Skepthical review: *Investigating Cognitive Resilience in Long-Lived Bats: Challenges in Integrating Epigenetic Age, Spatial Memory, and Brain Structure*

Summary

The manuscript sets out to test cognitive resilience in long-lived Egyptian fruit bats by integrating three modalities—epigenetic age (DNAmAge from skin), spatial cognition from a multi-phase foraging task, and brain structure from diffusion MRI—using a planned DNAmAge × brain-metric interaction framework (Sec. 1, Sec. 2, Sec. 2.5.2). In the delivered study, two critical technical failures prevented the core multimodal hypothesis from being evaluated: (i) MRI preprocessing could not produce brain masks because the presumed first $b = 0$ volume in all NIfTIs appeared empty/all-zero, so no diffusion/structural metrics were extracted (Sec. 2.4.1–2.4.2, Sec. 3.1, Sec. 3.4); and (ii) behavioral log parsing failed for most intended readouts (incorrect entries, perseveration, exploration), leaving only \text{Time_to_First_Food} (per phase) and a derived \text{Switch_Cost} as analyzable measures, with small and uneven Ns across phases and outcomes (Sec. 2.3, Sec. 3.2). The remaining analyses therefore reduce to correlations/regressions relating DNAmAge, sex, and origin colony to these limited behavioral measures (Sec. 2.5.1, Sec. 3.3). No DNAmAge associations are detected, while origin colony predicts \text{Switch_Cost} (borderline $p \approx 0.049$) in a small subset ($N \approx 17$) (Sec. 3.3.2). The manuscript's transparency about these failures is valuable, but as written the framing and conclusions still imply an achieved multimodal brain–age–cognition integration and broader claims about resilience that the data do not support. A substantial reframing toward a feasibility/QC/pipeline-diagnosis contribution, plus much more forensic documentation of the MRI and behavioral failure modes, stronger validation of the remaining behavioral metrics, and more robust/appropriately cautious statistics are needed for the work to be scientifically actionable and correctly scoped.

Strengths

- Ambitious, timely multimodal design in an unusually long-lived mammalian model, with a clearly articulated (intended) age-by-brain moderation hypothesis (Sec. 1, Sec. 2.5.2).
- Unusually candid reporting of major pipeline failures and their consequences, which can be valuable to the community if paired with deeper technical diagnosis and QC artifacts (Sec. 3.1–3.4, Conclusions).
- Clear definition (in plain language and units) of the two analyzable behavioral outcomes, \text{Time_to_First_Food} and \text{Switch_Cost}, enabling readers to follow the limited analyses that were possible (Sec. 2.3, Sec. 3.2).

- Use of broadly appropriate baseline statistical tools for exploratory analyses (Spearman correlations; regression with covariates) given the constrained dataset, and inclusion of basic diagnostics discussion (Sec. 2.5.1, Sec. 3.3).
- The exploratory origin-colony association with \text{Switch_Cost}—if treated as hypothesis-generating—suggests potentially interesting environmental/genetic influences worth targeted follow-up (Sec. 3.3.2, Sec. 3.4).

Major issues

1. **Core aim and framing mismatch: the manuscript repeatedly frames the study as integrating brain structure (DTI-MRI) with epigenetic age and spatial cognition, and as testing DNAmAge × brain-metric interactions (Sec. 1, Sec. 2.4, Sec. 2.5.2), but no usable MRI metrics were extracted due to the $b = 0$ failure (Sec. 3.1, Sec. 3.4).** The delivered results therefore do not test the central hypothesis about structural moderators of cognitive resilience, yet the Abstract/Introduction/Conclusions language still implies that neural structural correlates were analyzed.

   Recommendation: Reframe the paper explicitly as a feasibility/QC/pilot report (or a "lessons learned" pipeline paper) plus a limited DNAmAge–behavior exploration, rather than a completed multimodal brain–age–cognition study. Concretely: (i) revise the Abstract and Conclusions to state up front that MRI feature extraction failed and that the planned interaction analyses were not performed; (ii) in Sec. 1, add a short paragraph distinguishing planned vs achieved analyses and narrowing the contribution; (iii) move detailed unrealized DTI-metric/interaction model descriptions (Sec. 2.4, Sec. 2.5.2) to an "Intended analyses" subsection or Appendix, clearly labeled as not executed; and (iv) temper or remove claims about "structural signatures" and broad "cognitive resilience" if resilience is not operationalized with available data.

