Skepthical review: *3. RESULTS*

Summary

The manuscript examines "neuro-cognitive resilience" in the long-lived Egyptian fruit bat by integrating (i) a three-phase spatial foraging/re-learning task (Sec. 2.1.3, Sec. 2.2), (ii) epigenetic age (DNAmAge; Sec. 2.1.1, Sec. 2.3.1), and (iii) diffusion MRI mean diffusivity (MD) globally and across **24** atlas ROIs (Sec. 2.1.4, Sec. 2.3.2–2.3.3). The authors construct phase-wise behavioral scores and a composite Cognitive Flexibility Index (CFI; Sec. 2.2.2), then define a Cognitive Resilience Score (CRS) as residuals from $\mathbf{CFI} \sim \mathbf{DNAmAge} + \mathbf{Sex} + \mathbf{Origin}_{\text{colony}}$ (Sec. 2.3.1). In $N = 41$ bats with behavioral data (and an imaging subset reported as $N = 33$), the Results indicate (a) no significant association between CFI and DNAmAge (Sec. 3.2), and (b) no significant associations between CRS and global or regional MD after FDR correction (Sec. 3.4.1–3.4.2).

The study is timely and potentially impactful—multi-modal aging work in bats is rare—and the manuscript is generally well organized and commendably transparent about null findings. However, several core elements currently limit interpretability and reproducibility: key methodological details are missing for the epigenetic clock and the DTI pipeline; the behavioral metrics/CFI contain ambiguities (including a mathematically unclear adaptation-score formula) and likely suffer from restricted dynamic range/ceiling effects (Sec. 3.2, Fig. 3), undermining power to detect associations; the residualization approach (CRS) is under-justified and may obscure more interpretable brain–behavior relationships; and there are internal inconsistencies in sample sizes (e.g., "**28** MD maps" vs $N = 33$). Addressing these points would substantially strengthen the paper's ability to support (or appropriately qualify) claims about preserved cognitive flexibility and brain–cognition "decoupling" in this age range.

Strengths

- Ambitious multi-modal dataset in a non-traditional, longevity-relevant species, combining behavior, epigenetic aging, and diffusion MRI (Sec. 2.1.1).
- Clear overarching question and a task designed to probe re-learning/reversal-like components of spatial behavior (Sec. 1, Sec. 2.1.3, Sec. 2.2).
- Transparent reporting of null results and use of multiple-comparisons control for ROI analyses (Sec. 2.3.3, Sec. 3.4.2).
- The attempt to separate raw performance (CFI) from age-adjusted performance (CRS) is conceptually reasonable and common in resilience frameworks (Sec. 2.3.1), even if it needs clearer justification and implementation details here.
- Global and regional analyses are both provided, which is helpful for triangulating whether effects are diffuse or localized (Sec. 3.4.1–3.4.2).

Major issues

1. **Epigenetic age (DNAmAge) estimation is insufficiently documented and not validated in this cohort, limiting interpretability of both the "age" predictor and the CRS residualization that conditions on it (Sec. 2.1.1, Sec. 2.3.1, Sec. 3.2–3.3, Sec. 4).** The variable name suggests a skin-based clock (e.g., `DNAmAgeBat.Rousettus.aegyptiacus_Skin`), but tissue provenance, assay platform, clock training/coverage, expected error (MAE), preprocessing, and the DNAmAge–chronological age relationship in this sample are not reported.

   Recommendation: Add a dedicated Methods subsection describing: tissue source and collection timing relative to behavior/MRI; methylation assay and preprocessing/normalization; whether the clock is published or newly trained (training set/species, CpG count, model type); and clock performance (correlation, MAE) including a DNAmAge vs chronological-age plot for this cohort. Report chronological age summary stats (range/mean/SD) and its correlation with DNAmAge (Sec. 3.1), and discuss how clock uncertainty could attenuate associations (Sec. 4).

