

Inflammation-Linked Aging Signals in Frozen Single-Cell Foundation Models: Donor-Aware Detection and Robustness Testing

Ihor Kendiukhov

`ihor.kendiukhov@student.uni-tuebingen.de`

University of Tübingen

Research Article

Keywords: aging, inflammaging, single-cell transcriptomics, foundation models, interpretability, immune aging

Posted Date: April 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9289672/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Inflammation-Linked Aging Signals in Frozen Single-Cell Foundation Models: Donor-Aware Detection and Robustness Testing

Ihor Kendiukhov¹

¹Department of Computer Science, University of Tübingen, Tübingen, Germany
kendiukhov@gmail.com

Abstract

Single-cell foundation models offer a possible route to studying aging biology in latent space, but apparent age effects can be distorted by donor identity and cell-composition differences. We developed a donor-aware interpretability workflow to test whether frozen scGPT and Geneformer representations contain biologically coherent aging signals in age-labeled human single-cell datasets. The workflow combined donor-held-out age decoding, sparse autoencoder feature discovery, cross-model pathway matching, targeted latent-space interventions, donor-bootstrap confidence intervals, and progressively stricter confound controls.

Across five datasets, frozen representations contained detectable age information, with best donor-aware balanced accuracy of 0.384. Sparse autoencoders identified 132 donor-aware robust features, and cross-model pathway matching produced 193 paired features, with the strongest convergence in inflammation and NF-kappaB-related programs. The clearest intervention signal was observed in a global Geneformer inflammatory branch, where old-versus-random and old-versus-young contrasts remained positive after split expansion and donor-threshold tightening up to 400 donors. In a monocyte-restricted analysis, both scGPT and Geneformer also showed positive old-versus-random responses in one cohort.

The strongest global signal weakened under stricter controls. Fully composition-matched forward-pass reruns yielded 0 of 4 full strict replications. These results indicate that frozen single-cell foundation models do capture biologically plausible aging-related structure, especially around inflammatory programs, but also that donor-aware and composition-aware stress tests are necessary before interpreting such signals as robust mechanisms.

Keywords: aging; inflammaging; single-cell transcriptomics; foundation models; interpretability; immune aging

1 Introduction

Aging is now understood as a systems-level process rather than a single molecular clock, with recurring hallmarks that include chronic inflammation, senescence-associated secretory signaling, stem-cell dysfunction, impaired proteostasis, and altered intercellular communication [Lopez-Otin et al., 2013, 2023, Kennedy et al., 2014, Campisi et al., 2019]. In immune and stromal compartments, age-related inflammation is especially prominent. The inflammaging literature links late-life dysfunction to persistent innate immune activation, NF-kappaB signaling, senescence-associated secretory phenotypes, and age-related rewiring of adaptive immunity [FRANCESCHI et al., 2000, Franceschi and Campisi, 2014, Furman et al., 2019, Ferrucci and Fabbri, 2018, Nikolich-Zugich, 2018, Goronzy and Weyand, 2019, Salminen et al., 2008, Chien et al., 2011, Hernandez-Segura et al., 2018, Basisty et al., 2020, Zhang et al., 2013].

At the measurement level, aging can often be detected from molecular profiles, including DNA methylation clocks, blood transcriptomic signatures, and large atlas-scale single-cell references [Horvath, 2013, Lu et al., 2019, Peters et al., 2015, Almanzar et al., 2020, Schaum et al.,

2020, Consortium* et al., 2022, Yazar et al., 2022]. What remains unclear is whether compact latent representations learned by large models encode merely enough information for age prediction, or whether they contain interpretable biological programs that track conserved aging mechanisms.

Single-cell foundation models provide a concrete setting in which to test that distinction. Transformer architectures have become the dominant substrate for large representation learners [Vaswani et al., 2017], and recent biological variants such as Geneformer and scGPT show that frozen transformer backbones can organize rich cellular structure in transcriptomic data [Theodoris et al., 2023, Cui et al., 2024]. In parallel, mechanistic-interpretability work has developed increasingly structured ways to study latent circuits and sparse features in large models [Olah et al., 2020, Bau et al., 2017, Bricken et al., 2023, Templeton et al., 2024]. Together, these developments motivate a harder question for biogerontology: do frozen single-cell foundation models encode aging biology in a form that survives mechanistic scrutiny?

This distinction matters because age can be decodable without supporting a defensible biological mechanism. Donor identity, cell-type imbalance, and cohort-specific composition shifts can all create apparent aging signals that disappear under stronger controls. We therefore focused on a harder target: not just whether age can be decoded from frozen embeddings, but whether candidate aging programs remain directional, biologically coherent, and donor-aware under increasingly strict stress tests.

