CpGPT: a Foundation Model for DNA Methylation

Lucas Paulo de Lima Camillo*†
Shift Bioscience, University of Cambridge
Cambridge, UK
lpd27@cam.ac.uk, lucas_camillo@alumni.brown.edu

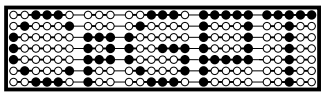
Raghav Sehgal†
Yale University School of Medicine
New Haven, CT, USA
raghav.sehgal@yale.edu

Jenel Armstrong
Yale University School of Medicine
New Haven, CT, USA
jenel.armstrong@yale.edu

Albert T. Higgins-Chen
Yale University School of Medicine
New Haven, CT, USA
a.higginschen@yale.edu

Steve Horvath
Altos Labs
Cambridge, UK
shorvath@altoslabs.com

Bo Wang
University of Toronto, Vector Institute, University Health Network
Toronto, Canada
bowang@vectorinstitute.ai



Abstract

DNA methylation is a critical epigenetic modification that regulates gene expression and plays a significant role in development and disease processes. Here, we present the Cytosine-phosphate-Guanine Pretrained Transformer (CpGPT), a novel foundation model pretrained on over 1,500 DNA methylation datasets encompassing over 100,000 samples from diverse tissues and conditions. CpGPT leverages an improved transformer architecture to learn comprehensive representations of methylation patterns, allowing it to impute and reconstruct genome-wide methylation profiles from limited input data. By capturing sequence, positional, and epigenetic contexts, CpGPT outperforms specialized models when finetuned for aging-related tasks, including chronological age prediction, mortality risk, and morbidity assessments. The model is highly adaptable across different methylation platforms and tissue types. Furthermore, analysis of sample-specific attention weights enables the identification of the most influential CpG sites for individual predictions. As a foundation model, CpGPT sets a new benchmark for DNA methylation analysis, achieving strong performance in the

*Main author
†Corresponding author

Biomarkers of Aging Challenge, where it placed second overall in chronological age estimation and first on the public leaderboard in methylation-based mortality prediction.

Highlights

- CpGPT is a novel foundation model for DNA methylation analysis, pretrained on over 1,500 datasets encompassing 100,000+ samples.
- The model demonstrates strong performance in zero-shot tasks including imputation, array conversion, and reference mapping.
- CpGPT achieves state-of-the-art results in mortality prediction and chronological age estimation.
- Sample-specific interpretability is enabled through analysis of attention weights.

1 Introduction

Since the introduction of the transformer architecture [1], artificial intelligence has undergone rapid advancements, particularly due to the development of foundation models and large language models (LLMs). Transformers leverage self-attention mechanisms to process sequential data more effectively, capturing long-range dependencies and complex patterns [2]. Pretrained on vast amounts of data in an unsupervised manner, these models have demonstrated exceptional performance across a variety of downstream tasks through transfer learning, making them highly versatile and effective.

Beyond natural language processing, transformers and foundation models have significantly impacted biology and medicine. Notably, they have advanced the analysis of single-cell transcriptomic data, uncovering previously unknown biology. Models such as scGPT [3], Geneformer [4], and Universal Cell Embeddings [5] display state-of-the-art performance for several tasks and even possess emergent behavior. The ability of transformers to integrate sequence, structural, and contextual information makes them particularly suited for biological data, which often involves complex interactions and hierarchies. The advent of these models has also begun to influence longevity research [6, 7].

In aging research, despite significant progress over the past decade, many widely used epigenetic aging clocks rely on relatively simple regularized linear models using Cytosine-phosphate-Guanine (CpG) DNA methylation data [8, 9, 10, 11, 12]. These models often do not consider the sequence context or genomic positions of CpG sites and cannot provide sample-specific interpretations of the epigenetic network. As a result, they may overlook complex interactions and the underlying biological mechanisms driving aging. A recent advancement involves the application of principal component decomposition before the linear model to enhance the reliability and performance of DNA methylation, blood chemistry, and histone mark clocks [13, 14, 15]. However, few age predictors, such as AltumAge [16] and DeepMAge [17], have utilized deep neural networks to model the complex relationships within methylation data.

Motivated by these advances, we have developed the CpG Pretrained Transformer (CpGPT), a transformer-based deep neural network that leverages the attention mechanism to effectively learn relationships between methylation sites by incorporating sequence, positional, and epigenetic information. As a foundation model, CpGPT is capable of performing a series of tasks in both zero-shot settings and when finetuned. For instance, the model can impute missing methylation values within a dataset, convert between different methylation platforms by reconstructing unmeasured CpG sites, perform zero-shot reference mapping to label samples without finetuning, and rank the importance of different CpG sites on a per-sample basis. Moreover, CpGPT excels when finetuned for specific tasks, achieving second

place overall for chronological age prediction and first place on the public leaderboard for methylation-based mortality prediction in the Biomarkers of Aging Challenge [18].

In this study, we focus on methylation-based mortality prediction by finetuning the proposed CpGPT model. It demonstrates robust performance across multiple cohorts, showing high accuracy and consistency across varied datasets and multiple metrics. The model effectively differentiated between individuals with high and low survival probabilities, revealing significant differences in survival curves and highlighting its ability to capture biologically meaningful variations in aging and mortality risk. Beyond mortality prediction, CpGPT exhibited strong predictive capabilities for a variety of morbidity outcomes, including multiple diseases and functional measures across different cohorts. The model also demonstrated associations with metabolic and lifestyle-related health assessments, cancer status, and psychosocial measures for depression, highlighting its broad applicability.

The CpGPT framework is highly generalizable and can be applied across various tasks, such as cancer prediction and classification, and even across different mammalian species. Overall, CpGPT establishes a new standard for DNA methylation analysis and offers a versatile tool for multiple applications in the field of aging and epigenetics.

2 Results

2.1 Developing a foundation model for DNA methylation

To fully harness the capabilities of foundation models, which often improve with increased data availability, we first curated a large-scale dataset named *CpGCorpus* (see Methods). This dataset comprises over 100,000 human DNA methylation samples collected from more than 1,500 studies, encompassing a diverse array of tissue types, developmental stages, and disease conditions. We preprocessed and harmonized this data to ensure consistency across different methylation array platforms, including Illumina 27k, 450k, EPIC, EPIC+, and EPICv2 arrays.