2. **Behavioral metric definitions and the construction of CFI are central but currently ambiguous/inconsistently specified, compromising reproducibility and potentially validity (Sec. 2.2.1–2.2.2).** In particular: (i) the adaptation-score equations are mathematically unclear because the operator with \text{Time_to_First_Correct} is missing (division vs multiplication; Sec. 2.2.1, p. 4); (ii) short- and long-term adaptation scores are repeatedly mislabeled (both as \text{Adaptation_Score_P1}); (iii) \text{Time_to_First_Correct_P2} is listed but it is unclear how/if it enters CFI; (iv) edge cases (\text{Total_Visits}=0; never finding correct; **time** = 0; capping at **10800** s) and censoring are not handled explicitly; and (v) mixing error fractions with time components creates unit/scale issues that may drive the z-scored composite.

   Recommendation: Rewrite Sec. 2.2.1–2.2.2 with formal, unambiguous definitions (symbols, units, allowed ranges) for every raw measure and derived score; explicitly state the implemented adaptation formula (e.g., (1-\text{PE}/\text{TV})/\text{Time_to_First_Correct} vs (1-\text{PE}/\text{TV})\times(1/\text{Time_to_First_Correct})) and ensure the text matches code. Fix naming/labeling (Adaptation_Score_STM vs _LTM consistently). Specify guardrails for division-by-zero and define how "no correct visit" is treated (censoring vs imputation to **10800** s), including counts per phase. Add a compact table summarizing all components and sign conventions (higher=better), and (ideally) provide pseudocode or a script in Supplementary Materials.

3. **Behavioral outcomes appear to have restricted dynamic range/ceiling or floor effects, which can easily produce null associations even if true effects exist (Sec. 3.2; Fig. 3).** The manuscript notes concentrations/limited variability

but does not quantify the extent (e.g., proportion of zero errors, tied times, or near-ceiling learning scores). This undermines the interpretability of null DNAmAge–CFI and CRS–MD results.

Recommendation: In Sec. 3.2 (and/or a Supplement), report descriptive statistics for each raw component and phase score (mean/SD, median/IQR), plus proportions at bounds (e.g., % with **0** perseverative errors; % with no incorrect visits; % censored at **10800** s). Provide histograms/density plots for key components and a correlation matrix among component scores. Consider reporting an internal-consistency/reliability check for the composite (e.g., correlations among components; a simple omega/alpha with appropriate caveats) to justify CFI as a unified construct.

4. **Use of CRS as in-sample regression residuals is under-motivated and can obscure more direct, interpretable brain–behavior relationships (Sec. 2.3.1–2.3.3, Sec. 3.3–3.4).** Stating CRS is "uncorrelated with predictors by design" is a mathematical property of OLS residuals (with intercept) rather than evidence of "age-independent cognition." If the $\mathbf{CFI} \sim \mathbf{DNAmAge}$ model is weak/non-significant, CRS may be nearly identical to CFI but noisier, and the two-stage procedure creates a generated-regressor setup that is not discussed.

   Recommendation: Clarify in Sec. 2.3.1 exactly how predictors were encoded (Sex/Origin reference levels; intercept; centering/scaling of DNAmAge), which N was used to fit the CRS model, and which diagnostics/outlier checks were applied. In Sec. 3.4, add a primary or parallel "single-step" model that directly tests brain–behavior links with covariates, e.g., $\mathbf{CFI} \sim \mathbf{MD} + \mathbf{DNAmAge} + \mathbf{Sex} + \mathbf{Origin}_{colony}$ (and similarly for ROI MD with multiplicity correction). If CRS remains central, reframe it as an "age-adjusted performance residual" with limitations, and consider out-of-sample residualization (e.g., cross-validation) as a robustness check.

5. **DTI acquisition, preprocessing, registration, and QC are too sparsely described to evaluate validity or enable replication (Sec. 2.1.4, Sec. 2.3.2).** The current description suggests matching atlas and MD-map dimensions, but dimension matching does not ensure anatomical alignment. MD is also sensitive to partial volume (CSF contamination), which is particularly relevant for small brains and ROI averages.

   Recommendation: Expand Sec. 2.1.4 to include scanner/sequence details (field strength; b-values; number of directions; voxel size; TR/TE; anesthesia/motion context), preprocessing steps (denoising, motion/eddy-current correction, susceptibility distortion correction, brain extraction), tensor-fitting software/algorithm, and explicit atlas registration (linear/nonlinear, reference space, QC criteria). Report QC/exclusion criteria and whether ROI erosion or tissue masking was used to mitigate partial volume. Provide enough detail that another lab could reproduce MD maps and ROI means.