2. **MRI failure diagnosis is under-documented and may be premature: the reported 'empty/all-zero first $b = 0$ volume' prevents masking and all downstream diffusion metrics (Sec. 2.4.1–2.4.2, Sec. 3.1), but the manuscript does not provide sufficient forensic QC to determine whether this is truly absent signal vs. conversion/scaling/indexing errors, nor whether alternative salvage routes exist. In addition, using reconstructed/assumed bvals/bvecs (generic 30-direction scheme) raises fundamental validity concerns even if masking were fixed (Sec. 2.4.1).**

   Recommendation: Add a dedicated MRI QC/forensics subsection (Sec. 2.4 and/or Sec. 3.1) with concrete evidence and diagnostics: (i) acquisition details (scanner/sequence, TE/TR, resolution, number and placement of b0s, gradient export availability); (ii) Bruker→NIfTI conversion tool, version, settings, and validation steps; (iii) volume-wise intensity summaries (min/max/mean, histograms) after NIfTI scaling is applied (e.g., verifying header scl_slope/scl_inter, dtype/NaNs), plus visual montages of

several slices; (iv) verify whether the first volume is truly a $b = 0$ (check bvals ordering vs actual volume signal; b0s can be interleaved or later); (v) attempt robust masking from the mean of all b0s or another reference image (and describe why alternatives failed); and (vi) explicitly justify or retract the use of generic bvecs—ideally recover true gradients from Bruker metadata (e.g., method file) or state clearly that quantitative DTI cannot be interpreted without correct bvecs/bvals. If the dataset is indeed unrecoverable, demonstrate this with the QC outputs above (e.g., a small set of representative subjects).

3. **Behavioral parsing failures are insufficiently specified and currently undermine interpretability and replicability: most planned behavioral metrics are reported as zero/NaN due to parsing/log-format issues (Sec. 2.3, Sec. 3.2), yet the paper does not provide enough detail about the raw log structure, event coding, phase rules, or parsing logic to (a) assess whether the failures could be corrected, (b) validate that the remaining outcomes (\text{Time_to_First_Food}, \text{Switch_Cost}) are correctly computed, or (c) reproduce the extraction.**

   Recommendation: Expand Sec. 2.3 and Sec. 3.2 into a more rigorous behavioral methods + parsing/QC section: (i) fully specify the task (arena/box layout, reward contingencies per phase, session duration/timeouts, phase transition rules, how 'correct box' is defined/stored); (ii) document the raw log/Excel schema with an anonymized example (column names, data types, header rows, where phase IDs and 'correct box' appear, allowed action codes such as E/F and variants); (iii) describe parsing rules precisely (handling of missing/duplicate timestamps, capitalization/whitespace variants, sheet naming differences, per-phase boundaries); (iv) perform and report a manual validation against ground truth on a subset (e.g., 5–10 sessions): compute entries/perseveration/exploration by hand and compare to script outputs; (v) if repair is feasible, re-extract and report the full intended metric set; if not feasible, provide a table of the observed file-format variants/error modes and a clear justification of why reliable recovery is impossible. Also clarify censoring: if some bats never find food within a session, state how \text{Time_to_First_Food} is defined/treated (timeout value vs missing), as this affects appropriate statistical modeling.

4. **Small, uneven, and potentially non-random sample sizes plus fragile inference: effective N varies widely across phases (e.g., P1$\approx$ 30, P2$\approx$ 21, P3$\approx$ 35) and \text{Switch_Cost} uses a much smaller paired subset ($\approx$ 17) (Sec. 3.2, Sec. 3.3). The manuscript does not quantify missingness causes, test whether inclusion is related to DNAmAge/sex/colony, or provide power/sensitivity estimates. As a result, the null DNAmAge results are not informative and the colony effect ($p \approx 0.049$) is likely highly unstable (Sec. 3.3.2, Sec. 3.4, Conclusions).**

Recommendation: Add a missingness and sensitivity section spanning Sec. 3.2–3.4: (i) provide a single summary table listing each metric/phase, the N used, and explicit exclusion reasons; (ii) test whether missingness/inclusion is associated with DNAmAge, sex, or colony (e.g., logistic regression or contingency analyses), and discuss implications; (iii) report power/sensitivity (e.g., minimum detectable Spearman $|\rho|$ at given N; detectable standardized effects in regression); and (iv) rephrase interpretation throughout Sec. 3.4 and Conclusions to emphasize low power for DNAmAge effects and to label the colony effect as exploratory/hypothesis-generating, not confirmatory.