Building upon this extensive dataset, we designed the CpGPT model architecture to capture three primary types of contextual information: (1) sequence-based context; (2) local and global positional context; and (3) epigenetic state. To encode the sequence context, we utilized embeddings of the flanking nucleotides around each CpG site derived from a pretrained DNA language model [19, 20, 21]. For global positional context, we sorted the sequence embeddings by genomic positions, grouped them by chromosomes, and applied stochastic shuffling to prevent positional biases. We incorporated modifications of the original positional encoding method from Vaswani et al. [1] along with rotary positional embeddings [22] to enable the model to understand both the overall genomic structure and the local relationships between methylation sites. Finally, we transformed the single-value methylation state (beta value) of each CpG site into an embedding representing its epigenetic status using a dedicated encoder. These embeddings were integrated to form the input to the model (Figure 1A).

At the core of CpGPT is an enhanced version of the transformer architecture, known as transformer++ [23], which includes several modifications to improve model expressivity and training stability. CpGPT learns to create meaningful sample-level embeddings that capture the comprehensive methylation profile of each sample. This is achieved by training the model in an unsupervised manner to predict methylation states (beta values) and their associated uncertainty for CpG sites. The model can utilize the sample embedding of any genomic position, including those not seen during training, as a query to reconstruct the methylation state of that specific locus.

The training process for CpGPT involves a multifaceted approach, employing various loss functions to optimize different aspects of the model's performance. These include losses for accurate beta value prediction, uncertainty estimation, and the quality of sample embeddings. The model is designed to handle missing data, enabling it to work effectively with incomplete methylation profiles, which are common due to varying array designs and experimental conditions. This comprehensive training strategy allows CpGPT to learn rich,

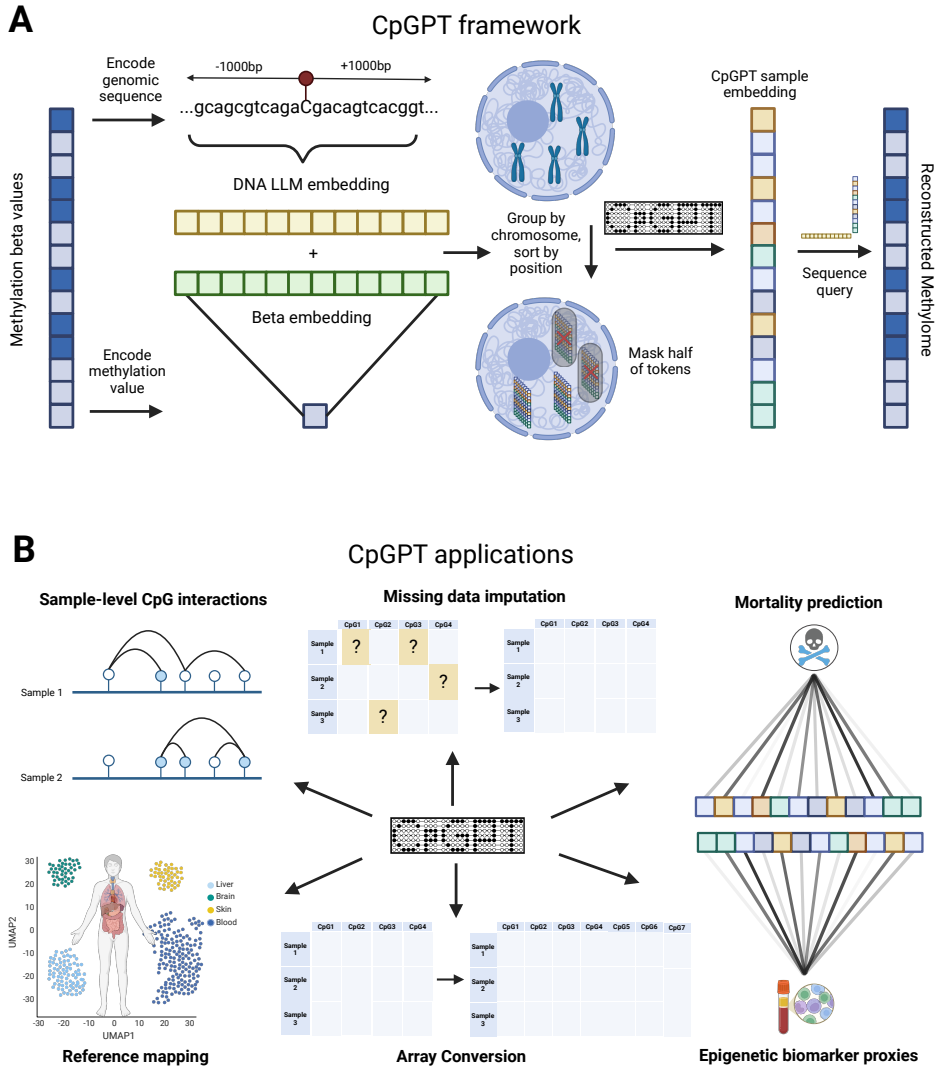


Figure 1: The CpGPT framework and downstream applications. (a) Epigenetic information is encoded by the model through several neural networks, including a projection of the methylation status, a transformation of the sequence embedding from a DNA LLM, and a positional addition to incorporate local and global proximity. The final CpG embeddings are then grouped by chromosome and sorted, with a certain proportion being masked. CpGPT, using just the transformed sequence embedding, can reconstruct the whole methylome by querying the sample embedding. (b) CpGPT can perform a series of downstream tasks, including zero-shot imputation, array conversion, reference mapping, and age and mortality predictions.