6. **ROI analysis strategy is likely underpowered and difficult to interpret with $N \approx 33$ and 24 ROIs, especially after FDR, and currently lacks effect-size uncertainty reporting (Sec. 3.4.2).** Additionally, the ROI models use \text{CRS} \sim \text{Regional_MD_Value} without scan-quality covariates; residualization handles DNAmAge/sex/colony but not imaging confounds (motion/SNR, session effects).

   Recommendation: For Sec. 3.4.1–3.4.2, report standardized effect sizes and confidence intervals for global and ROI regressions (not only p/q-values), and consider robust regression or bootstrap CIs as a sensitivity check. If available, include scan QC covariates (e.g., motion, SNR, outlier counts). To reduce multiplicity and increase interpretability, consider (i) a priori ROIs motivated by spatial-memory circuitry, and/or (ii) dimensionality reduction (e.g., PCA of ROI MD) with pre-specified components.

7. **Interpretation of null results as evidence for preserved cognitive flexibility and "decoupling" from microstructural integrity is currently stronger than warranted given limited N, multiple comparisons, restricted behavioral range, and potentially modest age span (Sec. 3.5, Sec. 4).** With the current design, null results are consistent with both biological resilience and insufficient sensitivity/type-II error.

   Recommendation: Temper claims in Sec. 3.5 and Sec. 4 and explicitly present detectable-effect considerations: provide approximate minimum detectable correlations for $N = 33$ (global MD) under the chosen thresholds, and discuss how ceiling/restricted range attenuates observed effects. Add a clearly labeled limitations paragraph distinguishing (i) what the study can conclude (within sampled ages and task sensitivity) from (ii) what it cannot. Optionally include a small simulation/post-hoc sensitivity analysis showing the effect size range that would likely be missed.

8. **Internal inconsistencies in sample sizes and analysis flow (and therefore in derived quantities such as CRS) reduce confidence in the reported results: (i) DTI processing mentions 28 MD NIfTI files while the main analysis states $N = 33$ with DTI; (ii) Methods imply CRS fit on "subjects with complete data," but Results report $F(3, 37)$ for CFI regression ($N = 41$), leaving unclear which sample generated the CRS used in DTI analyses (Sec. 2.3.1; Sec. 3.2–3.3).**

   Recommendation: Add a concise CONSORT-style flow (or table) specifying N at each stage (behavioral inclusion; DNAmAge availability; MRI availability; QC exclusions), and ensure all Ns match across text/figures. State explicitly whether the CRS residualization model is fit on $N = 41$ or $N = 33$, and (ideally) refit on the imaging subset if CRS is used as the dependent variable in imaging analyses (or justify using the full cohort and explain implications).

9. **The behavioral task description is too sparse to assess whether the operationalization truly reflects "cognitive flexibility" versus exploration strategy, motivation, or procedural artifacts (Sec. 2.1.3; Introduction; Sec. 4).** Key details are missing: apparatus geometry/box count and spacing, cue availability, phase durations and retention intervals (STM vs LTM), counterbalancing of correct locations, individual vs group testing, reward schedule, and motivational state (e.g., food restriction).

   Recommendation: Expand Sec. 2.1.3 with a task description from the animal's perspective: layout, number of options, cues, trial structure, phase start/stop rules, STM/LTM interval lengths, counterbalancing/randomization, and reward/motivation procedures. This will also help justify the mapping from perseverative errors and visits to "flexibility" in Sec. 1 and Sec. 4.

Minor issues

1. Atlas/ROI interpretability is limited because ROIs are referred to largely as "ROI 1–24" without anatomical names or lateralization (Sec. 2.1.4, Sec. 3.4.2; figures with ROI heatmaps). This prevents biological interpretation even of null findings (e.g., whether hippocampal formation analogs show any trend).

   Recommendation: Add a table (main text or Appendix) mapping ROI index → anatomical label, hemisphere, and brief functional relevance. Use anatomical names alongside indices in Sec. 3.4.2 and figure captions.

