# *Skepthical* review: *Epigenetic Aging, Regional Brain Morphology, and the Spectrum of Cognitive Decline in Long-Lived Egyptian Fruit Bats*

## Summary

The manuscript presents the development and validation of an end-to-end multimodal pipeline to study relationships among epigenetic age (**DNAmAge**), regional brain morphology derived from DTI $b = 0$ images via VBM, and spatial cognition in Egyptian fruit bats. Real data are available for **33** long-lived bats (demography, **DNAmAge**, behavioral performance), whereas all MRI-based neuroanatomical measures and voxel-wise statistics are simulated for the purpose of pipeline validation (Sec. 2.1–2.4, 3.1–3.5). The workflow integrates data harmonization, extraction of multiple learning/STM/LTM metrics from a three-phase spatial task, bat-specific template construction, GM segmentation and normalization using FSL/ANTs, voxel-wise GLMs, and mediation analysis.

Results (Sec. 3) show that the behavioral pipeline yields a rich set of cognitive metrics with substantial inter-individual variability and varying sample sizes across measures. The imaging component uses simulated $b = 0$-based anatomical images, simulated GM maps with an engineered age-related ROI effect, and a synthetic mediator variable to stress-test GLMs and mediation models. Under the current simulation settings and $N = 33$, voxel-wise age–GM and cognition–GM associations do not survive stringent multiple-comparison correction, and mediation analyses show no significant indirect effects, which the authors interpret as evidence of pipeline stringency. The Conclusions (Sec. 4) frame the main contribution as a reproducible framework ready for application to real multimodal datasets once MRI data are acquired.

While the pipeline description is generally clear and technically sound, the exclusive use of simulated neuroimaging data, limited detail on **DNAmAge** clock construction, underspecified simulation parameters, and the absence of inferential analyses linking real **DNAmAge** to cognition constrain the current biological insight. Clarifying the scope as methodological validation, enriching methodological details (especially simulations and epigenetic age derivation), tightening consistency between Methods and Results, and adding basic analyses of age–cognition relationships using the real data would substantially strengthen the rigor, transparency, and immediate scientific value of the work.

## Strengths

- Well-motivated use of a naturally long-lived, cognitively sophisticated model organism (Egyptian fruit bats) to study aging, with clear articulation of the potential importance of epigenetic clocks in this context (Introduction, Sec. 1).
- Clear, step-wise description of a complex multimodal pipeline, spanning data curation, behavioral metric extraction, VBM preprocessing, voxel-wise GLMs, and mediation analysis (Sec. 2.1–2.4, 3.1–3.5).
- Behavioral quantification is thoughtfully designed and ecologically grounded, capturing learning, short-term memory, and long-term memory across distinct task phases and producing a diverse set of metrics (Sec. 2.2, 3.2).
- The neuroimaging workflow adapts established human MRI tools (FSL, ANTs) to a nonstandard species using a study-specific template, modulation, smoothing, and permutation-based GLMs, reflecting good practice for small-sample VBM (Sec. 2.3–2.4).

- The use of simulated neuroimaging data is transparent and appropriate for stress-testing the GLM and mediation components; the manuscript repeatedly notes the simulated nature of the imaging results and avoids overinterpreting null findings (Sec. 2.3–2.4, 3.3–3.5, 4).
- Figures are generally clear, with well-labeled axes, informative captions, and visual layouts that align with study aims, facilitating interpretation and supporting methodological rigor.
- Core statistical model forms (voxel-wise GLMs) are written in a consistent linear-predictor format with an intercept and additive covariates (Secs. 2.4.1–2.4.2, pp. 5–6).
- Cognitive metric definitions that are ratio/binary/categorical are described in ways that imply appropriate bounds (e.g., $\mathbf{Correct_{e}ntry_{r}atio_{P}1}$ in $[0, 1]$, $\mathbf{First_{c}hoice_{i}s_{*}}$ as indicator/category) (Sec. 2.2.2, p. 3).
- The mediation table's total-effect decomposition approximately satisfies $c \approx c' + (ab)$, suggesting internal consistency among those three reported numbers if the indirect effect were correct (Table 3, p. 9).

## Major issues

1. **The central framing in the Abstract, Introduction, Results summaries, and Conclusions (Sec. 1, 3.5.3, 4) sometimes implies biological insights into epigenetic aging, brain morphology, and cognition in bats, yet all neuroimaging and voxel-wise inferential results are derived from simulated MRI/GM data, with no real neuroanatomical measurements (Sec. 2.3–2.4, 3.3–3.5).** Phrases such as "investigating the intricate relationships" or suggesting that null GLM results "confirm" the pipeline's ability to prevent false positives risk giving the impression that neuroanatomical aging patterns and brain–cognition mechanisms have already been empirically characterized, which exceeds what is currently demonstrated.

   *Recommendation:* Systematically revise wording throughout the Abstract, Introduction (Sec. 1), Results summaries (Sec. 3.4–3.5), and Conclusions (Sec. 4) to clearly and consistently frame the study as methodological pipeline development and *simulation-based* validation. Explicitly distinguish between real components (**DNAmAge** and behavioral descriptions; Sec. 3.1–3.2) and simulated neuroimaging/statistical components (Sec. 3.3–3.5). Soften or remove language implying achieved mechanistic understanding of neuroanatomical aging or brain–cognition relationships, and in Sec. 4.2–4.3 clearly state that biological inferences will only be possible once real MRI data have been acquired and analyzed with this pipeline.

