EPIGENETIC AGE PREDICTION WITH SEQUENCE-CONDITIONED CO-METHYLATION GRAPH NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Epigenetic clocks estimate biological age from DNA methylation profiles and are widely used in ageing research. Graph-based clocks model relationships among CpG sites but typically omit the local DNA sequence surrounding each site, despite its potential influence on methylation. Using sequence context requires careful interpretation because the sequence of each CpG site is fixed across individuals: improved prediction may reflect informative local context or merely provide a way to distinguish CpG sites. We introduce a sequence-conditioned graph clock that jointly models sample-specific methylation, local sequence context, and relationships among CpG sites, with sequence conditioning accounting for only 0.05% of the model's parameters. To determine what drives any improvement, we use matched comparisons that distinguish the contributions of sequence context and CpG relationships while testing whether gains depend on the correct sequence–site correspondence. Across 37 blood methylation studies, sequence conditioning consistently improves biological-age prediction, reducing mean test MAE from 3.550 to 3.405 years. This improvement largely disappears when sequences are assigned to the wrong CpG sites, showing that correct local context, rather than CpG-site identity alone, accounts for most of the gain. Prediction averaging further reduces MAE to 3.128 years, and the model performs favourably in an exploratory application to ovarian cancer samples. These findings show that local DNA sequence provides useful information for co-methylation graph clocks and highlight the importance of distinguishing sequence information from CpG-site identity.

1 INTRODUCTION

Aging alters DNA methylation in ways that recur across individuals and cohorts. DNA methylation, the addition of a methyl group at cytosine–guanine (CpG) dinucleotides, regulates transcription and supports development, differentiation, and homeostasis (Jaenisch & Bird, 2003). Because methylation changes with age alongside broader declines in cellular regulation (López-Otín et al., 2023), epigenetic clocks built on genome-wide methylation profiles (Horvath, 2013; Hannum et al., 2013) have become a widely used molecular readout of age in aging research, age-related disease studies, and the evaluation of longevity interventions (Horvath & Raj, 2018; Moqri et al., 2023).

Most clocks treat CpG sites as independent entries of a methylation vector, whether they use penalized linear models or deep neural networks (Horvath, 2013; Hannum et al., 2013; Galkin et al., 2021; de Lima Camillo et al., 2022; Ying et al., 2024a; Zhang et al., 2026). Graph-based clocks instead encode relations among sites, such as co-methylation, chromosomal proximity, and shared gene membership, and combine information across related CpGs through message passing (Ahmed et al., 2025), following a broader use of graph learning on biological data (Li et al., 2022). Their node representations, however, omit the sequence surrounding each CpG site, although GC content, CpG density, and local nucleotide composition are associated with methylation propensity (Schübeler, 2015; Santoni, 2021; Affinito et al., 2020), and deep sequence models predict gene regulatory activity directly from DNA (Avsec et al., 2021). Methylation foundation models do use sequence context, but, like foundation models in other omics (Cui et al., 2024), they rely on large-scale pretraining (de Lima Camillo et al., 2024; Ying et al., 2024b). It remains unclear whether sequence and graph message passing are complementary in a clock trained directly on methylation data.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Answering this question requires attributing any gain to the right cause. In a graph clock, the learned representation of each CpG site is shaped jointly by the transformation of its own features and by messages received from neighbouring sites. Because GNN performance depends on how these components and other architectural choices interact (You et al., 2020), an observed performance difference cannot by itself be attributed to graph propagation. The sequence surrounding a CpG site raises a further difficulty that is not specific to methylation. Because this sequence is identical in every sample, a sequence-derived input can also act as a site identifier, and node identity is known to increase the expressive power of GNNs (You et al., 2021). An accuracy gain therefore does not show that the model uses sequence content, because additional site-specific inputs can improve GNNs even when they carry no biological information (Sato et al., 2021; Abboud et al., 2021). Comparisons are further complicated by run-to-run variability, which can change the conclusions of benchmark comparisons when few training runs are reported (Bouthillier et al., 2021; Errica et al., 2020).

We develop a sequence-conditioned co-methylation graph clock that jointly models sample-specific methylation, local DNA sequence, and relationships among CpG sites. Each sample is represented on a graph whose nodes are CpG sites and whose edges encode co-methylation and genomic relationships. A shared convolutional encoder summarises the 122-bp context surrounding each site, using it both to condition the corresponding methylation measurement and to provide compact sequence features. Principal Neighbourhood Aggregation (PNA) then integrates information across related CpGs, and a prediction head maps the resulting site representations to biological age (Figure 1). Sequence conditioning accounts for only 0.05% of the complete model's parameters. To identify what drives predictive performance, we compare matched configurations with and without sequence conditioning and graph propagation, and reassign sequences among CpG sites to test whether gains depend on the correct sequence–site correspondence rather than CpG-site identity. Separately, we average independently trained predictions to assess how run-to-run variation affects performance.

Across 37 blood methylation studies, our results establish local sequence as a useful source of information for biological-age prediction: sequence conditioning improves both graph and no-graph models, reducing the graph model's mean test MAE from 3.550 to 3.405 years. This benefit largely disappears when sequences are reassigned among CpG sites, indicating that most of the gain comes from the correct local context. At the same time, variation across independently trained graph models exceeds the individual component effects, showing that reliable architectural attribution requires repeated, matched evaluation. After reducing this variability through prediction averaging, sequence conditioning and graph propagation provide complementary gains, and the complete model performs best among the matched configurations. An exploratory application to ovarian cancer samples further illustrates how this integrated representation can be used in a disease setting.

Our main contributions are as follows:

- **A sequence-conditioned graph clock.** We introduce the first epigenetic clock to jointly model sample-specific methylation, local DNA sequence, and explicit co-methylation relationships among CpG sites.
- **Local sequence provides information beyond CpG-site identity.** We find that most of its predictive value depends on the correct CpG–sequence correspondence rather than additional model capacity or an arbitrary fixed site identifier.
- **Training variance obscures architectural effects.** We find that run-to-run variation can obscure the complementary contributions of sequence conditioning and graph propagation, while matched prediction averaging reveals consistent gains from both.
- **A calibration-aware disease application.** Applying the model to ovarian cancer samples, we show how improved biological-age prediction can support more reliable age-acceleration analysis while highlighting the need to distinguish disease-associated deviations from age-dependent calibration bias.

2 Related Work

Epigenetic clocks and sequence-aware methylation models. Epigenetic clocks were established using penalised linear regression on age-associated CpG sites (Horvath, 2013; Hannum et al., 2013), and elastic-net clocks trained on large cohorts remain strong (Bernabeu et al., 2023). Later methylation biomarkers estimate the pace of biological ageing rather than biological age (Belsky et al., 2022).

Deep clocks model non-linear age effects using feedforward networks such as DeepMAge (Galkin et al., 2021) and AltumAge (de Lima Camillo et al., 2022), or residual convolutional networks such as ResNetAge (Shi et al., 2024). Despite their architectural differences, these models represent each sample as a methylation vector: CpG sites occupy fixed coordinates without explicit relations between sites or access to their local nucleotide context. Methylation foundation models such as CpGPT (de Lima Camillo et al., 2024) and MethylGPT (Ying et al., 2024b) learn dependencies among sites implicitly through self-attention, and CpGPT additionally incorporates sequence context. They do not, however, combine local CpG sequence with explicit co-methylation graphs or distinguish sequence information from CpG-site identity.

Graph-based clocks and local DNA sequence. Graph representation learning is increasingly used to model interacting biological entities (Li et al., 2022). CpG sites whose methylation covaries across individuals form co-methylation modules that change with age (Horvath et al., 2012). Graph-based clocks represent CpG sites as nodes and encode co-methylation, chromosomal proximity, and shared gene membership as edges (Ahmed et al., 2025); multi-relational extensions process different relations separately before combining them (Qing et al., 2026). These approaches make relationships among CpG sites part of the prediction, but comparisons with vector models often change site-wise processing and graph propagation together, so a performance difference cannot by itself be attributed to explicit graph structure. Local sequence nevertheless contains information relevant to CpG methylation and co-methylation (Santoni, 2021; Affinito et al., 2020), contributes to predicting single-cell methylation states (Angermueller et al., 2017), and can be represented by pretrained DNA language models (Nguyen et al., 2023; Zhou et al., 2024; Schiff et al., 2024; Dalla-Torre et al., 2025). Graph-based clocks and sequence-aware methylation models have therefore developed along largely separate lines. No previous epigenetic clock jointly models sample-specific methylation, the wider local sequence surrounding each CpG site, and explicit co-methylation relationships. This gap motivates a unified representation of methylation, sequence context, and CpG relationships.

Attributing fixed site features and controlling training variance. The difficulty of attributing predictive value to a fixed site feature is not specific to methylation. Inputs that make nodes distinguishable can improve GNNs, including positional encodings (Dwivedi et al., 2022) and random features without semantic content (Sato et al., 2021; Abboud et al., 2021). Because the sequence associated with each CpG site is fixed across individuals, an apparent benefit may similarly arise from informative local context or from CpG-site identity. Reassigning sequences among CpG sites preserves their distribution and the model's parameter count while breaking the correct CpG–sequence correspondence, allowing these explanations to be distinguished. Model comparisons are further complicated by run-to-run variability, and conclusions drawn from a small number of training runs may be unstable (Bouthillier et al., 2021; Agarwal et al., 2021). Under controlled evaluation, structure-agnostic baselines can match GNNs (Errica et al., 2020), while flawed benchmarks can overstate the benefits of specialised graph architectures (Platonov et al., 2023). Averaging independently trained networks can reduce variance and improve prediction (Lakshminarayanan et al., 2017). They motivate matched controls to separate sequence and graph effects, and repeated training with prediction averaging to assess stability.