2 Methods

2.1 Data

Core analyses centered on age-labeled human immune single-cell data already available in-project, with primary focus on two internal AIDA immune-aging cohorts. External and reference datasets included Tabula Sapiens-derived assets [Consortium* et al., 2022], the Yazar immune single-cell eQTL cohort [Yazar et al., 2022], and Allen immune aging-derived subsets. We treated atlas-scale external datasets primarily as stress-test context rather than as interchangeable pooled training data, because cross-cohort integration in single-cell studies is sensitive to normalization choices, batch effects, missingness, donor structure, and pseudoreplication [Luecken and Theis, 2019, Wolf et al., 2018, Hafemeister and Satija, 2019, Stuart et al., 2019, Haghverdi et al., 2018, Korsunsky et al., 2019, Tran et al., 2020, Hicks et al., 2018, Zimmerman et al., 2021, Squair et al., 2021].

The strongest-branch analyses were performed on the larger of the two AIDA cohorts, which had the greatest donor support in the deep-dive analysis ($n \approx 424$ in the baseline expanded-split run and $n \approx 333$ – 340 in the composition-matched reruns).

2.2 Models

We used frozen representations from:

- **scGPT** (contextual layer `scgpt_layer_09`) [Cui et al., 2024]
- **Geneformer** (contextual representation) [Theodoris et al., 2023]

Neither model was fine-tuned for age prediction in the core pipeline.

2.3 Interpretability workflow

The analysis proceeded in four blocks. First, we asked whether age-related signal was present at all, using donor-held-out linear probes and simple manifold diagnostics. Second, we searched for interpretable latent structure using sparse autoencoders and cross-model pathway matching,

motivated by sparse dictionary learning frameworks [Bricken et al., 2023, Templeton et al., 2024]. Third, we tested candidate pathways with targeted interventions, donor-bootstrap confidence intervals, and conservative promotion criteria. Finally, we stress-tested the strongest surviving branch with donor-threshold sweeps, donor-composition reweighting, fully composition-matched reruns, and multi-seed matched reruns.

Bootstrap confidence intervals used nonparametric donor resampling [Efron, 1979].

2.4 Statistical safeguards

Single-cell datasets are unusually vulnerable to optimistic results when cells are treated as independent evidence, when donor effects leak across splits, or when batch correction silently redefines the tested biological contrast [Crowell et al., 2020, Zimmerman et al., 2021, Squair et al., 2021]. For that reason, the pipeline was built around donor-aware data partitioning, donor-bootstrap uncertainty estimates, and contrast checks that remained interpretable after composition reweighting or composition-matched reruns. The goal was not to maximize nominal detection rate, but to suppress exactly the kinds of confounded positives that are common in atlas-scale analyses [Luecken and Theis, 2019, Tran et al., 2020, Hicks et al., 2018].

2.5 Strict gate policy

Claims were promoted only when multiple directional checks agreed: expected-age contrasts had to point in the correct direction, marginal effects had to remain supportive, and old-probability shifts had to be consistent. This policy was intentionally stricter than contrast-only screening because the project goal was mechanism-grade evidence rather than loose signal detection.

3 Results

3.1 Global trajectory across the analysis pipeline

Table 1 summarizes the main outcomes across the pipeline.

Table 1: Analysis outcomes across the full pipeline.

Block	Objective	Key outcome
1	Frozen age probe	Best balanced accuracy range: 0.299–0.384 across 5 datasets
2	Manifold geometry	Geometry gate passes: 0/5 datasets
3	SAE features	132 robust donor-aware sparse features retained
4	Cross-model convergence	193 pathway-matched cross-model feature pairs
6 (strict)	Model-specific claim promotion	0/4 strict claim passes
8 (strict audit)	Contrast + marginal strict gate	1/19 runs with at least one full strict row
9	Deep-dive robustness	Surviving row stable up to min-donors=400; fails at 500
10	Composition controls	Reweighting preserved strict pass; full compmatched forward re-run removed full strict pass
11	Compmatched seed panel	0/4 full strict replications under composition-matched seeds

3.2 Age can be decoded, but global age geometry is not robust

Frozen probes reached moderate performance (best BA = 0.3838 in AIDA phase 1 v2), indicating that age-associated signal is present in the frozen representations. However, the manifold-alignment criteria did not pass robustness gates in any dataset (0/5), even after donor-aware within-cell-type checks. Figure 1 shows this separation clearly: age is detectable, but it does not organize the global latent geometry in a stable way.