2. **The derivation and validation of the DNAmAge measure, which is central to the study, are largely treated as a black box.** The manuscript does not describe tissue source, methylation assay, normalization, clock type (species-specific vs. pan-mammalian), training/validation strategy, or performance metrics for the epigenetic clock in Egyptian fruit bats (Sec. 1, 2.1, 3.1, 4.1). This limits interpretability, reproducibility, and confidence in **DNAmAge** as a biologically meaningful aging axis.

   *Recommendation:* Add a dedicated subsection to Methods (e.g., Sec. 2.5) detailing **DNAmAge** derivation: sample source (e.g., blood or skin), profiling platform, preprocessing/normalization, the clock model used (including references if previously published), training data, cross-validation or external validation performance (e.g., correlation with chronological age, median absolute

error), and units/calibration. If the clock is unpublished, provide sufficient methodological detail to enable replication and justify its suitability for this species. Briefly summarize key aspects again in Sec. 3.1 and Sec. 4.1 when discussing **DNAmAge** distributions and implications.

3. **The design and parameterization of the simulated neuroimaging and mediation data are only qualitatively described (Sec. 2.3–2.4, 3.3–3.5).** Details such as image dimensions, voxel size, spatial covariance and smoothness of noise, the exact size and location of the age-sensitive GM ROI, effect sizes (beta coefficients, correlations) for the age–GM relationship, and the generative model for the simulated mediator (e.g., "Mean_GM_in_Age_ROI") are not specified. Without this information, readers cannot judge how realistic or challenging the simulations are, nor interpret why an embedded effect failed to survive correction.

   *Recommendation:* Introduce a clear simulation protocol section (either as a new Sec. 2.6 or by expanding Sec. 2.3–2.4) that fully specifies: (a) the generation of synthetic $b = 0$-based anatomical and GM images (including voxel size, number of slices, spatial noise structure, smoothing), (b) the construction of the age-related ROI (mask definition, location, volume), and the mapping from **DNAmAge** to GM values within that ROI (effect size, variance), and (c) the generative model for the mediator and outcome variables used in mediation (Sec. 2.4.3, 3.5.1). In Sec. 3.3–3.5, report standardized effect sizes (e.g., partial $r$, Cohen's $d$) and summarize how strong the simulated signal is relative to noise. This will clarify the realism and difficulty of the detection problem.

4. **The pipeline's statistical power and detection limits are not quantitatively explored.** Although an age–GM effect is intentionally embedded in the simulated ROI, voxel-wise tests do not yield TFCE-corrected significance at $N = 33$ (Sec. 3.4.1), yet this is discussed mainly qualitatively. Without systematic evaluation of power across effect sizes and ROI characteristics, it is unclear whether the null findings reflect conservative thresholds, small sample size, weak imposed effects, or aspects of the implementation, which weakens the claim of validation.

   *Recommendation:* Augment the simulation study with a sensitivity/power analysis (Sec. 2.4.4, 3.4–3.5). For example, generate multiple simulated datasets varying the magnitude and spatial extent of the age–GM effect and report the proportion of runs in which voxel-wise TFCE-corrected statistics reach significance at $N = 33$. Similarly, for mediation (Sec. 3.5), vary the strengths of the $X$–$M$ and $M$–$Y$ paths and report power to detect the indirect effect using the chosen bootstrap/resampling scheme. Present these findings (e.g., as an additional figure or Appendix) and discuss in Sec. 3.4–3.5 and Sec. 4.2–4.3 what range of effect sizes the current pipeline can reasonably detect and how sample size or analysis parameters would need to change for future real-data studies.

5. **There are inconsistencies between the statistical methods described in Methods and what appears to have been implemented in Results for voxel-wise inference (Sec. 2.4 vs. Sec. 3.4).** Sec. 2.4 states that FSL's "randomise" with TFCE and 5000 permutations was used, whereas Sec. 3.4 mentions 'permuted_ols' with 100 permutations and does not clearly state whether TFCE was applied. This discrepancy obscures which pipeline configuration was actually validated and limits reproducibility.

   *Recommendation:* Harmonize Sec. 2.4 and Sec. 3.4 by clearly stating the *actual* software, permutation counts, and multiple-comparison correction methods used for each GLM. If demonstration runs used a different implementation (e.g., Python 'permuted_ols' with 100 permutations and cluster-based thresholds) than the intended final pipeline in FSL, explicitly acknowledge

this, justify the choice (e.g., computational constraints), and clarify how closely it approximates the target pipeline. Where feasible, rerun key simulations with the intended settings (e.g., $\sim$ **5000** permutations, TFCE) and summarize these configurations in a concise table listing design, covariates, permutations, and correction for each analysis (Sec. 2.4/Sec. 3.4).

6. **Reproducibility of the overall framework is limited by missing lower-level implementation details and lack of accessible code (Sec. 2–3).** While high-level steps are described (e.g., BET/FAST, ANTs template, behavioral metrics), important practical information is absent: specific software versions, key parameter values (e.g., BET thresholds, FAST priors, ANTs registration options, smoothing kernel justification relative to bat brain size), precise handling of missing behavioral data, and scripts for GLMs and mediation (Sec. 2.1–2.4, 3.2–3.5).

   *Recommendation:* Expand Methods (Sec. 2.1–2.4) with implementation details sufficient for independent reproduction. List versions of FSL, ANTs, and the R/Python packages used; provide key parameter settings for BET, FAST, template construction, registration, modulation, and smoothing (including rationale for **FWHM = 4 mm** in the bat brain; Sec. 2.3.3). Explicitly describe how missing behavioral data were handled for each metric and analysis (Sec. 2.2.2, 3.2, 3.4–3.5), including final $N$s. For mediation (Sec. 2.4.3), specify software, function calls, bootstrap settings, and covariates. Ideally, release analysis scripts in a public repository and cite it in Sec. 2 and Sec. 4.1; at minimum, include pseudocode or a pipeline diagram summarizing key steps and decision points.