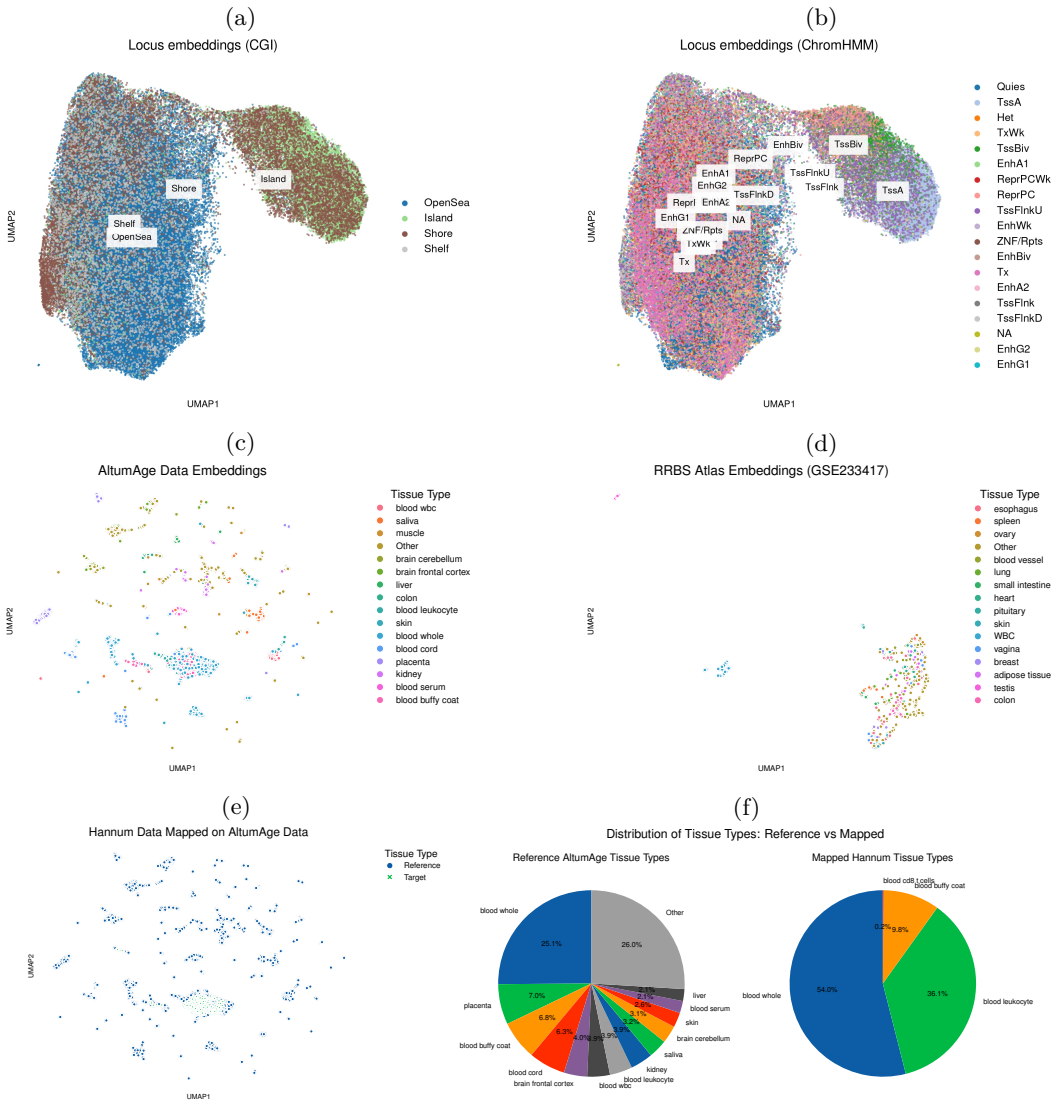


Figure 2: CpGPT CpG and sample embeddings contain meaningful information. UMAP plots of CpG locus embeddings colored by (a) CpG island annotation and (b) ChromHMM annotation. Labels are shown in the centroid of the points for that particular annotation. UMAP plots of the sample embeddings for (a) the AltumAge dataset and (b) reduced representation bisulfite sequencing atlas (GSE233417). The top 15 most represented tissue types are shown with different colors, and the other labels in the data are aggregated under "Other". (e) Hannum data (GSE40279) mapped onto the AltumAge reference data. (f) Pie plots showing the tissue distribution of the reference data vs the queried data.

nuanced representations of DNA methylation patterns across diverse genomic contexts and biological conditions, allowing it to perform a series of downstream tasks (Figure 1B).

2.2 CpGPT learns meaningful methylation and sample embeddings

An essential measure of a foundation model's understanding of a specific domain is its ability to learn meaningful embeddings in an unsupervised manner. If the model effectively captures the underlying structure of the data, it should naturally cluster features and samples with similar attributes, even without explicit labels. Therefore, we generated low-dimensional

representations of the embeddings using Uniform Manifold Approximation and Projection (UMAP) [24] at both the feature (CpG site) and sample levels.

CpGPT processes transformed embeddings of the nucleotide sequences surrounding each CpG site We hypothesized that the model's locus embeddings would reflect functional genomic annotations. To test this, we focused on approximately one million CpG sites present on the Illumina EPICv2 array. We extracted the locus embeddings from CpGPT from 10% of the data and applied UMAP to reduce the high-dimensional embeddings to two dimensions for visualization.

Regarding CpG island annotations, CpG islands are genomic regions with high CpG density, often located near gene promoters and associated with gene regulation. Surrounding these islands are shores (regions up to 2 kilobases from CpG islands), shelves (regions 2–4 kb from CpG islands), and open sea regions (CpG sites located outside of islands, shores, and shelves, typically isolated across the genome). These annotations indicate different regulatory potentials and methylation patterns. In the UMAP visualization colored by CpG island annotations (Figure 2A), we observed distinct clustering of CpG sites according to their island status. CpG sites within islands formed a separate, cohesive cluster, clearly segregated from those in shores, shelves, and open sea regions. Shores appeared in an intermediate position between islands and shelves/open sea, reflecting their transitional genomic context and intermediate methylation levels. This segregation suggests that CpGPT's locus embeddings inherently capture the genomic features and regulatory significance associated with different CpG island regions, even though the model was not explicitly trained on these annotations.