5. **Statistical modeling does not match distributional features and small-N uncertainty: \text{Time_to_First_Food} is likely right-skewed and potentially censored; \text{Switch_Cost} can be heavy-tailed. The manuscript notes assumption violations in diagnostics (Sec. 3.3) but largely proceeds with standard linear regression, and the key colony result is based on a very small N with multiple predictors and limited robustness checks.**

   Recommendation: Strengthen Sec. 2.5 and Sec. 3.3 with analyses appropriate to the outcomes and sample size: (i) consider log/log1p transforms for \text{Time_to_First_Food} and report whether conclusions change; (ii) use heteroskedasticity-robust SEs (e.g., HC3/HC4) and/or robust regression as a sensitivity analysis; (iii) for \text{Switch_Cost} and small N, add non-parametric or permutation/bootstrapped inference (bootstrapped CIs for coefficients; permutation test for colony effect), and report full coefficient tables with CIs (not only p-values); (iv) explicitly address multiple testing across phases/outcomes/predictors (even if limited) or justify why not; and (v) if timeouts/censoring exist, consider survival/TOBIT-style approaches or at minimum clearly define how censoring was handled and its potential bias.

6. **Reproducibility materials and reporting are not yet aligned with the paper's implicit 'pipeline cautionary tale' contribution: without code, data dictionaries, and QC outputs, readers cannot learn from or verify the failure modes and the limited analyses (Sec. 2.2–2.5, Sec. 3.1–3.3). Ethical/animal welfare details are also missing or too sparse for animal research (Sec. 2.1, Sec. 2.4).**

   Recommendation: Add a Reproducibility and Ethics package: (i) provide versioned code (behavior parsing + MRI preprocessing attempts + stats), software versions, and a brief runbook; (ii) include an anonymized behavioral log schema/example and a data dictionary for all variables; (iii) include representative MRI header dumps and QC figures (volume-wise mean intensity plots; slice montages showing the 'empty $b0$' issue); (iv) document random seeds (if any) and exact inclusion/exclusion criteria; and (v) add an Ethics/Animal Welfare subsection in Sec. 2.1 describing approvals, housing/handling, and MRI procedures (e.g., anesthesia/restraint, stress mitigation), with approval IDs or explicit regulatory justification.

Minor issues

1. Abstract/title/keywords misrepresent achieved scope: the paper still reads like a completed multimodal neuroimaging study, and the keyword list includes unrelated astronomy terms (Abstract, Title).

   Recommendation: Rewrite the Abstract and (if needed) the title to reflect the achieved analyses (DNAmAge + limited behavioral outcomes; MRI metrics unavailable). Replace keywords with domain-appropriate terms (e.g., epigenetic clock, DNA methylation, bats, spatial learning, cognitive flexibility, diffusion MRI QC).

2. Apparent placeholder or non-scientific affiliation/author block text in the unstructured materials (e.g., 'Anthropic, Gemini & OpenAI servers. Planet Earth.').

   Recommendation: Replace any placeholder author/affiliation text with correct institutional affiliations and corresponding author details, consistent with journal requirements.

3. DNAmAge description is too thin to interpret null results: the manuscript does not adequately describe how the DNAmAge clock was constructed/validated, its error, correlation with chronological age, or limitations of skin DNAmAge as a proxy for brain aging (Sec. 2.1, Sec. 3.1).

   Recommendation: In Sec. 2.1, provide clock construction/validation details (training set, model type, performance metrics such as MAE and r vs chronological age) with citations; state whether chronological age is available and how it relates to DNAmAge here; discuss tissue-specific limitations (skin vs brain) in Sec. 3.4.