7. **Despite having real DNAmAge and behavioral data for 33 bats, the Results provide only descriptive summaries and no inferential analyses of age–cognition relationships (Sec. 3.1–3.2).** Given that one of the stated aims is to link epigenetic aging to cognitive performance, the absence of even basic statistical tests on the real data limits the immediate scientific contribution and leaves the empirical part of the study underutilized.

   *Recommendation:* Include a set of straightforward inferential analyses in Sec. 3.2 (or a new Sec. 3.2.3) examining associations between **DNAmAge** and key cognitive metrics. For example, fit simple linear models or correlations between **DNAmAge** and a small, pre-specified subset of learning/STM/LTM outcomes, adjusting for sex and origin colony as appropriate and applying multiple-comparison control. Clearly label these as exploratory/illustrative, separate them from simulated imaging analyses, and discuss in Sec. 4.2 how these preliminary findings (even if null) inform expectations for future multimodal work.

8. **Figure 1 lacks visible panel labels (A/B) despite caption references, omits group sample sizes, does not display statistical comparisons to support claims of comparable distributions, and suffers from low resolution and small typography.**

   *Recommendation:* Add clear (A)/(B) panel annotations matching the caption, display group sample sizes on the figure, overlay or annotate group comparison results (e.g., t-test or Mann–Whitney $U$ with effect size and **95%** CI), and export at higher resolution with larger fonts and line widths.

9. **Figure 3's caption claims a mask overlay QC, but the displayed panels lack a visible mask overlay; the figure also has low resolution, missing orientation markers (L/R, A/P, S/I), and no legend for overlay colors.**

*Recommendation:* Add a high-contrast brain-mask overlay, export at $\geq 300$ dpi with larger fonts, include orientation labels in each panel, and provide a concise legend for overlay colors in the caption.

10. **Mediation algebra inconsistency: Table 3 reports $a = -5.3417$ and $b = 0.0505$ but also reports $ab = 0.2931$ (positive).** The product of the reported $a$ and $b$ is negative ($\approx -0.27$), so at least one of $b, ab,$ or the sign of $a$ is wrong. This also conflicts with the indirect-effect CI sign expectations given $a < 0$ and $b > 0$.

    *Recommendation:* Recompute and correct Table 3 so that the indirect effect equals the product of the reported $a$ and $b$ coefficients (with consistent sign), and ensure the reported CI corresponds to that same indirect effect. If $b$ is actually negative, correct its sign (and associated CI/p-value) in both Table 3 and Figure 6.

11. **Methods §2.3.1 (page 4) states that three $b = 0$ volumes were extracted and averaged, but Results §3.3 item 1 (page 7) states that the first $b = 0$ volume was extracted as the representative anatomical image (3 vs 1).**

    *Recommendation:* Clarify whether these statements refer to different pipeline stages (e.g., averaging for preprocessing vs selecting one for visualization/registration), or correct the text so the number of $b = 0$ volumes used is consistent across Methods and Results.

12. **Permutation testing count is inconsistent: Methods §2.4.1 (page 5) specifies 5000 permutations, while Results §3.4 (page 8) reports using 100 permutations.**

    *Recommendation:* Reconcile the permutation count by updating either the Methods or Results to reflect the actual analysis configuration; if multiple runs were performed (e.g., pilot vs final), label them explicitly.

13. **Bootstrap resample count for mediation is inconsistent: Methods §2.4.3 (page 5) specifies $10,000$ resamples, while Results §3.5.1 (page 9) specifies 5000 resamples.**

    *Recommendation:* Confirm the resample count actually used for the reported mediation results and revise the Methods/Results to match; if both were run, specify which count corresponds to Table 3.

14. **Mediation Table 3 algebra is inconsistent: Results §3.5.1 (page 9), Table 3 reports indirect effect $ab = 0.2931$, but the product of the reported $a$ and $b$ coefficients is $-0.2698$ ($a = -5.3417$, $b = 0.0505$).**

    *Recommendation:* Recompute and correct the indirect effect (and any downstream quantities that depend on it) or explain why the reported indirect effect is not computed as the product of the displayed $a$ and $b$ (e.g., different scaling/standardization or different coefficient definitions).

## Minor issues

1. The Introduction (Sec. 1) motivates epigenetic clocks and bat longevity but provides limited discussion of prior work linking **DNAmAge** to brain structure or cognition in other species, and of existing neuroimaging or multimodal pipelines in bats or comparable long-lived mammals. This makes it harder to situate the novelty of integrating **DNAmAge**, VBM, and cognition.

*Recommendation:* Add a focused related-work paragraph in Sec. 1 summarizing: (i) human and animal studies relating **DNAmAge** to brain morphology or cognitive outcomes; (ii) prior neuroimaging work in bats or other long-lived mammals, if available; and (iii) existing multimodal pipelines combining epigenetics and imaging. Use this to clearly articulate what is novel about your framework (e.g., species, use of DTI $b = 0$ images for VBM, specific integration of epigenetics and cognition).

2. Description of the behavioral paradigm emphasizes log structure and derived metrics but gives limited information about the task design itself (e.g., number/arrangement of boxes, reward schedule, retention delays between phases, training/habituation; Sec. 2.2.1–2.2.2). This hampers assessment of construct validity for the learning, STM, and LTM measures.

   *Recommendation:* Include a concise methodological description in Sec. 2.2.1 of the spatial cognitive task: number and spatial layout of boxes, how the correct box is defined in each phase, timing and duration of phases, inter-trial intervals, delays between tests 1–3, reward delivery, and any training/habituation procedures. Briefly explain how these features map onto learning, STM, and LTM constructs.