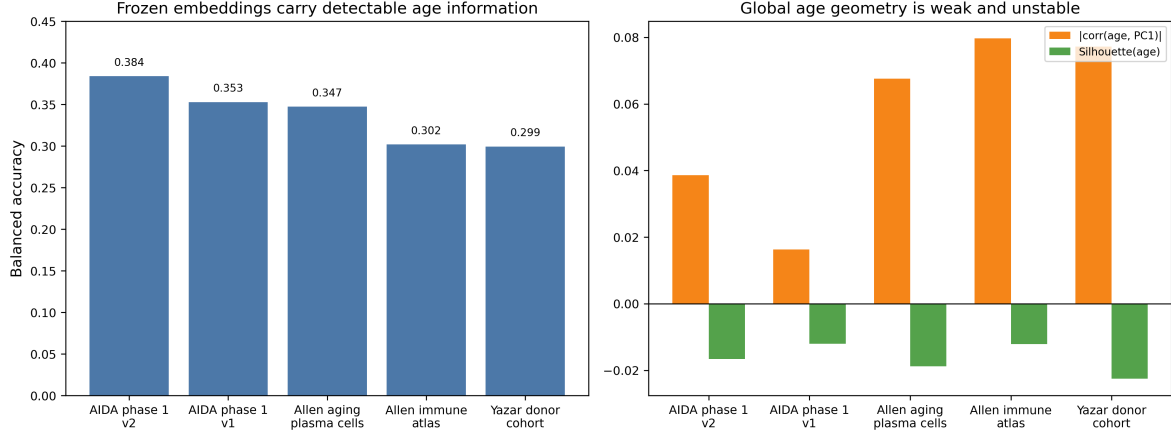


Figure 1: Detection versus geometry. Left: donor-aware age-decoding performance by dataset. Right: age-alignment geometry statistics remain weak and inconsistent, with no dataset passing the geometry gate.

3.3 Sparse features and cross-model overlap point to inflammatory programs

Sparse autoencoder analysis identified many donor-aware robust features (91 in scGPT core runs and 41 in Geneformer core runs). Figure 2 summarizes these counts.

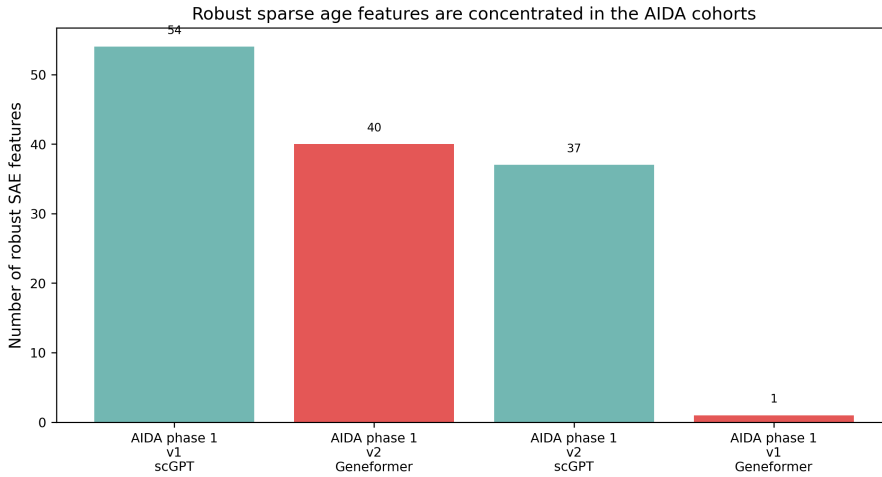


Figure 2: Donor-aware robust sparse features retained after filtering. Most of the signal comes from the two AIDA cohorts.

Cross-model pathway matching produced 193 pairs in aggregate. This overlap was dominated by inflammation / NF-kappaB in AIDA phase 1 v2 (175 pairs; best convergence score 0.0957), suggesting that both models capture a shared inflammatory latent program rather than completely unrelated features (Figure 3).

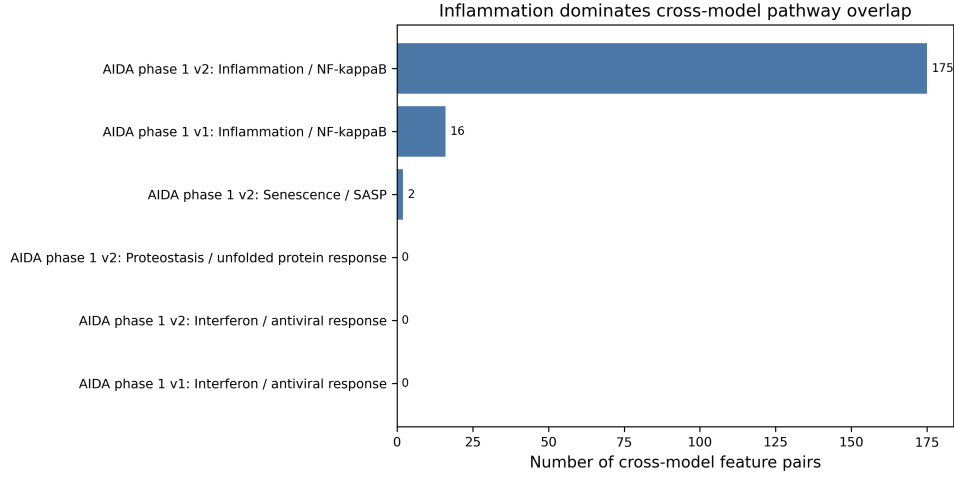


Figure 3: Cross-model pathway convergence. Inflammation / NF-kappaB dominates the shared signal between scGPT and Geneformer.