2. Figures are frequently hard to read and/or under-annotated (e.g., small fonts, unclear units, generic labels like "Age Distribution," ROI indices only), and some content appears redundant (e.g., overlap between Figures **7** and **8**). Several figures/captions lack sample sizes per panel and do not show uncertainty or effect sizes (Sec. 3.2–3.4; multiple figures).

   Recommendation: Increase resolution and font sizes (prefer vector exports), standardize axis units/labels, include N per panel, and add uncertainty visualizations (CIs) for key associations. Consolidate redundant figures or clearly differentiate their roles.

3. The manuscript does not clearly report correlations among CFI components, nor whether z-scoring was performed using the full $N = 41$ or only complete cases (Sec. 2.2.2, Sec. 3.2). This affects reproducibility and interpretation of the composite metric.

   Recommendation: State explicitly the reference sample used for z-scoring and provide a component correlation matrix. Consider reporting results for each component score alongside CFI (exploratory) to show whether null results are driven by a specific phase.

4. Statistical reporting would be more informative with standardized predictors and effect sizes. For example, the large coefficient reported for **CRS Global MD** suggests MD is on a very small numerical scale; units for MD are not clearly reported (Sec. 3.4.1).

   Recommendation: Report MD units (typically $\mathbf{mm^2/s}$) and typical value ranges; consider standardizing MD (z-score) so coefficients are interpretable and comparable across ROIs; report β, SE, CI, and p/q.

5. Data/code availability is unclear despite references to specific filenames and variable names (Sec. 2.1.1–2.1.5, Sec. 2.3).

   Recommendation: Add a Data and Code Availability statement specifying what will be shared (behavioral data, DNAmAge values, atlas labels, analysis scripts) and where; if sharing is restricted, provide a minimal reproducible scaffold (synthetic data + code) and explicit access procedure.

6. The framing around "microstructure" and "neural correlates" may overpromise relative to analyzing only MD (a single diffusion scalar) (Abstract, Introduction, Sec. 4).

   Recommendation: Explicitly justify the focus on MD and note that other diffusion scalars (FA, AD/RD) or higher-order models could detect effects MD misses. Align title/abstract wording with the actual scope.

Very minor issues

1. Apparent placeholder or inappropriate affiliation/header text ("Anthropic, Gemini & OpenAI servers. Planet Earth.") is not suitable for a scientific manuscript.

   Recommendation: Replace with correct author affiliations and remove placeholder text.

2. Typos/encoding artifacts and inconsistent naming reduce professionalism and can confuse readers (e.g., "Mean Diffusiv ity", "Re silience"; "Herzeliya" vs "Herzliya"; "Perseverative" vs misspellings; "Origin, colony" vs "Origin_colony"; broken line breaks such as "as\npects"; inconsistent "CF I" vs "CFI") (Sec. 1–3 and figures).

   Recommendation: Thoroughly proofread and fix encoding/hyphenation issues; standardize variable naming and place names across text, equations, and figures; re-render equations with consistent LaTeX.

3. Some descriptive percentages appear inconsistent with stated counts (e.g., colony and sex percentages vs $N = 41$), and epigenetic age summary statistics differ between sections beyond simple rounding (Sec. 2.1.1, Sec. 3.1, Conclusions).

   Recommendation: Recompute and harmonize all descriptive statistics across Methods/Results/Conclusions, and clearly indicate which subset each statistic refers to (full cohort vs imaging subset).

4. Model notation uses formula shorthand (tilde syntax) but does not always explicitly state intercept inclusion and categorical reference levels; some interpretive statements rely on OLS properties (Sec. 2.3, Sec. 3.3).

   Recommendation: Explicitly state intercept inclusion and categorical encodings/reference categories wherever models are introduced.

Mathematical consistency audit

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

Maths relevance: light

The paper's mathematics is primarily the definition of composite behavioral scores (phase scores, CFI via z-scored averaging) and residualization via linear regression to define CRS, followed by simple univariate regressions of CRS on MD measures (global and ROI-wise) with FDR correction. There are no extended derivations; internal consistency hinges on clear, unambiguous score formulas and consistent definitions of which cohort is used for each regression/residualization step.

