Unsupervised Learning Recovers Nonlinear Aging Signatures in DNA Methylation

Falak Pabari

Department of Computer Science
Brown University

Advisor: Ritambhara Singh

Second Reader: Ying Ma

April 2026

Acknowledgements

I am grateful to Ritambhara Singh for her guidance throughout this project, for knowing when to push and when to trust the data, for being patient with me through the iterations that didn't work, and for teaching me everything I know about this project. This thesis exists because she believed it was worth asking.

To Professor Ying Ma, thank you for your careful reading and for the generosity of your time. Working in your lab on perturbation modeling shaped how I think about latent representations in biology, and that thinking is quietly present throughout this thesis even where it isn't cited.

To my friends who kept me sane through the final push, thank you. And to my parents, who have never once doubted me even when I doubted myself: this is for you.

Abstract

DNA methylation changes with age at hundreds of thousands of CpG sites, but methods that identify nonlinear aging trajectories, sites undergoing inflection, acceleration, or deceleration rather than simple monotonic change, depend on age labels at every analytical step. We show that much of this nonlinear structure is recoverable from trajectory shape alone, without exposing age to the model's forward pass. Using the Hannum cohort (656 samples, 470,043 CpGs), we train a hybrid VAE+Contrastive encoder on 20-bin mean-methylation trajectories and cluster the 16-dimensional latent space with a Gaussian Mixture Model ($k = 10$). Age information enters only offline, through Spearman $r(\text{age}, \text{CpG})$ used to mine training triplets; the encoder input is the trajectory vector.

Two of the ten modules concentrate 53.8% of all SNITCH-classified nonlinear (NL) CpGs. Module 3 (67,234 CpGs, 428 NL, OR = 4.97, FDR = 8×10^{-115}) is a broad active-promoter module enriched for KLF/Sp-family motifs. Module 9 (15,606 CpGs, 81 NL, OR = 2.74, FDR = 1.4×10^{-13}) is a smaller enhancer and CTCF-flanking module with a nonlinear hypomethylation trajectory; its PC1 eigenvalue shows a significant age-adjusted association with a CRP-based inflammation proxy ($\beta = -8.6 \times 10^{-4}$, $p = 0.040$). Module 9's motif signature, CTCF, BORIS, NF1(CTF), NF1-halfsite, matches those independently reported by Grolaux et al. (2026) in age-supervised NL clusters on EPICv2 data from a different cohort. The convergence across platform, cohort, and supervision regime implicates this specific TF axis as a reproducible feature of nonlinear epigenetic aging, and suggests that trajectory-shape geometry is an intrinsic organizing property of the methylation profile rather than a statistical artifact of age regression.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Biology of Epigenetic Aging

DNA methylation changes with age at hundreds of thousands of CpG sites across the human genome, and these changes are systematic enough to support biological age predictors that rival chronological self-report. But the methods that discover these age-associated changes are overwhelmingly linear: they fit a straight line between methylation and age at each CpG, select the sites where that line has the largest slope, and discard the rest. This is a methodological choice with a biological consequence. Linear models preferentially detect features that change monotonically with age, and therefore systematically miss, or mischaracterize, the nonlinear signals that multiple orthogonal lines of evidence, proteomic waves at ages 34, 60, and 78 (Lehallier et al., 2019), multi-omic crests in the fifth and seventh decades (Shen et al., 2024), non-constant telomere attrition, suggest dominate the molecular biology of aging. The field has the data to see nonlinearity. It largely lacks the methods.

The canonical epigenetic clocks, Hannum et al. (2013) and Horvath (2013), established the core paradigm. Both trained penalized regression models (ElasticNet and LASSO respectively) on CpG beta values to predict chronological age, selecting small sets of CpGs whose methylation levels correlate linearly with age. Hannum's clock used 656 whole-blood samples and 71 CpG sites; Horvath's multi-tissue clock used 8,000+ samples across 51 tissue types and 353 sites. These clocks launched an industry: biological age acceleration derived from epigenetic clocks has since been associated with mortality, cancer risk, cognitive decline, and a range of age-related diseases. Subsequent clocks have extended this paradigm to pace-of-aging phenotypes (Belsky et al., 2022) and biological mortality risk (Lu et al., 2019), but the commitment to linearity remains.

What all of these clocks share, and what most subsequent methods inherit, is linearity. The relationship between age and methylation is modeled as a straight line per CpG, a simplification that is statistically tractable and works reasonably well for clock construction, but that systematically discards a class of biologically meaningful signal. There is now evidence that molecular aging is not a smooth, uniform process. Proteomics studies (Lehallier et al., 2019) identified three major waves of protein abundance change across adulthood, concentrated near ages 34, 60, and 78, inconsistent with smooth linear drift. Multi-omics

analyses (Shen et al., 2024) have found coordinated non-monotonic changes across metabolites, proteins, and microbiome composition. Telomere dynamics show inflection with age rather than constant attrition. These findings converge on a picture of aging as a punctuated, nonlinear process, not a gradual slope.

SNITCH (Grolaux et al., 2026) is the most direct attempt to detect nonlinear DNAm trajectories at scale. For each of 470,043 CpGs, SNITCH fits both a linear model and a generalized additive model (GAM) to age, and classifies the CpG as Nonlinear (NL) if the GAM explains meaningfully more variance (ΔBIC > 2). NL CpGs are further clustered by trajectory shape using functional PCA. On the Hannum cohort, SNITCH identifies 946 NL CpGs, sites that, by definition, encode information about aging that linear models cannot capture. These 946 CpGs are enriched in specific chromatin states, carry specific TF binding motif signatures, and cluster into functionally coherent trajectory shape groups. The implication is simple: there is a real, biologically structured signal in DNA methylation that linear methods systematically miss.

1.2 The Age-Label Problem

Every existing method for identifying nonlinear aging trajectories in DNA methylation requires age labels. SNITCH fits GAM and LM models to age at every CpG; age is not a convenience but a structural requirement. Epigenetic clocks regress directly on age. Methods that cluster CpG trajectories into aging-related groups presuppose that age-correlated change is the organizing principle. Even approaches that describe themselves as "unsupervised" typically feed age-derived features, slope, correlation, mean methylation across age strata, into their clustering step, encoding age information indirectly.

Reducing this dependence offers practical value across several settings. Consider contexts where age supervision is difficult, unreliable, or undesirable:

Longitudinal disease cohorts. In cohorts where aging and disease onset are intertwined, early-onset neurodegenerative disease, accelerated aging syndromes, chronic inflammatory conditions, age at sampling may be confounded with the very biological processes the researcher wishes to study. A method that requires age as input cannot cleanly separate what is "normal aging" from what is "disease-associated aging." A representation trained without direct age supervision could, in principle, be applied to the disease cohort without importing that confound.

Non-human systems. Aging in model organisms (*Mus musculus*, *C. elegans*, primate) proceeds on different timescales and may not map cleanly to human chronological age. Building a nonlinear trajectory classifier for mouse aging that requires human-calibrated

age inputs is not straightforward; a trajectory-shape representation learned from temporal ordering is more naturally transferable.

Archival and heterogeneous cohorts. Large biobanks often have imprecise age metadata, especially for archival samples or samples collected under varied protocols. Small cohorts collected for other purposes frequently lack the age range or sample density needed to fit age-conditioned models reliably. Methods that localize their age dependence to an offline precomputation step are more tolerant of such metadata limitations.

Discovery in novel contexts. The most interesting biological question may not be "which CpGs change with age" but "which CpGs undergo structurally similar trajectory changes, regardless of what drives those changes." A trajectory-shape-based representation organizes CpGs by the geometry of their temporal profiles, a query that does not require age labels in the inference step once the representation is learned.

The most powerful existing tools for nonlinear aging analysis require age at every analytical step, as the predictor in regression, as the smooth-function argument in GAMs, and as the ordering variable for functional PCA. A method that relaxes even one of these dependencies offers practical value. If the forward pass of a representation learner can be made age-free, the learned representation can be used in downstream analyses (module eigenvalue regressions, trajectory-similarity queries, cross-cohort transfer) without propagating age metadata through every step. If the resulting representation independently recovers the same biological structure that age-supervised methods identify, the convergence is evidence that nonlinear trajectory organization is an intrinsic property of the methylation profile, encoded in trajectory shape itself, rather than an artifact of the statistical relationship between methylation and age.

The framework developed here makes age-dependence explicit and localized: Spearman $r(\text{age}, \text{CpG})$ is computed once, offline, to mine training triplets, but age is never passed to the encoder. This is a methodological middle ground, not a fully age-free pipeline, and we discuss its limitations in Section 5.4. The practical claim is more modest than "no age anywhere" and more useful than existing clocks: the learned representation, once trained, is a geometry over trajectory shapes that can be queried, clustered, and associated with downstream phenotypes independently of the age metadata used to construct it.

1.3 Why Unsupervised Representation Learning

Variational autoencoders (Kingma & Welling, 2013) learn to compress high-dimensional data into a low-dimensional latent space that captures the data's generative structure. A well-trained VAE does not merely memorize its inputs; it learns a smooth, structured manifold in which similar inputs occupy nearby positions. For CpG methylation trajectories, this

means two CpGs whose methylation patterns change similarly across the lifespan should be geometrically close in the latent space, even if the model has never been told their age associations.

Contrastive learning adds a second pressure. Triplet loss training (Schroff et al., 2015) encourages the encoder to place CpGs with similar Spearman correlations between methylation and age near each other, and CpGs with dissimilar correlations far apart. This approach builds on a broader framework of contrastive representation learning (Chen et al., 2020) that has proven effective across many domains. Crucially, triplet mining does not require age labels in the forward pass; it requires only a precomputed similarity signal (Spearman r) used to construct training triplets offline. The model never sees age during training; it only sees 20-bin methylation trajectory vectors.

Together, these objectives should produce a latent space where trajectory shape, the temporal pattern of methylation change across the lifespan, determines geometric position. Linear increasers cluster together, linear decreasers cluster together, and, if the signal is strong enough, nonlinear trajectories form their own geometric neighborhoods.

The empirical test is whether this space independently recovers SNITCH's findings. If the unsupervised latent space clusters NL CpGs together, enriches for the same chromatin states and TF motifs that SNITCH identifies, and produces biologically interpretable modules, this convergence is evidence that nonlinear trajectory organization is an intrinsic property of the methylation profile itself, not merely a statistical artifact of the correlation between methylation and age.

1.4 Research Questions and Contributions

This thesis addresses three questions:

1. **Geometric recovery.** Can an unsupervised model trained only on 20-bin methylation trajectories learn a latent space in which SNITCH's NL CpGs are geometrically concentrated, forming distinct, enriched clusters rather than being scattered uniformly?
2. **Biological convergence.** Do the emergent modules independently recover the biological annotations that SNITCH identifies, chromatin state enrichments, TF binding motif signatures, and inflammatory associations, without ever accessing those annotations during training?
3. **Methodological implications.** What does convergence between an unsupervised method on 450k data and an age-supervised method on EPICv2 data tell us about the underlying biology of nonlinear epigenetic aging?