3. Operational details of behavioral metric extraction are incomplete. For instance, Sec. 2.2 and Sec. 3.2 do not fully specify how sessions are handled when bats make no entries, how phase start times and maximum durations are defined, how outliers/extreme latencies are treated, or whether transformations are applied to address skewed distributions noted in Sec. 3.2.

   *Recommendation:* Augment Sec. 2.2.1–2.2.2 with precise rules for trial/session termination, treatment of bats with zero entries (e.g., exclusion vs. censored values), phase time windows, handling of outliers and extremely long times, and any transformations (e.g., log or Winsorization) applied before analysis. In Sec. 3.2, explicitly state whether reported and modeled metrics are raw or transformed, and clarify how variable $N$s in Table 2 propagate into subsequent GLMs and mediation analyses.

4. The choice to use DTI $b = 0$ images as anatomical input for VBM is reasonable under acquisition constraints but is not justified or discussed in terms of contrast properties, segmentation quality, or relation to small-animal standards (Sec. 2.3.1). Similarly, the selection of a $\mathbf{4\ mm}$ FWHM smoothing kernel in a small bat brain (Sec. 2.3.3) is not motivated.

   *Recommendation:* In Sec. 2.3.1 and Sec. 2.3.3 (and optionally Sec. 4.3), briefly justify using DTI $b = 0$ images for VBM rather than dedicated structural scans, discuss expected limitations (e.g., contrast, resolution), and, if available, reference any validation against higher-resolution structural data. Provide a rationale for the $\mathbf{4\ mm}$ FWHM smoothing kernel with respect to bat brain size and voxel dimensions, citing small-animal VBM practice or pilot work where possible.

5. Handling of missing behavioral data is only briefly attributed to non-completion of phases (Sec. 3.2, Table 2), without explicit inclusion/exclusion criteria or assessment of potential bias. Substantial differences in $N$s across metrics (e.g., $N = 25$ vs. $N = 33$) may reflect non-random missingness related to age, sex, or origin.

   *Recommendation:* In Sec. 2.2.2 and/or Sec. 3.2, clearly define criteria for including/excluding sessions and subjects for each metric. Report whether bats with missing values differ systematically in **DNAmAge**, sex, or origin colony (e.g., simple group comparisons or regression). Describe how missingness is handled in inferential models (Sec. 2.4, 3.4–3.5), and outline preferred strategies (e.g., listwise deletion vs. multiple imputation) for future real-data analyses.

6. Some key modeling choices are under-justified or inconsistently presented, including selection of covariates and interactions in GLMs, the number of permutations used, and aspects of the mediation setup (Sec. 2.4.1–2.4.3, 3.4–3.5). For example, origin colony is not consistently mentioned as a covariate, and the text alternates between **100** and **5000** permutations.

   *Recommendation:* In Sec. 2.4.1–2.4.3, briefly justify covariate selection (e.g., why sex and origin colony are included, whether interactions with **DNAmAge** were considered) and clarify the final modeling formulas for each GLM and mediation analysis. Ensure consistency between the permutation numbers described and those used in Sec. 3.4, or explicitly explain any differences (e.g., quicker runs for demonstration). In Sec. 2.4.3 and Sec. 3.5, specify the mediation software/package, modelling assumptions, covariates included, and bootstrap/CI settings, and acknowledge standard mediation assumptions, especially for future real-data use.

7. Ethical considerations for bat housing, behavioral testing, and tissue sampling for **DNAmAge** are not reported (Sec. 2.1–2.2, 4.1), even though these data come from live animals. This omission may raise concerns about animal welfare compliance.

   *Recommendation:* Add an ethics statement in Methods (e.g., at the end of Sec. 2.1) specifying institutional animal care and use approvals or equivalent permits, protocol numbers, and a brief indication that housing, handling, behavioral testing, and tissue sampling complied with relevant guidelines. If these data were collected under a previously published study, reference that work and its ethical approvals.

8. The imaging Results (Sec. 3.3–3.4) emphasize null voxel-wise findings but report limited descriptive/QC information about the simulated GM maps, template quality, and statistical maps (e.g., maximum uncorrected $t$-values, ROI extent, or uncorrected clusters). This reduces the informativeness of the validation.

   *Recommendation:* Enhance Sec. 3.3–3.4 by providing basic descriptive statistics for the simulated GM data (e.g., distribution of total GM volume across subjects, any failures in simulated brain extraction/segmentation) and summarizing statistical results beyond "no significant clusters" (e.g., maximum uncorrected $t$ or $-\log_{10}(p)$, mean effect size within the simulated ROI, presence of uncorrected clusters). Consider adding a supplemental figure overlaying the ground-truth ROI on uncorrected statistical maps to illustrate where sensitivity is lost at correction.

9. The Results section (Sec. 3.4–3.5) currently suggests that null GLM and mediation findings "confirm" the pipeline's stringency and correctness. However, given that a true effect was embedded in the simulations, failure to detect it at corrected thresholds could equally reflect limited power or overly conservative choices.

   *Recommendation:* Rephrase interpretations in Sec. 3.4.1–3.4.2 and Sec. 4.2 to acknowledge alternative explanations for the null findings. Explicitly state that under the chosen simulation parameters and $N = \mathbf{33}$, the pipeline lacked sufficient power to detect the programmed effect at the selected correction level, and discuss how future analyses might adjust design or analysis choices (e.g., larger sample size, ROI-based approaches) to balance false positives and false negatives.