Checked items

1. ✔ **Phase-1 learning score definition** (Learning_Score_P1 definition, Sec. 2.2.1, p. 4)
   - **Claim:** Defines initial learning efficiency as \text{Learning_Score_P1}=1 - \frac{\text{Incorrect_Visits_P1}}{\text{Total_Visits_P1}}, with higher values better.
   - **Checks:** definition consistency, range/bounds sanity check
   - **Verdict:** PASS; confidence: medium; impact: moderate
   - **Assumptions/inputs:** \text{Total_Visits_P1} > 0, \text{Incorrect_Visits_P1} counts only visits to incorrect boxes, \text{Total_Visits_P1} counts all visits (correct + incorrect) in Phase 1
   - **Notes:** The formula is coherent and yields a dimensionless score; it is bounded in $[0, 1]$ if \text{Incorrect_Visits_P1}\leq\text{Total_Visits_P1}, but the paper does not explicitly state the constraint or handle \text{Total_Visits_P1}=0.
2. ⚠ **Short-term adaptation score operator ambiguity** (Adaptation_Score_STM definition, Sec. 2.2.1, p. 4)
   - **Claim:** Defines short-term adaptation score using a perseveration penalty term and \text{Time_to_First_Correct_P2}.
   - **Checks:** algebra/syntax unambiguity, dimensional consistency
   - **Verdict:** UNCERTAIN; confidence: high; impact: critical

- **Assumptions/inputs:** \text{Perseverative_Errors_P2} and \text{Total_Visits_P2} are defined as counts in Phase **2**, \text{Time_to_First_Correct_P2} is in seconds
- **Notes:** As rendered, the equation is missing an explicit operator between (1 - \frac{\text{Perseverative_Errors_P2}}{\text{Total_Visits_P2}}) and \text{Time_to_First_Correct_P2}. It is unclear whether the intended formula divides by time, multiplies by time, or multiplies by inverse time. This prevents verifying the CFI construction and interpretation.

3. ⚠ **Long-term adaptation score operator ambiguity** (Adaptation_Score_LTM definition, Sec. 2.2.1, p. 4)
    - **Claim:** Defines long-term adaptation score analogously for Phase **3** using perseveration and \text{Time_to_First_Correct_P3}.
    - **Checks:** algebra/syntax unambiguity, dimensional consistency
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** \text{Perseverative_Errors_P3} and \text{Total_Visits_P3} are defined as counts in Phase **3**, \text{Time_to_First_Correct_P3} is in seconds
    - **Notes:** Same missing/unclear operator problem as STM adaptation score; cannot confirm whether the score is intended as a rate, a time-weighted fraction, or something else.

4. ✔ **CFI construction from z-scores** (CFI formula, Sec. 2.2.2, p. 4)
    - **Claim:** CFI equals the mean of z-scored phase-specific scores: (Z(\text{Learning_Score_P1}) + Z(\text{Adaptation_Score_STM}) + Z(\text{Adaptation_Score_LTM}))/3.
    - **Checks:** algebra, definition consistency, dimensional consistency
    - **Verdict:** PASS; confidence: medium; impact: critical
    - **Assumptions/inputs:** Each component score is computed for each subject, $Z(\cdot)$ means subtracting the cohort mean and dividing by cohort standard deviation, Standard deviations are non-zero for each component
    - **Notes:** Averaging standardized scores is internally consistent and yields a dimensionless composite. However, since the adaptation-score formulas are ambiguous, the upstream values feeding the z-scores are not fully verifiable.

5. ✔ **CRS definition as regression residuals** (CRS definition, Sec. 2.3.1, p. 4; reiterated Sec. 3.3, pp. 6–7)
    - **Claim:** CRS is defined as residuals from a multiple linear regression predicting CFI from epigenetic age, sex, and origin colony.
    - **Checks:** definition consistency, logical derivation
    - **Verdict:** PASS; confidence: medium; impact: critical

- **Assumptions/inputs:** Linear regression includes an intercept, Sex and origin are encoded as predictors in the design matrix, Residuals are computed as observed CFI minus fitted CFI
- **Notes:** Residualization is a coherent way to define an age-adjusted score. The paper does not specify intercept inclusion or categorical encoding explicitly.