3.4 The clearest cell-type-specific signal appears in AIDA v1 monocytes

A clear positive result appeared in AIDA phase 1 v1 CD14 monocytes for inflammation / NF-kappaB at intervention scale 2.0:

- Geneformer old-minus-random expected-age contrast: 0.0357 [0.0314, 0.0398], positive.
- scGPT old-minus-random expected-age contrast: 0.0419 [0.0368, 0.0474], positive.

In contrast, AIDA phase 1 v2 monocytes were weak or discordant (Geneformer uncertain; scGPT negative), emphasizing that the signal is real but cohort-dependent (Figure 4).

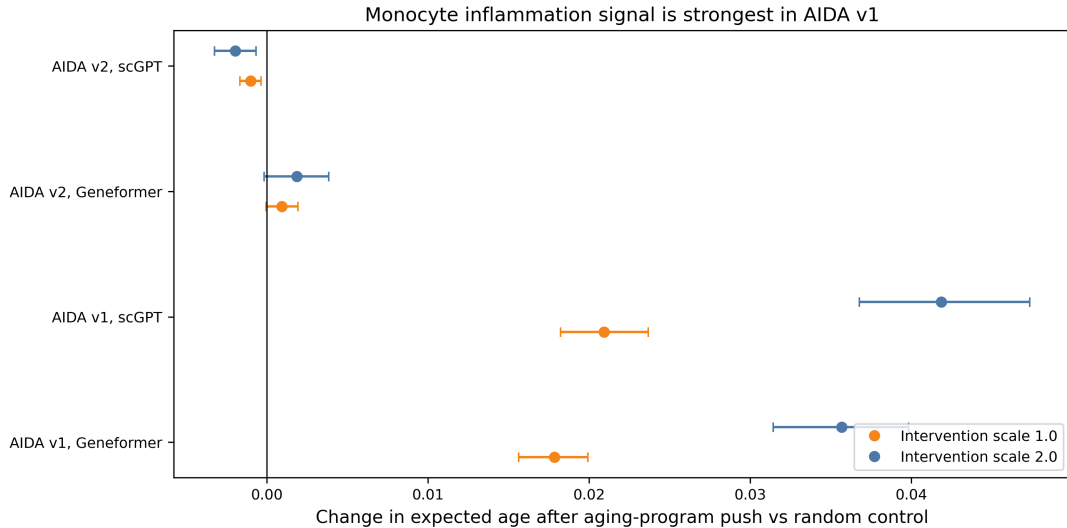


Figure 4: Monocyte-restricted inflammation intervention results. Positive old-versus-random effects are strongest in AIDA phase 1 v1, especially at intervention scale 2.0.

3.5 A single global Geneformer branch survives the deep-dive stress test

Under strict gating, one global branch survived: the AIDA phase 1 v2 Geneformer inflammation / NF-kappaB branch evaluated at global scope and intervention scale 2.0. After expanding split

seeds (60 splits per model), this branch remained full-strict positive up to a minimum donor threshold of 400, failing only at 500 because donor coverage became too sparse (Figure 5).

The key Geneformer expected-age contrasts in this deep-dive run were:

- old-minus-random: 0.1494 [0.1423, 0.1564], positive
- young-minus-random: -0.1324 [-0.1379, -0.1267], negative
- old-minus-young: 0.2817 [0.2698, 0.2940], positive

This was the strongest positive mechanistic candidate in the full project trajectory.