3 METHOD

3.1 SEQUENCE-CONDITIONED CO-METHYLATION GRAPH CLOCK

Sample k is represented by methylation beta values $\mathbf{m}^{(k)} \in [0,1]^N$ over $N = 20{,}318$ CpG sites. Each site i has fixed genomic annotations $\mathbf{f}_i \in \mathbb{R}^9$ and a one-hot sequence context $\mathbf{S}_i \in \{0,1\}^{122\times 4}$, both shared across samples (Appendix A).

Fold-specific graph. For each train–validation fold f, we build a graph $\mathcal{G}^{(f)} = (\mathcal{V}, \mathcal{E}^{(f)})$ over the CpG sites from training samples only. Let $r_{ij}^{(f)}$ denote the Pearson correlation between sites i and j in the training partition, c_i their chromosome labels, and d_{ij} their genomic distance. Following Ahmed et al. (2025), we set $\tau = 0.70$ and connect distinct sites when

$$\left|r_{ij}^{(f)}\right| > \tau \;\vee\; \left(c_i = c_j \wedge \left|r_{ij}^{(f)}\right| > \tau - 0.02\right) \;\vee\; \left(c_i = c_j \wedge d_{ij} < 10^5 \wedge \left|r_{ij}^{(f)}\right| > \tau - 0.04\right), \quad (1)$$

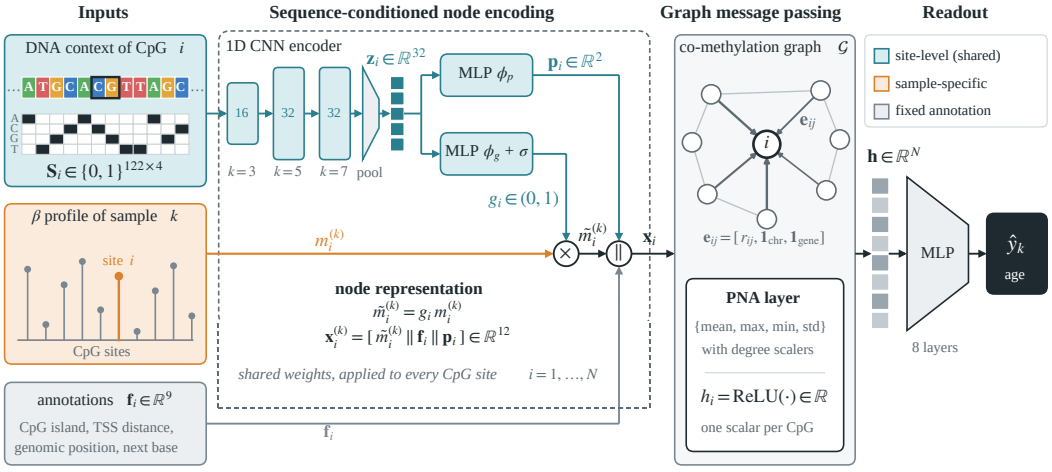


Figure 1: **Sequence-conditioned co-methylation graph clock.** A shared CNN encodes the 122-bp sequence $\mathbf{S}_i$ of each CpG site into a gate g_i and a projection $\mathbf{p}_i$. The gated value $\tilde{m}_i^{(k)} = g_i m_i^{(k)}$, the annotations $\mathbf{f}_i$, and $\mathbf{p}_i$ form the node features of a co-methylation graph built from training samples only. A PNA layer maps each node to a scalar, and an MLP predicts age.

so that nearby sites on the same chromosome may connect at lower correlation. Each edge carries $\mathbf{e}_{ij}^{(f)} = [\, r_{ij}^{(f)},\ \mathbf{1}(c_i = c_j),\ \mathbf{1}(\text{gene}(i) = \text{gene}(j))\,]$. Between 17% and 23% of sites have no neighbors in a given fold and therefore receive no messages. Additional information are provided in Appendix B.

Sequence conditioning. A shared three-layer 1D CNN encodes the sequence context of each site,

$$\mathbf{z}_i = \text{GlobalMaxPool}\big(\text{CNN}_\theta(\mathbf{S}_i)\big) \in \mathbb{R}^{32}, \tag{2}$$

and two small MLPs map $\mathbf{z}_i$ to a scalar gate and a two-dimensional projection,

$$g_i = \sigma\big(\phi_g(\mathbf{z}_i)\big), \qquad \mathbf{p}_i = \phi_p(\mathbf{z}_i) \in \mathbb{R}^2. \tag{3}$$

The gate rescales the methylation value to $\tilde{m}_i^{(k)} = g_i m_i^{(k)}$, which gives the node representation

$$\mathbf{x}_i^{(k)} = \big[\, \tilde{m}_i^{(k)} \,\|\, \mathbf{f}_i \,\|\, \mathbf{p}_i \,\big] \in \mathbb{R}^{12}. \tag{4}$$

Appendix C provides additional information about this sequence encoder.

Message passing and readout. A single PNA layer (Corso et al., 2020) forms edge-aware messages and combines them with the node's own representation,

$$\boldsymbol{\mu}_{ij}^{(k)} = \psi_\Theta\big([\,\mathbf{x}_i^{(k)} \| \mathbf{x}_j^{(k)} \| \mathbf{e}_{ij}^{(f)}\,]\big), \qquad h_i^{(k)} = \text{ReLU}\Big(\rho_\Theta\big([\,\mathbf{x}_i^{(k)} \| \mathbf{a}_i^{(k)}\,]\big)\Big), \tag{5}$$

where $\mathbf{a}_i^{(k)}$ concatenates mean, maximum, minimum, and standard-deviation aggregations of $\{\boldsymbol{\mu}_{ij}^{(k)}\}$ over the neighbors of i, each rescaled by identity, amplification, and attenuation degree scalers. The N scalar node outputs are concatenated into $\mathbf{h}^{(k)} \in \mathbb{R}^N$ and mapped to biological age by an eight-layer MLP.

3.2 Controlled Component Dissection

The proposed model jointly incorporates sequence conditioning and graph propagation, but aggregate performance alone cannot identify their respective contributions. Matched controls isolate these effects and test whether sequence information depends on the correct CpG–sequence correspondence. All configurations share data partitions, seeds, annotations, readout, and training settings, varying one component at a time (Appendix D).

Factorial design. We cross sequence conditioning with graph message passing. Sequence-agnostic variants remove the sequence pathway and use

$$\mathbf{x}_{i,\text{base}}^{(k)} = \left[\, m_i^{(k)} \,\|\, \mathbf{f}_i \,\right] \in \mathbb{R}^{10} \tag{6}$$

in place of Equation 4, and no-graph variants replace Equation 5 with a site-wise map $h_i^{(k)} = \mathrm{ReLU}\big(\mathbf{w}^\top \mathbf{x}_i^{(k)} + b\big)$ whose weights are shared across sites. This yields the sequence-conditioned PNA (our model), the sequence-agnostic PNA, the sequence-conditioned MLP, and the sequence-agnostic MLP, so each contrast along one factor changes a single component.

Permuted-sequence and operator controls. To distinguish correct local sequence context from CpG-site identity, the parameter-matched sequence-reassignment control uses the same architecture but replaces $\mathbf{S}_i$ with $\mathbf{S}_{\pi(i)}$, where π is a random permutation of the N sites drawn once per seed. Separately, as single-fold diagnostics, we replace PNA with a Graph Convolutional Network (GCN), Graph Attention Network (GAT), GraphSAGE, or Graph Isomorphism Network (GIN) to assess dependence on the message-passing operator. Elastic Net provides a linear baseline trained on the same partitions. Additional details are provided in Appendix E.

3.3 Training and Evaluation

We hold out 20% of the samples of each study as a fixed test set and divide the remaining samples into five train–validation folds. Each configuration is trained with three pre-specified seeds per fold, giving 15 runs. Our variants share identical fixed hyperparameters, graphs are estimated from training samples only, and checkpoints are selected by validation MAE (Appendix F).

Matched comparisons. All comparisons are matched by fold and seed. For models A and B, the paired difference in test MAE and its seed average within fold f are

$$d_{f,s}(A,B) = \mathrm{MAE}_A^{(f,s)} - \mathrm{MAE}_B^{(f,s)}, \qquad \bar{d}_f = \tfrac{1}{3}\textstyle\sum_s d_{f,s}(A,B), \tag{7}$$

so positive values favor B. We report $\bar{d} = \frac{1}{5}\sum_f \bar{d}_f$ with the interval $\bar{d} \pm t_{0.975,4}\,\mathrm{sd}(\bar{d}_f)/\sqrt{5}$, together with the numbers of folds and runs that favor each model. These intervals are approximate because folds share development samples and the test set.

Matched ensembles. Because the graph models showed run-to-run variability exceeding the component effects being measured (Section 5.1), we also evaluate matched ensembles as a secondary analysis. Ensembles average predicted ages rather than MAEs, so for sample k the three-seed ensemble of fold f and the 15-model ensemble predict

$$\hat{y}_k^{(f)} = \tfrac{1}{3}\textstyle\sum_s \hat{y}_k^{(f,s)}, \qquad \hat{y}_k = \tfrac{1}{15}\textstyle\sum_{f,s} \hat{y}_k^{(f,s)}, \tag{8}$$

and the five fold-level ensemble MAEs are compared with Equation 7. Formal definitions are given in Appendix G.