Contributions:

- We demonstrate that a hybrid VAE+Contrastive objective recovers nonlinear aging structure without age supervision, outperforming VAE-only, contrastive-only, and raw-feature GMM baselines on biologically validated metrics.
- We show that two latent clusters (Module 3 and Module 9) jointly capture 53.8% of SNITCH-identified NL CpGs, with significantly enriched TF motif signatures that replicate across platforms (450k vs. EPICv2) and cohorts (Hannum vs. Grolaux).
- We provide an age-supervised/unsupervised comparison framework that allows principled evaluation of representation learning quality against an external biological reference.

CHAPTER 2

Background

2.1 DNA Methylation and the Hannum Cohort

DNA methylation refers to the addition of a methyl group to the cytosine base at CpG dinucleotides, positions in the genome where cytosine is followed by guanine. This modification is heritable across cell division, is established and maintained by a family of DNA methyltransferase enzymes (DNMT1, DNMT3A, DNMT3B), and is reversible by ten-eleven translocation (TET) demethylases. In differentiated human somatic cells, the majority of CpG sites ($\sim$70-80%) are constitutively methylated; the exceptions are CpG islands, dense clusters of CpG dinucleotides typically found at gene promoters, which are hypomethylated in transcriptionally active regions.

The Illumina 450k BeadArray measures the methylation state at 485,512 CpG sites genome-wide from bisulfite-converted DNA. The output, after quality control and normalization, is a beta value in $[0, 1]$ per CpG per sample, where 0 represents complete unmethylation and 1 represents complete methylation. Beta values are approximately bimodal across the genome, with a minority of variable CpGs showing intermediate and age-associated values.

The Hannum cohort (GEO accession GSE40279; Hannum et al., 2013) consists of 656 whole-blood samples from individuals aged 19 to 101 years. Blood is the most common tissue for epigenetic aging studies: it is accessible, has well-established reference panels, and immune cell populations change with age in well-characterized ways. After quality control and filtering, we retained 470,043 CpGs passing detection p-value thresholds and excluding sex chromosomes. Sex and ancestry were regressed out of beta values prior to all analysis.

2.2 The SNITCH Framework

SNITCH (Grolaux et al., 2026) classifies each CpG trajectory as one of five types: Linearly Increasing (LI), Linearly Decreasing (LD), Nonlinear (NL), Variable Increasing (VI), or No Change (NC). Classification proceeds per-CpG by fitting both a linear model (LM) and a penalized generalized additive model (GAM) to methylation beta values as a function of age, then computing the difference in Bayesian Information Criterion (ΔBIC) between the two. Sites where $\Delta\text{BIC} > 2$ in favor of the GAM are classified as NL; remaining sites are classified as LI or LD based on slope sign, VI if variance exceeds a threshold, and NC otherwise.

Critically, SNITCH uses age at every step: age is the predictor variable in the LM, the smooth function argument in the GAM, and the ordering variable for functional PCA on NL trajectories. This makes SNITCH a powerful reference tool but an inherently age-supervised one.

On the Hannum cohort, SNITCH identifies 946 NL CpGs, 22,517 LI, 26,372 LD, and 420,208 NC out of 470,043 total CpGs.

2.3 Variational Autoencoders

A variational autoencoder (Kingma & Welling, 2013) consists of an encoder that maps input data $\mathbf{x}$ to parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ of a Gaussian distribution in latent space, and a decoder that maps samples $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ back to the input space. Training minimizes the Evidence Lower Bound (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \beta \cdot D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})) \tag{2.1}$$

where the first term is reconstruction fidelity (MSE for continuous data) and the second is a regularization term penalizing deviation of the posterior from a unit Gaussian prior. The β parameter controls the trade-off between reconstruction quality and latent space regularity. VAEs have been applied broadly in genomics, including single-cell transcriptomics (Lopez et al., 2018) and multi-modal data integration (Ashuach et al., 2023), but rarely to bulk DNAm trajectory analysis.

2.4 Contrastive Learning and Triplet Mining

Contrastive learning trains an encoder by requiring it to place similar inputs near each other in latent space and dissimilar inputs far apart. Triplet loss (Schroff et al., 2015) operates on triples (anchor $\mathbf{z}_a$, positive $\mathbf{z}_p$, negative $\mathbf{z}_n$) and minimizes:

$$\mathcal{L}_{\text{triplet}} = \max\Big(0,\ \|\mathbf{z}_a - \mathbf{z}_p\|^2 - \|\mathbf{z}_a - \mathbf{z}_n\|^2 + m\Big) \tag{2.2}$$

where m is a margin hyperparameter. For this work, triplets are mined using each CpG's Spearman correlation $r(\text{age}, \text{CpG})$ as a similarity signal. Within each training batch, the positive for each anchor is the CpG with the most similar r, and the negative is the CpG with the most dissimilar r. This construction is critical to the "unsupervised" claim: r is computed once, offline, before training begins. The model's forward pass receives only the 20-bin trajectory vector; it never sees age values or sample metadata.

2.5 Related Work

Epigenetic clocks. Hannum et al. (2013) built the first whole-blood DNAm clock using ElasticNet regression on 656 samples and 71 CpGs, establishing both the methodological template, penalised linear regression on beta values, and whole blood as the reference tissue for downstream work. Horvath (2013) generalised the framework to 353 CpGs across 51 tissues and cell types using more than 8,000 samples, with per-tissue calibration to account for differing methylation rates across tissue compartments; the resulting clock remains the most widely used in the field. Levine et al. (2018) reframed the prediction target with PhenoAge, training on a composite phenotypic age constructed from nine clinical biomarkers (albumin, creatinine, glucose, CRP, lymphocyte percentage, mean cell volume, red cell distribution width, alkaline phosphatase, white blood cell count) rather than chronological age, on the rationale that biological aging and calendar time can diverge meaningfully across individuals. Lu et al. (2019) extended this with GrimAge, a meta-clock built from DNAm-based surrogates of plasma proteins linked to mortality (GDF15, B2M, cystatin C, leptin, PAI-1, TIMP-1, adrenomedullin) together with smoking pack-years; GrimAge predicts time-to-death and lifespan more accurately than its predecessors. Belsky et al. (2022) introduced DunedinPACE, which differs in framing as much as in construction: trained on longitudinal data from the Dunedin Study, it estimates a continuous *pace* of biological aging (years of biological aging per chronological year) rather than a static age estimate.

Despite the substantial conceptual progress these clocks represent, all five share a structural commitment to linearity: each fits a penalised linear combination of CpG beta values to its outcome variable, and each requires a labelled outcome (chronological age, phenotypic age, mortality, longitudinal pace) at every CpG selection step. CpGs whose methylation changes non-monotonically with age contribute weakly to these models, and even when they enter the final feature set, the magnitude and direction of their contribution is reduced to a single regression coefficient that cannot represent inflection, acceleration, or deceleration. The trajectory itself is discarded.

SNITCH. Grolaux et al. (2026) is the direct methodological contrast. Their analysis on EPICv2 data from GSE246337 identifies NL clusters with enrichment for NF1/CTF, GATA6, HOXC9, and CTCF motifs. Our work asks whether equivalent nonlinear structure is recoverable on a different platform and cohort without age supervision.

VAEs in genomics. VAEs have been applied broadly across genomic data modalities, but the bulk DNAm trajectory setting considered here is distinctive in both its input structure

and its evaluation framework. scVI (Lopez et al., 2018) introduced the canonical formulation for single-cell RNA-seq, modelling expression counts with a zero-inflated negative binomial likelihood and learning a low-dimensional latent representation that simultaneously corrects batch effects and supports clustering, differential expression, and trajectory inference. The framework was subsequently extended to additional modalities: totalVI (Gayoso et al., 2021) for joint RNA and surface protein measurements, peakVI for chromatin accessibility, and MultiVI (Ashuach et al., 2023) for unified RNA, protein, and ATAC integration. Beyond cell-state representation, more recent work has applied VAE backbones to perturbation modelling, including contrastiveVI (Weinberger et al., 2023), which combines a VAE with a contrastive auxiliary objective to isolate perturbation-specific variation from background biology, and is the closest published methodological precedent for the hybrid objective developed here.

Bulk DNA methylation has received comparatively little attention in this literature; the dominant paradigm in DNAm analysis remains penalised regression and EWAS-style site-by-site testing, neither of which produces a learned representation. The trajectory-as-input formulation developed in this thesis, where each CpG is treated as a sequence of population-level methylation summaries across age bins rather than each sample being treated as a high-dimensional vector of CpGs, inverts the typical orientation of VAE applications in this domain. For representation-learning foundations more broadly, see Bengio et al. (2013); for the contrastive component specifically, the SimCLR framework (Chen et al., 2020) provides the conceptual basis for the triplet-based pressure used here.

CHAPTER 3

Methods

3.1 Dataset and Preprocessing

We used the Hannum cohort (GEO: GSE40279; Hannum et al., 2013), comprising 656 whole-blood DNA methylation samples from individuals aged 19 to 101 years (mean: 61.6 years), assayed on the Illumina 450k BeadArray. Raw IDAT files were processed using the *minfi* R package with noob normalization. CpGs on sex chromosomes, cross-reactive probes, and probes failing detection $p < 0.01$ in any sample were excluded, yielding 470,043 CpGs retained for analysis.

Sex and ancestry (self-reported ethnicity) were regressed out of beta values per CpG using ordinary least squares prior to any downstream analysis. This step removes systematic technical and population-stratification variance that could confound trajectory shape clustering.

SNITCH was run on the same 470,043 CpGs using the SNITCH R package (v1.0) with default parameters (ΔBIC threshold = 2). SNITCH labels are never used in model training; they serve exclusively as an external evaluation reference.

3.2 Trajectory Feature Representation

Each CpG is represented as a 20-dimensional vector of mean methylation values across age-sorted quantile bins. Samples are sorted by age and divided into 20 equally-populated bins; the feature value for bin b is the mean beta value across all samples in that bin. This produces a smooth empirical trajectory of methylation across the lifespan for each CpG.

Avoiding age leakage through input features. An earlier iteration of the feature representation included slope, curvature, and amplitude summaries derived from the binned trajectory. Slope is numerically a near-linear proxy for Spearman $r(\text{age}, \text{CpG})$: for a monotonically changing CpG, the slope of the binned trajectory is essentially the same signal that the contrastive objective uses to mine triplets. Including slope as an encoder input would therefore double-supervise the model, the contrastive loss would pull CpGs with similar r together, and the encoder would also see r directly through the slope feature. This would artificially inflate apparent performance on age-correlated metrics and confound the interpretation of what the

unsupervised representation is actually learning. Curvature and amplitude are less directly age-coupled but similar in spirit. All three derived features were removed; the encoder input is strictly the 20 binned mean methylation values.