4. Internal inconsistencies in cohort summaries: sex counts differ between Methods and Results/Table 1 for the same $N = 41$ (23M/18F vs 24M/17F), and DNAmAge mean/SD differs between sections while min/max match (Sec. 2.1, Sec. 3.1, Table 1).

   Recommendation: Recompute and harmonize cohort descriptives from a single clearly defined dataset (or explicitly label pre- vs post-exclusion datasets). Update Sec. 2.1, Sec. 3.1, and Table 1 to be consistent.

5. \text{Switch_Cost} definition vs summary statistics mismatch: \text{Switch_Cost} is defined as P2_\text{Time_to_First_Food} - P1_\text{Time_to_First_Food}, but the reported mean \text{Switch_Cost} does not equal the difference of the reported $P2$ and $P1$ means (Sec. 3.2, Table 2).

   Recommendation: Explicitly state that \text{Switch_Cost} is computed on a paired subset and report P1/P2 summary statistics on that same paired subset (and N). Alternatively, correct any reporting error so the summaries are mathematically consistent.

6. Figures and statistical reporting: several figures are hard to read (size/font), mix adjusted and unadjusted views without clear labeling, omit per-panel Ns and key statistics, and include panels reflecting failed extraction without strong visual annotation (Figures 1–7; Sec. 3.1–3.3).

   Recommendation: Increase figure resolution and font sizes; add per-panel N, units, and (where relevant) regression coefficients/CIs/p-values; clearly mark unavailable/failed metrics panels; and consider supplementing bivariate plots with adjusted-effect or partial-regression visualizations aligned to Sec. 3.3 models.

7. Terminology and spelling inconsistencies (e.g., 'Herzeliya' vs 'Herzliya'; 'Switch Cost' vs 'Switch_Cost') and dispersed sample-size reporting (Sec. 3.2–3.3, Tables/Figures).

   Recommendation: Standardize colony naming and variable labels across text/figures/tables, and restate effective N next to each reported analysis (and in table captions) to reduce reader confusion.

Very minor issues

1. Typographical/formatting problems: spurious line breaks within words (e.g., 'Sec\nondly'), inconsistent spacing around units and statistical notation, truncated table headers/values (e.g., 'Mi' for 'Min'), and minor caption redundancy (Sec. 1, Sec. 3.1–3.4, Tables, Figure captions).

   Recommendation: Proofread from source to remove hard line breaks, standardize statistical formatting (e.g., "$M = \ldots$, $SD = \ldots$"), fix truncated headers/entries, and tighten captions to avoid repetition while keeping Ns/units and panel references.

2. Section/heading style inconsistencies (mixed Markdown-like levels and numbering; inconsistent math spacing) (Sec. 3–4).

   Recommendation: Standardize headings and numbering to the venue style and clean up inline math formatting for consistency.

3. Event-code definition ambiguity in the behavioral metrics (e.g., counting 'F' as an incorrect entry if 'F' denotes food retrieval) (Sec. 2.3; metric definitions).

   Recommendation: Clarify the event coding schema and ensure metric definitions align with the semantics of each code; adjust the metric definition or parsing accordingly.

Key statements and references

- • **The Egyptian fruit bat, Rousettus aegyptiacus, has a maximum recorded lifespan of over 25 years in captivity, making it an exceptionally long-lived mammalian model relative to its body size and a suitable species for studying neural correlates of cognitive resilience.**
- *Reference(s):* Denario [11]

- • **Biological age for each bat in this study was quantified using a DNA-methylation-based epigenetic clock specifically developed for Rousettus aegyptiacus from skin samples (DNAmAgeBat.Rousettus.aegyptiacus_Skin), providing a more precise measure of aging status than chronological age.**
- *Reference(s):* Denario [11]
- • **A standard electrostatically optimized 30-direction diffusion-weighting scheme, commonly used in diffusion tensor imaging (DTI) acquisitions, was adopted to generate the b-vector file (bvecs.txt) for all bats in the absence of the exact Bruker scanner gradient scheme, under the assumption that this scheme provides a robust representation of diffusion directions for tensor fitting.**
- *Reference(s):* Denario [11]
- • **Spearman correlation analysis showed that epigenetic age (DNAmAge) was not significantly associated with initial spatial learning efficiency or cognitive flexibility in this cohort, with DNAmAge–P1_Time_to_First_Food $\rho = -0.10$ $(p > 0.05)$ and DNAmAge–Switch_Cost $\rho = 0.17$ $(p > 0.05)$, indicating that DNAmAge alone did not predict these specific cognitive performance measures.**
- *Reference(s):* Denario [11]

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: definitions of behavioral metrics (including a difference score), a voxel-count-based brain volume computation, and linear regression model forms (including an interaction-term model described but not executed due to missing MRI-derived predictors). No multi-step algebraic derivations are presented; most checks are definition/notation consistency and dimensional sanity checks.

