_2 + I_3)/3$
    - **Notes:** Averaging is mathematically straightforward, but the document provides no acquisition table or explicit indicator that the first three volumes are $b = 0$; this is an empirical/metadata dependency not verifiable from the provided text.

6. ✔ **Behavioral metric: Entries_to_First_Correct definition** (Sec. 2.2.2, p.3)
    - **Claim:** Entries_to_First_Correct counts incorrect box entries before the first correct entry in Phase 1.
    - **Checks:** definition consistency, sanity bounds
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** A well-defined ordering of entries exists, The first correct entry exists for included subjects (else missing)
    - **Notes:** Count is nonnegative integer; definition is internally consistent and matches intended learning-efficiency notion.

7. ✔ **Behavioral metric: Perseverative_Errors_STM bounds** (Sec. 2.2.2, p.3)

- **Claim:** Perseverative_Errors_STM is the number of entries into Phase-1 correct box within the first five total entries of Phase 2.
- **Checks:** sanity bounds, definition consistency
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** At least one entry occurs in Phase 2 for included subjects; otherwise metric missing
- **Notes:** This count is bounded between 0 and 5 inclusive (or $\mathbf{0..}n$ if fewer than 5 entries exist; handling of fewer-than-5 is not stated but does not create a contradiction).

8. ✔ **Behavioral metric: Perseverative_Errors_LTM bounds** (Sec. 2.2.2, p.3)
    - **Claim:** Perseverative_Errors_LTM counts entries into either Phase-1 or Phase-2 correct boxes within the first five entries of Phase 3.
    - **Checks:** sanity bounds, definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Phase 3 entries exist for included subjects, else missing
    - **Notes:** Also bounded by $\mathbf{0..5}$ (or $\mathbf{0..}n$ if fewer than 5 entries). Definition is consistent with a longer-term perseveration construct.
9. ⚠ **Learning consolidation (Correct_Box_Preference) and log transform** (Sec. 2.2.2, p.4; Results Sec. 3.2.2, p.7)
    - **Claim:** Correct_Box_Preference is a proportion of correct-box entries relative to total entries, computed only after first correct entry; sometimes log-transformed.
    - **Checks:** definition consistency, domain/transform validity
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** After first correct entry, at least one entry is counted in numerator and denominator (so preference $> \mathbf{0}$), No zeros are logged
    - **Notes:** If computed 'after the first successful visit' but excluding that first visit, it is possible to have zero post-discovery entries, making the proportion undefined ($\mathbf{0/0}$) or $\mathbf{0/0}$-like depending on implementation. The text does not clarify inclusion/exclusion of the first correct entry in the post-discovery window and does not specify edge-case handling. The log transform requires strictly positive values; likely satisfied if at least one correct entry is included, but not provable from the description.
10. ✔ **FDR correction family for brain metrics** (Sec. 2.4.2, p.5; Results Sec. 3.3.1, p.7)
    - **Claim:** FDR correction is applied across 'all 24 atlas-defined regions and both morphometric measures' when testing age effects on brain metrics.
    - **Checks:** definition consistency

- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** Two measures per ROI: volume and mean intensity (48 tests)
- **Notes:** The described correction universe ($\mathbf{24 \times 2}$) is internally consistent with the feature matrix description. Exact details (whether additional global metrics are included) are not specified but no direct contradiction appears.

11. ✖ **Cohort size and demographic summary consistency** (Methods Sec. 2.1/Table 1 (p.3) vs Results Sec. 3.1 (p.5–6) and Abstract (p.1))
    - **Claim:** The analytic cohort is described consistently across the paper.
    - **Checks:** definition consistency, cross-section consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Single final cohort is used for all downstream analyses unless otherwise stated
    - **Notes:** Methods reports final matched cohort $N = \mathbf{31}$ with specific sex/origin counts and DNAm age range up to 15.1 years, whereas Results states $N = \mathbf{33}$ with different counts and a different DNAm age range (max 13.84). Abstract also states 33 bats and range 6.6–13.8. This is an internal consistency break affecting all subsequent statistical claims and the meaning of reported coefficients/p-values.
12. ✖ **Moderation-testing scope vs FDR statement consistency** (Sec. 2.4.3, p.5 vs Results Sec. 3.4.1, p.8)
    - **Claim:** Moderation analyses were run on a clearly specified set of brain–behavior pairs and FDR was applied accordingly.
    - **Checks:** cross-section consistency, multiple-testing definition consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Either all pairs were tested, or a subset was tested based on a pre-screen rule
    - **Notes:** Methods claims moderation applied to 'brain–behavior pairs that showed promising trends in the initial analyses', while Results describes an extensive series 'across all brain-behavior pairs'. These are materially different analysis plans and imply different FDR correction families; the adjusted-significance conclusion depends on which is true.

Limitations

- The provided content contains no explicit numbered equations, no full regression-output tables (Tables 2–4 not included in the parsed text), and no formal derivation steps, limiting verification to stated formulas/definitions and cross-section consistency.
- Several verifications depend on implementation choices (e.g., exact computation window for Correct_Box_Preference, handling of phases with $< \mathbf{5}$ entries, exact set of tests included in each FDR family) that are not fully specified in the text; these are

marked UNCERTAIN where they block analytic validation.