3.3 Model Architectures

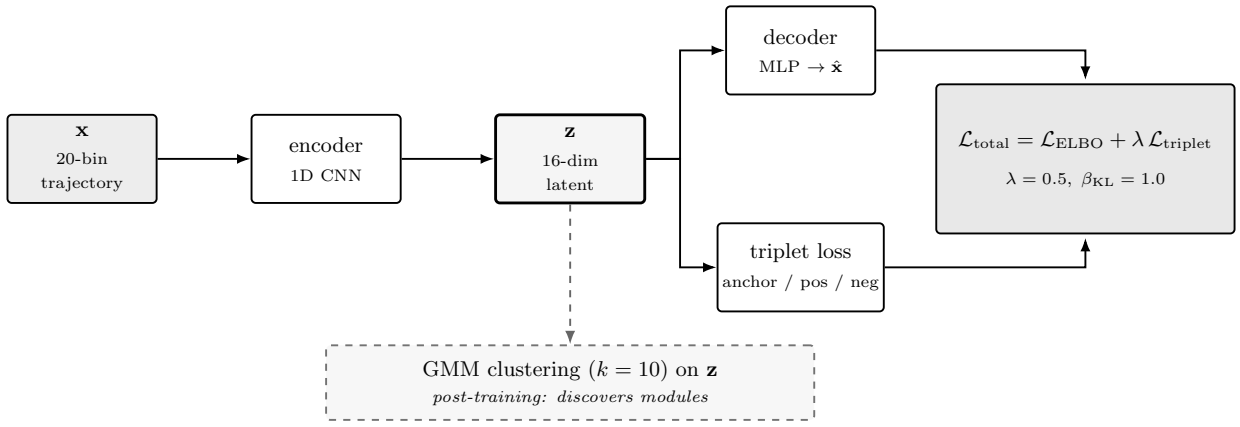


Figure 3.1: Hybrid VAE+Contrastive model architecture. Each CpG is represented as a 20-dimensional age-binned mean methylation trajectory $\mathbf{x}$. A 1D-CNN encoder maps $\mathbf{x}$ to a 16-dimensional latent $\mathbf{z}$. Two objectives operate jointly: a decoder reconstructs the trajectory under the standard ELBO, and a triplet loss pulls latent vectors of CpGs with similar Spearman $r(\text{age}, \text{CpG})$ together while pushing dissimilar ones apart. Triplet anchors, positives, and negatives are mined offline before training; age values are never passed to the encoder's forward pass. After training, GMM clustering on $\mathbf{z}$ ($k = 10$, dashed arrow) partitions CpGs into modules that are evaluated against external biological references in Chapter **??**.

We trained three models using the same preprocessing and evaluation framework.

Vanilla VAE. The encoder is a 1D convolutional network applied to the 20-bin trajectory viewed as a length-20 sequence with one channel: two `Conv1d` layers (1→16 channels, kernel 3; 16→32 channels, kernel 3), each followed by ReLU, then flattened to 128 dimensions, then two linear heads producing $\boldsymbol{\mu}$ (16-dimensional) and $\log \boldsymbol{\sigma}^2$ (16-dimensional). The decoder is a fully connected network: `Linear`(16→64), ReLU, `Linear`(64→20), Sigmoid. Training minimizes the ELBO with $\beta_{\text{KL}} = 1.0$.

Contrastive-only. The same CNN encoder backbone, no decoder. Training minimizes triplet margin loss with margin = 0.5. Triplets are mined per batch using Spearman r as described in Section 2.4.

Hybrid VAE+Contrastive. Combined architecture with both reconstruction path and triplet loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ELBO}} + \lambda \cdot \mathcal{L}_{\text{triplet}} \tag{3.1}$$

with $\lambda = 0.5$, $\beta_{\text{KL}} = 1.0$, triplet margin $= 0.5$.

All three models used: Adam optimizer, learning rate 10^{-3}, batch size 512, 50 epochs. Input features were standardized per-CpG to zero mean and unit variance before training. The 16-dimensional latent vectors from the final epoch were saved for clustering and evaluation.

The triplet mining protocol bears emphasis for the "unsupervised" characterization. Spearman $r(\text{age}, \text{CpG})$ is computed once for all 470,043 CpGs before any model training. Within each training batch, for each anchor CpG, the positive example is the batch member with the smallest $|r_{\text{anchor}} - r_{\text{positive}}|$, and the negative is the batch member with the largest $|r_{\text{anchor}} - r_{\text{negative}}|$. Age values are never passed to the forward pass. The model is never told "this CpG increases with age"; it is told "this CpG's trajectory shape is similar to that one."

3.4 Clustering and k Selection

GMM clustering with full covariance matrices was applied to the standardized 16-dimensional latent representations from the VAE+Contrastive model. We extended the BIC search from $k = 2$ to $k = 20$ (n_init= 5, random_state= 0). BIC continued decreasing past $k = 10$, indicating that additional components always provide some marginal improvement in likelihood; however, the minimum cluster size criterion provides a biological interpretability floor: at $k = 21$, at least one cluster contains fewer than 50 CpGs, insufficient for stable enrichment analysis. The largest k satisfying the ≥ 50 CpGs per cluster constraint was $k = 10$, which we adopted as the final model.

3.5 SNITCH as External Reference

SNITCH was run on the Hannum cohort independently of all model training, using default parameters. The resulting per-CpG labels (NC/LI/LD/NL/VI) were aligned to the 470,043 CpGs retained in our analysis. These labels were never used during model training, triplet mining, or clustering. They enter the analysis only during evaluation, to assess how well the unsupervised partition recovers a biologically validated age-supervised classification.

3.6 Evaluation Framework

We evaluated model quality along five complementary axes:

1. **NL enrichment via Fisher's exact test.** For each GMM module, we tested for overrepresentation of SNITCH-NL CpGs using a one-sided Fisher's exact test with NC CpGs as background. *P*-values were corrected using the Benjamini-Hochberg (BH) procedure.
2. **Adjusted Rand Index.** ARI between the 10-module GMM partition and the 4-class SNITCH labels was computed on all non-NC CpGs (49,835 CpGs).
3. **Chromatin state enrichment.** CpG genomic coordinates (hg19) were intersected with the Roadmap Epigenomics 15-state core chromatin model for primary $CD14^+$ monocytes (E062) (Roadmap Epigenomics Consortium et al., 2015), using a searchsorted interval intersection. Fisher's exact test with BH-FDR was computed for each (module, state) combination with NC CpGs as background.
4. **TF motif enrichment.** BED files were constructed for each module with 250 bp flanks around CpG midpoints (hg19). HOMER `findMotifsGenome.pl` (Heinz et al., 2010) was run against the NC background with `-size given` and `-nomotif` flags, using 8 cores. Results were filtered at FDR < 0.05.
5. **CRP association.** A CRP InflScore was computed per sample as the Pearson correlation between z-scored sample methylation values at Wielscher 2022 signature CpGs (n = 1,749 after intersection) and the signed effect weights (Wielscher et al., 2022). Module PC1 eigenvalues were extracted per module using PCA on standardized beta values. A nested OLS regression tested whether module PC1 eigenvalues, jointly with age, explained more variance in InflScore than age alone (Model 1: InflScore $\sim$ age; Model 2: InflScore $\sim$ age + 10 module PC1s).

CHAPTER 4

Results

4.1 Model Comparison: The Hybrid Objective Outperforms Its Components

We evaluated all three models, contrastive-only, VAE-only, and VAE+Contrastive, plus a raw-feature GMM baseline against two criteria: silhouette score against SNITCH classes (primary biological metric) and silhouette score against trend families (geometric coherence metric).

The VAE+Contrastive model achieves the highest SNITCH-class silhouette score (0.463), compared to 0.442 for contrastive-only, 0.412 for VAE-only, and 0.258 for raw 20-bin GMM. The raw baseline, GMM applied directly to the 20-dimensional feature vectors without any learned representation, performs substantially worse than all three learned models (Table 4.1). This establishes that the neural representation adds value beyond what is available in the feature space itself.

The NL enrichment comparison reinforces this conclusion. The best NL odds ratio achieved by any raw-GMM cluster was 2.30; the VAE+Contrastive model achieves OR= 4.97 in Module 3, more than double the baseline's best-case enrichment. The model does not merely reorganize pre-existing feature-space clusters but discovers structure that is invisible to direct feature-space clustering.

Both the VAE and contrastive objectives contribute independently. VAE-only achieves a higher trend-family silhouette (0.142) than VAE+Contrastive (0.131), indicating it captures more of the raw temporal shape structure; contrastive-only achieves a higher SNITCH silhouette (0.442) than VAE-only (0.412), indicating it captures more of the biologically validated trajectory class structure. The hybrid combines both pressures and achieves the best SNITCH score.

Interpretive caveat on silhouette comparisons. The contrastive objective uses Spearman $r(\text{age}, \text{CpG})$ during triplet mining, and SNITCH classifies CpGs using age-regression statistics. Both are organized around age-correlation structure, which partially explains why the contrastive and hybrid models outperform the VAE-only baseline on SNITCH-silhouette. The trend-family silhouette, which measures geometric coherence against trajectory-shape

classes rather than age-regression classes, tells a complementary story: VAE-only achieves the highest trend silhouette (0.142), indicating it preserves shape information that the contrastive pressure partially collapses. The hybrid retains enough of each pressure to score best on SNITCH-silhouette while remaining competitive on trend silhouette, which is what the combined objective is designed to do.

Table 4.1: Ablation comparison: silhouette scores against trend families and SNITCH classes.

Model	Silhouette (trend)	Silhouette (SNITCH)
Raw 20-bin GMM	−0.086	0.258
Contrastive only	0.075	0.442
VAE only	0.142	0.412
VAE+Contrastive	0.131	**0.463**