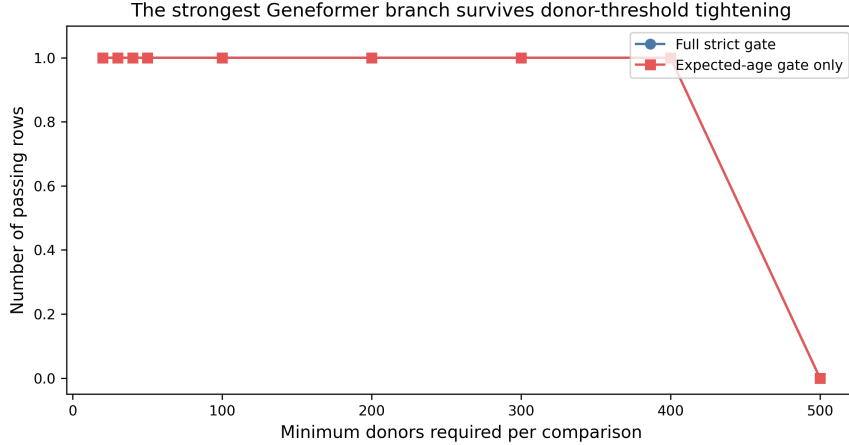


Figure 5: Donor-threshold stress test for the strongest surviving global branch. The full strict gate remains satisfied until the minimum donor threshold reaches 500.

3.6 Composition matching weakens the strongest global signal

Positive control result (reweighting): donor-composition reweighting preserved strict pass status across ridge penalties of 0.1, 1, 10, and 100, with near-zero effect shifts.

Stronger control result (full composition-matched forward rerun): the full strict pass disappeared ($n_{\text{full strict}} = 0$) despite one expected-age contrast-strict row. Figure 6 shows both points: simple reweighting barely changes the estimated effect sizes, but fully matched reruns do not preserve the baseline claim. Figure 7 compares the donor-bootstrap intervention intervals before and after full composition matching.

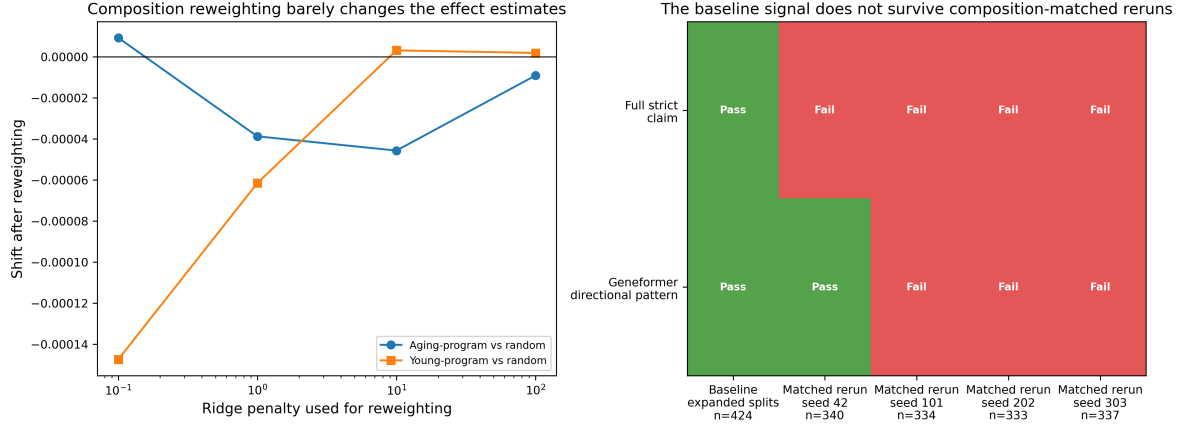


Figure 6: Composition controls. Left: donor-composition reweighting produces only negligible changes in the effect estimates. Right: the baseline strict claim does not replicate under composition-matched reruns, although one matched rerun retains the weaker Geneformer directional pattern.

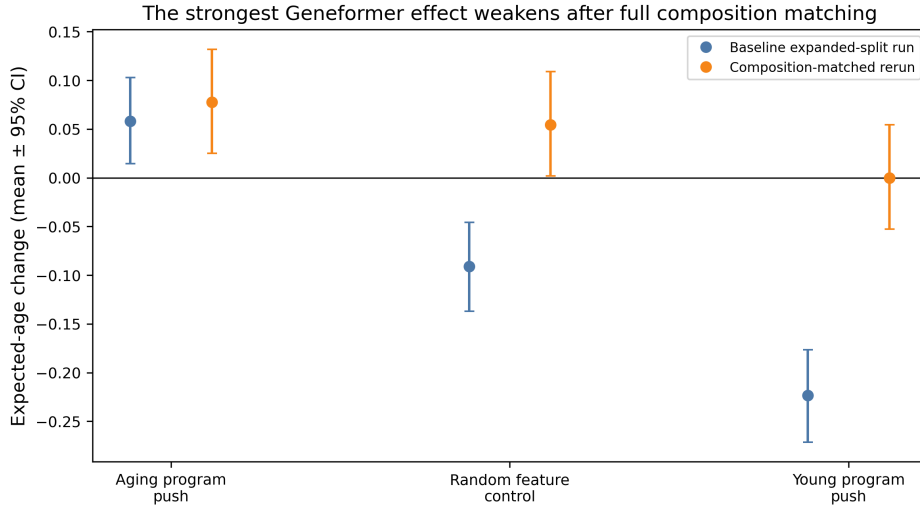


Figure 7: Geneformer donor-bootstrap intervention effects before and after full composition matching. The aging-program push remains positive, but the overall directional structure becomes less clean once full matching is enforced.