4 Experimental Setup

We use the curated dataset released by Ahmed et al. (2025), which contains 3,707 blood methylation profiles from 37 public studies in the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2013) and EBI ArrayExpress (Athar et al., 2019), listed in Table 3. The dataset separates healthy samples, which we use for training and evaluation, from case samples such as those of GSE19711, which we use only in Section 5.6. We reserve 20% of the samples of each study for a fixed held-out test set, producing 756 test samples, and split the remaining 2,951 samples into five folds that each use four partitions for training and one for validation, giving approximately 2,361 training and 590 validation samples per fold. Because the test set samples each study, our evaluation measures within-study. We compare against Elastic Net and ResnetAge (Shi et al., 2024), trained on the same folds and seeds as our variants. The published Horvath (Horvath, 2013) and AltumAge (de Lima Camillo et al., 2022) clocks are applied without refitting and reported as external references, because their training data may overlap with our studies. MAE is the primary metric, with median absolute error (MedAE), root mean squared error (RMSE), and R^2 as secondary metrics, all computed on the held-out test set. All experiments are conducted on one NVIDIA A100 GPU. Details are given in Appendices A and C.

5 RESULTS

5.1 OVERALL PREDICTIVE PERFORMANCE

Table 1: **Overall performance on the held-out test set (years).** Mean ± standard deviation over 15 runs, or five folds for the deterministic Elastic Net. Published clocks are evaluated once, and bold marks the best matched model.

Model	MAE ↓	MedAE ↓	RMSE ↓	R^2 ↑
Trained on our partitions				
Sequence-conditioned PNA (ours)	$\mathbf{3.405 \pm 0.210}$	$\mathbf{2.122 \pm 0.193}$	$\mathbf{5.434 \pm 0.244}$	$\mathbf{0.9570 \pm 0.0039}$
Sequence-agnostic PNA	3.550 ± 0.162	2.242 ± 0.169	5.584 ± 0.258	0.9546 ± 0.0042
Sequence-agnostic MLP	3.693 ± 0.139	2.326 ± 0.162	5.759 ± 0.225	0.9517 ± 0.0037
Elastic Net	3.789 ± 0.037	2.771 ± 0.064	5.595 ± 0.060	0.9545 ± 0.0010
ResnetAge	5.373 ± 0.354	3.726 ± 0.349	7.728 ± 0.425	0.9129 ± 0.0098
Published clocks, external reference				
AltumAge	3.710	1.375	6.905	0.9307
Horvath	5.823	3.892	8.838	0.8865

Table 1 compares our model with matched baselines and published clocks. Among models trained on our partitions, the sequence-conditioned PNA attains the lowest MAE, MedAE, and RMSE and the highest R^2. The sequence-agnostic PNA is the strongest baseline, and our model improves on it by 0.144 years ($[0.027, 0.261]$), on the sequence-agnostic MLP, which uses neither sequence nor graph, by 0.288 years ($[0.079, 0.497]$), on Elastic Net by 0.384 years ($[0.233, 0.535]$), and on ResnetAge by 1.967 years ($[1.765, 2.170]$), in all five folds. The neural models and Elastic Net differ in the shape of their error distributions rather than only in magnitude. Elastic Net has a higher MAE than the sequence-agnostic MLP (3.789 versus 3.693 years) but a lower RMSE (5.595 versus 5.759 years), so the neural models are more accurate on typical samples at the cost of larger errors on the hardest ones. The run-to-run standard deviation of our model, 0.210 years, however, exceeds the component effects measured in Section 5.2, which we analyze in Section 5.4. Among the published clocks, AltumAge has the lowest MedAE but a much larger RMSE (6.905 years), and for both the three worst studies contribute 38–39% of the total absolute error.

5.2 SEQUENCE AND GRAPH CONTRIBUTIONS

Table 2: **Factorial analysis of sequence conditioning and graph message passing.** Cells report test MAE (years) as the mean ± standard deviation over 15 runs (top) or five three-seed ensembles (bottom). Effects are paired MAE reductions, with 95% t-intervals over folds.

	Backbone	Sequence-agnostic	Sequence-conditioned	Sequence effect
Single runs	No graph	3.693 ± 0.139	3.487 ± 0.178	0.206 $[-0.010, 0.423]$
	PNA	3.550 ± 0.162	$\mathbf{3.405 \pm 0.210}$	0.144 $[0.027, 0.261]$
	Graph effect	0.144 $[-0.098, 0.386]$	0.082 $[-0.066, 0.229]$	
Three-seed ensembles	No graph	3.637 ± 0.123	3.391 ± 0.125	0.247 $[0.049, 0.445]$
	PNA	3.455 ± 0.114	$\mathbf{3.128 \pm 0.085}$	0.327 $[0.206, 0.448]$
	Graph effect	0.182 $[-0.066, 0.430]$	0.263 $[0.133, 0.392]$	

Table 2 (top) reports the 2×2 design with single runs. Sequence conditioning reduces MAE with and without graph propagation. With PNA, it reduces MAE from 3.550 to 3.405 years, a paired reduction of 0.144 years ($[0.027, 0.261]$) that favors sequence conditioning in every fold. Without a graph, it reduces MAE from 3.693 to 3.487 years (0.206 years) in every fold, although the fold-level interval, $[-0.010, 0.423]$, includes zero because this gain varies across folds.

The effect of graph message passing is smaller and less stable. For sequence-conditioned nodes, PNA reduces MAE from 3.487 to 3.405 years (0.082 years) in four of five folds, with an interval of $[-0.066, 0.229]$. For sequence-agnostic nodes, PNA reduces MAE by 0.144 years, but in only

three of five folds ($[-0.098, 0.386]$). In the single-fold operator diagnostic, GCN, GAT, GraphSAGE, and GIN all yield higher MAE than PNA (3.898 to 4.579 versus 3.459 years, Appendix E). At the single-run level, sequence conditioning therefore provides the larger and more reliable gain, and the fold-level analysis does not establish a contribution of graph propagation. Section 5.4 shows that this conclusion changes once run-to-run variability is averaged out.

5.3 Sequence-specific Controls

Because the sequence pathway produces the same g_i and $\mathbf{p}_i$ for a site in every sample, its benefit could reflect node identity rather than sequence content. The parameter-matched permuted-sequence control separates these explanations. Over the 15 fold–seed runs, the mean MAE is 3.550 years for the sequence-agnostic PNA, 3.542 years for the permuted-sequence model, and 3.405 years for the correctly matched model. The correctly matched model outperforms the permuted-sequence model in all 15 runs and in all five folds, by 0.137 years on average ($[0.025, 0.249]$). Permutation therefore removes about 95% of the 0.144-year sequence benefit at the single-run level, and the permuted-sequence model performs on par with the sequence-agnostic PNA, improving on it in only 7 of 15 runs. With three-seed ensembles, the correctly matched model remains better than the permuted-sequence model by 0.199 years ($[0.078, 0.321]$, five of five folds), which corresponds to about 60% of the ensemble sequence benefit, whereas the permuted-sequence model improves on the sequence-agnostic PNA by a non-significant 0.128 years ($[-0.053, 0.309]$). The sequence benefit thus depends mainly on the correspondence between each CpG site and its own sequence, while a smaller part, visible only after ensembling, is consistent with node identity alone. This control applies to any node feature that is constant across samples, and without it the sequence gain could not be distinguished from node identity. Per-run and age-stratified results are given in Appendix H.

5.4 Run-to-run Variability and Matched Ensembles

Graph-model runs vary considerably. Our model's run-to-run standard deviation is 0.210 years, which exceeds both the sequence effect of 0.144 years and the graph effect of 0.082 years that Section 5.2 measures, with fold 2, seed 1 reaching 3.97 years. A single run therefore cannot resolve either component. Averaging the 15 fold–seed predictions reduces MAE by 0.395 years for our model, 0.219 years for the sequence-agnostic PNA, 0.161 years for the sequence-conditioned MLP, and 0.131 years for the sequence-agnostic MLP. The configurations that combine more components gain most from averaging, indicating that a large share of their single-run error is variance.

Table 2 (bottom) repeats the factorial analysis with three-seed ensembles within each fold, giving every model the same number of members. Removing either component from the combined model now increases error in all five folds. Sequence conditioning reduces MAE by 0.327 years with PNA ($[0.206, 0.448]$) and by 0.247 years without a graph ($[0.049, 0.445]$). Graph message passing reduces MAE by 0.263 years for sequence-conditioned nodes ($[0.133, 0.392]$, five of five folds) but by 0.182 years for sequence-agnostic nodes ($[-0.066, 0.430]$, four of five folds). Graph propagation is therefore consistently beneficial only when nodes are sequence-conditioned, although the difference between the two graph effects, 0.081 years, is not significant ($[-0.137, 0.298]$).

The combined model attains the lowest ensemble MAE, 3.128 ± 0.085 years, and averaging also reduces its variability, from 0.210 years across single runs to 0.085 years across ensembles. With all 15 runs averaged, its MAE is 3.010 years. Single-run comparisons therefore understate the benefit of combining sequence and graph structure, and the two components are complementary only in an analysis that averages over training runs. More details are shown in Appendix I.