Checked items

1. ✔ **Planned interaction regression model form** (Sec. 1 (end of Introduction), p.2: "Cognitive Metric ∼ DNAmAge+Brain Metric+DNAmAge×Brain Metric+Covariates")
    - **Claim:** Cognitive outcome is modeled as a linear function of DNAmAge, a brain metric, their interaction, and covariates.
    - **Checks:** notation consistency, model specification sanity
    - **Verdict:** PASS; confidence: high; impact: minor

- **Assumptions/inputs:** "∼" denotes a linear model formula interface (not probabilistic distributional equality)., Interaction term corresponds to product of the two predictors in the design matrix.
- **Notes:** The model form is standard and internally consistent with later description in Sec. 2.5.2.

2. ✔ **Expanded regression model with covariates** (Sec. 2.5.2, p.4: "Cognitive_Metric ∼ DNAmAge+Brain_Metric+DNAmAge×Brain_Metric+Sex+Origin_colony")
   - **Claim:** Regression includes main effects, interaction, and categorical covariates for sex and origin colony.
   - **Checks:** definition consistency, notation consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Sex and Origin_colony are coded as categorical (dummy/indicator) variables.
   - **Notes:** Consistent with the earlier general form and with the later simplified models when MRI metrics were unavailable.

3. ✔ **Mean-centering predictors for interaction interpretation** (Sec. 2.5.2, p.4: "all continuous predictor variables … were mean-centered")
   - **Claim:** Mean-centering DNAmAge and Brain_Metric makes main-effect coefficients interpretable at the mean of the interacting variable.
   - **Checks:** algebraic interpretation check
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Model includes an intercept term., Mean-centering is applied as $x_c = x - \mathbf{mean}(x)$.
   - **Notes:** With interaction $x * z$, centering makes the main-effect coefficient correspond to the effect when the other variable is at its mean; statement is consistent.

4. ✔ **Time_to_First_Food definition** (Sec. 2.3, p.3)
   - **Claim:** \text{Time_to_First_Food} is elapsed time in seconds from phase start to first food retrieval ("F").
   - **Checks:** unit consistency, definition clarity
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** Absolute_Time is available and can be differenced to yield seconds from phase start.
   - **Notes:** Definition is unit-consistent (seconds) and later tables label the metric in seconds, matching the definition.

5. ✔ **Switch_Cost definition as a difference score** (Sec. 2.3, p.3 and reiterated Sec. 3.2, p.5–6)

- **Claim:** \text{Switch_Cost} = P2_\text{Time_to_First_Food} - P1_\text{Time_to_First_Food}; negative implies faster adaptation in Phase 2 than Phase 1.
- **Checks:** algebra, unit consistency, sanity/limiting case
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** P1 and P2 times are in the same units (seconds) and computed comparably.
- **Notes:** Difference of times yields seconds. If P2=P1 then \text{Switch_Cost}=0; if P2<P1 then negative, consistent with the stated interpretation.

6. ⚠ **Incorrect_Entries_Before_First_Food event-type inclusion** (Sec. 2.3, p.3)

   - **Claim:** \text{Incorrect_Entries_Before_First_Food} counts "E" or "F" actions into incorrect boxes prior to the first food event.
   - **Checks:** definition consistency, logical coherence of event coding
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
   - **Assumptions/inputs:** Action log may record "F" events that could occur in locations later judged incorrect, or "F" is being used as a generic entry marker as well as success marker.
   - **Notes:** If "F" is strictly a successful retrieval at the correct box, then "F into incorrect boxes" is contradictory. The paper does not define the action-code semantics tightly enough to verify this metric definition.