Similarly, when examining chromatin state annotations from ChromHMM [25], which segments the genome into different chromatin states based on combinations of histone modifications, distinct patterns emerged. Chromatin states include active transcription start sites (TssA), associated with transcription initiation; bivalent promoters (TssBiv), which have both activating and repressive marks and are often in developmentally regulated genes; enhancers (Enh), which enhance transcription of distant genes; and repressed or inactive states associated with gene silencing. When we colored the UMAP plot by ChromHMM chromatin states (Figure 2B), CpG sites associated with active transcription start sites formed a separate cluster, isolated from other chromatin states. Bivalent promoters and bivalent enhancers were positioned between the TssA cluster and other states, reflecting their intermediate functional status between active and repressed chromatin. The ability of CpGPT's locus embeddings to segregate CpG sites based on chromatin states underscores the model's capacity to capture complex epigenetic information purely from sequence context. Despite ChromHMM being heavily biased towards embryonic and developmental tissues, this indicates that the model has learned functionally relevant representations of chromatin states.

To evaluate whether CpGPT's sample embeddings capture meaningful biological variation, we analyzed embeddings from three publicly available datasets. The first is a multi-tissue DNA methylation dataset used to develop the AltumAge clock [16], which includes DNA methylation profiles from various human tissues. A portion of this dataset overlaps with *CpGCorpus*, allowing us to assess the model's representation of seen data. The second dataset is a Reduced Representation Bisulfite Sequencing (RRBS) atlas (GSE233417) [26], comprising RRBS data from multiple human cell types and providing a diverse set of methylation profiles across different lineages. This dataset was not included in CpGCorpus, serving as an independent test of the model's generalization. The third dataset is a blood-based methylation cohort (GSE40279) [9], containing blood methylation profiles from individuals of varying ages and used to develop the Hannum epigenetic clock. This dataset was also not part of *CpGCorpus*.

For each dataset, we extracted the sample embeddings from CpGPT and reduced them to two dimensions using UMAP. In the AltumAge dataset (Figure 2C), the sample embeddings formed distinct clusters corresponding to different tissues such as placenta, brain, and blood. This indicates that CpGPT captures tissue-specific methylation signatures in its sample embeddings. In the RRBS atlas (Figure 2D), samples clustered according to cell type, with a clear separation between divergent lineages such as testis, white blood cells, and the

pituitary gland. The embeddings reflect the underlying cellular identity and differentiation status. These observations demonstrate that CpGPT's sample embeddings encode global methylation patterns characteristic of specific tissues and cell types, despite the absence of explicit labels during training.

An important application of learning meaningful sample embeddings is the capacity for zero-shot reference mapping, which enables the transfer of labels from a well-annotated reference dataset to an unlabeled target dataset without the need for additional model training. To evaluate this capability, we utilized CpGPT to project the sample embeddings of the Hannum dataset onto those of the AltumAge dataset, which contains comprehensive tissue-specific annotations (Figure 2E). Our objective was to assess the extent to which the blood-based samples from the Hannum cohort could be accurately classified using the tissue categories defined in the AltumAge atlas. The mapping results were significant: out of the 656 samples in the Hannum dataset, 354 were classified as "blood whole," 237 as "blood leukocyte", 64 as "blood buffy coat", and 1 as "blood cd8 t cells" (Figure 2F). The successful classification of these samples without explicit retraining indicates that CpGPT's sample embeddings effectively capture subtle epigenetic differences associated with specific blood cell types.

2.3 CpGPT performs zero-shot imputation and array conversion

During its extensive pretraining, CpGPT implicitly learns the relationships between different methylation sites, effectively compressing the entire methylome into a compact numerical representation, i.e., the sample embedding. This embedding can be queried with any locus embedding to predict the beta value at that site, analogous to how language models predict words in natural language processing. To practically assess the utility of CpGPT in imputing methylation data, we evaluated its ability to augment the latest Illumina MSA array [27] with probes present in earlier platforms used for training various epigenetic clocks.

Due to the absence of paired datasets measured with both the MSA array and previous Illumina methylation arrays, we utilized the Hannum dataset [9], which comprises blood-based methylation profiles obtained using the Illumina 450k array. From the original set of 473,034 probes on the 450k array, we retained only the 113,585 probes that are also present on the MSA array. This approach simulates a scenario where methylation data are measured using the MSA array, and the goal is to reconstruct the beta values for the remaining probes from the 450k array that are commonly used in various epigenetic clocks.

Despite being provided with only a subset of the methylation data, CpGPT was able to accurately reconstruct the beta values for the missing probes (Figure 3A). Across all 656 samples in the Hannum dataset, the median Pearson correlation coefficient between the imputed and actual beta values was 0.990, the median Spearman correlation was 0.964, and the median mean absolute error per sample was 0.028 (Figure 3B). These results demonstrate that CpGPT can effectively reconstruct comprehensive methylation patterns from partial data. In contrast to the R package *mLiftOver* [28], which is limited to specific tissues, CpGPT's pretraining on the multi-tissue *CpGCorpus* enables it to perform tissue-agnostic imputation.

To evaluate the practical impact of CpGPT's array conversion on downstream analyses, we compared the performance of several epigenetic clocks calculated using the CpGPT-imputed data to those calculated using the full set of 450k probes (considered the ground truth) utilizing `pyaging` [29]. For several recently developed epigenetic clocks, such as *encen40* [30], *intrinclock* [31], and *stemtoc* [32], multiple probes required for their calculations are absent from the MSA array. Our results indicate that the clock estimates derived from the CpGPT-imputed data closely align with those obtained from the complete 450k data (Figure 3C). The original version of the *intrinclock* demonstrates a Pearson correlation of 0.923 with age, whereas the CpGPT-enhanced version of the same *intrinclock* achieves an improved Pearson correlation of 0.987 (Figure 3c). Moreover, the normalized mean absolute error (scaled by the range of the clock predictions) was consistently lower across various clocks when using CpGPT's imputed values (Figure 3D). These findings suggest that CpGPT's extensive pretraining across diverse conditions positions it as a robust tool for array imputation across a wide range of samples.