5.5 Failure Modes across Age and Calibration

Figure 2 compares the three factorial variants that differ in one component at a time, so the age-resolved effect of each component can be read directly. Error grows with age for all models. For our model, MAE rises from 0.19 ± 0.08 years at ages ≤ 0 and 1.86 ± 0.18 years at 0–20 to 5.61 ± 0.57, 7.30 ± 1.25, and 8.83 ± 0.98 years at 65–75, 75–80, and above 80 years. Sequence conditioning lowers the MAE of the PNA clock in all eight strata, including the two oldest (8.11 to 7.30 and 9.31 to 8.83 years), although these strata contain only 20 and 10 test samples. Adding PNA to the sequence-conditioned MLP lowers MAE in seven of eight strata, the exception being 65–75

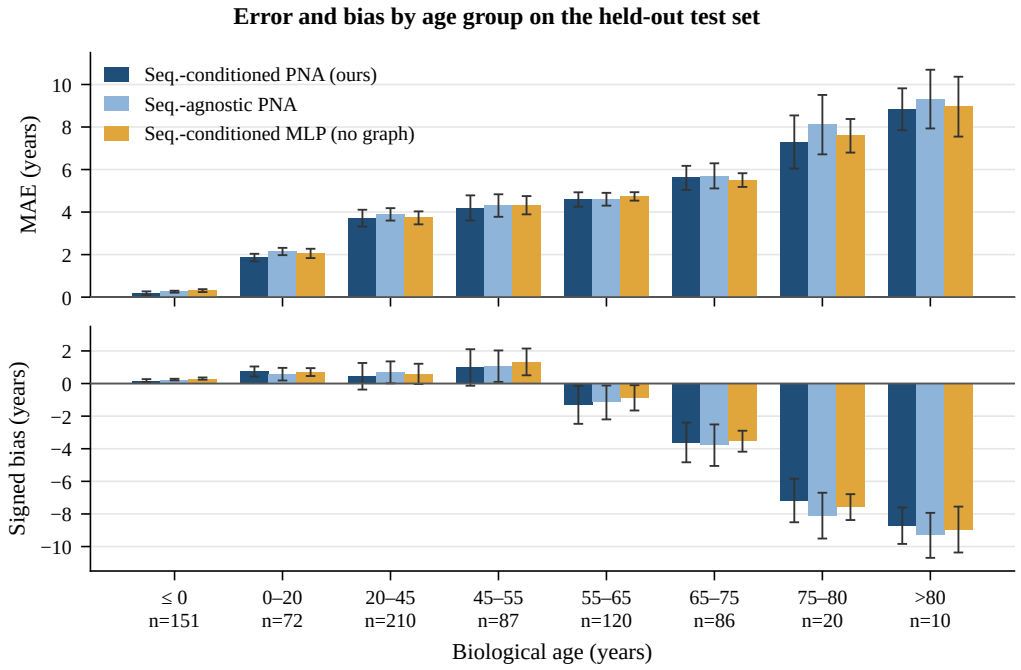


Figure 2: Top: MAE in years. Bottom: signed bias (predicted minus biological age). Bars show means over 15 fold–seed runs with one standard deviation.

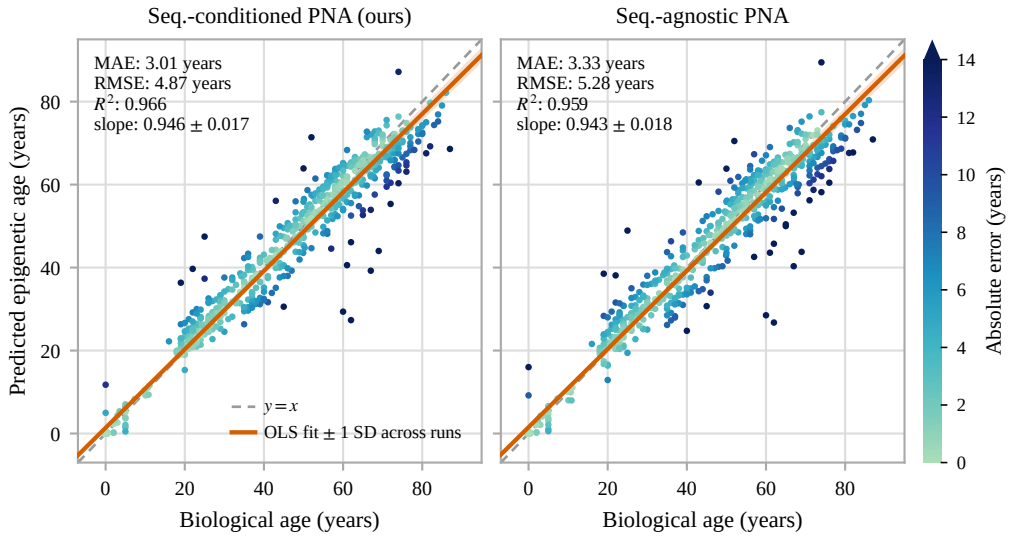
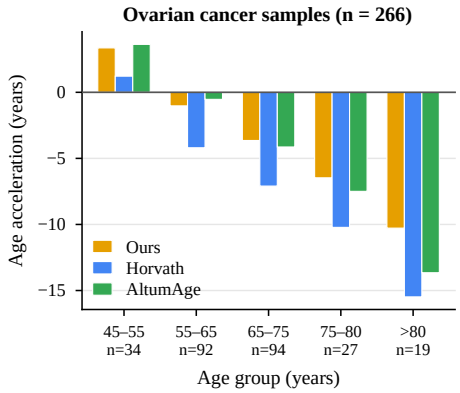


Figure 3: **Calibration on the held-out test set.** Each point shows a test-sample prediction averaged over 15 fold–seed runs, coloured by absolute error. The dashed line marks $y = x$, and the orange line and band show the least-squares fit and ± 1 standard deviation across the 15 run-specific fits.

years. The main failure mode is age compression. All three models overpredict younger ages and increasingly underpredict older ones, and for our model the signed bias reaches -7.18 ± 1.33 and -8.72 ± 1.12 years in the two oldest strata.



Metric	Ours	Horvath	AltumAge
MAE ↓	4.82	7.97	5.64
MedAE ↓	3.59	6.25	4.24
RMSE ↓	6.84	11.34	7.69
AA	−2.60	−5.95	−2.92

Figure 4: Age acceleration and prediction error on the 266 ovarian cancer samples. Left: mean age acceleration by age group. Right: prediction error and mean age acceleration, AA, defined as the predicted minus the biological age. Lower is better for the three error metrics.

Figure 3 plots the sample-averaged predictions of our model and the sequence-agnostic PNA. Our model has lower MAE (3.01 versus 3.33 years) and RMSE (4.87 versus 5.28 years) and higher R^2 (0.9656 versus 0.9595). Both models have regression slopes below one (0.946 ± 0.017 and 0.943 ± 0.018 across runs), consistent with the age compression above, so sequence conditioning reduces error without removing the compression. Corresponding diagnostics for the sequence-conditioned MLP and sequence-agnostic MLP are given in Appendix J.

5.6 Application to Ovarian Cancer

In an application, we apply our model and the published Horvath and AltumAge clocks to the 266 ovarian cancer samples of GSE19711, defining age acceleration as the predicted minus the biological age. Our model is averaged over 15 runs, and the two published clocks serve as external references. As shown in Figure 4 (right), our model has the lowest MAE, MedAE, and RMSE of the three clocks, followed by AltumAge, whereas the Horvath clock has the largest errors on all three metrics. The mean age acceleration in the table is dominated by the 55–75 groups, which contain 186 of the 266 samples, so it summarizes mainly middle-aged and early-old individuals.

Figure 4 (left) breaks age acceleration down by age group. All three clocks yield positive age acceleration in the 45–55 group and increasingly negative values at older ages. This pattern is consistent with the age compression observed on the healthy test set (Section 5.5), in which older individuals are systematically underpredicted, so the negative values at older ages mainly reflect clock calibration rather than decelerated aging. The clocks differ most where compression is strongest. Across the three clocks, the range of mean age acceleration is 2.41 years in the 45–55 group and 5.20 years above 80 years, and our model has the smallest underprediction in the three oldest groups, reaching -10.28 years above 80 compared with -13.65 for AltumAge and -15.48 for Horvath.

6 Conclusion

In this work, we introduced a sequence-conditioned co-methylation graph clock that jointly models sample-specific methylation, local DNA sequence, and relationships among CpG sites for biological-age prediction. Matched controls show that local sequence improves prediction primarily when paired with its corresponding CpG site, demonstrating value beyond CpG-site identity alone. Graph propagation provides an additional, consistent benefit when training variance is reduced through matched prediction averaging, indicating complementary roles for sequence context and co-methylation structure. Run-to-run variability nevertheless exceeds the individual component effects, explaining why the contribution of graph propagation is difficult to resolve from single runs. More broadly, these findings show that fixed site-level features require appropriate reassignment controls for reliable attribution and that high-variance graph models require repeated, matched evaluation.