3.7 Matched reruns do not reproduce the baseline claim

A four-run composition-matched seed panel (seed42, 101, 202, 303) yielded:

- 0/4 full strict replications,
- Geneformer directional-pattern criterion retained in 1/4 runs,
- unstable random/young intervention directional structure across seeds.

Table 2 summarizes the panel-level strict outcomes.

Table 2: Composition-matched seed panel summary under the strict outcome criteria.

Run	$n_{\text{full strict}}$	Aging push	Random control	Young-program push
Baseline expanded-split run	1	positive	negative	negative
Composition-matched rerun, seed 42	0	positive	positive	uncertain
Composition-matched rerun, seed 101	0	uncertain	positive	positive
Composition-matched rerun, seed 202	0	positive	positive	positive
Composition-matched rerun, seed 303	0	positive	uncertain	uncertain

4 Discussion

A clear but mixed pattern emerged from the study. Frozen single-cell foundation models do contain aging-related structure, and some of that structure is biologically interpretable. We repeatedly recovered donor-aware sparse features, saw substantial cross-model overlap in inflammatory programs, and observed a strong cell-type-local monocyte signal in one of the two main immune-aging cohorts. This biological direction is plausible: inflammatory drift, NF-kappaB-linked signaling, senescence-associated secretory phenotypes, and immune remodeling are well-established features of aging biology [FRANCESCHI et al., 2000, Furman et al., 2019, Ferrucci and Fabbri, 2018, Goronzy and Weyand, 2019, Salminen et al., 2008, Chien et al., 2011, Basisty et al., 2020]. These findings are therefore more than generic age decoding. They indicate that frozen models can organize biologically meaningful aging-associated structure even without age-specific fine-tuning.

At the same time, the strongest global signal did not survive the most demanding control. The larger AIDA cohort yielded a Geneformer inflammation branch that remained convincing after split expansion, donor-threshold tightening, and donor-composition reweighting. However, once full composition matching was enforced and the analysis was repeated across multiple matched seeds, the full strict claim did not replicate. That result is not incidental. It is the central methodological message of the study.

4.1 Interpretation for longevity-focused mechanistic research

For longevity-focused mechanistic work, these results suggest a practical strategy:

- Use frozen models first to test signal existence and feature interpretability.
- Promote candidate mechanisms only after donor-aware, cell-type-aware, and composition-aware replication.
- Distinguish *model-specific interpretable signal* from *cross-model robust mechanism*; these are not equivalent.

4.2 Limitations

Key limitations include limited donor support in some strata and external subsets, reduced power under strict within-cell-type constraints, sensitivity of final strict status to the specific composition-matched sample realization, and incomplete higher-budget composition-matched reruns due runtime constraints.

5 Conclusion

This study shows that frozen scGPT and Geneformer representations contain meaningful and interpretable aging-associated latent programs, with the clearest evidence concentrated in inflammatory biology. However, under repeated fully composition-matched forward-pass controls, the strongest global signal did not replicate at the strictest decision threshold. The most accurate summary is therefore: **promising mechanistic signal, but not yet a robust mechanism-level claim.**

The highest-value next step is a higher-power composition-matched consensus protocol, with larger matched cell budgets and pooled multi-resample decision rules, to determine whether the deep-dive inflammatory signal reflects a stable mechanism or a sampling-sensitive artifact.

Statements and Declarations

Funding: The author received no specific funding for this work.

Competing interests: The author declares no competing interests.

Author contributions: The sole author conceived the study, designed the analyses, executed the computational work, interpreted the results, and wrote the manuscript.

Ethics approval: This study analyzed previously generated datasets and did not involve new recruitment or intervention on human participants within the present work.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Data availability: The public project repository includes the processed summary artifacts needed to reproduce the manuscript figures and tables. Access to the raw underlying single-cell datasets is governed by the original data providers; the repository documents dataset provenance and contains derived summary outputs only.

Code availability: The analysis code, manuscript source, figure-generation script, and reproducibility notes are available at <https://github.com/Biodyn-AI/longevity-mechinterp>.