7. ✔ **bvals length and composition** (Sec. 2.4.1, p.3)

   - **Claim:** bvals.txt contains 33 values: 3 zeros ($b = 0$) followed by 30 values of 1000 s/mm^2.
   - **Checks:** counting/algebra, unit consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Each DTI 4D volume has 33 diffusion volumes.
   - **Notes:** $3 + 30 = 33$ is consistent, and b-value units (s/mm^2) are stated.

8. ✔ **Voxel volume computation for brain volume** (Sec. 2.4.2, p.4)

   - **Claim:** Voxel volume is 0.5 mm × 0.5 mm × 1.0 mm = 0.25 mm^3/voxel; \text{Brain_Volume} = (#\textrm{ mask voxels}) \times \textrm{voxel volume}.
   - **Checks:** dimensional analysis, arithmetic sanity
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Voxel dimensions are as stated and constant across the image.
   - **Notes:** Product gives mm^3; $0.5 \times 0.5 \times 1.0 = 0.25$, consistent with stated voxel volume.

9. ✔ **Regression df consistency (P1_Time_to_First_Food model)** (Sec. 3.3.2, p.7: "$F(3, 26)$" with $N = 30$ (Table 2))
    - **Claim:** The reported F-statistic degrees of freedom match a linear regression with 3 predictors (excluding intercept) and $N = 30$.
    - **Checks:** model df algebra
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Model includes intercept + 3 predictors: DNAmAge(centered), Sex, Origin_colony., No additional predictors or dropped rows beyond $N = 30$.
    - **Notes:** For $N = 30$, $k = 3$ predictors (excluding intercept): $df_1 = k = 3$ and $df_2 = N - k - 1 = 30 - 3 - 1 = 26$, matching $F(3, 26)$.
10. ✔ **Regression df consistency (Switch_Cost model)** (Sec. 3.3.2, p.7: "$F(3, 13)$" with $N = 17$ (Table 2))
    - **Claim:** The reported F-statistic degrees of freedom match a linear regression with 3 predictors (excluding intercept) and $N = 17$.
    - **Checks:** model df algebra
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Model includes intercept + 3 predictors: DNAmAge(centered), Sex, Origin_colony., No additional predictors or dropped rows beyond $N = 17$.
    - **Notes:** For $N = 17$, $k = 3$ predictors: $df_2 = 17 - 3 - 1 = 13$, matching $F(3, 13)$.
11. ✖ **Cohort statistic consistency across sections** (Sec. 2.1 (Methods), p.2 vs Sec. 3.1 (Results), p.5 and Table 1)
    - **Claim:** The paper reports a single cohort of 41 bats; demographic counts and DNAmAge summary statistics should be consistent unless a different subset is stated.
    - **Checks:** internal consistency of definitions/summaries
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** Same cohort definition is intended in Methods and Results when both state $N = 41$.
    - **Notes:** Methods: 18 females/23 males and DNAmAge mean 9.78 SD 1.83. Results/Table 1: 17 females/24 males and DNAmAge mean 9.60 SD 1.74. No explanation (e.g., data correction or recoding) is given while N remains 41, so the summaries are internally inconsistent.

Limitations

- The provided PDF content contains essentially no step-by-step mathematical derivations; most 'checks' are limited to verifying definitions, dimensional consistency, and internal consistency of stated model forms.
- No explicit equations are numbered in the provided text, so locations are referenced by section and page as shown in the parsed PDF.
- Where verification would require precise definitions of behavioral log codes (e.g., semantics of "E" and "F"), the paper does not provide enough detail to fully validate metric definitions symbolically.

Numerical results audit

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

Of 19 candidate numeric checks, 14 passed, 3 failed, and 1 was uncertain. The main problems are cross-section inconsistencies in cohort sex counts and DNAmAge mean/SD, plus a mismatch between the reported \text{Switch_Cost} mean and the difference of reported phase means (not necessarily an error if computed on a paired subset with different N). Reported regression t-statistics and p-values from F/t distributions were internally consistent within stated tolerances.