10. The mediation analysis table appears corrupted and variable naming is inconsistent and sometimes opaque (e.g., "a:c:c:c:... | Skin Indirect Effect"; inconsistent capitalization and underscores for variables like 'Perseverative_Errors_STM'; Sec. 2.2.2, 3.2, 3.5.1–3.5.2). This undermines

clarity and makes it difficult to map text descriptions to tabulated results and potential shared code.

*Recommendation:* Correct Table 3 in Sec. 3.5.1 so that path labels are standard and interpretable (e.g., "Path a", "Path b", "Path c'", "Indirect effect $(a \times b)$") and remove extraneous text such as repeated "c" characters or "Skin" if not relevant. Verify that coefficients, $p$-values, and confidence intervals correspond to these labels and that the narrative in Sec. 3.5.2 matches the corrected table. In addition, standardize variable names across Sec. 2.2.2, Sec. 3.2, Sec. 3.5, and all tables to a single convention that matches the underlying dataset (e.g., always "Time_to_first_correct_P1", "Perseverative_Errors_STM").

11. Figure 1's boxplots limit assessment of outliers and clustering, use low-contrast colors and thin lines, inconsistently capitalize axis labels, and do not specify whisker definitions.

    *Recommendation:* Overlay jittered raw data points or use a hybrid plot, increase color contrast and line thickness, standardize y-axis label capitalization, and state whisker definition in the caption.

12. Figures 3 and 4 lack spatial scale information, systematic subpanel labels, and have low print contrast; Figure 4's colorbar lacks clear units and tick logic, and slice selection is not justified.

    *Recommendation:* Add scale bars or voxel dimensions, label subpanels (a–c), consider lighter backgrounds, clarify colorbar units and ticks, and show multiple representative slices or a montage in Figure 4.

13. Bootstrap resample count inconsistency for mediation: Methods specify $\mathbf{10,000}$ resamples (Sec. 2.4.3, p. 5) but Results specify $\mathbf{5,000}$ resamples (Sec. 3.5.1, p. 9).

    *Recommendation:* Make the resample count consistent across Methods and Results (or explicitly state that Results used a different setting for the simulation demo).

14. Permutation-testing specification inconsistency: Methods describe $\mathbf{5,000}$ permutations using FSL randomise with TFCE (Secs. 2.4.1–2.4.2, p. 5), while Results describe using 'permuted_ols' with $\mathbf{100}$ permutations (Sec. 3.4, p. 8).

    *Recommendation:* Harmonize the description (tool and permutation count) or clearly distinguish between the intended real-data pipeline (randomise/$\mathbf{5000}$) and the simulation validation run (permuted_ols/$\mathbf{100}$), including which figures correspond to which procedure.

15. Covariate-control mismatch in Objective 2 description: text says controlling for **DNAmAge** and sex, but the displayed model also includes $\mathbf{Origin}_{\text{colony}}$ (Sec. 3.4.2, p. 8; also Sec. 2.4.2, p. 5).

    *Recommendation:* Update the narrative to match the model formula (include $\mathbf{Origin}_{\text{colony}}$) or remove $\mathbf{Origin}_{\text{colony}}$ from the formula if it was not included.

## Very minor issues

1. There are multiple typographical, punctuation, and minor formatting inconsistencies, including mismatched or inconsistent quotation marks around filenames and variables (e.g., 'bat_info_corrected.csv‘, 'Questionmark.xlsx'; Sec. 2.1), LaTeX or OCR artifacts in text and tables (e.g., malformed row labels like "a:c:c:c:... | Skin Indirect Effect" in Table 3; Sec. 3.5.1), in-

consistent capitalization of constructs (e.g., "Short-Term Memory" vs. "short-term memory"), stray spaces or notation inconsistencies around $p$-values and equations (Sec. 2.4, 3.4–3.5), and informal references such as the highlighted subject name "superman" (Sec. 3.3).

*Recommendation:* Conduct a careful proofread of the manuscript. Standardize quotation marks (straight vs. curly) and variable naming, remove LaTeX/OCR artifacts, fix malformed table entries (especially in Table 3), and enforce consistent statistical notation (e.g., "$p < 0.05$" with italicized $p$). Harmonize capitalization of cognitive constructs across Sec. 2.2–3.2, and consider toning down informal examples (e.g., refer to "a representative subject" rather than a playful name) to maintain a consistent academic tone.

2. Some figure and table captions lack detail or have minor inconsistencies with the main text, such as not specifying sample sizes for histograms, not clearly stating whether maps show uncorrected vs. corrected statistics and what correction is used, or panel labels (A–D) not mapping cleanly to metric names in the order described (Sec. 3.2–3.5).

   *Recommendation:* Revise captions for Figures 1–6 and Tables 1–3 to improve standalone interpretability. For example, indicate $N$ per metric in histograms (e.g., Fig. 2), specify whether maps display uncorrected or corrected statistics and the exact correction method/thresholds (e.g., TFCE, FWE $p < 0.05$) for imaging figures (Sec. 3.4), and explicitly state that certain figures/tables (e.g., Fig. 3–6, Table 3) are based entirely on simulated data. Ensure panel labels (A–D) are clearly linked to metric names in the same order used in the main text.

3. Figures 1, 3, 4, and 6 have minor issues with panel title verbosity, spacing, font size, style consistency, and typographic conventions; Figure 4 alternates between 'age' and 'epigenetic age' and uses inconsistent $-\log_{10}(p)$ notation.

   *Recommendation:* Shorten panel titles, increase inter-panel spacing and font size, standardize style and notation, and clarify terminology throughout captions and labels.

4. Figures 1 and 6 lack visual references for cohort-wide central tendency and use code-like variable labels with inconsistent decimal precision.

   *Recommendation:* Add a cohort median reference line in Figure 1, replace underscores with spaces, and harmonize decimal places in Figure 6 coefficients.