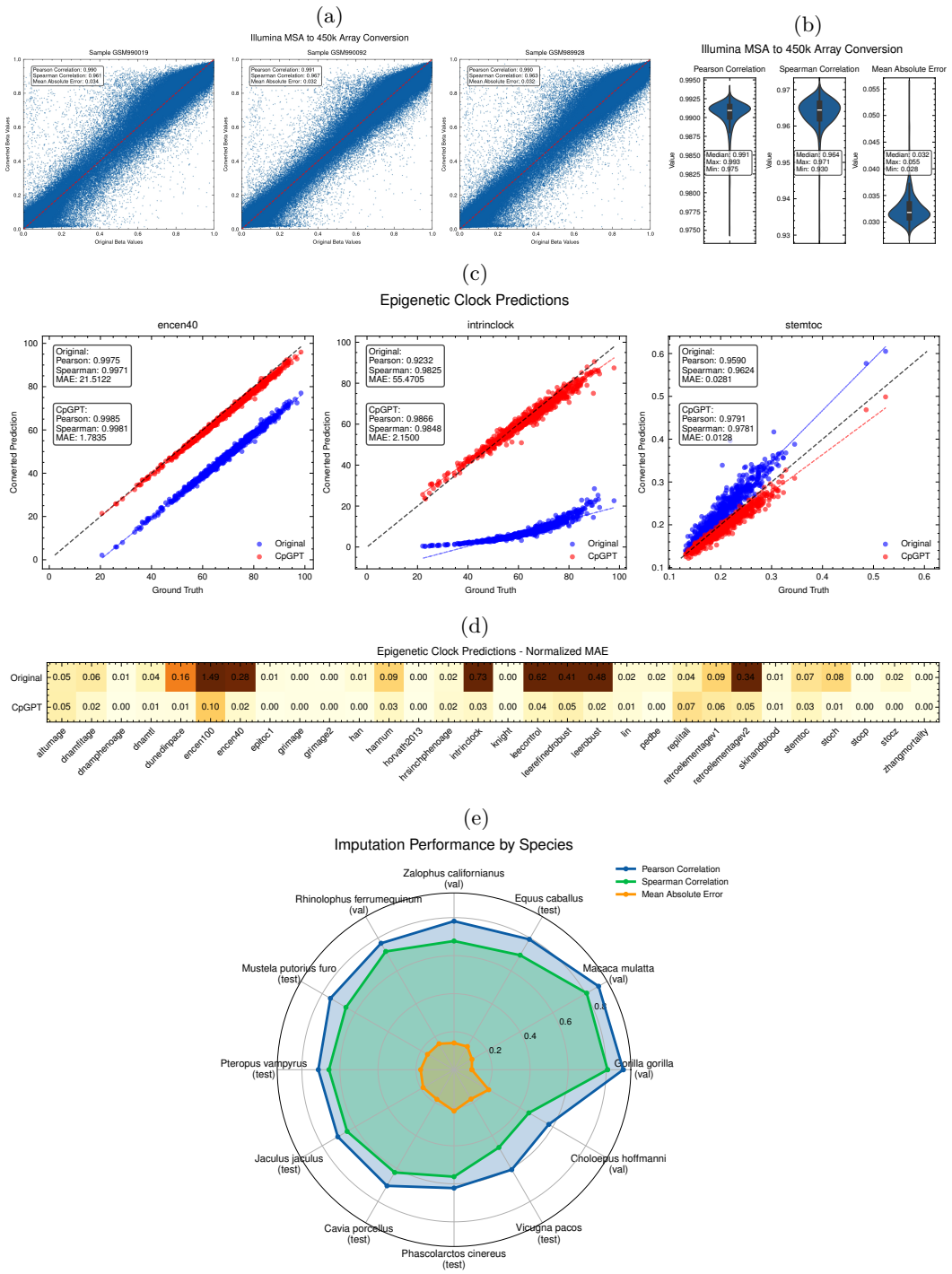


Figure 3: CpGPT effectively imputes unseen methylation sites. (a) Scatter plot showing the relationship between CpGPT-imputed and ground truth beta values for the 450k probes that are missing in the MSA array for three random samples of the Hannum dataset (GSE40279). (b) Violin plots showing CpGPT's Pearson correlation, Spearman correlation, and average mean absolute error (MAE) for the imputation of unseen probes across the whole Hannum dataset. (c) Scatter plots of three epigenetic clocks predicted with either the reconstructed CpGPT 450k or with only the probes present in the Illumina MSA array. (d) Heatmap showing the normalized MAE (scaled by the maximum and minimum predictions for te specific biomarker) of the ground truth predictions against the CpGPT reconstructed array or the original data with only the MSA probes. (e) Imputation of half of the Horvath mammalian array for 12 unseen species with a finetuned CpGPT on multiple mammalian species.

To evaluate the generalization capabilities of CpGPT across mammalian species, we fine-tuned the main model using data from multiple species [33, 34]. From the 55 species with available genomic annotation of the Horvath mammalian array obtained from [35], we divided the dataset into training (n=43), validation (n=5), and testing (n=7) sets. Subsequently, we masked half of the beta values per species and attempted to reconstruct them with CpGPT. Remarkably, for some unseen species, CpGPT achieved Pearson correlation coefficients approaching 0.9 (Figure **??**E). Species phylogenetically closer to humans, such as Gorilla gorilla and Macaca mulatta, exhibited the best performance, likely reflecting the extensive pretraining of CpGPT on human data. These results indicate that the CpGPT framework effectively learns the intrincacies of the epigenetic network across diverse mammalian species.

2.4 CpGPT enables sample-specific interpretation of the methylome

While achieving high performance across multiple tasks is crucial for foundation models, their interpretability is equally important. The attention mechanism inherent in the transformer architecture allows for dynamic weighting of features based on the context of a specific sample. This mechanism enables us to assign importance scores to individual methylation sites, offering insights into which CpG sites are most influential in a given sample's methylation profile.

To obtain CpGPT's importance scores, we analyzed the mean attention weights from the last transformer block. This results in a CpG site-by-CpG site matrix that reflects how much "attention" each site pays to every other site within the context of a particular sample. This matrix serves as a proxy for the influence or importance of each CpG site within the sample's epigenetic network. We hypothesized that CpGPT would be capable of identifying the most relevant methylation sites integral to tissue-specific epigenetic regulation. To test this, we examined individual samples from the multi-tissue AltumAge dataset [16]. As a case study, we constructed graphs from the attention scores, retaining only the edges corresponding to the top 0.01% of attention weights to focus on the most significant interactions. We analyzed samples from heart tissue and brain tissue to assess whether the model highlights biologically relevant genes.