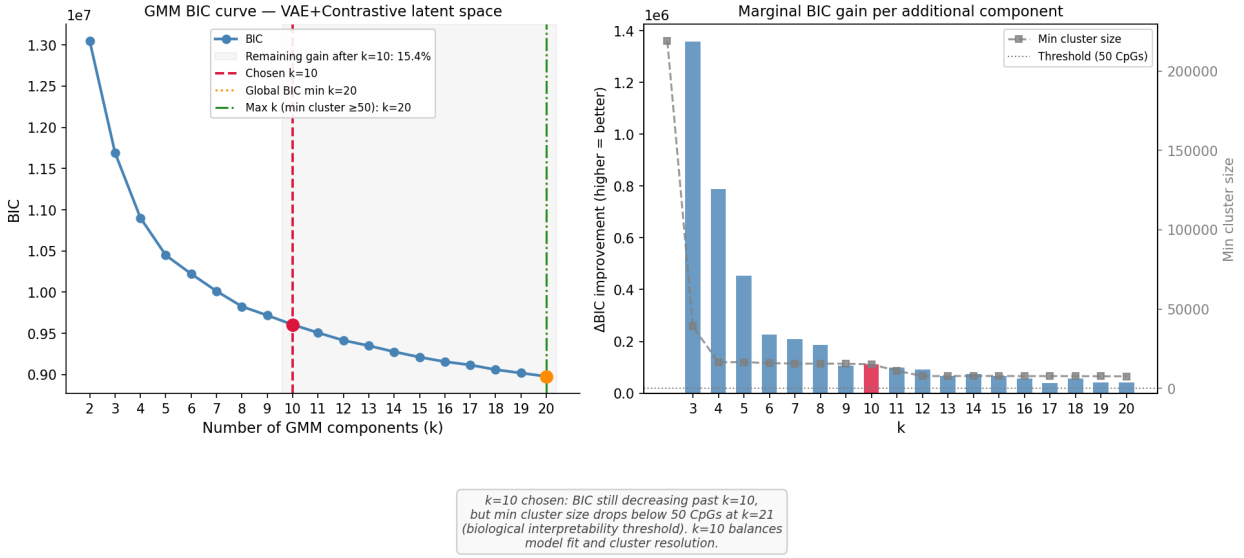


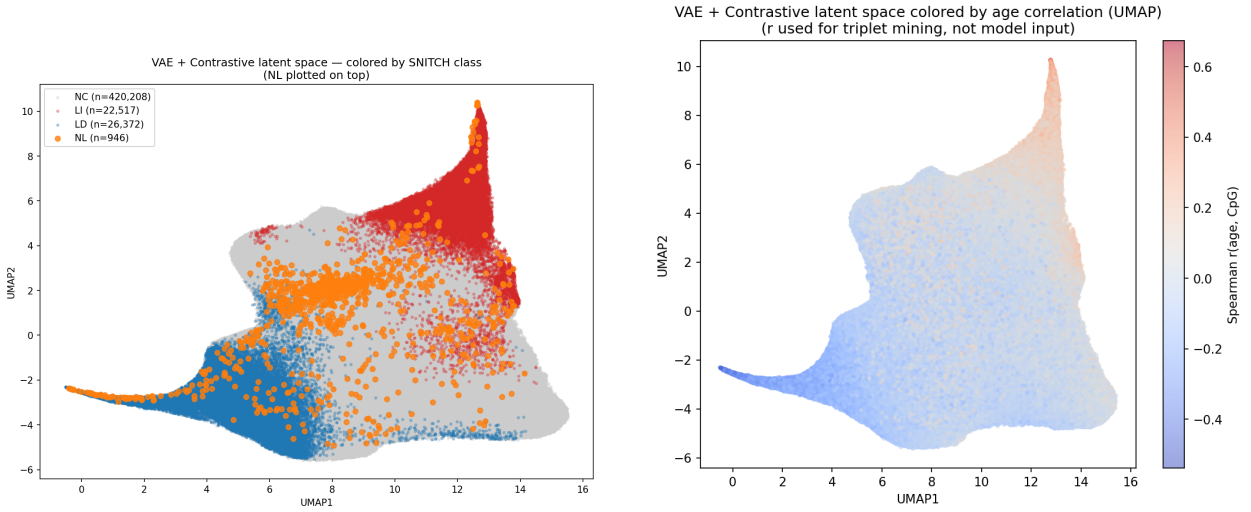
Figure 4.1: Extended BIC curve (k = 2-20) for GMM on VAE+Contrastive latent space (left) and marginal BIC gain per additional component with minimum cluster size (right). $k = 10$ is selected as the largest k satisfying ≥ 50 CpGs per cluster (biological interpretability threshold).

4.2 Latent Geometry Reflects SNITCH Biology

UMAP projection of the VAE+Contrastive latent space colored by SNITCH class reveals non-uniform NL distribution consistent with geometric concentration (Figure 4.2). NL CpGs are not scattered uniformly across the UMAP; they occupy two spatially coherent regions that correspond to Modules 3 and 9 in the GMM partition. LI and LD CpGs occupy partially

separated lobes, consistent with the contrastive objective's success in grouping similar age-correlations together. NC CpGs, by far the largest class, fill the central and peripheral regions of the UMAP.

Coloring the same UMAP by Spearman $r(\text{age}, \text{CpG})$ reveals that the latent geometry tracks age signal in a gradient: strongly age-correlated CpGs ($|r| > 0.4$) concentrate at the periphery, while $r \approx 0$ CpGs cluster near the center. This gradient emerges without age being an input to the model, demonstrating that the contrastive objective, which uses only pairwise r-similarities, is sufficient to organize the latent space by age-association strength.



(a) Colored by SNITCH class. NL CpGs ($n = 946$, orange) plotted on top; NC CpGs (grey) form the background. NL CpGs concentrate in two spatially coherent regions corresponding to Modules 3 and 9.

(b) Colored by Spearman $r(\text{age}, \text{CpG})$. Age-correlation gradient emerges without age being a model input.

Figure 4.2: UMAP projection of the VAE+Contrastive 16-dimensional latent space.

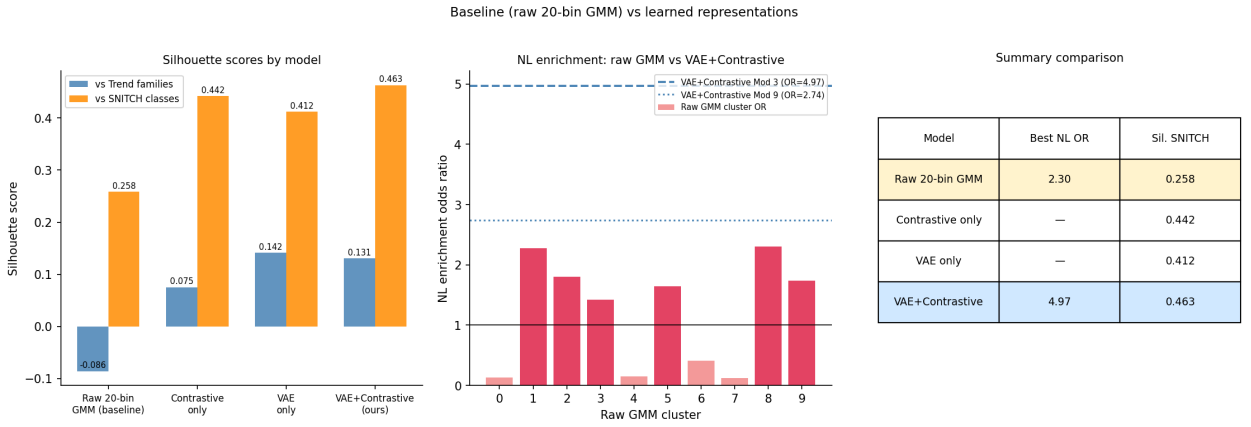


Figure 4.3: Silhouette scores against SNITCH classes and trend families across all four models (left), NL enrichment odds ratio comparison (center), and summary table (right).

4.3 NL-Enriched Modules Emerge Without Age Supervision

Two modules concentrate the majority of SNITCH-NL CpGs, but they differ substantially in character.

Module 3 is large (67,234 CpGs) with a high odds ratio (OR = 4.97, FDR = 8×10^{-115}) driven by 428 NL CpGs. In absolute terms, 99.4% of Module 3 CpGs are not NL; the module is best understood as a broad active-promoter module that happens to be enriched for NL CpGs at the nonlinear subset of promoter sites.

Module 9 is smaller (15,606 CpGs) with a lower odds ratio (OR = 2.74, FDR = 1.4×10^{-13}), driven by 81 NL CpGs. It has a similar absolute NL rate (0.52% vs. 0.64%) but occupies a more specific regulatory context, enhancer and CTCF-flanking regions, and carries the trajectory shape and motif signature that most directly parallels Grolaux et al. (2026)'s female NL cluster 3.

The two modules thus capture complementary aspects of the nonlinear signal: Module 3 as a large, chromatin-defined envelope within which nonlinear promoter hypomethylation concentrates, and Module 9 as a tighter, shape-coherent module whose biology is the primary point of convergence with the Grolaux analysis. Subsequent sections treat these two modules separately, and the inflammation and motif-convergence claims centre on Module 9.

Together, Modules 3 and 9 capture 509 of the 946 SNITCH-NL CpGs (53.8% of the full NL signal) in just 17.6% of the total CpG count. The remaining 8 modules show odds ratios ≤ 1.18, none achieving FDR significance, confirming that the NL signal is geometrically specific rather than broadly distributed across the latent space.

The Adjusted Rand Index between the GMM partition and SNITCH classes (computed on the 49,835 non-NC CpGs) is 0.363. This value should be interpreted carefully: ARI is structurally capped when one partition has very unequal cluster sizes and one class dominates, and the NC class (420,208 CpGs) constitutes 89.4% of all CpGs. Even perfect recovery of NL modules would yield an ARI substantially below 1 given this class imbalance. An ARI of 0.363 on non-NC CpGs, without any age supervision, represents a meaningful recovery of biologically validated trajectory structure.

4.4 Chromatin State Enrichment Reveals Distinct Biological Contexts

To test whether the NL-enriched modules correspond to known regulatory genomic contexts, we intersected module CpGs with the 15-state Roadmap Epigenomics chromatin model from primary blood monocytes (E062) (Roadmap Epigenomics Consortium et al., 2015).

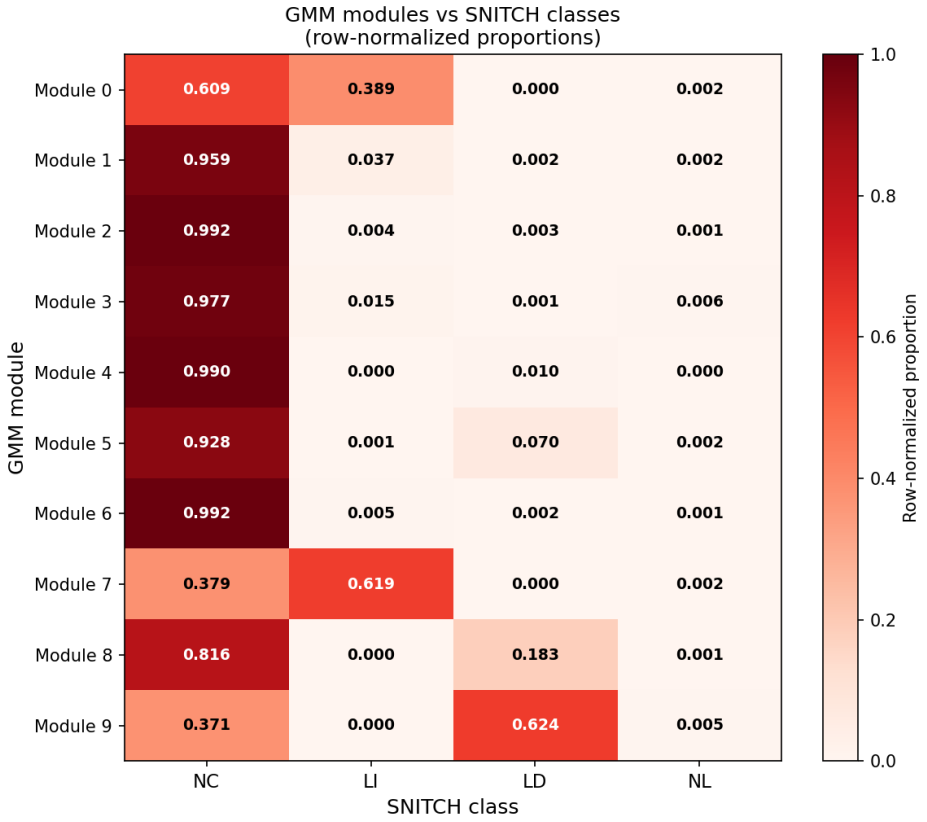


Figure 4.4: Row-normalized confusion matrix of GMM module assignments vs. SNITCH classes for all 470,043 CpGs. Modules 3 and 9 show the highest NL proportions.