AI USE STATEMENT

In this work, we used generative AI tools only to edit manuscript text for grammar, clarity, concision, and readability. We did not use generative AI tools for any task for which disclosure is required under the ICLR 2027 AI policy. All AI-assisted edits were reviewed and, where necessary, revised by the authors against the underlying analyses, experimental results, and cited sources. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. All DNA methylation profiles used in this study are publicly available from the Gene Expression Omnibus and ArrayExpress and were generated by previously published studies. The data are de-identified, we did not collect new data or interact with human subjects, and we made no attempt to re-identify individuals. Epigenetic age predictors can reveal age-related information from molecular data and could be misused to infer such information about individuals without their consent. Our models estimate biological age for research purposes only and have not been validated for clinical, forensic, or other individual-level decisions. Public methylation cohorts are also unevenly distributed across age groups and populations, and we report age-stratified errors so that differences in performance across groups are visible.

REPRODUCIBILITY STATEMENT

We have taken several steps to support reproducibility. The dataset was released by Ahmed et al. (2025) and derives from 37 publicly available studies, whose accessions and sample counts are listed in Appendix A. The model, the co-methylation graph construction, and all controlled variants are specified in Section 3, with architectural details, parameter counts, and variant definitions in Appendices B, C, and D. The data partitions, pre-specified seeds, optimization settings, and checkpoint selection are described in Section 3.3 and Appendix F, and all statistical procedures in Appendix G. Per-run results for the permutation control and the ensemble analyses are reported in Appendices H and I. We provide anonymised code in an anonymous repository, including scripts for preprocessing, graph construction, training, and evaluation, the fixed split indices and seeds, and the per-sample test predictions of every run, from which all reported statistics can be recomputed. The code is available at `https://anonymous.4open.science/r/official_code-FE3B`.

REFERENCES

Ralph Abboud, İsmail İlkan Ceylan, Martin Grohe, and Thomas Lukasiewicz. The surprising power of graph neural networks with random node initialization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2112–2118. International Joint Conferences on Artificial Intelligence Organization, 2021. doi: 10.24963/ijcai.2021/291. URL https://doi.org/10.24963/ijcai.2021/291.

Ornella Affinito, Domenico Palumbo, Annalisa Fierro, Mariella Cuomo, Giulia De Riso, Antonella Monticelli, Gennaro Miele, Lorenzo Chiariotti, and Sergio Cocozza. Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics*, 112(1):144–150, 2020. doi: 10.1016/j.ygeno.2019.05.007.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f514cec81cb148559cf475e7426eed5e-Paper.pdf.

Saleh Sakib Ahmed, Nahian Shabab, Md Abul Hassan Samee, and M. Sohel Rahman. GraphAge: Unleashing the power of graph neural network to decode epigenetic aging. *PNAS Nexus*, 4(6): pgaf177, 2025. doi: 10.1093/pnasnexus/pgaf177.

Christof Angermueller, Heather J. Lee, Wolf Reik, and Oliver Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18:67, 2017. doi: 10.1186/s13059-017-1189-z.

Awais Athar, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, Nuno A. Fonseca, Robert Petryszak, Irene Papatheodorou, Ugis Sarkans, and Alvis Brazma. ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Research*, 47(D1): D711–D715, 2019. doi: 10.1093/nar/gky964.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x.

Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013. doi: 10.1093/nar/gks1193.

Daniel W Belsky, Avshalom Caspi, David L Corcoran, Karen Sugden, Richie Poulton, Louise Arseneault, Andrea Baccarelli, Kartik Chamarti, Xu Gao, Eilis Hannon, Hona Lee Harrington, Renate Houts, Meeraj Kothari, Dayoon Kwon, Jonathan Mill, Joel Schwartz, Pantel Vokonas, Cuicui Wang, Benjamin S Williams, and Terrie E Moffitt. DunedinPACE, a DNA methylation biomarker of the pace of aging. *eLife*, 11:e73420, 2022. doi: 10.7554/eLife.73420.

Elena Bernabeu et al. Refining epigenetic prediction of chronological and biological age. *Genome Medicine*, 15(1):12, 2023. doi: 10.1186/s13073-023-01161-y.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13260–13271, 2020.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 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.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025. doi: 10.1038/s41592-024-02523-z.

Lucas Paulo de Lima Camillo, Louis R. Lapierre, and Ritambhara Singh. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging*, 8(1):4, 2022. doi: 10.1038/s41514-022-00085-y.

Lucas Paulo de Lima Camillo et al. CpGPT: A foundation model for DNA methylation. *bioRxiv*, 2024. doi: 10.1101/2024.10.24.619766. Version 3, August 2025.

Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations. In *10th International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=wTTjnvGphYj.

Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. A fair comparison of graph neural networks for graph classification. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HygDF6NFPB.

Fedor Galkin, Polina Mamoshina, Kirill Kochetov, Denis Sidorenko, and Alex Zhavoronkov. DeepMAge: A methylation aging clock developed with deep learning. *Aging and Disease*, 12(5): 1252–1262, 2021. doi: 10.14336/AD.2020.1202.

Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359–367, 2013. doi: 10.1016/j.molcel.2012.10.016.

Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10): R115, 2013. doi: 10.1186/gb-2013-14-10-r115.

Steve Horvath and Kenneth Raj. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*, 19(6):371–384, 2018. doi: 10.1038/s41576-018-0004-3.

Steve Horvath, Yafeng Zhang, Peter Langfelder, René S. Kahn, Marco P. M. Boks, Kristel van Eijk, Leonard H. van den Berg, and Roel A. Ophoff. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biology*, 13(10):R97, 2012. doi: 10.1186/gb-2012-13-10-r97.

Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(Suppl):245–254, 2003. doi: 10.1038/ng1089.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Michelle M. Li, Kexin Huang, and Marinka Zitnik. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6(12):1353–1369, 2022. doi: 10.1038/s41551-022-00942-x.

Carlos López-Otín, 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.

Mahdi Moqri, Chiara Herzog, Jesse R. Poganik, Biomarkers of Aging Consortium, Jamie Justice, Daniel W. Belsky, Albert Higgins-Chen, Alexey Moskalev, Georg Fuellen, Alan A. Cohen, Ivan Bautmans, Martin Widschwendter, Jingzhong Ding, Alexander Fleming, Joan Mannick, Jing-Dong Jackie Han, Alex Zhavoronkov, Nir Barzilai, Matt Kaeberlein, Steven Cummings, Brian K. Kennedy, Luigi Ferrucci, Steve Horvath, Eric Verdin, Andrea B. Maier, Michael P. Snyder, Vittorio Sebastiano, and Vadim N. Gladyshev. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell*, 186(18):3758–3775, 2023. doi: 10.1016/j.cell.2023.08.003.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin W. Thomas, Callum Birch Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Christopher Ré. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *11th International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=tJbbQfw-5wv.

Qing Qing, Xikun Zhang, Zhongyuan Zhang, Jiarui Liu, Xingtong Yu, Xiaotao Shen, Ziqi Xu, Qixin Zhang, Zhe Wang, and Renqiang Luo. Learning multi-relational graph representations for DNA methylation-based biological age estimation. *arXiv preprint arXiv:2605.07175*, 2026.

Daniele Santoni. The impact of flanking sequence features on DNA CpG methylation. *Computational Biology and Chemistry*, 92:107480, 2021. doi: 10.1016/j.compbiolchem.2021.107480.

Ryoma Sato, Makoto Yamada, and Hisashi Kashima. Random features strengthen graph neural networks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 333–341. Society for Industrial and Applied Mathematics, 2021. doi: 10.1137/1.9781611976700.38. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611976700.38.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: bi-directional equivariant long-range DNA sequence modeling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Dirk Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, 2015. doi: 10.1038/nature14192.

Lijuan Shi, Boquan Hai, Zhejun Kuang, Han Wang, and Jian Zhao. ResnetAge: A resnet-based DNA methylation age prediction method. *Bioengineering*, 11(1):34, 2024. doi: 10.3390/bioengineering11010034.

Kejun Ying, Hanna Liu, Andrei E. Tarkhov, Marie C. Sadler, Ake T. Lu, Mahdi Moqri, Steve Horvath, Zoltán Kutalik, Xia Shen, and Vadim N. Gladyshev. Causality-enriched epigenetic age uncouples damage and adaptation. *Nature Aging*, 4(2):231–246, 2024a. doi: 10.1038/s43587-023-00557-0.

Kejun Ying et al. MethylGPT: A foundation model for the DNA methylome. *bioRxiv*, 2024b. doi: 10.1101/2024.10.30.621013.

Jiaxuan You, Zhitao Ying, and Jure Leskovec. Design space for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17009–17021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c5c3d4fe6b2cc463c7d7ecba17cc9de7-Paper.pdf.

Jiaxuan You, Jonathan M. Gomes-Selman, Rex Ying, and Jure Leskovec. Identity-aware graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10737–10745, 2021. doi: 10.1609/aaai.v35i12.17283.

Yu Zhang, Yichen Yao, Yuanhao Tang, Yuan Cheng, Yinghui Xu, Ying He, Yuan Qi, and Li Jin. A robust computational framework for methylation age and disease-risk prediction based on pairwise learning. *Nature Computational Science*, 6(4):388–403, 2026. doi: 10.1038/s43588-025-00939-x.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V. Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *International Conference on Learning Representations*, 2024.