6. ✖ **Which N is used for CRS residualization model** (Methods Sec. 2.3.1 (p. 4) vs Results Sec. 3.2–3.3 (pp. 6–7))

   - **Claim:** CRS residualization model is fit on 'subjects with complete data', but results discuss CFI model with df implying $N = 41$ and then define CRS from 'the' model without stating sample.
   - **Checks:** definition consistency, cohort/df logical consistency
   - **Verdict:** FAIL; confidence: high; impact: critical
   - **Assumptions/inputs:** Complete multimodal data subset is $N = 33$ (demographics + behavior + DTI), Behavioral-only cohort is $N = 41$
   - **Notes:** There is an internal mismatch: Methods imply CRS is based on the complete-data subset (likely $N = 33$), but the Results first present the CFI regression with $F(3, 37)$ consistent with $N = 41$ and then proceed to CRS without clarifying whether CRS uses $N = 41$ or $N = 33$. This affects CRS values and properties in the subset used for neuroimaging regressions.

7. ✔ **Claim that CRS is uncorrelated with predictors** (Sec. 3.3, p. 6–7)

   - **Claim:** States CRS is uncorrelated with epigenetic age, sex, and colony origin 'as designed'.
   - **Checks:** linear-model identity check
   - **Verdict:** PASS; confidence: medium; impact: moderate
   - **Assumptions/inputs:** CRS residuals are from OLS with intercept using those predictors, Statement pertains to the same sample used to fit the residualization
   - **Notes:** For OLS residuals, orthogonality to the design matrix holds in-sample (including continuous predictors and dummy-coded categorical predictors), so the claim is mathematically valid if CRS is evaluated on the same data used to fit the CRS regression. If CRS is computed from a model fit on a different N than plotted/used downstream, the claim becomes ambiguous.

8. ✔ **CFI regression degrees-of-freedom coherence** (Model summary statement in Sec. 3.2, p. 6)

   - **Claim:** Reports an overall F-statistic $F(3, 37) = 1.576$ for the regression of CFI on DNAmAge, sex, and origin.
   - **Checks:** symbolic df consistency
   - **Verdict:** PASS; confidence: high; impact: minor

- **Assumptions/inputs:** Model has **3** predictors plus intercept (**4** parameters total), Residual df $= n - 4$
- **Notes:** Residual df **37** implies $n = 41$, which matches the stated behavioral cohort size. This is internally consistent (analytic check only; not validating numeric values).

9. ✔ **ROI-level association model specification** (Sec. 2.3.2, p. 4; Sec. 3.4.2, p. 8)
   - **Claim:** For each ROI, runs a regression model \text{CRS} \sim \text{Regional_MD_Value}.
   - **Checks:** model specification consistency
   - **Verdict:** PASS; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Each model includes an intercept, CRS and regional MD are scalar per subject
   - **Notes:** The stated model is coherent as a univariate linear regression. The directionality hypothesis (higher CRS associated with lower MD) is logically compatible with a negative slope expectation.
10. ✔ **Global association model specification** (Sec. 2.3.3, p. 5; Sec. 3.4.1, p. 8)
    - **Claim:** Runs regression \text{CRS} \sim \text{Global_Mean_MD} to test a global brain association.
    - **Checks:** model specification consistency, dimensional sanity
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Intercept included, Global_Mean_MD is a scalar mean over atlas-defined brain mask
    - **Notes:** Mathematically coherent univariate regression. Units of MD do not affect regression validity; they only scale the coefficient.
11. ✔ **Edge-case handling for no-correct-visit imputation** (Time_to_First_Correct definition, Sec. 2.1.3, p. 3; used in Sec. 2.2.1, p. 4)
    - **Claim:** If no correct visit occurs, \text{Time_to_First_Correct} is set to phase duration (**10800** s) and then used in phase scores.
    - **Checks:** logical consistency, units consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Phase duration is constant across individuals, \text{Time_to_First_Correct} is later used in adaptation score formula
    - **Notes:** The imputation rule is internally consistent. Its analytic implications depend on whether adaptation scores divide or multiply by time (currently ambiguous).