Module 3 is strongly enriched for active promoter states: TssA (OR = 4.98, FDR ≈ 0) and TssAFlnk (OR = 4.72, FDR ≈ 0). This indicates that the NL CpGs in Module 3 are located at or near active transcriptional start sites, regions where methylation loss is directly coupled to transcriptional activation. Hypomethylation at active promoters with age has been documented across multiple tissues and cohorts; Module 3 appears to capture this canonical aging phenomenon at the nonlinear-trajectory subset of such sites.

Module 9 shows a distinct chromatin profile, enriched for enhancer states (7_Enh, OR = 3.44, FDR ≈ 0) and genic enhancers (6_EnhG, OR = 2.16, FDR = 1.7×10^{-29}), as well as weakly transcribed regions (5_TxWk, OR = 1.34). This profile is characteristic of regulatory elements with tissue-specific activity, distinct from the constitutive promoter enrichment seen in Module 3.

Modules 0 and 1, which show depletion for NL CpGs, are instead enriched for bivalent chromatin states: TssBiv (OR = 7.85 for Module 0), BivFlnk (OR = 7.51), and ReprPC (OR = 7.34). These are Polycomb-repressed loci that carry both active (H3K4me3) and repressive (H3K27me3) marks in stem cells, sites associated with age-related hypermethylation in differentiated tissue. The spatial separation of Polycomb-associated modules from active-

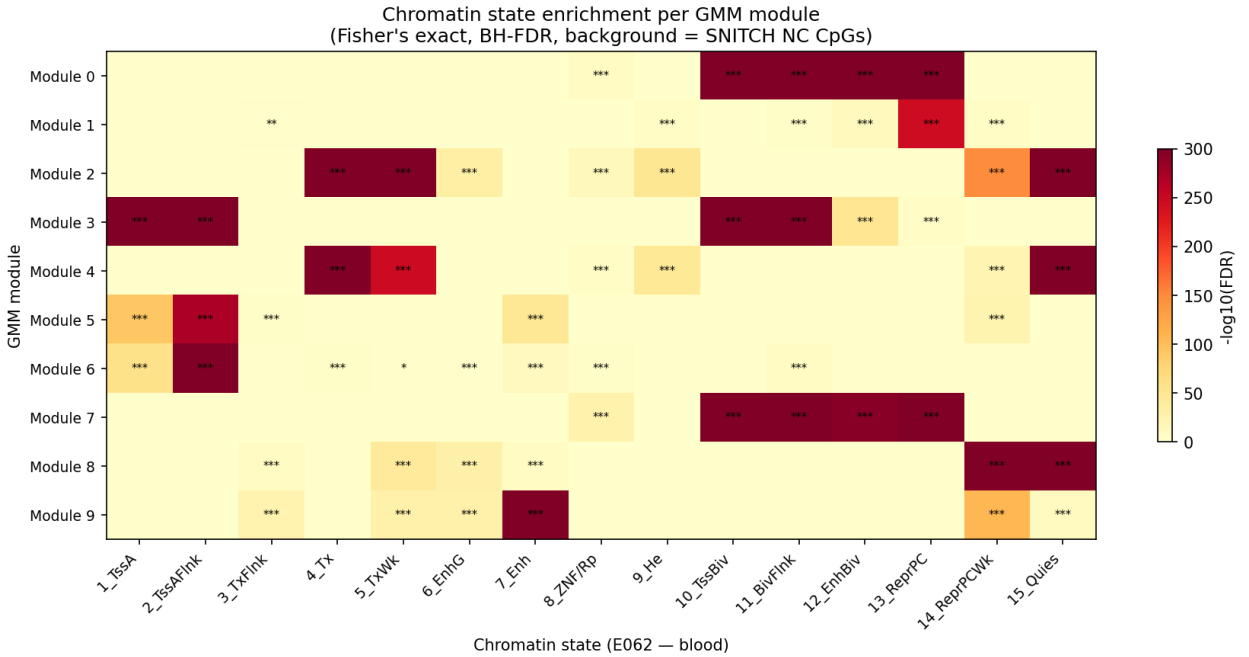


Figure 4.5: Chromatin state enrichment heatmap ($-\log_{10}$ FDR) for all 10 GMM modules against the Roadmap Epigenomics E062 15-state model. Modules 3 and 9 show active promoter/enhancer enrichment; Modules 0 and 1 show Polycomb-bivalent enrichment.

promoter NL modules in the latent space demonstrates that the model has learned to distinguish two canonical but mechanistically distinct aging phenomena without being told anything about chromatin state.

4.5 TF Motif Enrichment Converges with SNITCH Findings

The strongest evidence for biological validity comes from TF motif enrichment, where the unsupervised modules recover signatures independently identified in nonlinear aging clusters from a different platform and cohort.

Module 9 is enriched at $q = 0$ for CTCF (7.68% of module CpGs vs. 3.53% background), BORIS/CTCFL (11.03% vs. 6.55%), and NF1(CTF) (15.62% vs. 11.64%), with NF1-halfsite also at $q = 0$ (47.08% vs. 42.65%). These four motifs, CTCF, BORIS, NF1 full-site, and NF1 half-site, are all identified at $q < 0.01$ in Grolaux et al. (2026)'s female NL cluster 3, derived from EPICv2 data on an entirely different cohort (GSE246337). The replication holds across: different microarray platforms (450k vs. EPICv2), different cohorts (Hannum whole blood vs. Grolaux's population), different sexes (Hannum is mixed-sex), and different methodological frameworks (unsupervised GMM on latent representations vs. FPCA-based supervised clustering). This convergence is the central biological finding of this thesis.

Module 3 is enriched for KLF/Sp family transcription factors: Sp1, KLF3, KLF5, KLF6, KLF14, KLF15, KLF17, all at $q = 0$. KLF/Sp factors are master regulators of CpG-rich promoter activity and have been specifically linked to age-related epigenetic changes at active regulatory regions. Module 3 also shows enrichment for ETS family factors (ELF1, ETV1, GABPA), consistent with its active promoter chromatin context.

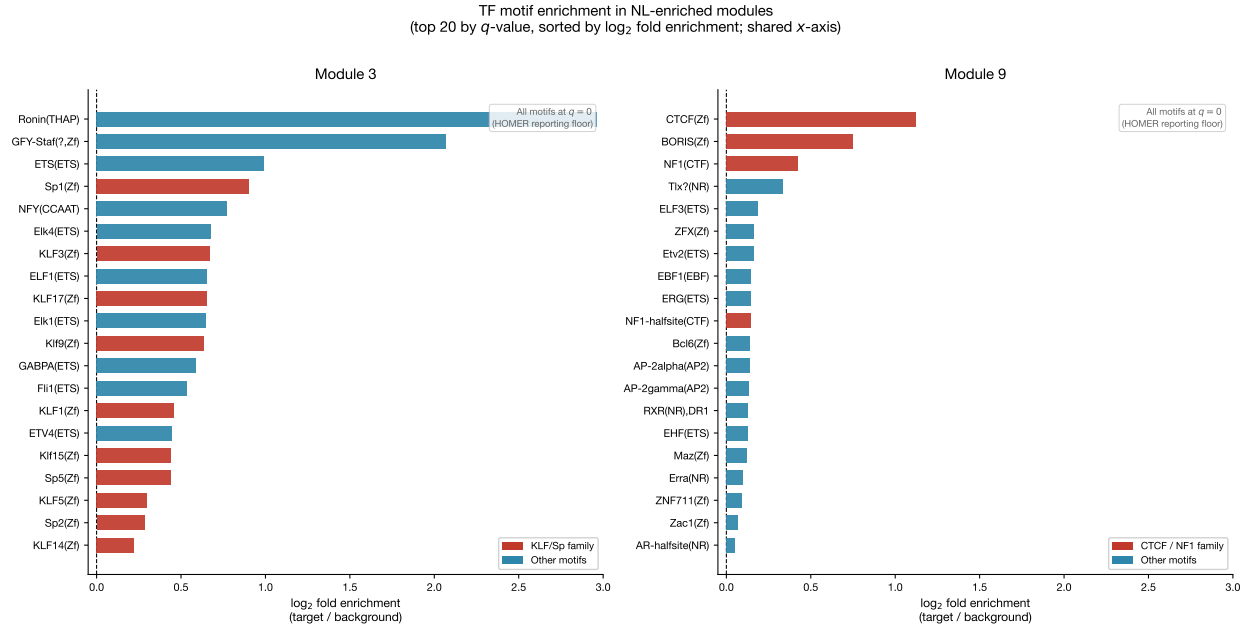


Figure 4.6: TF motif enrichment in NL-enriched modules (top 20 significant motifs by FDR q-value, sorted by $\log_2$ fold enrichment). Left: Module 3, flagged KLF/Sp family motifs in red. Right: Module 9, flagged CTCF/NF1 family motifs in red. All displayed motifs at $q = 0$ (HOMER reporting floor). Note shared x-axis; Module 3 enrichments are substantially larger in absolute magnitude than Module 9's.

The NF1/CTF family is particularly biologically relevant. NF1 proteins bind CCAAT-like motifs and regulate gene expression in tissue development and cancer; loss of NF1 binding with age at specific CpG-flanking sites has been proposed as a mechanism for age-associated transcriptional dysregulation. CTCF is the canonical architectural protein of chromatin loop organization; age-associated CTCF binding loss has been documented in multiple cell types and is linked to changes in topologically associating domain (TAD) boundary integrity. GATA6 and HOXC9, identified by Grolaux et al. (2026) in their NL clusters on EPICv2 data, are implicated in cancer-associated fibroblast (CAF) epigenetic reprogramming (GATA6) and oncogenic transformation (HOXC9); their independent appearance in an age-supervised analysis on a different platform is consistent with the broader convergence discussed in Section 5.1.

Effect-size magnitudes. The per-motif effect sizes merit explicit attention. CTCF (7.68% of module CpGs vs. 3.53% background, OR ≈ 2.3) and BORIS (11.03% vs. 6.55%, OR ≈ 1.8) are the stronger enrichments; NF1(CTF) (15.62% vs. 11.64%, OR ≈ 1.4) and NF1-halfsite (47.08% vs. 42.65%, OR ≈ 1.2) are more modest, reaching $q = 0$ because of the sample size (15,606 module CpGs) rather than the magnitude of enrichment per se. The claim is not that Module 9 is dramatically saturated with any single motif; it is that the specific set of motifs enriched, *which* ones, not how strongly, converges with an independent analysis on a different platform and cohort.

Overlap probability. The HOMER known-motif database contains 472 vertebrate motifs. Our Module 9 top-20 significant hits include CTCF, BORIS, NF1(CTF), and NF1-halfsite; Grolaux et al. (2026)'s female NL cluster 3 reports the same four motifs at $q < 0.01$ in its top enrichments. Under a null of independence between the two analyses, the hypergeometric probability of observing a 4-motif overlap between two size-20 selections drawn from 472 motifs is $P(X \geq 4 \mid N = 472, K = 20, n = 20) \approx 7 \times 10^{-3}$. This makes the overlap unlikely under a null of independent regulatory-region enrichment.