A Data

Source. We use the curated dataset released by Ahmed et al. (2025), who selected the blood-derived studies from the dataset list of AltumAge (de Lima Camillo et al., 2022) in the Gene Expression Omnibus (Barrett et al., 2013) and ArrayExpress (Athar et al., 2019). It contains 3,707 blood methylation profiles from 37 studies annotated as healthy, and the 624 case samples of these studies are used only in Section 5.6. Profiles come from two Illumina array platforms and are restricted to the $N = 20{,}318$ CpG sites shared across Illumina methylation platforms (de Lima Camillo et al., 2022), so all samples are described by the same site set and the co-methylation graph is defined over a fixed node set. Biological age is the regression target. Table 3 lists the source studies.

Composition. Biological ages span -0.3 to 93 years (mean 36.3, median 37.0), where values at or below zero denote prenatal and newborn samples recorded relative to birth. Of all profiles, 62.6% are annotated female. Studies contain 16 to 719 samples (median 60) and cover whole blood, buffy coat, peripheral leukocytes, sorted leukocyte populations, and umbilical cord blood. Age coverage is uneven: 19.4% of samples are neonatal and 1.4% are older than 80 years.

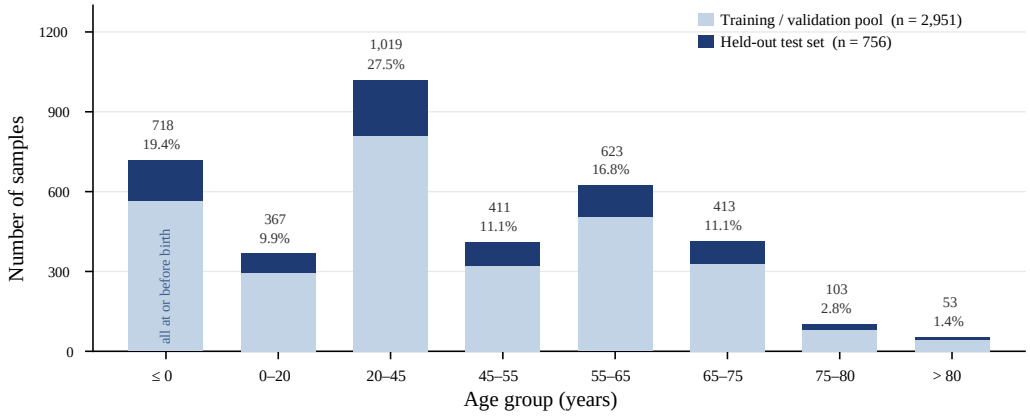


Figure 5: **Age distribution of the dataset.** Stacked bars show the development pool ($n = 2{,}951$) and the test set ($n = 756$). Numbers above the bars give the total count and its share of all samples.

Site annotations and sequence. Each CpG site i has a nine-dimensional fixed annotation vector $\mathbf{f}_i$, comprising a CpG-island indicator, CpG-island length, CpG-island start and end coordinates, distance to the transcription start site, three indicators encoding whether the next nucleotide is A, C, or T, and a normalized genomic coordinate, together with a 122-bp sequence context centred on the CpG dinucleotide and represented as a 122×4 one-hot matrix over A, C, G, and T.

Preprocessing and partitions. We use the beta values as provided in this dataset as described by Ahmed et al. (2025). Before training, 20% of the samples of each study are reserved for a fixed held-out test set, producing 756 test samples. The remaining $2{,}951$ samples form the development set, which is split into five folds, each using four partitions for training and one for validation, giving approximately $2{,}361$ training and 590 validation samples per fold. Figure 5 shows the age distribution of both partitions. Because the split is stratified by study, each age group contributes between 19% and 21% of its samples to the test set, so the test set has nearly the same age composition as the full dataset. The test set contains unseen individuals from studies represented in the development data, so the evaluation measures within-study rather than unseen-study generalization.

B Co-methylation Graph Details

A separate co-methylation graph is constructed for each train–validation fold from the training partition only, using the edge rule of Equation 1. Each qualifying undirected CpG pair is represented by two directed edges carrying the attributes $\mathbf{e}_{ij}^{(f)}$ defined in Section 3.1. The topology is fixed after construction and reused for all training, validation, and held-out test samples of that fold, and the PNA degree statistics are computed from this training-derived graph alone. Table 4 reports the resulting statistics. Folds are numbered from 1 to 5 throughout the paper.

The five graphs are similar in size, with 2.13 to 2.38 million undirected edges and mean degrees between 209.9 and 233.8. The degree distribution is strongly right-skewed. The median degree is only 33 to 54, whereas the most connected sites have more than 1,500 neighbors, which motivates the degree scalers of PNA. The 3,552 to 4,628 isolated sites receive no messages, so for them the PNA layer reduces to a site-wise transformation of their own features. Fold 2 has the sparsest graph, with the fewest edges and the most isolated sites.

Table 3: Source studies. GSE accessions are from the Gene Expression Omnibus and E-GEOD or E-MTAB accessions from ArrayExpress, where E-GEOD-n mirrors GSEn. Tissue annotations are study-level labels from the source records, and all samples listed are blood-derived. Test counts and age ranges refer to the held-out test set.

Accession	Tissue annotation	Total	Test	Test age range
E-GEOD-51388	Blood	60	12	24 to 74
E-GEOD-52588	Blood	58	12	19 to 61
E-GEOD-53128	Blood (whole)	43	9	47.3 to 69.3
E-GEOD-53740	Blood	165	33	45 to 86
E-GEOD-54399	Umbilical cord blood, placenta	48	10	0
E-GEOD-54690	Blood	20	4	30 to 48
E-GEOD-56553	Blood (mononuclear cells)	31	7	19.3 to 28.5
E-GEOD-57484	Blood (whole)	22	5	9.8 to 11.1
E-GEOD-58045	Blood	172	35	33.0 to 68.3
E-GEOD-59509	Blood, saliva, menstrual blood, vaginal swab	36	8	20 to 59
E-GEOD-59592	Blood	120	24	0.38
E-GEOD-62219	Blood (peripheral leukocytes)	60	12	0 to 5
E-GEOD-64495	Blood (whole)	106	22	2.4 to 61
E-GEOD-64940	Umbilical cord blood	216	44	0
E-GEOD-65638	Blood (whole)	16	4	22 to 32
E-GEOD-67444	Blood (neonatal)	70	14	0 to 5
E-GEOD-67705	Blood	44	9	27 to 60
E-GEOD-71245	Blood	24	5	30 to 52
E-GEOD-71955	Blood (CD4+ and CD8+ T cells)	61	13	36 to 74
E-GEOD-72338	Blood (neutrophils and monocytes)	36	8	19 to 49
E-GEOD-77445	Blood	85	17	20 to 67
E-GEOD-79056	Umbilical cord blood	36	8	−0.30 to 0.00
E-GEOD-83334	Blood (whole and umbilical cord)	30	6	0 to 5
E-MTAB-2344	White blood cells	24	5	69 to 80
E-MTAB-2372	White blood cells	47	10	23 to 62
GSE19711	Blood	536	108	50 to 84
GSE20236	Blood	93	19	53 to 73
GSE20242	Blood (T cells)	40	8	18 to 44
GSE34257	Umbilical cord, blood	84	17	0
GSE36642	Blood, endothelial cells, placenta	123	25	−0.15 to −0.04
GSE37008	Blood	99	20	26 to 42
GSE41037	Blood	719	144	18 to 81
GSE49904	Blood (buffy coat)	71	15	36 to 85
GSE56606	Blood (CD14+ monocytes)	74	15	16 to 52
GSE57285	Blood	42	9	29 to 69
GSE69176	Umbilical cord blood	150	30	0
GSE99624	Blood	46	10	57 to 87
Total		3,707	756	−0.30 to 87

Table 4: Co-methylation graph statistics for each fold. Edges are undirected.

Fold	Edges	Mean degree	Median degree	Maximum degree	Isolated sites
1	2,375,554	233.84	53	1,616	3,628
2	2,132,119	209.87	33	1,730	4,628
3	2,315,741	227.95	42	1,581	4,185
4	2,338,385	230.18	54	1,651	3,552
5	2,359,753	232.28	49	1,628	3,873

C Architecture and Implementation Details

C.1 Sequence Encoder

The sequence encoder contains three one-dimensional convolutional layers with 16, 32, and 32 output channels and kernel sizes 3, 5, and 7, respectively. Each convolution is followed by batch normalization and a ReLU activation. The stacked convolutional layers have an effective receptive field of 13 bp. Global max pooling over the sequence dimension produces $\mathbf{z}_i \in \mathbb{R}^{32}$. The encoder weights are shared across all CpG sites.

C.2 Gate and Sequence Projection

The gate network ϕ_g has dimensions

$$32 \to 16 \to 1,$$

with a ReLU hidden activation and sigmoid output. The projection network ϕ_p has dimensions

$$32 \to 16 \to 2,$$

with a ReLU hidden activation.

The CNN contains $10{,}160$ trainable parameters, the gate network contains 545, and the projection network contains 562, giving

$$10{,}160 + 545 + 562 = 11{,}267$$

trainable parameters in the complete sequence pathway.

C.3 PNA Layer

We use one PNA layer with the four aggregators $\{\mathrm{mean}, \max, \min, \mathrm{std}\}$ and identity, amplification, and attenuation degree scalers, as defined in Equation 5. For the 12-dimensional sequence-conditioned node features, the concatenated aggregate $\mathbf{a}_i^{(k)}$ has dimension 144. We use one pre-layer and one post-layer in the PNA module, with three-dimensional edge attributes.