References

- Nicole Almanzar, Jane Antony, Ankit S. Baghel, Isaac Bakerman, Ishita Bansal, et al. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583 (7817):590–595, 2020. doi: 10.1038/s41586-020-2496-1. URL <https://doi.org/10.1038/s41586-020-2496-1>.
- Nathan Basisty, Abhijit Kale, Ok Hee Jeon, Chisaka Kuehnemann, Therese Payne, Chirag Rao, et al. A proteomic atlas of senescence-associated secretomes for aging biomarker development. *PLOS Biology*, 18(1):e3000599, 2020. doi: 10.1371/journal.pbio.3000599. URL <https://doi.org/10.1371/journal.pbio.3000599>.
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, 2017. doi: 10.1109/CVPR.2017.354. URL <https://doi.org/10.1109/CVPR.2017.354>.
- Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning, 2023. URL <https://arxiv.org/abs/2305.01610>. Validated against the transformer-circuits publication page.
- Judith Campisi, Pankaj Kapahi, Gordon J. Lithgow, Simon Melov, John C. Newman, and Eric Verdin. From discoveries in ageing research to therapeutics for healthy ageing. *Nature*, 571

- (7764):183–192, 2019. doi: 10.1038/s41586-019-1365-2. URL <https://doi.org/10.1038/s41586-019-1365-2>.
- Yuchen Chien, Claudio Sciuoppo, Xiaowo Wang, Xueping Fang, Brian Balgley, Jessica E. Bolden, et al. Control of the senescence-associated secretory phenotype by nf-b promotes senescence and enhances chemosensitivity. *Genes & Development*, 25(20):2125–2136, 2011. doi: 10.1101/gad.17276711. URL <https://doi.org/10.1101/gad.17276711>.
- The Tabula Sapiens Consortium*, Robert C. Jones, Jim Karkanias, Mark A. Krasnow, Angela Oliveira Pisco, Stephen R. Quake, et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594), 2022. doi: 10.1126/science.abl4896. URL <https://doi.org/10.1126/science.abl4896>.
- Helena L. Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11(1), 2020. doi: 10.1038/s41467-020-19894-4. URL <https://doi.org/10.1038/s41467-020-19894-4>.
- Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL <https://doi.org/10.1038/s41592-024-02201-0>.
- B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1979. doi: 10.1214/aos/1176344552. URL <https://doi.org/10.1214/aos/1176344552>.
- Luigi Ferrucci and Elisa Fabbri. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. *Nature Reviews Cardiology*, 15(9):505–522, 2018. doi: 10.1038/s41569-018-0064-2. URL <https://doi.org/10.1038/s41569-018-0064-2>.
- C. Franceschi and J. Campisi. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 69(Suppl 1):S4–S9, 2014. doi: 10.1093/gerona/glu057. URL <https://doi.org/10.1093/gerona/glu057>.
- CLAUDIO FRANCESCHI, MASSIMILIANO BONAFE, SILVANA VALENSIN, FABIOLA OLIVIERI, MARIA DE LUCA, ENZO OTTAVIANI, et al. Inflammaging: An evolutionary perspective on immunosenescence. *Annals of the New York Academy of Sciences*, 908(1): 244–254, 2000. doi: 10.1111/j.1749-6632.2000.tb06651.x. URL <https://doi.org/10.1111/j.1749-6632.2000.tb06651.x>.
- David Furman, Judith Campisi, Eric Verdin, Pedro Carrera-Bastos, Sasha Targ, Claudio Franceschi, et al. Chronic inflammation in the etiology of disease across the life span. *Nature Medicine*, 25(12):1822–1832, 2019. doi: 10.1038/s41591-019-0675-0. URL <https://doi.org/10.1038/s41591-019-0675-0>.
- Jorg J. Goronzy and Cornelia M. Weyand. Mechanisms underlying t cell ageing. *Nature Reviews Immunology*, 19(9):573–583, 2019. doi: 10.1038/s41577-019-0180-1. URL <https://doi.org/10.1038/s41577-019-0180-1>.
- Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1), 2019. doi: 10.1186/s13059-019-1874-1. URL <https://doi.org/10.1186/s13059-019-1874-1>.

- Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018. doi: 10.1038/nbt.4091. URL <https://doi.org/10.1038/nbt.4091>.
- Alejandra Hernandez-Segura, Jamil Nehme, and Marco Demaria. Hallmarks of cellular senescence. *Trends in Cell Biology*, 28(6):436–453, 2018. doi: 10.1016/j.tcb.2018.02.001. URL <https://doi.org/10.1016/j.tcb.2018.02.001>.
- Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018. doi: 10.1093/biostatistics/kxx053. URL <https://doi.org/10.1093/biostatistics/kxx053>.
- Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10), 2013. doi: 10.1186/gb-2013-14-10-r115. URL <https://doi.org/10.1186/gb-2013-14-10-r115>.
- Brian K. Kennedy, Shelley L. Berger, Anne Brunet, Judith Campisi, Ana Maria Cuervo, Elissa S. Epel, et al. Geroscience: Linking aging to chronic disease. *Cell*, 159(4):709–713, 2014. doi: 10.1016/j.cell.2014.10.039. URL <https://doi.org/10.1016/j.cell.2014.10.039>.
- Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019. doi: 10.1038/s41592-019-0619-0. URL <https://doi.org/10.1038/s41592-019-0619-0>.
- Carlos Lopez-Otin, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013. doi: 10.1016/j.cell.2013.05.039. URL <https://doi.org/10.1016/j.cell.2013.05.039>.
- Carlos Lopez-Otin, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. Hallmarks of aging: An expanding universe. *Cell*, 186(2):243–278, 2023. doi: 10.1016/j.cell.2022.11.001. URL <https://doi.org/10.1016/j.cell.2022.11.001>.
- Ake T. Lu, Austin Quach, James G. Wilson, Alex P. Reiner, Abraham Aviv, Kenneth Raj, et al. DNA methylation grimage strongly predicts lifespan and healthspan. *Aging*, 11(2):303–327, 2019. doi: 10.18632/aging.101684. URL <https://doi.org/10.18632/aging.101684>.
- Malte D Luecken and Fabian J Theis. Current best practices in singlecell RNAseq analysis: a tutorial. *Molecular Systems Biology*, 15(6), 2019. doi: 10.15252/msb.20188746. URL <https://doi.org/10.15252/msb.20188746>.
- Janko Nikolich-Zugich. The twilight of immunity: emerging concepts in aging of the immune system. *Nature Immunology*, 19(1):10–19, 2018. doi: 10.1038/s41590-017-0006-x. URL <https://doi.org/10.1038/s41590-017-0006-x>.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3), 2020. doi: 10.23915/distill.00024.001. URL <https://doi.org/10.23915/distill.00024.001>.
- Marjolein J. Peters, Roby Joehanes, Luke C. Pilling, Claudia Schurmann, Karen N. Conneely, Joseph Powell, et al. The transcriptional landscape of age in human peripheral blood. *Nature Communications*, 6(1), 2015. doi: 10.1038/ncomms9570. URL <https://doi.org/10.1038/ncomms9570>.

- Antero Salminen, Jari Huuskonen, Johanna Ojala, Anu Kauppinen, Kai Kaarniranta, and Tiina Suuronen. Activation of innate immunity system during aging: NF- κ B signaling is the molecular culprit of inflamm-aging. *Ageing Research Reviews*, 7(2):83–105, 2008. doi: 10.1016/j.arr.2007.09.002. URL <https://doi.org/10.1016/j.arr.2007.09.002>.
- Nicholas Schaum, Benoit Lehallier, Oliver Hahn, Robert Palovics, Shayan Hosseinzadeh, Song E. Lee, et al. Ageing hallmarks exhibit organ-specific temporal signatures. *Nature*, 583(7817):596–602, 2020. doi: 10.1038/s41586-020-2499-y. URL <https://doi.org/10.1038/s41586-020-2499-y>.
- Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1), 2021. doi: 10.1038/s41467-021-25960-2. URL <https://doi.org/10.1038/s41467-021-25960-2>.
- Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, et al. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019. doi: 10.1016/j.cell.2019.05.031. URL <https://doi.org/10.1016/j.cell.2019.05.031>.
- Adly Templeton, Trenton Bricken, Joshua Batson, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2024. URL <https://arxiv.org/abs/2406.04093>. Validated against the transformer-circuits publication page.
- Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9. URL <https://doi.org/10.1038/s41586-023-06139-9>.
- Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, 21(1), 2020. doi: 10.1186/s13059-019-1850-9. URL <https://doi.org/10.1186/s13059-019-1850-9>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 2017. URL <https://papers.neurips.cc/paper/7181-attention-is-all-you-need>.
- F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 2018. doi: 10.1186/s13059-017-1382-0. URL <https://doi.org/10.1186/s13059-017-1382-0>.
- Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, et al. Single-cell eqtl mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376(6589), 2022. doi: 10.1126/science.abf3041. URL <https://doi.org/10.1126/science.abf3041>.
- Guo Zhang, Juxue Li, Sudarshana Purkayastha, Yizhe Tang, Hai Zhang, Ye Yin, et al. Hypothalamic programming of systemic ageing involving ikk-, nf-b and GnRH. *Nature*, 497(7448): 211–216, 2013. doi: 10.1038/nature12143. URL <https://doi.org/10.1038/nature12143>.
- Kip D. Zimmerman, Mark A. Espeland, and Carl D. Langefeld. A practical solution to pseudoreplication bias in single-cell studies. *Nature Communications*, 12(1), 2021. doi: 10.1038/s41467-021-21038-1. URL <https://doi.org/10.1038/s41467-021-21038-1>.