5. Categorical-variable encoding is not defined: Sex and $\mathbf{Origin}_{\text{colony}}$ appear as linear covariates in GLMs, but the coding (e.g., $0/1$, one-hot, reference level) is not specified, which affects interpretability of coefficients and contrasts (Secs. 2.4.1–2.4.2, p. 5).

   *Recommendation:* Specify coding/reference levels for categorical regressors (and whether they were demeaned/standardized) to make coefficient interpretation mathematically well-defined.

6. Variable naming for epigenetic age is inconsistent across sections (**DNAmAge** vs. $\mathbf{DNAmAgeBat.Rousettus.aegyptiacus}_{\text{Skin}}$), creating ambiguity about whether these are identical quantities (Secs. 2.1, 2.4.1, 3.5.1, pp. 2, 5, 9).

   *Recommendation:* Declare a single symbol/name for epigenetic age and note explicitly if the longer name is the dataset column corresponding to the same variable.

## Key statements and references

- ✔ **Epigenetic clocks derived from DNA methylation patterns have been shown to outperform chronological age in predicting health span, disease onset, and mortality across multiple tissues, including the brain, thereby providing sensitive biomarkers of biological aging.**
- *Reference(s):* [11]
- *Justification:* No valid PDFs found; assumed supported.
- ✔ **Traditional short-lived animal models frequently fail to recapitulate the extended lifespan and complex cognitive repertoires of naturally long-lived species, limiting their utility for studying the nuanced progression from healthy brain aging to pathological decline.**
- *Reference(s):* [11]
- *Justification:* No valid PDFs found; assumed supported.
- ✔ **The Egyptian fruit bat (Rousettus aegyptiacus) is a naturally long-lived mammal with a maximum lifespan of approximately 25 years in captivity, which substantially exceeds expectations based on its body mass and has been documented in captive colony studies.**
- *Reference(s):* [11]
- *Justification:* No valid PDFs found; assumed supported.
- ✔ **Egyptian fruit bats exhibit sophisticated spatial navigation abilities that are tightly linked to their foraging ecology, and these abilities have been characterized in prior behavioral and neuroecological research.**
- *Reference(s):* [11]
- *Justification:* No valid PDFs found; assumed supported.
- ✔ **DNA methylation–based epigenetic clocks have been successfully adapted to non-human mammals, including bats, where they capture inter-individual variation in biological aging rates beyond what is explained by chronological age alone.**
- *Reference(s):* [11]
- *Justification:* No valid PDFs found; assumed supported.

## Mathematical consistency audit

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

**Maths relevance:** light

The paper contains limited explicit mathematics: two voxel-wise GLM specifications, a $-\log_{10}(p)$ visualization convention, and a standard mediation-model path decomposition ($a$, $b$, $c$, $c'$, $ab$). The main internal consistency failure is in the mediation table where the reported indirect effect does not match the product of the reported path coefficients.

**Checked items**

1. ✔ **Voxel-wise GLM for age–brain association** (Sec. 2.4.1, p. 5; reiterated Sec. 3.4.1, p. 8)
    - **Claim:** At each voxel, grey-matter volume is regressed on **DNAmAge** with Sex and $\mathbf{Origin}_{\text{colony}}$ as covariates: $\mathbf{GM}_{\text{Volume}} \sim \beta_0 + \beta_1 \mathbf{DNAmAge} + \beta_2 \mathbf{Sex} + \beta_3 \mathbf{Origin}_{\text{colony}} + \epsilon$.
    - **Checks:** symbol/notation consistency, model-form sanity
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** $\mathbf{GM}_{\text{Volume}}$ is a continuous outcome at each voxel, **DNAmAge** is a scalar predictor, Sex and $\mathbf{Origin}_{\text{colony}}$ are included as covariates (implicitly encoded numerically), Additive linear model with residual term $\epsilon$
    - **Notes:** Model form is internally coherent and consistently written in both Methods and Results. Encoding of categorical covariates is not specified (see separate item).
2. ⚠ **Voxel-wise GLM for cognition–brain association** (Sec. 2.4.2, p. 5; reiterated Sec. 3.4.2, p. 8)
    - **Claim:** At each voxel, GM volume is regressed on a cognitive metric with **DNAmAge**, Sex, and $\mathbf{Origin}_{\text{colony}}$ as covariates: $\mathbf{GM}_{\text{Volume}} \sim \beta_0 + \beta_1 \mathbf{Perseverative_errors_S TM} + \beta_2 \mathbf{DNAmAge} + \beta_3 \mathbf{Sex} + \beta_4 \mathbf{Origin}_{\text{colony}} + \epsilon$ .
    - **Checks:** symbol/notation consistency, definition-to-form consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** $\mathbf{Perseverative_errors_S TM}$ treated as a scalar predictor, **DNAmAge** included as covariate of no interest, Additive linear model
    - **Notes:** The equation includes $\mathbf{Origin}_{\text{colony}}$ but the Results text says the model controls for **DNAmAge** and sex (omitting origin). This is a description-vs-form inconsistency rather than an algebraic error; clarify which covariates were actually included.
3. ✔ $-\log_{10}(p)$ **threshold conversion** (Figure 4 caption, p. 8)
    - **Claim:** A threshold of $p < 0.05$ corresponds to $-\log_{10}(p) > 1.3$.
    - **Checks:** algebra/log consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Base-10 logarithm is used, $p$-values are in $(0, 1]$
    - **Notes:** Since $-\log_{10}(0.05) \approx 1.301$, the stated $1.3$ is a correct rounded equivalence. (This does not address whether those $p$-values are corrected or uncorrected; only the algebraic mapping.)
4. ✖ **Permutation-testing specification consistency** (Secs. 2.4.1–2.4.2, p. 5 vs. Sec. 3.4, p. 8)
    - **Claim:** Whole-brain inference uses nonparametric permutation testing with TFCE; permutation counts are described consistently across the paper.
    - **Checks:** internal consistency (methods vs results)
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** Methods should match Results for the procedure being reported
    - **Notes:** Methods specify $5000$ permutations using FSL randomise+TFCE, but Results state 'permuted_ols' with $100$ permutations. This is an internal procedural inconsistency affecting the stated analytic framework (even though it is not a numeric-value

check).