Checked items

1. ✔ **C1** (p.2 (Methods 2.1 Subjects))
    - **Claim:** Cohort comprised 18 females (43.9%) and 23 males (56.1%) out of 41 bats.
    - **Checks:** parts_vs_total_and_percentages
    - **Verdict:** PASS
    - **Notes:** Counts sum to 41; computed percentages match within 0.1 percentage points; percents sum to 100.
2. ✔ **C2** (p.2 (Methods 2.1 Subjects))
    - **Claim:** Bats originated from two colonies: 23 individuals (56.1%) Aseret and 18 individuals (43.9%) Herzeliya out of 41.
    - **Checks:** parts_vs_total_and_percentages
    - **Verdict:** PASS
    - **Notes:** Counts sum to 41; computed percentages match within 0.1 percentage points; percents sum to 100.
3. ✔ **C3** (p.2 (Methods 2.1 Subjects))
    - **Claim:** DNAmAge ranged from 6.62 to 15.07 years, mean 9.78 years, SD 1.83 years.

- **Checks:** range_and_summary_consistency
- **Verdict:** PASS
- **Notes:** Mean lies within [min,max]; SD is non-negative; max>min.

4. ✖ **C4** (p.5 (Results 3.1) and Table 1)
   - **Claim:** Results report cohort has 24 males (58.5%) and 17 females (41.5%) out of 41; differs from Methods (23 males, 18 females).
   - **Checks:** cross_section_repeated_numbers_consistency
   - **Verdict:** FAIL
   - **Notes:** Methods vs Results/Table 1 sex counts differ by 1 in both categories for $N = 41$.

5. ✔ **C5** (p.5 (Results 3.1) and Table 1)
   - **Claim:** Sex percentages in Results/Table 1: 17 (41.5%) female and 24 (58.5 %) male out of 41.
   - **Checks:** percentages_from_counts
   - **Verdict:** PASS
   - **Notes:** Percentages consistent with counts within rounding; percents sum to 100.

6. ✖ **C6** (p.5 (Results 3.1) and Table 1)
   - **Claim:** DNAmAge summary in Results/Table 1: mean 9.60 years (SD 1.74), min 6.62, max 15.07; differs from Methods mean 9.78 (SD 1.83) with same min/max.
   - **Checks:** cross_section_repeated_numbers_consistency
   - **Verdict:** FAIL
   - **Notes:** Methods vs Results/Table 1 mean differs by 0.18 years and SD differs by 0.09 years.

7. ✔ **C7** (p.5 (Results 3.1) and Table 1)
   - **Claim:** Origin colony counts/percentages: Aseret $N = 23$ (56.1%), Herzliya $N = 18$ (43.9%) out of 41.
   - **Checks:** parts_vs_total_and_percentages
   - **Verdict:** PASS
   - **Notes:** Counts sum to 41; computed percentages match within 0.1 percentage points; percents sum to 100.

8. ✔ **C8** (p.3 (Methods 2.4.1) and p.4 (Methods 2.4.1/2.4.2))
   - **Claim:** bvals.txt contains a sequence of 33 values: three '0's then thirty '1000's.
   - **Checks:** count_and_composition_consistency
   - **Verdict:** PASS

- **Notes:** Verified $3 + 30 = 33$.

9. ✔ **C9** (p.4 (Methods 2.4.2 Brain Volume calculation))
   - **Claim:** Voxel dimensions 0.5 mm × 0.5 mm × 1.0 mm imply voxel volume = 0.25 mm^3/voxel.
   - **Checks:** unit_volume_multiplication
   - **Verdict:** PASS
   - **Notes:** $0.5 \times 0.5 \times 1.0 = 0.25$ exactly within tight tolerance.
10. ✔ **C10** (p.5 (Results 3.1 Cohort demographics and data completeness))
    - **Claim:** MRI availability: initially 8 subjects lacked DTI files; remaining 33 DTI files had processing error; totals should match $N = 41$.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Verified $8 + 33 = 41$.
11. ⚠ **C11** (p.6 (Results 3.2 text) and Table 2)
    - **Claim:** Phase sample sizes reported: $N = 30$ for P1, $N = 21$ for P2, $N = 35$ for P3, and $N = 17$ for Switch Cost; should match Table 2 N column.
    - **Checks:** cross_reference_table_text_consistency
    - **Verdict:** UNCERTAIN
    - **Notes:** Table 2 Ns were not provided in the check payload, so the text-vs-table match could not be tested.
12. ✖ **C12** (p.6 (Table 2) and p.5-6 (Results 3.2 text defining Switch_Cost))
    - **Claim:** \text{Switch_Cost} is defined as P2_\text{Time_to_First_Food} - P1_\text{Time_to_First_Food}; check that reported \text{Switch_Cost} mean equals difference of reported $P2$ and $P1$ means (not generally true unless same paired sample, but worth flagging).
    - **Checks:** derived_metric_mean_difference_sanity_check
    - **Verdict:** FAIL
    - **Notes:** Computed ($P2$ mean $-P1$ mean) = 514.28 s vs reported \text{Switch_Cost} mean 985.59 s; mismatch may be expected given differing Ns (P1=30, P2=21, Switch_Cost=17) and paired-subset computation.
13. ✔ **C13** (p.6 (Table 2))
    - **Claim:** For each behavioral metric in Table 2, mean should lie between min and max.
    - **Checks:** range_check_mean_within_bounds
    - **Verdict:** PASS
    - **Notes:** For P1, P2, P3, and Switch_Cost, each mean is within its reported [min,max].