In the heart tissue sample, the CpG site with the highest number of strong connections (edges) was located at position 2,572,626 on chromosome 11, within the *KCNQ1* gene (Figure 4A). *KCNQ1* encodes a voltage-gated potassium channel alpha subunit critical for cardiac repolarization during the cardiac action potential [36, 37]. Mutations in *KCNQ1* are known to cause Long QT Syndrome Type 1 (LQT1), a disorder affecting the heart's electrical activity and potentially leading to arrhythmias and sudden cardiac death. The prominence of *KCNQ1* in the attention graph suggests that CpGPT recognizes its importance in heart tissue, highlighting its role in cardiac electrophysiology and function. Similarly, in the brain tissue sample, the CpG site with the most significant number of strong connections was located at position 37,839,666 on chromosome 5, within the *GDNF* gene (Figure 4B). *GDNF* (Glial cell line-derived neurotrophic factor) is a potent neurotrophic factor that promotes the survival and differentiation of various neuronal subpopulations [38]. It plays a crucial role in neurodevelopment, neuroprotection, and synaptic plasticity. The identification of *GDNF* as a key node in the brain sample's attention graph indicates that CpGPT captures critical elements of neural epigenetic regulation.

These findings demonstrate that CpGPT's attention mechanisms can provide sample-specific interpretations of the methylome, pinpointing genes highly relevant to the tissue type being analyzed. In contrast to traditional linear model epigenetic clocks, which assign fixed importance weights to CpG sites based on their overall contribution across all samples, CpGPT's attention-based approach allows for dynamic, context-dependent importance scoring. This means that the significance of a particular CpG site can vary between samples, reflecting the sample-specific epigenetic landscape.

2.5 CpGPT strongly predicts mortality

While CpGPT demonstrates strong zero-shot performance, we sought to evaluate its effectiveness when fine-tuned for a specific objective, namely mortality prediction. To this

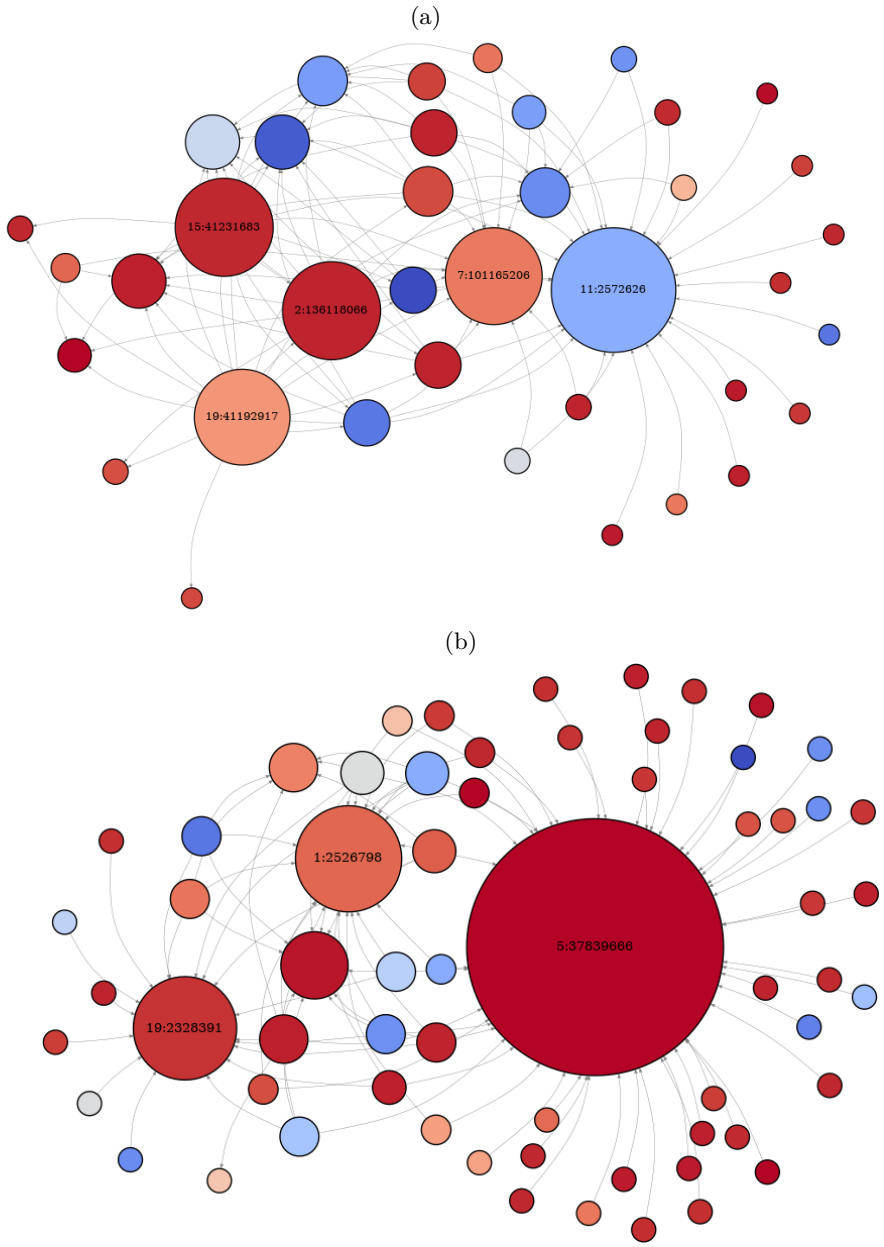


Figure 4: CpGPT can effectively understand the epigenetic network on a per-sample basis. Networks showing the connections between different genomic loci (nodes) showing the top 0.01% of attention weights (edges) for a heart sample (a) and brain sample (b). Blue represents a high beta value (hypermethylation) and red represents a low beta value (hypomethylation).