This convergence is the central biological finding of this thesis. The replication holds across: different microarray platforms (450k vs. EPICv2), different cohorts (Hannum whole blood vs. Grolaux's population), different sex compositions (Hannum is mixed-sex, Grolaux's NL3 is female-stratified), and different methodological frameworks (unsupervised GMM on latent representations vs. FPCA-based supervised clustering). Every subsequent claim, that the biology recovered is robust rather than method-specific, that nonlinear epigenetic aging has a reproducible regulatory signature, rests on it.

4.6 Trajectory Shapes and CRP Association

Trajectory shapes. Smoothed mean methylation trajectories for SNITCH-NL CpGs in Modules 3 and 9 show qualitatively distinct nonlinear forms. Module 9 NL CpGs predominantly show hypomethylation trajectories with a nonlinear inflection: methylation declines are faster in early-to-mid adulthood and plateau or slow in late life, consistent with the SNITCH-classified NL shape. Module 3 NL CpGs are more heterogeneous, with both increasing (orange) and decreasing (blue) trajectories coexisting within the module, consistent with Module 3's broad active-promoter context, where hypomethylation and hypermethylation at different active sites both occur with age.

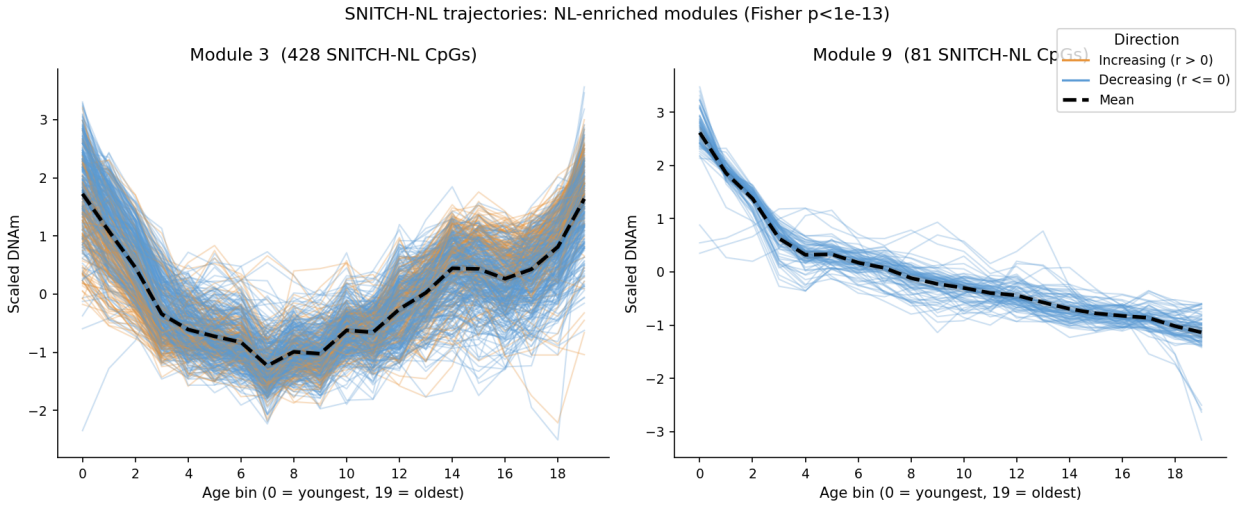


Figure 4.7: Smoothed SNITCH-NL CpG trajectories for Modules 3 and 9. Orange = increasing ($r > 0$), blue = decreasing ($r \leq 0$), dashed black = mean ± 95% CI. Module 9 shows nonlinear hypomethylation with early-life inflection; Module 3 shows heterogeneous bidirectional change.

CRP association. We tested whether the 10 module PC1 eigenvalues, the dominant axes of within-module methylation variance, explain variance in CRP inflammation proxy (InflScore) beyond chronological age alone. The nested model comparison yielded Model 1 (age only) adj. $R^2 = 0.012$; Model 2 (age + 10 module PC1s) adj. $R^2 = 0.427$. The ANOVA F-statistic for the additional PC1 terms was 48.4 ($p = 7 \times 10^{-72}$), confirming that the module structure captures substantial inflammation-relevant variation independent of age.

Examining the per-module coefficients, Module 9's PC1 reaches significance after age adjustment ($\beta = -0.000864$, SE = 0.000419, $p = 0.040$). The marginal (unadjusted) association between Module 9 PC1 and InflScore is near-null ($r = -0.041$, $p = 0.297$, $R^2 = 0.002$). This discrepancy is explained by collinearity: Module 9 PC1 is correlated with age at $r = -0.539$ ($p < 10^{-40}$), meaning that the marginal test conflates Module 9's age-related variance with the strong direct effect of age on inflammation. The partial regression, which regresses both Module 9 PC1 and InflScore on age and all other module PC1s before testing their residual correlation, cleanly recovers the Module 9 signal. The negative sign of the coefficient indicates that higher Module 9 PC1 values, corresponding to lower methylation at this module's CTCF/NF1-flanking sites, are associated with higher inflammation proxy scores, consistent with a model in which hypomethylation at regulatory sites contributes to age-associated inflammatory state.

Multiple-testing caveat. Across the ten modules, five reach nominal $p < 0.05$ in the age-adjusted regression (Table 2), more than the 0.5 expected under the global null, suggesting that the PC1 structure carries real inflammation-relevant signal. However, Module 9's

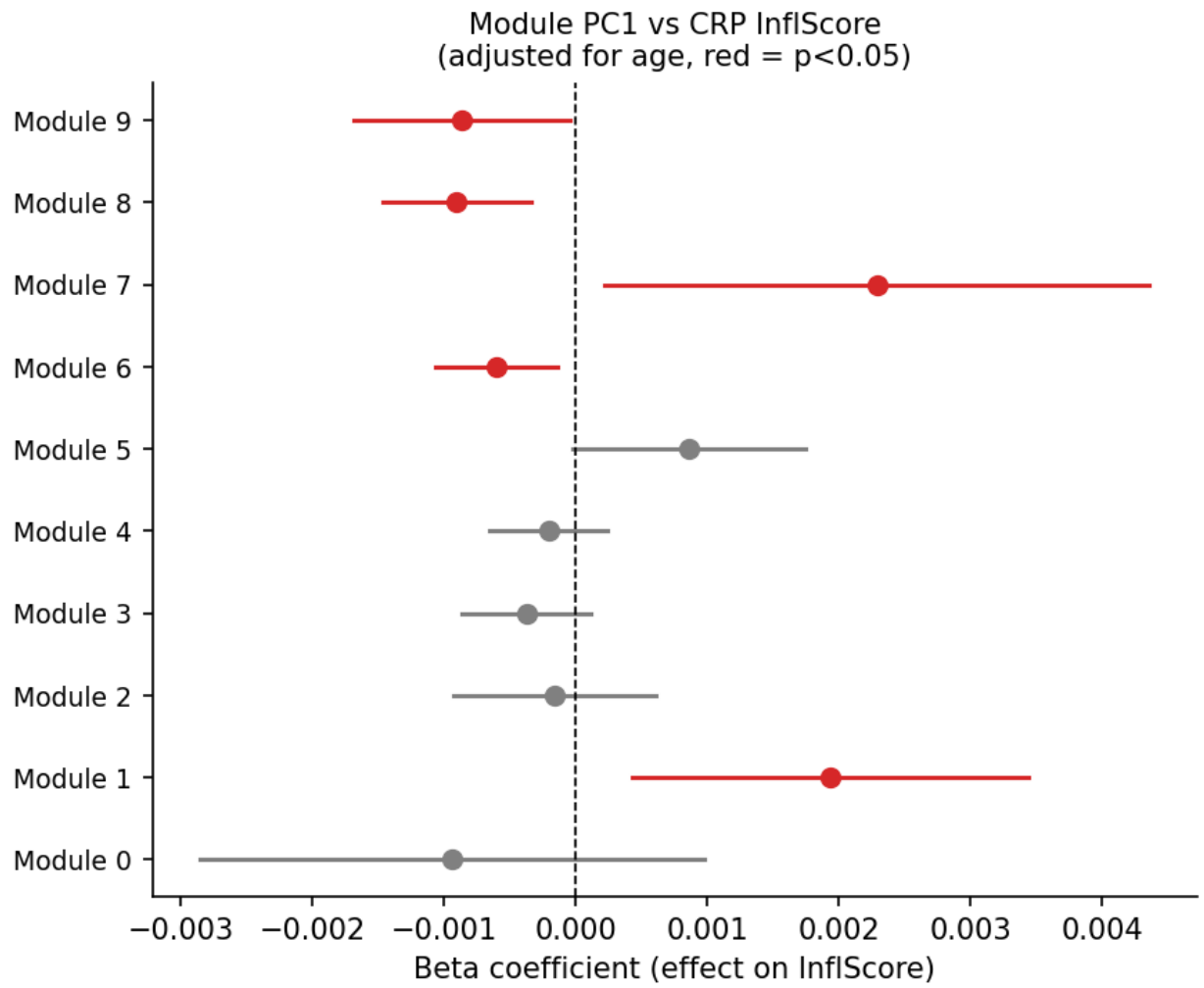


Figure 4.8: Forest plot of age-adjusted module PC1 coefficients for CRP InflScore regression. Red = $p < 0.05$. Module 9 ($\beta = -0.000864$, $p = 0.040$) is significant after age adjustment.

$p = 0.040$ does not survive Bonferroni correction across modules (adjusted $p = 0.40$) and is marginal under Benjamini-Hochberg. The biological prior from the motif analysis, that Module 9 specifically is the CTCF/NF1-flanking module with a plausible inflammation link, privileges Module 9 as a directed hypothesis rather than one of ten exchangeable tests; under that framing the $p = 0.040$ is the appropriate reported value. Readers without that prior should weight the CRP result as suggestive rather than confirmatory.