Limitations

- Equations are not numbered, and some formulas are corrupted/spacing-distorted in the PDF rendering; this limits precise symbolic verification.

- The audit cannot verify omitted implementation details (e.g., exact z-score definition, intercept inclusion, categorical encoding, handling of missing/zero denominators) beyond what is explicitly stated.
- No closed-form derivations are provided for statistical procedures (e.g., Benjamini–Hochberg), so only consistency of stated intent can be checked, not step-by-step proof.

Numerical results audit

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

Executed **17** internal numeric consistency checks: **13** PASS, **4** FAIL. Failures include a DTI sample-size mismatch (**28** processed files vs **33** subjects analyzed), two percentage-from-count inconsistencies (sex and colony percentages vs $N = 41$ counts), and a cross-section discrepancy in epigenetic age summary (notably SD **1.91** vs **1.7**). Regression identity checks (Adjusted R^2 and F from R^2/df) and several logical/unit checks passed.

Checked items

1. ✔ **C1** (p.2 §2.1.1 (Cohort characterization))
    - **Claim:** Cohort exhibited a balanced sex distribution (**23** males, **18** females) out of **41** bats.
    - **Checks:** parts_sum_to_total
    - **Verdict:** PASS
    - **Notes:** $23 + 18 = 41$.
2. ✔ **C2** (p.2 §2.1.1 (Cohort characterization))
    - **Claim:** Cohort colony representation (**22** from Aseret, **19** from Herzeliya) out of **41** bats.
    - **Checks:** parts_sum_to_total
    - **Verdict:** PASS
    - **Notes:** $22 + 19 = 41$.
3. ✔ **C3** (p.2 §2.1.1 (Cohort characterization))
    - **Claim:** Epigenetic age ranged from **6.62** to **15.07** years with mean **9.74** years and SD **1.91** years.
    - **Checks:** range_ordering_and_nonneg_sd
    - **Verdict:** PASS
    - **Notes:** Min $\leq$ mean $\leq$ max and SD nonnegative.
4. ✔ **C4** (p.3 §2.1.4 (DTI data processing))
    - **Claim:** Atlas contained **24** distinct brain regions (ROIs), and extraction computed mean MD for each of **24** ROI labels.

- **Checks:** repeated_constant_consistency
- **Verdict:** PASS
- **Notes:** ROI count (**24**) matches implied index end (MD_ROI_24).

5. ✖ **C5** (p.3 §2.1.4 (DTI data processing) and p.5 §3.1)
   - **Claim:** DTI MD maps were processed for **28** individual bat MD map NIfTI files, but later DTI data were available for **33** individuals ($N = 33$ for primary analyses).
   - **Checks:** cross_section_count_consistency
   - **Verdict:** FAIL
   - **Notes:** Counts differ (**28** vs **33**); not internally reconciled by the stated values.

6. ✖ **C6** (p.5 §3.1 (Cohort characteristics) vs p.2 §2.1.1)
   - **Claim:** Sex percentages (**58.5%** male, **41.5%** female) correspond to **23** males and 18 females out of $N = 41$.
   - **Checks:** percentage_from_counts
   - **Verdict:** FAIL
   - **Notes:** Check flagged under ±**0.1** percentage-point tolerance.

7. ✖ **C7** (p.5 §3.1 (Cohort characteristics) vs p.2 §2.1.1)
   - **Claim:** Colony percentages (**56.1%** Aseret, **43.9%** Herzliya) correspond to **22** Aseret and **19** Herzliya out of $N = 41$.
   - **Checks:** percentage_from_counts
   - **Verdict:** FAIL
   - **Notes:** Computed $22/41 = 53.66\%$ vs reported **56.1%** (and $19/41 = 46.34\%$ vs **43.9%**).

8. ✔ **C8** (p.5 §3.1 (Cohort characteristics))
   - **Claim:** Epigenetic age spanned **6.6** to **15.1** years with mean **9.6** and SD **1.7** years.
   - **Checks:** range_ordering_and_nonneg_sd
   - **Verdict:** PASS
   - **Notes:** Min ≤ mean ≤ max and SD nonnegative.