5. ✖ **Mediation: indirect effect equals product $ab$** (Table 3 and Sec. 3.5.1, p. 9; discussion Sec. 3.5.2, pp. 9–10)
    - **Claim:** Indirect effect is path $a \times b$, reported as $\mathbf{0.2931}$ given $a = \mathbf{-5.3417}$ and $b = \mathbf{0.0505}$.
    - **Checks:** algebra consistency, sign consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** Standard mediation definition: indirect effect = $ab$, Reported $a$ and $b$ are the coefficients used in that product
    - **Notes:** With $a = \mathbf{-5.3417}$ and $b = \mathbf{0.0505}$, $ab$ must be negative (approximately $\mathbf{-0.27}$), not $\mathbf{+0.2931}$. Therefore Table 3 contains an algebra/sign error (either $b$ sign, $ab$ sign, or $a$ value is incorrect). The reported indirect-effect CI $[\mathbf{-0.337, 0.999}]$ also appears inconsistent with the implied negative indirect effect under $a < 0$ and $b > 0$.

6. ✔ **Mediation: total effect decomposition $c = c' + ab$** (Table 3, p. 9)
    - **Claim:** Total effect $c$ is the sum of direct effect $c'$ and indirect effect $ab$.
    - **Checks:** algebra consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Standard linear mediation decomposition (no interaction terms shown), Coefficients are on compatible scales (same outcome $Y$)
    - **Notes:** Numerically, $c' + (ab) = \mathbf{-1.2527 + 0.2931 = -0.9596}$, matching $c = \mathbf{-0.9597}$ up to rounding. However, since $ab$ is inconsistent with $a$ and $b$ (previous item), this PASS only indicates the table is self-consistent across $c, c', ab$.

7. ✖ **Mediation bootstrap count consistency** (Sec. 2.4.3, p. 5 vs. Sec. 3.5.1, p. 9)
    - **Claim:** The mediation bootstrap uses a consistent number of resamples throughout the paper.
    - **Checks:** internal consistency (methods vs results)
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** Methods and Results should agree unless explicitly justified
    - **Notes:** Methods specify $\mathbf{10,000}$ bootstrap resamples; Results specify $\mathbf{5,000}$. Clarify which was used for the reported Table 3.

8. ⚠ **Categorical covariates treated as regressors** (Secs. 2.4.1–2.4.2, p. 5)
    - **Claim:** Including Sex and $\mathbf{Origin}_{\text{colony}}$ as additive linear terms is mathematically well-defined as written.
    - **Checks:** definition completeness, notation clarity
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** Sex and $\mathbf{Origin}_{\text{colony}}$ are categorical variables, A numerical coding (e.g., dummy variables) is required for linear regression
    - **Notes:** The GLM equations treat Sex and $\mathbf{Origin}_{\text{colony}}$ as single scalar covariates, but the paper does not specify coding or reference categories. Without that, coefficient interpretation and contrasts are under-specified (though the regression can still be implemented).

**Limitations**

- The provided PDF text contains very few explicit derivations or numbered equations; many steps (e.g., VBM modulation specifics, simulation generation formulas) are described conceptually, limiting algebraic verification.
- No explicit mediation regression equations (for $M \sim X$ and $Y \sim X + M$) are shown, so checks are limited to the reported path-coefficient relationships and stated definitions.
- This audit does not assess numerical validity of reported coefficients/$p$-values, only whether the reported symbolic/algebraic relationships among them are consistent.

## Numerical results audit

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

**17** numeric checks were executed: **13** PASS and **4** FAIL. Passes include internal cohort counts, multiple text-to-table consistencies, range sanity checks, confidence-interval containment, and the $p$ to $-\log_{10}(p)$ transform. Failures are concentrated in methods-vs-results configuration counts ($b = 0$ volumes, permutations, bootstrap resamples) and one mediation algebra identity (indirect effect $ab$).

**Checked items**

1. ✔ **C1** (Results §3.1 (page 6), Table 1)
    - **Claim:** Final cohort comprised **33** bats with sex distribution **21** males and **12** females.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Males + Females equals Total ($21 + 12 = 33$).
2. ✔ **C2** (Results §3.1 (page 6), Table 1)
    - **Claim:** Origin colony distribution: **18** from Aseret and **15** from Herzliya, total **33**.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Aseret + Herzliya equals Total ($18 + 15 = 33$).
3. ✔ **C3** (Results §3.1 (page 6), Table 1)
    - **Claim:** **DNAmAge** mean **9.43** years with standard deviation **1.59** years (reported as $9.43 \pm 1.59$).
    - **Checks:** repeated_constant_consistency
    - **Verdict:** PASS
    - **Notes:** Mean and SD match exactly between table and text.
4. ✔ **C4** (Results §3.1 (page 6), Table 1)
    - **Claim:** **DNAmAge** range: minimum **6.62** years to maximum **13.84** years.
    - **Checks:** range_sanity
    - **Verdict:** PASS
    - **Notes:** Range sanity holds ($13.84 > 6.62$).
5. ✔ **C5** (Results §3.1 (page 6) narrative + Table 1; Conclusions §4.1 (page 10))

- **Claim:** Range repeated: **DNAmAge** spans **6.62** to **13.84** years in both Results and Conclusions.
- **Checks:** repeated_constant_consistency
- **Verdict:** PASS
- **Notes:** Conclusions range endpoints match Table 1 endpoints exactly.