- Image-based figures are present but do not add verifiable symbolic mathematics beyond captions; numeric values in plots were not audited per scope.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

11 numeric checks were executed: 9 PASS and 2 FAIL. Failures were cross-section consistency issues (cohort N mismatch; DNAm age range mismatch between Table 1 and Results). Internal arithmetic/sanity checks (category totals, voxel-volume multiplication, mean-within-range) and one logical multiple-testing check (FDR $p \geq$ raw p) passed.

Checked items

1. ✔ **C1** (Page 3, Table 1 (Characteristics of the Final Matched Cohort))
    - **Claim:** Sex Distribution: 17 Males, 14 Females; Total Subjects (N) 31.
    - **Checks:** parts_sum_to_total
    - **Verdict:** PASS
    - **Notes:** Checked sum(parts) equals total.
2. ✔ **C2** (Page 3, Table 1 (Characteristics of the Final Matched Cohort))
    - **Claim:** Origin Colony: 18 Aseret, 13 Herzeliya; Total Subjects (N) 31.
    - **Checks:** parts_sum_to_total
    - **Verdict:** PASS
    - **Notes:** Checked sum(parts) equals total.
3. ✔ **C3** (Page 4, Methods 2.3.2 (Regional Volume calculation))
    - **Claim:** Voxel volume computed as $0.5\,\mathrm{mm} \times 0.5\,\mathrm{mm} \times 1.0\,\mathrm{mm} = 0.25\,\mathrm{mm}^3$.
    - **Checks:** unit_arithmetic
    - **Verdict:** PASS
    - **Notes:** Computed xyz and compared to stated.
4. ✖ **C4** (Page 5, Results 3.1 vs Page 3, Table 1 and Page 1, Abstract)
    - **Claim:** Cohort size inconsistently stated: Table 1 says final matched cohort $N = 31$, Results 3.1 says $N = 33$, Abstract says study investigated 33 bats.
    - **Checks:** cross_section_constant_consistency
    - **Verdict:** FAIL
    - **Notes:** Checked all reported constants match exactly.
5. ✖ **C5** (Page 5, Results 3.1 vs Page 3, Table 1)

- **Claim:** DNAm age range differs: Table 1 gives range 6.6–15.1 years; Results 3.1 gives range 6.62–13.84 years.
- **Checks:** cross_section_range_consistency
- **Verdict:** FAIL
- **Notes:** Compared range endpoints between sections using abs_tol.

6. ✔ **C6** (Page 1, Abstract vs Page 10, Conclusions 4.2 vs Page 5, Results 3.1)
   - **Claim:** DNAm age range wording: Abstract says 6.6–13.8 years; Conclusions say 6.6–13.8 years; Results 3.1 says 6.62–13.84 years.
   - **Checks:** cross_section_range_consistency
   - **Verdict:** PASS
   - **Notes:** Rounded Results range to 1 decimal and compared to Abstract/Conclusions.
7. ✔ **C7** (Page 5, Results 3.1)
   - **Claim:** Sex distribution included 21 males and 12 females for cohort $N = 33$.
   - **Checks:** parts_sum_to_total
   - **Verdict:** PASS
   - **Notes:** Checked sum(parts) equals total.
8. ✔ **C8** (Page 6, Results 3.1)
   - **Claim:** Origin colonies: 18 from Aseret and 15 from Herzeliya for cohort $N = 33$.
   - **Checks:** parts_sum_to_total
   - **Verdict:** PASS
   - **Notes:** Checked sum(parts) equals total.
9. ✔ **C9** (Page 5, Results 3.1 (DNAm age summary))
   - **Claim:** DNAm age: mean 9.43 years, SD 1.59 years, range 6.62 to 13.84 years.
   - **Checks:** range_contains_mean_plausibility
   - **Verdict:** PASS
   - **Notes:** Sanity check: mean within range; SD>0; max$\geq$min.
10. ✔ **C10** (Page 3, Table 1 (DNAm Age summary))
    - **Claim:** DNAm Age (Years): Mean 9.7, SD 1.9, Range 6.6–15.1.
    - **Checks:** range_contains_mean_plausibility
    - **Verdict:** PASS
    - **Notes:** Sanity check: mean within range; SD>0; max$\geq$min.
11. ✔ **C11** (Page 6, Results 3.2.1)

- **Claim:** Perseverative_Errors_STM: $\beta = -0.91$, $p = 0.004$, FDR-adjusted $p = 0.043$.
- **Checks:** p_value_ordering
- **Verdict:** PASS
- **Notes:** Checked adjusted $p \geq$ raw p (typical BH/FDR).

Limitations

- Tables 2–4 are referenced but not included in the provided parsed text, preventing fast recomputation/verification of most statistical claims.
- No raw subject-level behavioral or neuroimaging data are present, so regression coefficients, p-values, and FDR adjustments cannot be audited beyond basic logical checks (e.g., adjusted p not smaller than raw p).
- Figures are present as images, but numeric extraction from plots/pixels is out of scope per instructions.