9. ✖ **C9** (p.2 §2.1.1 vs p.5 §3.1 vs p.9 Conclusions)
   - **Claim:** Epigenetic age summary differs between sections: Methods: range **6.62–15.07**, mean **9.74**, SD **1.91**; Results/Conclusions: range **6.6–15.1**, mean **9.6**, SD **1.7**.
   - **Checks:** cross_section_numeric_discrepancy
   - **Verdict:** FAIL

- **Notes:** Largest absolute difference is SD: **1.91** vs **1.7** (diff **0.21**), exceeding the **0.15** threshold.

10. ✔ **C10** (p.3 §2.1.5 and p.5 §3.1)
    - **Claim:** Complete multi-modal dataset size: final master DataFrame contained **33** subjects with complete data; primary analyses conducted on $N = 33$.
    - **Checks:** repeated_constant_consistency
    - **Verdict:** PASS
    - **Notes:** **33** matches **33**.
11. ✔ **C11** (p.3 §2.1.3 (Behavioral data extraction))
    - **Claim:** If no correct visit occurred within the phase, \text{Time_to_First_Correct} was set to the phase duration (**10800** seconds).
    - **Checks:** unit_conversion
    - **Verdict:** PASS
    - **Notes:** **10800** seconds equals **3** hours (3×3600).
12. ✔ **C12** (p.6 §3.2 (Regression results))
    - **Claim:** Model F-statistic reported as $F(3, 37) = 1.576$ with $p = 0.211$; $R^2 = 0.113$; Adjusted $R^2 = 0.041$.
    - **Checks:** adjusted_r2_identity_check
    - **Verdict:** PASS
    - **Notes:** Adjusted R^2 computed as **0.041081...**, consistent with reported **0.041** within tolerance.
13. ✔ **C13** (p.6 §3.2 (Regression results))
    - **Claim:** F-statistic $F(3, 37) = 1.576$ is consistent with $R^2 = 0.113$ for $n = 41$ and $p = 3$.
    - **Checks:** f_stat_from_r2_identity_check
    - **Verdict:** PASS
    - **Notes:** F computed from R^2 and df equals **1.5712**, within stated absolute tolerance of reported **1.576**.
14. ✔ **C14** (p.5 §3.2 (CFI distribution))
    - **Claim:** CFI mean = -0.016 and standard deviation = 0.622.
    - **Checks:** plausibility_check_sd_nonnegative
    - **Verdict:** PASS
    - **Notes:** SD is nonnegative.
15. ✔ **C15** (p.6 §3.3 (CRS characteristics))

- **Claim:** CRS distribution centered at zero with standard deviation 0.586.
- **Checks:** plausibility_check_sd_nonnegative
- **Verdict:** PASS
- **Notes:** SD is nonnegative.

16. ✔ **C16** (p.8 §3.4.1 (Global MD regression))
    - **Claim:** Global regression: coefficient = 2848.16, $p = 0.305$, $R^2 = 0.034$; 'accounting for only 3.4% of the variance'.
    - **Checks:** percent_from_r2
    - **Verdict:** PASS
    - **Notes:** $100 \times R^2 = 3.4\%$, matches the reported variance percent.
17. ✔ **C17** (p.8 §3.4.2 (ROI analysis, multiple comparisons))
    - **Claim:** 24 ROIs tested; lowest uncorrected $p = 0.174$ became FDR-adjusted $p = 0.880$.
    - **Checks:** bh_fdr_single_value_sanity
    - **Verdict:** PASS
    - **Notes:** BH necessary bounds satisfied: $0.174 \leq 0.880 \leq \min(1, 0.174 \times 24) = 1$.

Limitations

- Checks are constrained to internal arithmetic/logical consistency of reported numbers; the PDF does not include per-subject datasets or full regression/ROI result tables needed for recomputation.
- Figure-derived numeric verification is avoided (no pixel/plot extraction), per instructions; only text-reported values are used.
- BH-FDR adjustment can only be sanity-checked via necessary bounds without the complete vector of p-values.