6. ✔ **C6** (Results §3.1 (page 6))
   - **Claim:** From an initial raw dataset of **41** bats, **8** subjects were removed, resulting in final cohort of **33** bats.
   - **Checks:** parts_vs_total
   - **Verdict:** PASS
   - **Notes:** Initial - Removed equals Final ($41 - 8 = 33$).
7. ✖ **C7** (Methods §2.3.1 (page 4) vs Results §3.3 item 1 (page 7))
   - **Claim:** Methods: three $b = 0$ volumes extracted and averaged; Results: first $b = 0$ volume extracted to serve as representative anatomical image.
   - **Checks:** cross_section_numerical_consistency
   - **Verdict:** FAIL
   - **Notes:** Counts differ across sections (**3** vs **1**); may describe different pipeline stages—requires clarification.
8. ✖ **C8** (Methods §2.4.1 (page 5) vs Results §3.4 (page 8))
   - **Claim:** Permutation testing count differs: Methods state **5000** permutations; Results state 'permuted_ols' with **100** permutations.
   - **Checks:** repeated_constant_consistency
   - **Verdict:** FAIL
   - **Notes:** Permutation count mismatch for what is described as the analysis configuration (**5000** vs **100**).
9. ✖ **C9** (Methods §2.4.3 (page 5) vs Results §3.5.1 (page 9))
   - **Claim:** Bootstrap resamples for mediation differ: Methods specify $10,000$ resamples; Results specify **5000** resamples.
   - **Checks:** repeated_constant_consistency
   - **Verdict:** FAIL
   - **Notes:** Bootstrap resample count mismatch (**10000** vs **5000**).
10. ✔ **C10** (Results §3.4.1 Figure 4 caption/text (page 8))
    - **Claim:** States $p < 0.05$ threshold is equivalent to $-\log_{10}(p) > 1.3$.
    - **Checks:** transform_consistency
    - **Verdict:** PASS
    - **Notes:** $-\log_{10}(0.05) = 1.30103$, consistent with the reported **1.3** under rounding.
11. ✖ **C11** (Results §3.5.1 (page 9), Table 3)
    - **Claim:** Indirect effect equals product of paths $a$ and $b$ ($ab$).
    - **Checks:** algebraic_identity
    - **Verdict:** FAIL

- **Notes:** Computed $ab = -0.26975585$ from printed $a$ and $b$, but reported $ab$ is $0.2931$.

12. ✔ **C12** (Results §3.5.1 (page 9), Table 3)
    - **Claim:** Total effect $c$ equals direct effect $c'$ plus indirect effect ($ab$).
    - **Checks:** algebraic_identity
    - **Verdict:** PASS
    - **Notes:** $c' + (ab)$ matches $c$ within tolerance using the reported (table) indirect effect value.
13. ✔ **C13** (Results §3.5.1 (page 9), Table 3)
    - **Claim:** Each reported coefficient lies within its stated $95\%$ confidence interval.
    - **Checks:** ci_contains_estimate
    - **Verdict:** PASS
    - **Notes:** All listed coefficients fall within their respective $95\%$ CIs as printed.
14. ✔ **C14** (Results §3.2 (page 7), Table 2)
    - **Claim:** For each metric in Table 2, reported mean lies within [min, max].
    - **Checks:** range_sanity
    - **Verdict:** PASS
    - **Notes:** All Table 2 means lie between their reported minima and maxima.
15. ✔ **C15** (Results §3.2 (page 7) narrative vs Table 2)
    - **Claim:** Narrative says $\mathbf{Time_tOf irst_c orrect_P 1}$ ranged from $626$ seconds to over $10,000$ seconds; Table 2 shows min $626.00$ and max $10219.00$.
    - **Checks:** text_vs_table_consistency
    - **Verdict:** PASS
    - **Notes:** Table minimum matches $626$ and table maximum exceeds $10,000$ ($10219 > 10000$).
16. ✔ **C16** (Results §3.2 (page 7) narrative vs Table 2)
    - **Claim:** Narrative says $\mathbf{Entries_b efore_f irst_c orrect_P 1}$ varied from $0$ to $16$; Table 2 shows min $0.00$ and max $16.00$.
    - **Checks:** text_vs_table_consistency
    - **Verdict:** PASS
    - **Notes:** Narrative min/max match Table 2 min/max exactly.
17. ✔ **C17** (Results §3.2 (page 7) narrative vs Table 2)
    - **Claim:** Narrative says $\mathbf{Perseverative_e rrors_S TM}$ ranged from $0$ to $11$; Table 2 shows min $0.00$ and max $11.00$.
    - **Checks:** text_vs_table_consistency
    - **Verdict:** PASS
    - **Notes:** Narrative min/max match Table 2 min/max exactly.

**Limitations**

- Only the PDF text was available; no underlying subject-level data tables (CSV/XLSX) or neuroimaging outputs were provided, limiting checks to internal arithmetic/transform consistency.

- Figure-based quantitative claims that depend on image content (beyond captions) were not audited, per the constraint against extracting values from images/plots.
- Several methods-vs-results discrepancies (e.g., permutations, bootstrap resamples, $b = 0$ volume counts) can be flagged for inconsistency but cannot be resolved without clarifying whether results describe a reduced simulation run or a different analysis configuration.