C.4 Readout Network

The scalar outputs of all $N = 20{,}318$ CpG sites are concatenated into $\mathbf{h}^{(k)} \in \mathbb{R}^N$ and passed to an eight-layer MLP:

$$20{,}318 \to 1024 \to 656 \to 256 \to 124 \to 64 \to 32 \to 8 \to 1.$$

ReLU activations are used by default, with SELU after the 656- and 64-unit layers. Dropout with probability 0.2 is applied after the 256- and 124-unit layers. The final layer is linear.

C.5 Baselines

Baselines fall into two groups that differ in the data on which they were trained. The first group consists of models trained on exactly the partitions described in Section 3.3, namely ResnetAge and Elastic Net. They use the same five folds as our models, ResnetAge uses the same three seeds, and any selection uses validation MAE within each fold and never test data. The second group consists of the published Horvath and AltumAge clocks, applied without refitting.

ResnetAge. ResnetAge (Shi et al., 2024) applies a residual convolutional network to the beta values of all $20{,}318$ sites as a one-dimensional signal, with fixed hyperparameters, the checkpoint with the lowest validation MAE, and channel widths reduced relative to the original work to suit the size of our training partitions. Sites are ordered by Illumina probe identifier rather than by genomic coordinate, so adjacent positions rarely lie on the same chromosome.

Elastic Net. Elastic Net is fitted to the raw beta values of all $20{,}318$ sites, with α and the ℓ_1 ratio selected from 24 combinations by validation MAE within each fold. All five folds select $\alpha = 0.003$ and an ℓ_1 ratio of 0.1. Because the fit is deterministic, Elastic Net is summarized over the five folds.

Published clocks. The Horvath multi-tissue clock (Horvath, 2013) and AltumAge (de Lima Camillo et al., 2022) are applied to the held-out test samples with their released coefficients and weights, without refitting or additional normalization. For the Horvath clock, 336 of the 353 clock sites are present among our 20,318 sites and carry 96.4% of the total absolute coefficient weight, and the missing sites are set to zero. AltumAge is defined on the same sites, so its input is complete. Because the 37 studies were selected from the dataset list of AltumAge (Ahmed et al., 2025), the test samples likely overlap with its development data, which may explain its low median absolute error. Horvath was also developed on public datasets, so overlap cannot be excluded for either clock.

D Controlled Model Variants

Table 5 summarizes all variants. Every variant shares the fixed annotations $\mathbf{f}_i$, a learned site-wise node transformation, the readout network, and the training settings of Appendix F. The variants differ only in the sequence input and the graph operator.

Table 5: Controlled model variants. Correct denotes each site's own sequence and permuted sequences reassigned across sites. The alternative operators are evaluated on one fold as a diagnostic.

Model	Sequence input	Graph propagation	Evaluation
Sequence-conditioned PNA (ours)	correct	PNA	5 folds $\times$ 3 seeds
Sequence-agnostic PNA	–	PNA	5 folds $\times$ 3 seeds
Sequence-conditioned MLP	correct	–	5 folds $\times$ 3 seeds
Sequence-agnostic MLP	–	–	5 folds $\times$ 3 seeds
Permuted-sequence PNA	permuted	PNA	5 folds $\times$ 3 seeds
GCN, GAT, GraphSAGE, GIN	–	alternative	fold 3 $\times$ 3 seeds

D.1 Factorial Variants

The sequence-conditioned variants use the node representation $\mathbf{x}_i^{(k)}$ of Equation 4, and the sequence-agnostic variants remove the complete sequence pathway and use $\mathbf{x}_{i,\mathrm{base}}^{(k)}$ of Equation 6. The PNA variants process these representations with the PNA layer of Appendix C. The no-graph variants remove neighborhood aggregation and map each node representation to a scalar independently with ϕ_{node}, a single linear layer whose weights are shared across all sites, followed by a ReLU. No graph topology, edge attributes, or neighborhood aggregation are used in these variants. In all four configurations, the N scalar site outputs are concatenated into $\mathbf{h}^{(k)} \in \mathbb{R}^N$ and passed to the same eight-layer readout.

D.2 Permuted-sequence Control

The permuted-sequence model retains the complete CNN, gate, projection, PNA layer, and readout. A random permutation π of the N CpG sites replaces $\mathbf{S}_i$ with $\mathbf{S}_{\pi(i)}$. The permutation is determined by the random seed, so the five folds of a seed share the same permutation and the three seeds use three different permutations. This preserves the sequence distribution, architecture, and parameter count while breaking the correspondence between each CpG site and its own local sequence. Comparing this model with the correctly matched sequence model tests whether predictive differences depend on the site–sequence correspondence rather than on node identity alone.

D.3 Alternative Graph Operators

To examine whether graph-model performance depends on the message-passing operator, the PNA layer of the sequence-agnostic graph model is replaced by GCN, GAT, GraphSAGE, or GIN. Each operator is a single message-passing layer that maps the 10-dimensional node representation to one scalar per site, and graph construction, partitions, readout, optimization, and checkpoint selection are kept fixed. The operators differ in their use of edge attributes. GCN uses $|r_{ij}|$ as a scalar edge weight, GAT uses the full three-dimensional edge attributes with a single attention head, and GraphSAGE and GIN use no edge attributes. GIN applies an internal MLP with one hidden layer of 16 units.

E Alternative Graph Operators

All operators use sequence-agnostic node representations and are evaluated on fold 3 with three seeds.

Table 6: Test MAE (years) of alternative graph operators on fold 3, as mean ± standard deviation over three seeds.

Operator	MAE
PNA	3.459 ± 0.134
GraphSAGE	3.898 ± 0.074
GAT	4.342 ± 0.213
GIN	4.426 ± 0.609
GCN	4.579 ± 0.624

All four alternative operators perform worse than PNA on this fold. GraphSAGE is closest, with an MAE 0.439 years higher than PNA, followed by GAT ($+0.883$ years), GIN ($+0.967$ years), and GCN ($+1.120$ years). The alternatives are also worse than the sequence-agnostic MLP, which uses no graph and reaches 3.523 years on the same fold. Replacing PNA with a standard message-passing operator therefore does not merely remove the benefit of the graph but lowers accuracy below that of a model without neighborhood aggregation. GCN and GIN are also markedly less stable across seeds, with standard deviations of 0.624 and 0.609 years compared with 0.134 years for PNA, whereas GraphSAGE has the lowest seed variability of all operators (0.074 years).

In their standard formulations, PNA and GraphSAGE both combine the own representation of a node with its aggregated neighborhood through separate transformations, whereas GCN, GAT, and GIN mix the two before the node update. The observed ordering is consistent with the interpretation that preserving the own methylation signal of each site matters when the readout operates on per-site scalar outputs. This diagnostic uses a single fold and sequence-agnostic nodes, however, and does not test this mechanism directly. We therefore use PNA throughout the main analysis without claiming that the ordering generalizes to other folds or to sequence-conditioned nodes.

F Training Details

Every neural configuration is trained with the three pre-specified random seeds $s \in \{0, 1, 2\}$ in each of the five folds of Appendix A, from a fresh random initialization. Model parameters, optimizer state, and checkpoints are not transferred between folds or seeds.

All variants of our model are optimized for 150 epochs using Adam with learning rate 8×10^{-5} and weight decay 6×10^{-4}. A ReduceLROnPlateau scheduler monitors validation MAE with factor 0.4, patience 10, and minimum learning rate 10^{-7}. Each batch contains one sample.

Learning-rate scheduling and checkpoint selection use validation MAE only, and each run retains the checkpoint with the lowest validation MAE. Test predictions are generated only afterwards and never influence scheduling or model selection.

G Statistical Analysis Details

Fold-level intervals. The paired difference $d_{f,s}$, its seed average $\bar{d}_f$, and the t-interval over the five folds are defined in Equation 7. The intervals are approximate because the folds share development samples and the same held-out test set, so the five fold-level differences are not independent.

Study-level counts. For each of the 37 studies represented in the test set, we compute the MAE of each model over the test samples of that study, averaged over the 15 fold–seed runs, and report the number of studies in which each model has the lower value. Studies differ in size, from 4 to 144 test samples, so this count weights each study equally rather than each sample.

Matched ensembles. Ensembles average predicted ages rather than MAEs, as defined in Equation 8. The five three-seed ensemble MAEs are compared with the fold-level analysis above, and the 15-model ensemble is used for sample-level diagnostics, with its study-level counts computed from

the absolute errors of the ensemble predictions. We also report five-fold ensembles within each seed, $\hat{y}_k^{(s)} = \frac{1}{5}\sum_f \hat{y}_k^{(f,s)}$, in Appendix I. Because averaging cancels errors of opposite sign across members, the MAE of an ensemble never exceeds the mean MAE of its members, and the two are equal only when all members err in the same direction on every sample. The reduction from ensembling therefore reflects how differently the members err, and we report it as the difference between the mean single-run MAE and the ensemble MAE.

H PERMUTED-SEQUENCE CONTROL RESULTS

Table 7: Test MAE (years) of the permuted-sequence control and the matched PNA models for each fold–seed run. The last row averages the 15 runs.