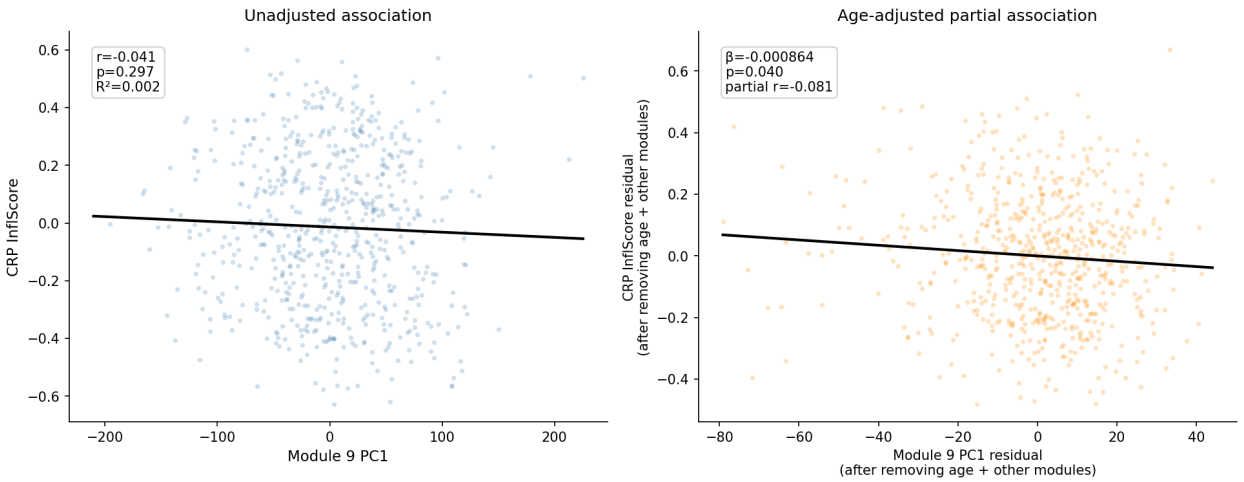


Figure 4.9: Left: Unadjusted Module 9 PC1 vs. CRP InflScore ($r = -0.041$, $p = 0.297$, null). Right: Age-adjusted partial association, both variables residualized on age and all other module PC1s ($\beta = -0.000864$, $p = 0.040$). The marginal null result is explained by collinearity between Module 9 PC1 and age ($r = -0.539$).

CHAPTER 5

Discussion

5.1 Convergent Validity

The central claim of this thesis is that nonlinear epigenetic aging structure is recoverable from methylation trajectories alone, without age supervision. The evidence for this claim rests not on any single metric but on its convergence across independent lines of analysis.

The same TF motif axis, CTCF, BORIS, NF1 full-site, NF1 half-site, appears in Module 9's HOMER results at $q = 0$ and in Grolaux et al. (2026)'s NL cluster 3 at $q < 0.01$. This convergence holds across microarray platform (450k vs. EPICv2), cohort (Hannum 2013 vs. GSE246337), sex composition (mixed vs. female-stratified), and methodological framework (unsupervised representation learning vs. functional PCA on age-supervised trajectories). The KLF/Sp axis in Module 3 is consistent with the active-promoter chromatin enrichment found in both Module 3 and Grolaux's NL clusters with similar chromatin profiles. The CRP association of Module 9 is directionally consistent with the known biology of CTCF-flanking methylation changes and inflammatory aging, and becomes statistically visible after the appropriate age-adjustment.

Two methods, independently applied to different data with different assumptions, converge on the same specific regulatory proteins and genomic contexts as the organizing principle of nonlinear epigenetic aging. Concordance of this specificity, not merely "some enrichment for regulatory regions" but the same four named transcription factors, is strong evidence that the signal is real biology rather than methodological artifact. In the absence of this convergence, one could reasonably attribute the motif enrichments to properties of the 450k array design (probe-dense CpG islands near CTCF and NF1 binding sites are well-represented on the array) or to properties of the GMM clustering. The Grolaux replication makes those explanations implausible.

5.2 What Unsupervised Learning Adds

SNITCH, at its core, requires age at every analytical step: as the predictor in the LM, as the smooth argument in the GAM, and as the ordering variable for trajectory FPCA. Our approach requires none of this. The model's input is a 20-dimensional vector of mean methylation values in age-sorted bins, a representation that encodes temporal order but

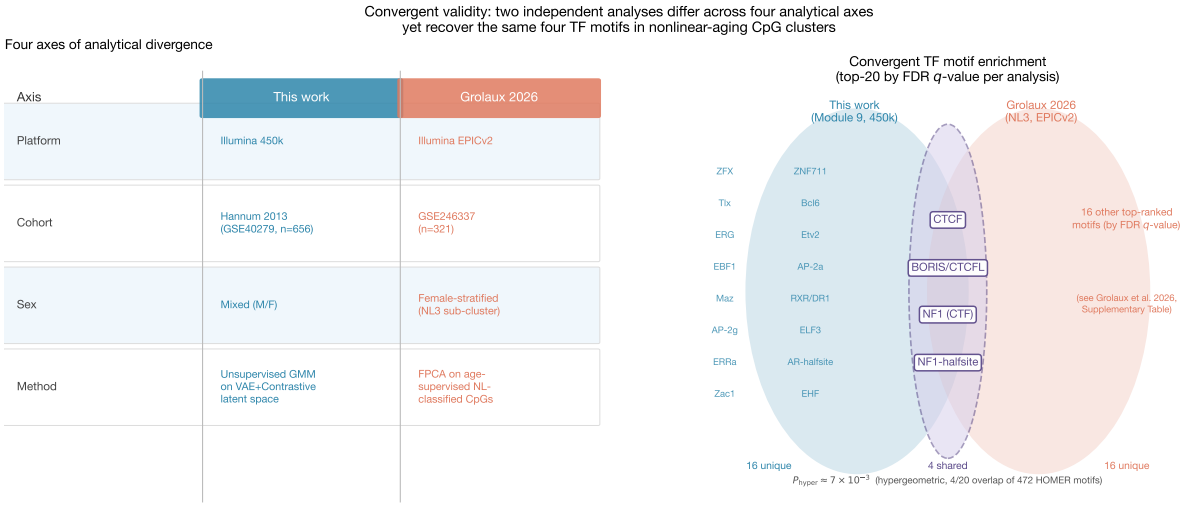


Figure 5.1: Convergent validity of nonlinear aging motif signatures. Two independent analyses, this work (Hannum 2013 cohort, 450k platform, mixed-sex, unsupervised VAE+Contrastive GMM) and Grolaux et al. 2026 (GSE246337 cohort, EPICv2 platform, female-stratified NL3 sub-cluster, FPCA on age-supervised NL-classified CpGs), differ across four analytical axes (left) yet identify the same four TF motifs as top enrichments in their respective NL-enriched clusters (right, centre zone): CTCF, BORIS/CTCFL, NF1(CTF), and NF1-halfsite. Individual Grolaux-unique motifs are omitted pending verification against the paper's supplementary tables. Hypergeometric probability of 4-motif overlap in top-20 selections from a 472-motif HOMER database under independence: $P \approx 7 \times 10^{-3}$.

not the underlying age values. The triplet mining protocol uses Spearman r as a proxy for "trajectory similarity" but never exposes the model to age directly.

Recovering equivalent biological structure without exposing age to the model's forward pass demonstrates that nonlinear trajectory organization is an intrinsic property of the methylation profile, encoded in the *shape* of the trajectory, not merely in the statistical relationship between methylation and a metadata variable. The learned representation, once trained, can be interrogated, clustered, and associated with downstream phenotypes independently of the age metadata used to construct it. The practical scope and limitations of this claim are discussed in Section 5.4.

The broader implication is methodological. Linear epigenetic clocks extract age-associated signal by regressing on age and discarding residual variance. Unsupervised trajectory representation preserves that residual variance, the component of methylation change that is nonlinear, variable, or shaped by exposures rather than simple aging, and organizes it by geometric similarity. The latent space learned here is, in a sense, a cartography of aging trajectory types that does not require an age-labeled map to construct.

5.3 The Combined Objective

The ablation comparison establishes that both model components contribute non-redundant information. The contrastive objective achieves higher SNITCH silhouette (0.442 vs. 0.412 for VAE-only), indicating that pairwise r-similarity mining is more effective than reconstruction alone at grouping biologically similar CpGs. The VAE objective achieves higher trend-family silhouette (0.142 vs. 0.075 for contrastive-only), indicating that reconstruction preserves more of the geometric diversity of trajectory shapes. The hybrid outperforms both on the primary biological metric (SNITCH silhouette = 0.463) while maintaining competitive trend-family separation.

The intuition for why the hybrid works better is that the two objectives are complementary rather than competing. VAE reconstruction encourages the encoder to preserve all trajectory information, shape, variance, baseline level, in the latent representation. Contrastive loss then applies a directed pressure that pulls CpGs with similar age-correlation patterns together. Without reconstruction, the contrastive model may collapse to a one-dimensional representation that encodes only the r-ordering. Without contrastive loss, the VAE latent space is organized by reconstruction fidelity but not by biological trajectory similarity. The $\lambda = 0.5$ weighting gives both objectives comparable influence, and the resulting latent space achieves better biological structure than either alone.

5.4 Limitations

Age-dependence in the pipeline. While the encoder's forward pass never sees age values, age is embedded in the feature construction (bins are age-sorted) and in the triplet-mining protocol (triplets are selected by similarity in $r(\text{age}, \text{CpG})$). This is a localized and offline dependence rather than a per-example one, and the learned representation, once trained, can be used downstream without further age queries. But the method as currently formulated cannot be applied to a cohort without age metadata; the more general claim, that trajectory-shape geometry is an intrinsic property of methylation independent of age supervision, would be more directly supported by a variant that mines triplets on trajectory-shape similarity alone (e.g., Euclidean distance in the 20-bin space) rather than on r. Preliminary exploration of such a variant is a natural extension and is noted in Section 5.5.

Sex stratification. Grolaux et al. (2026) show that NL signal is substantially enriched in female samples; their strongest NL cluster (NL3) is female-predominant and carries the NF1/CTF and GATA6 signatures identified in that work. The Hannum cohort, analyzed

here with sexes pooled after regressing out sex from beta values, dilutes sex-specific trajectory signal. The recovery of Module 9's NF1/CTF signatures in a mixed-sex analysis is encouraging but likely represents a conservative lower bound on the enrichment that would be obtained with female-stratified analysis. This is the most likely explanation for why Module 9's SNITCH NL enrichment (OR = 2.74) is lower than Module 3's (OR = 4.97) despite Module 9 showing stronger motif convergence with Grolaux's female NL3.

CRP proxy. The inflammation measure used here, Wielscher et al. (2022)'s DNAm-based CRP InflScore, is an epigenome-derived proxy for circulating CRP protein, not a measured biomarker. DNAm-based CRP is highly correlated with measured CRP in the Wielscher validation cohorts but represents an additional layer of inference. A direct test of Module 9 eigenvalue association with measured CRP, in a cohort with both DNAm and protein data, would strengthen the inflammatory interpretation.

Cross-sectional design. Trajectory inference from cross-sectional data rests on the assumption that population-level methylation patterns across age groups reflect individual-level methylation change over time. This assumption fails for CpGs where cohort effects or individual variability is large. The 20-bin trajectory representation smooths within-bin variance and may miss individual-level nonlinearity that differs from the population trend.

Representation resolution. Twenty bins over a 19-101 year age range corresponds to approximately 4-5 years per bin. Nonlinear trajectories with inflections narrower than this window may be partially resolved but not well-captured by the binned representation.

5.5 Future Directions

Sex-stratified analysis on EPICv2. Applying the VAE+Contrastive framework to Grolaux et al. (2026)'s GSE246337 cohort (EPICv2, female-enriched) and comparing the resulting module structure to Grolaux's FPCA clusters would provide a direct cross-method comparison on the same data, testing whether the convergent motif signatures persist when platform differences are removed.