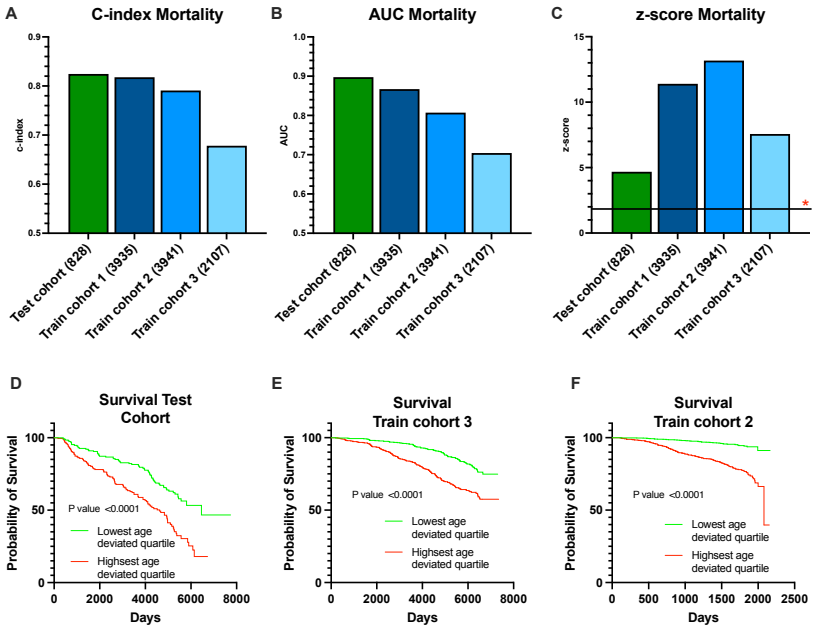


Figure 5: Predictive performance of CpGPT for mortality across multiple cohorts. (A) C-index values for mortality prediction in three training cohorts (n = 3,935; n = 3,941; n = 2,107) and one test cohort (n = 828) range from 0.68 to 0.82, indicating strong model accuracy in predicting survival. (B) AUC values for mortality prediction across the same cohorts range from 0.70 to 0.90, further confirming the model's classification performance. (C) Z-scores for mortality prediction, representing the strength of association between CpGPT scores and mortality, vary across cohorts, with higher scores reflecting stronger associations. (D-F) Kaplan-Meier survival curves comparing the most age-decelerated (lowest quartile) and most age-accelerated (highest quartile) groups in the test cohort (D) and training cohorts 3 (E) and 2 (F). Significant differences in survival probabilities ($p < 0.0001$) were observed between the two groups in all three cohorts, demonstrating CpGPT's ability to stratify individuals based on survival risk.

end, we finetuned a smaller version of CpGPT using mortality data from three independent training cohorts, one validation cohort, and one test cohort, employing a modified Cox proportional hazards loss function (see Methods).

We assessed the model's ability to predict mortality by performing association analyses with time-to-mortality in the three training cohorts (Cohort 1: N = 3,935, deaths n = 319; Cohort 2: N = 3,941, n = 443; Cohort 3: N = 2,107, n = 563) and one test cohort (N = 828, n = 333). Cox proportional hazards models and receiver operating characteristic (ROC) analyses were adjusted for age, and we evaluated three key metrics: the concordance index (C-index) for the Cox model, the z-score of the predictor coefficient, and the area under the ROC curve (AUC) for mortality (Figures 5A–C). In the training cohorts, the C-index ranged from 0.68 to 0.82 (Cohort 1: 0.82; Cohort 2: 0.80; Cohort 3: 0.68), indicating good predictive performance, and the test cohort achieved a C-index of 0.82 (Figure 5A). Similarly, AUC values for mortality prediction in the training cohorts ranged from 0.70 to 0.87 (Cohort 1: 0.87; Cohort 2: 0.81; Cohort 3: 0.70), with the test cohort showing an AUC of 0.90 (Figure 5B). The z-scores of the predictor coefficients, influenced by sample size and effect magnitude, varied in the training cohorts from 7.57 to 13.17 (Cohort 1: 11.40; Cohort 2: 13.17; Cohort 3: 7.57), while the test cohort had a z-score of 4.68, reflecting its smaller size (Figure 5C).

Given the comparatively weaker association in Training Cohort 3, we further analyzed the model's ability to differentiate between individuals with high and low survival probabilities. We stratified participants into quartiles based on age-residualized CpGPT risk scores and

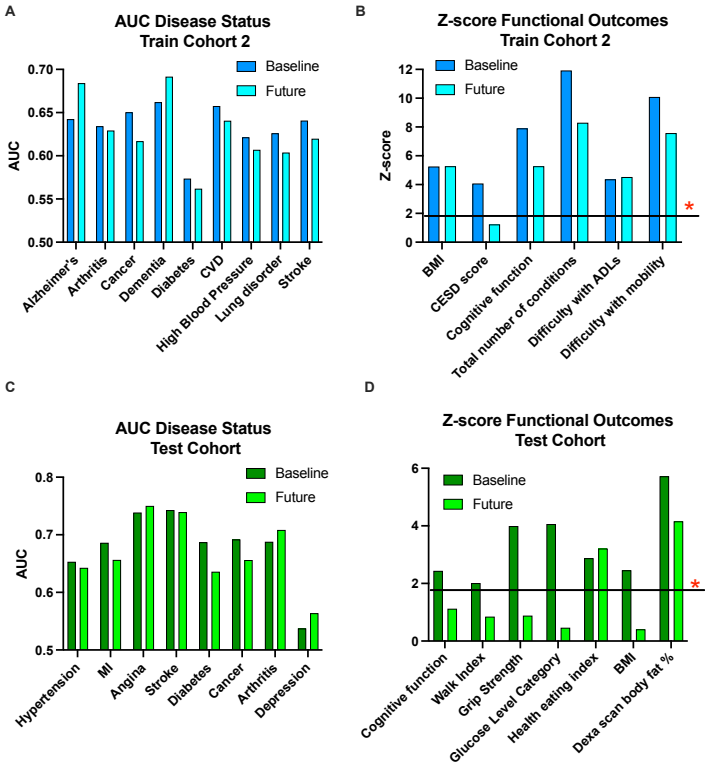


Figure 6: CpGPT predictive performance for disease status and functional outcomes in the test and training cohorts. (A) AUC values for disease status in Training Cohort 2 at baseline and follow-up demonstrate CpGPT's ability to predict various diseases like dementia, Alzheimer's, cancer, and cardiovascular disorders. (B) Z-scores for functional outcomes in Training Cohort 2, such as total number of conditions and difficulty with mobility, reflect strong associations at baseline and future time points (C) AUC values for disease status prediction in the test cohort at baseline and future time points show strong performance across conditions such as hypertension, stroke, diabetes, and cancer (D) Z-scores for functional outcomes in the test cohort, including cognitive function, BMI, and DEXA scan body fat percentage, highlight the model's predictive accuracy at both baseline and future assessments.