Fold	Seed	Sequence-agnostic PNA	Permuted-sequence PNA	Sequence-conditioned PNA
1	0	3.6635	3.2287	3.0900
1	1	3.4652	3.7039	3.4480
1	2	3.2661	3.2953	3.2821
2	0	3.7558	3.6177	3.3435
2	1	3.6419	4.0608	3.9715
2	2	3.6646	3.4593	3.4494
3	0	3.3851	3.1455	3.1358
3	1	3.6139	3.5582	3.4771
3	2	3.3784	3.4503	3.4462
4	0	3.8456	3.4738	3.3388
4	1	3.6745	3.7227	3.6274
4	2	3.5813	3.3928	3.2883
5	0	3.4431	3.8049	3.2629
5	1	3.4419	3.7616	3.4863
5	2	3.4226	3.4594	3.4339
Mean		3.550	3.542	3.405

The correctly matched model has a lower MAE than the permuted-sequence model in all 15 runs. The differences range from 0.004 to 0.542 years, with a mean of 0.137 years, and the fold means are 0.136, 0.124, 0.032, 0.112, and 0.281 years for folds 1 to 5. By contrast, the permuted-sequence model improves on the sequence-agnostic PNA in only 7 of 15 runs, and these differences range from -0.419 to 0.435 years around a mean of 0.007 years. Relative to the sequence-agnostic PNA, the correctly matched model is better in 10 of 15 runs, by 0.144 years on average, so permutation removes about 95% of this single-run benefit and the gain cannot be attributed to node identity alone. Because the permutation is tied to the training seed, the control rests on three permutations and provides a matched comparison rather than a permutation test.

The permuted-sequence and correctly matched models track each other across runs, with a correlation of 0.80 between their run-level MAEs, compared with 0.19 between the sequence-agnostic and correctly matched models. The two models share the complete architecture, including the sequence pathway, and the random seed, and differ only in the assignment of sequences to sites. Their paired differences therefore have a smaller standard deviation (0.147 years, compared with 0.240 years for the comparison with the sequence-agnostic PNA). The permuted-sequence model is at least as variable as the correctly matched model, with run-level standard deviations of 0.242 and 0.210 years compared with 0.162 years for the sequence-agnostic PNA, and within-fold seed standard deviations of 0.229 and 0.201 years compared with 0.108 years. The additional variability is therefore associated with the site-conditioning pathway itself rather than with the sequence content it receives. Fold 2, seed 1 illustrates this, giving the highest MAE of both models (4.061 and 3.972 years).

Because the permuted-sequence model is also highly variable, it benefits from ensembling as well. Averaging all 15 runs reduces its MAE by 0.309 years, compared with 0.395 years for the correctly matched model and 0.219 years for the sequence-agnostic PNA. Its three-seed ensemble MAE is 3.327 ± 0.135 years (folds 1 to 5: 3.178, 3.426, 3.194, 3.366, and 3.473), between the sequence-agnostic PNA (3.455 ± 0.114) and the correctly matched model (3.128 ± 0.085). The correctly matched model is better by 0.199 years ($[0.078, 0.321]$, five of five folds), whereas the advantage

of the permuted-sequence model over the sequence-agnostic PNA, 0.128 years, is not significant ($[-0.053, 0.309]$, four of five folds).

The benefit of correct sequence assignment is concentrated at older ages. Below 65 years, the correctly matched model is better than the permuted-sequence model by 0.057 to 0.089 years in every stratum. The difference grows to 0.333 years at 65–75, 0.906 years at 75–80, and 0.892 years above 80, and in these three strata the permuted-sequence model is worse than the sequence-agnostic PNA. These strata contain 86, 20, and 10 test samples, so we interpret this pattern descriptively. Finally, differences between the correctly matched and permuted-sequence models are smallest in fold 3 (0.032 years), so conclusions drawn from any single fold, including the single-fold operator diagnostic in Appendix E, should be read with this variation in mind.

I Ensemble Results

Table 8: Test MAE (years) of matched ensembles. Top: three-seed ensembles within each fold. Bottom: five-fold ensembles within each seed. Seed 1 contains our highest single-run error (fold 2).

	Seq.-cond. PNA (ours)	Seq.-agn. PNA	Seq.-cond. MLP	Seq.-agn. MLP
Fold 1	3.015	3.398	3.384	3.669
Fold 2	3.252	3.540	3.588	3.617
Fold 3	3.107	3.370	3.352	3.509
Fold 4	3.143	3.613	3.244	3.562
Fold 5	3.123	3.356	3.385	3.830
Mean $\pm$ sd	3.128 ± 0.085	3.455 ± 0.114	3.391 ± 0.125	3.637 ± 0.123
Seed 0	3.068	3.444	3.260	3.615
Seed 1	3.356	3.381	3.497	3.568
Seed 2	2.997	3.324	3.384	3.532

Table 8 lists the ensemble MAE of each configuration by fold and by seed. Three patterns stand out. First, the ranking within the factorial design is stable across folds. The combined model has a lower ensemble MAE than both single-component variants, the sequence-agnostic PNA and the sequence-conditioned MLP, in every fold, and sequence conditioning improves the no-graph model in every fold. The only factorial contrast that reverses is graph propagation without sequence, for which the sequence-agnostic MLP is better in fold 4 (3.562 versus 3.613 years).

Second, ensembling helps our model because its variability is dominated by the random seed rather than by the training fold. Averaged over folds, the standard deviation across the three seeds of a fold is 0.201 years for our model, compared with 0.137 years for the sequence-conditioned MLP, 0.108 years for the sequence-agnostic PNA, and 0.072 years for the sequence-agnostic MLP. The standard deviation of the fold means, in contrast, is similar across the four configurations (0.115 to 0.132 years). Seed variability thus grows as components are combined, and averaging over seeds removes most of it, which reduces the standard deviation of our model from 0.210 years across single runs to 0.085 years across three-seed ensembles.

Third, the per-seed five-fold ensembles show that the advantage of the combined model does not depend on a single fortunate seed. Our model has the lowest MAE of the four configurations for all three seeds (3.068, 3.356, and 2.997 years), including seed 1, which contains the run with the highest single-run error of our model (fold 2, 3.97 years). The ordering within the factorial design is therefore stable across both folds and seeds.

J Additional Prediction Diagnostics

Two observations follow from Table 9. First, each single component yields almost the same sample-level accuracy. The sequence-conditioned MLP and the sequence-agnostic PNA both reach an ensemble MAE of 3.33 years, with RMSEs of 5.29 and 5.28 years and R^2 values of 0.9594 and 0.9595, and both improve on the variant with neither component (3.56 years). Only their combination lowers these values to 3.01 years, 4.87 years, and 0.9656. At the sample level, the gain of the full

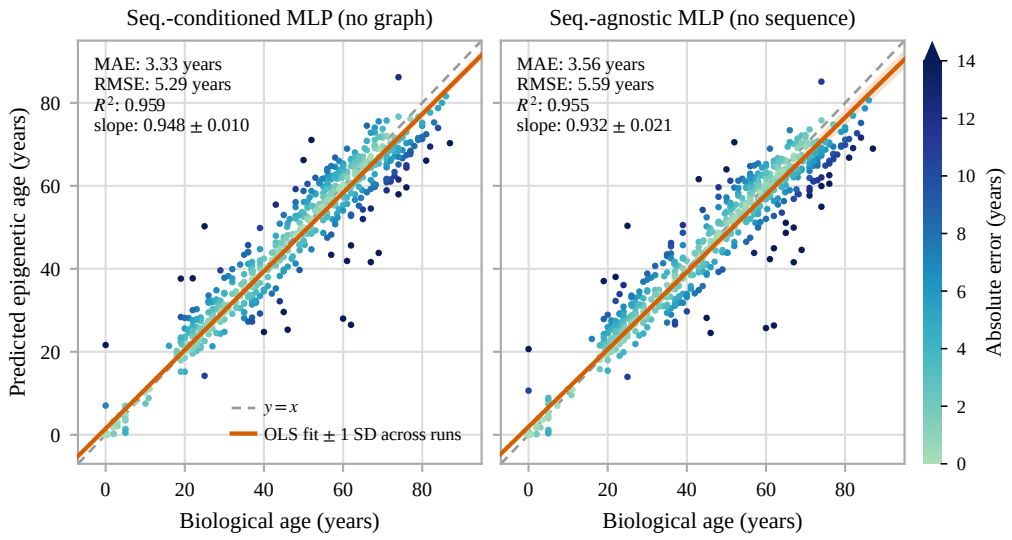


Figure 6: **Calibration of the two no-graph variants**, in the same format as Figure 3.

Table 9: Sample-level diagnostics of the 15-model ensembles. MAE, RMSE, and R^2 are computed from the sample-averaged predictions. The slope of the predicted-versus-biological regression is reported as the mean ± standard deviation across the 15 runs.

Model	MAE	RMSE	R^2	Slope
Sequence-conditioned PNA (ours)	3.01	4.87	0.9656	0.946 ± 0.017
Sequence-agnostic PNA	3.33	5.28	0.9595	0.943 ± 0.018
Sequence-conditioned MLP	3.33	5.29	0.9594	0.948 ± 0.010
Sequence-agnostic MLP	3.56	5.59	0.9545	0.932 ± 0.021

model therefore comes from combining sequence conditioning with graph propagation rather than from either component alone, in line with the ensemble factorial analysis in Section 5.4.

Second, calibration and accuracy diverge. All four configurations have regression slopes between 0.932 and 0.948, so neither sequence conditioning nor graph propagation reduces the age compression described in Section 5.5. The combined model lowers MAE and RMSE without moving the slope closer to one, which indicates that the two components reduce scatter around the regression line rather than the systematic underprediction of older individuals.