14. ✔ **C14** (p.7 (Results 3.3.2) model for P1_Time_to_First_Food)
    - **Claim:** Overall model: $F(3, 26) = 0.047$, $p = 0.986$. Verify p-value matches F-statistic and degrees of freedom.
    - **Checks:** test_statistic_to_pvalue
    - **Verdict:** PASS
    - **Notes:** Computed $p = 0.986180\ldots$, consistent with reported 0.986.

15. ✔ **C15** (p.7 (Results 3.3.2) model for P1_Time_to_First_Food)
    - **Claim:** Table 3 t-statistics should equal coefficient / std. error (e.g., Intercept 5185.96 / 844.66 = 6.140).
    - **Checks:** t_stat_recomputation
    - **Verdict:** PASS
    - **Notes:** All checked rows match $t \approx \beta/\mathrm{SE}$ within $abs_tol\ 0.002$.

16. ✔ **C16** (p.7 (Results 3.3.2) model for Switch_Cost)
    - **Claim:** Overall model: $F(3, 13) = 2.056$, $p = 0.156$. Verify p-value matches F-statistic and degrees of freedom.
    - **Checks:** test_statistic_to_pvalue
    - **Verdict:** PASS
    - **Notes:** Computed $p = 0.155729\ldots$, consistent with reported 0.156.

17. ✔ **C17** (p.7 (Table 4 Regression Results for Switch_Cost))
    - **Claim:** Table 4 t-statistics should equal coefficient / std. error (e.g., Origin[Herzliya] -3930.51 / 1808.18 = -2.174).
    - **Checks:** t_stat_recomputation
    - **Verdict:** PASS
    - **Notes:** All checked rows match $t \approx \beta/\mathrm{SE}$ within $abs_tol\ 0.002$.

18. ✔ **C18** (p.7 (Table 4, Origin[Herzliya]))
    - **Claim:** For Switch_Cost model, Origin[Herzliya] has $t = -2.174$ and $p = 0.049$; verify p-value matches two-sided t-test with $df = 13$ (from $F(3, 13)$).
    - **Checks:** t_stat_to_pvalue
    - **Verdict:** PASS
    - **Notes:** Computed two-sided $p = 0.048768\ldots$, consistent with reported 0.049.

19. ✔ **C19** (p.7 (Table 3, Intercept))
    - **Claim:** For P1_Time_to_First_Food model, Intercept has $t = 6.140$ and $p < 0.001$; verify that two-sided p-value for $t = 6.140$ with $df = 26$ is indeed < 0.001.
    - **Checks:** t_stat_to_pvalue_inequality

- **Verdict:** PASS
- **Notes:** Computed two-sided $p = 1.7175 \times 10^{-6}$, satisfying $p < 0.001$.

Limitations

- Checks are restricted to arithmetic/logical consistency and distribution-based p-value recomputations using reported test statistics; underlying raw data are not available in the PDF text.
- No values were extracted from plots/images beyond what is explicitly written in the text/tables; pixel-based extraction is excluded.
- Some potentially meaningful checks (e.g., verifying means/SDs from raw observations, recomputing correlations, validating regression R^2/adjusted R^2 from data) cannot be performed with the information provided.