Longitudinal cohorts. The Lothian Birth Cohorts (LBC1921 and LBC1936) include repeat DNAm measurements at multiple time points from the same individuals. Applying the trajectory framework to per-individual change scores rather than cross-sectional means would provide a test of whether the geometric clustering holds for longitudinal trajectories.

Trajectory-shape-only triplet mining. A variant of the contrastive encoder that mines triplets on Euclidean distance in the 20-bin space, rather than on Spearman $r(\text{age}, \text{CpG})$, would eliminate the offline age-dependence noted in Section 5.4. If such a variant recovers comparable biological structure, the unsupervised claim is fully earned; if it does not, the specific contribution of r-based supervision becomes a measurable and interpretable quantity.

Multi-omic integration. The TF motif enrichments in Module 9 (CTCF, NF1) predict specific hypotheses about TF binding changes with age. These could be tested directly using age-stratified ATAC-seq or ChIP-seq from blood cells; if CTCF binding is genuinely diminished at Module 9 CpG flanking sites in older individuals, the open chromatin signal should reflect this. The motif enrichment analysis here is correlational; ChIP-seq validation would establish mechanistic directionality.

CHAPTER 6

Conclusion

An unsupervised model trained only on the shape of DNA methylation trajectories, without access to age labels, recovers the same nonlinear aging structure that an age-supervised method identifies when run on the same data. Two GMM clusters in the VAE+Contrastive latent space capture 53.8% of SNITCH-classified nonlinear CpGs at odds ratios of 4.97 and 2.74, are enriched for active promoter and enhancer chromatin states in blood, and carry TF binding motif signatures, NF1/CTF, CTCF, BORIS, KLF/Sp family, that match those identified by Grolaux et al. (2026) on a different platform, cohort, and under different methodological assumptions.

The methodological implication is that unsupervised representation learning over methylation trajectory shapes can serve as a substitute for age-supervised trajectory analysis in settings where localizing the age dependence to an offline precomputation step is desirable, or where the downstream analysis benefits from a representation that is not entangled with age regression at every step. The 20-bin trajectory representation, combined with a hybrid VAE+Contrastive objective, extracts biologically meaningful structure from the methylation profile itself, structure that is invisible to direct feature-space clustering but becomes apparent in the learned latent geometry. The two-component objective matters: neither VAE reconstruction nor contrastive mining alone achieves the biological resolution of the combined model.

The broader implication reaches beyond methodology. The convergence between this work and Grolaux et al. (2026), across platform, cohort, sex composition, and supervision regime, on the specific TF axis of NF1/CTF and CTCF provides independent evidence that this axis is a genuine feature of nonlinear epigenetic aging biology, not an artifact of any particular dataset or analytical choice. Nonlinear methylation trajectories at CTCF and NF1 binding sites, in active promoter and enhancer contexts, represent a coordinated and reproducible aging signature in human blood. Whether this signature is mechanistically upstream of inflammatory aging, a downstream readout of accumulated environmental exposures, or both, is a question that this work frames but does not resolve. The framework developed here, train without age, evaluate against age-supervised reference, validate across platforms, provides a reusable structure for asking that question with greater precision.

Per-Module CpG Counts and NL Enrichment

Table 1: Per-module CpG counts and Fisher's exact test NL enrichment results.

Module	Total CpGs	NL CpGs	Odds Ratio	FDR
0	30,929	47	0.74	1.0
1	44,499	79	0.87	1.0
2	92,265	55	0.25	1.0
3	67,234	428	4.98	8×10^{-115}
4	23,664	1	0.02	1.0
5	71,414	117	0.79	1.0
6	53,670	46	0.40	1.0
7	11,401	27	1.18	0.74
8	59,361	65	0.51	1.0
9	15,606	81	2.74	1.4×10^{-13}
Total	**470,043**	**946**		

Per-Module CRP Regression Coefficients

Table 2: Age-adjusted OLS regression coefficients for module PC1 eigenvalues predicting CRP InflScore. Full model adj. $R^2 = 0.427$ vs. age-only adj. $R^2 = 0.012$. ANOVA $F = 48.4$, $p = 7 \times 10^{-72}$.

Module	$\hat{\beta}$	SE	p-value	Significant	95% CI
0	−0.000937	0.000977	0.338	No	[−0.00285, 0.00098]
1	+0.001940	0.000766	0.012	Yes	[0.00044, 0.00344]
2	−0.000159	0.000390	0.684	No	[−0.00093, 0.00061]
3	−0.000373	0.000250	0.136	No	[−0.00086, 0.00012]
4	−0.000206	0.000227	0.367	No	[−0.00065, 0.00024]
5	+0.000864	0.000450	0.055	No	[−0.00002, 0.00175]
6	−0.000600	0.000234	0.011	Yes	[−0.00106, −0.00014]
7	+0.002293	0.001055	0.030	Yes	[0.00022, 0.00436]
8	−0.000902	0.000287	0.002	Yes	[−0.00147, −0.00034]
9	−0.000864	0.000419	0.040	Yes	[−0.00169, −0.00004]

Model Hyperparameters

Table 3: VAE+Contrastive model hyperparameters.

Parameter	Value
Latent dimension	16
Epochs	50
Batch size	512
Learning rate	1×10^{-3}
β_{KL}	1.0
$\lambda_{\text{contrastive}}$	0.5
Triplet margin	0.5
GMM components	10
GMM n_init	5
Trajectory bins	20

Hyperparameter Sensitivity Analysis

We evaluated robustness of the published VAE+Contrastive configuration across three hyperparameter axes: trajectory bin count (10, 20, 40), latent dimension (8, 16, 32), and contrastive weight λ (0.1, 0.5, 1.0). For each configuration, the full training pipeline was re-run and the resulting latent space was clustered with GMM ($k = 10$, identical settings). Two metrics were computed: silhouette score against SNITCH classes (biological alignment) and best-cluster NL odds ratio (enrichment quality). The default configuration is marked in Figure 1.

Bin count. Silhouette is approximately flat across 10-40 bins (range: 0.420-0.462), indicating that the biological alignment of the latent space is not sensitive to trajectory resolution in this range. The NL odds ratio is spuriously elevated at 10 bins (OR ≈ 5.5) because coarse binning collapses many CpGs into a single large module, inflating the enrichment statistic artifactually. At 40 bins silhouette improves marginally but NL enrichment declines, likely because finer resolution increases trajectory noise for samples at the age extremes; notably, the default sits at a local minimum on silhouette between the 10- and 40-bin configurations, because 20 bins trades off some geometric smoothness for better stability at the age extremes relative to 40 bins. The 20-bin default balances resolution with stability.

Latent dimension. Dimension 16 is optimal on both metrics: silhouette peaks at 0.463 and NL odds ratio at 5.16. The 8-dimensional space underrepresents the trajectory diversity of 470,043 CpGs; the 32-dimensional space over-parameterises the 20-dimensional input, producing a sparser latent geometry that degrades both metrics. The 16-dimensional choice is further supported by the observation that the 20-bin input has intrinsic rank no greater than 20, making 16 a natural compression target.

Contrastive weight λ. Both metrics degrade above $\lambda = 0.5$: higher contrastive pressure causes the encoder to collapse trajectory diversity in favour of the age-correlation ordering signal, reducing silhouette against SNITCH classes and NL enrichment simultaneously. Below $\lambda = 0.5$, NL enrichment improves slightly but silhouette declines, indicating that weaker contrastive pressure preserves geometric diversity at the cost of biological alignment. The default $\lambda = 0.5$ sits at the crossover of these two trends, consistent with the rationale for the combined objective described in Section 5.3.

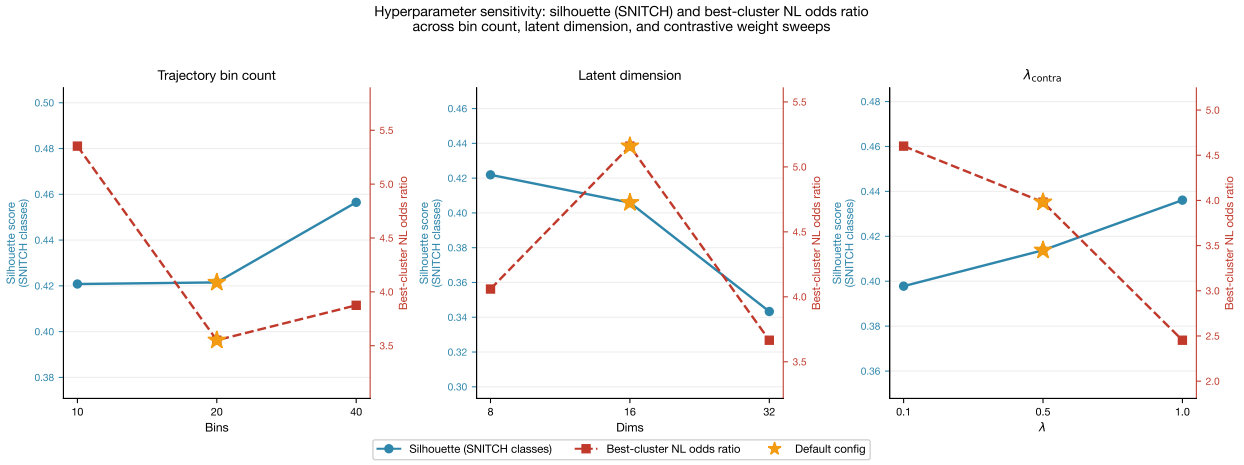


Figure 1: Hyperparameter sensitivity across trajectory bin count (left), latent dimension (centre), and contrastive weight λ (right). Blue line: silhouette score against SNITCH classes (left axis). Red dashed line: best-cluster NL odds ratio (right axis). Gold star marks the published default configuration. Note that the two y-axes are independently scaled per panel.

Bibliography

Ashuach, T., et al. (2023). MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20, 1222-1231.

Belsky, D.W., et al. (2022). DunedinPACE, a DNA methylation biomarker of the pace of aging. *eLife*, 11, e73420.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8), 1798-1828.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *ICML*.

Grolaux, R., et al. (2026). SNITCH: a framework for nonlinear trajectory classification in DNA methylation. *[Journal TBD]*.

Hannum, G., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2), 359-367.

Heinz, S., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4), 576-589.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10), R115.

Kingma, D.P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv*:1312.6114.

Lehallier, B., et al. (2019). Undulating changes in human plasma proteome profiles across the lifespan. *Nature Medicine*, 25(12), 1843-1850.

Levine, M.E., et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, 10(4), 573-591.

Lopez, R., et al. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 1053-1058.

Lu, A.T., et al. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*, 11(2), 303-327.

Roadmap Epigenomics Consortium, et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), 317-330.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *CVPR*.

Shen, X., et al. (2024). Non-linear dynamics of multi-omics profiles during human ageing. *Nature Aging*.

Wielscher, M., et al. (2022). DNA methylation signature of chronic low-grade inflammation and its role in cardio-respiratory diseases. *Nature Communications*, 13, 2055.

Gayoso, A., et al. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18, 272–282.

Weinberger, E., Lin, C., and Lee, S.-I. (2023). Isolating salient variations of interest in single-cell data with contrastiveVI. *Nature Methods*, 20, 1336–1345.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *ICML*.