compared survival curves between the most age-decelerated group (lowest quartile) and the most age-accelerated group (highest quartile). In the test cohort, the median survival was significantly different between these groups, with the most age-decelerated individuals having a median survival of 6,462 days, compared to 4,625 days in the most age-accelerated group (Figure 5D). In Training Cohort 3, where median survival could not be calculated due to insufficient events within the follow-up period, we calculated the restricted mean survival time (RMST). The RMST was 6,677 days for the most age-decelerated group and 5,886 days for the most age-accelerated group (Figure 5E). In both cohorts, the survival curves between the most age-decelerated and most age-accelerated groups were significantly different ($p < 0.0001$). Additionally, in Training Cohort 2, we observed a significant difference in survival curves between the most and least age-accelerated groups ($p < 0.0001$). These results demonstrate that CpGPT robustly predicts mortality across different cohorts, effectively stratifying individuals based on their biological aging profiles.

2.6 CpGPT strongly predicts morbidity

We investigated the association of CpGPT with various morbidity outcomes to assess its potential utility in predicting disease onset and functional decline. In Training Cohort 2, we performed ROC analyses on nine diseases at baseline and at a four-year follow-up, including

Alzheimer's disease, arthritis, cancer, dementia, diabetes, cardiovascular disease (CVD), high blood pressure, lung disorders, and stroke (Figure 6A). At baseline, the AUC values ranged from 0.57 to 0.66, with dementia and CVD exhibiting the highest AUCs of 0.66, indicating moderate discriminative ability, while diabetes had the lowest AUC of 0.57. At the four-year follow-up, the AUCs ranged from 0.56 to 0.70, with dementia again showing the highest AUC of 0.70. These results suggest that CpGPT has predictive capability for these diseases, particularly for neurodegenerative conditions like dementia.

We also assessed functional outcomes in Training Cohort 2, specifically body mass index (BMI), Center for Epidemiologic Studies Depression Scale (CESD) score, cognitive function, total number of chronic conditions, difficulty with activities of daily living (ADLs), and difficulty with mobility. Linear regression models yielded z-scores ranging from 4.07 to 11.93 at baseline, with the strongest associations observed for total number of chronic conditions ($z = 11.93$) and difficulty with mobility ($z = 10.08$). At the four-year follow-up, the associations remained significant but were somewhat attenuated, with z-scores ranging from 4.52 to 8.29, and the strongest associations again being total number of chronic conditions ($z = 8.29$) and difficulty with mobility ($z = 7.58$) (Figure 6B). These findings indicate that higher CpGPT scores are associated with greater morbidity burden and functional impairment, both at baseline and prospectively.

In the test cohort, we conducted ROC analyses on eight conditions at baseline and at a two-year follow-up, including hypertension, myocardial infarction (MI), angina, stroke, diabetes, cancer, arthritis, and depression (Figure 6C). At baseline, the AUC values ranged from 0.54 to 0.74, with angina and stroke achieving the highest AUCs of 0.74, suggesting good discriminative ability, while depression had the lowest AUC of 0.54. At the two-year follow-up, the AUCs ranged from 0.56 to 0.75, with angina again showing the highest AUC of 0.75. We further examined functional outcomes in the test cohort, including cognitive function, walk index (a measure of gait speed and mobility), grip strength, glucose level category, healthy eating index, BMI, and dual-energy X-ray absorptiometry (DEXA) scan body fat percentage (Figure 6D). At baseline, the z-scores ranged from 2.01 to 5.72, with DEXA scan body fat percentage showing the strongest association ($z = 5.72$), indicating that higher CpGPT scores correlate with increased adiposity. At the two-year follow-up, only two significant associations remained: healthy eating index ($z = 3.21$) and DEXA scan body fat percentage ($z = 4.15$). These results suggest that CpGPT may reflect aspects of metabolic health and lifestyle factors over time.

Overall, these findings demonstrate that CpGPT predicts a range of diseases and functional morbidity outcomes across diverse cohorts. The model's associations with neurodegenerative diseases, cardiovascular conditions, and measures of physical function highlight its potential utility in identifying individuals at risk for significant health declines, reinforcing its applicability in clinical and research settings.

3 Discussion

In this study, we introduce CpGPT, a foundation model specifically designed for DNA methylation analysis. By leveraging the comprehensive dataset *CpGCorpus* comprising over 100,000 samples from diverse tissues and conditions, CpGPT captures the intricate relationships inherent in the methylome. Our model effectively integrates sequence context, positional information, and epigenetic state to learn rich, meaningful embeddings at both the CpG site and sample levels. This multifaceted approach allows CpGPT to excel in various tasks, including imputation of missing methylation values, array conversion, zero-shot reference mapping, and age and mortality prediction.

The development of CpGPT addresses a significant gap in DNA methylation research. Traditional epigenetic clocks and methylation analyses often rely on linear models that may not fully capture the complex dependencies among CpG sites [39, 16]. Although newer two-layer explainable perceptron models for DNA methylation to predict mortality have been developed, they still fail to capture the full complexity of the aging methylome [40, 41]. By adopting a transformer architecture [1], CpGPT leverages self-attention mechanisms to model long-range interactions and contextual relationships within the methylome.

Our findings demonstrate that the model's feature embeddings reflect functional genomic annotations, such as CpG islands and chromatin states, without explicit supervision. This indicates that CpGPT has internalized biologically relevant patterns purely from the data, highlighting the power of unsupervised deep learning in uncovering epigenetic information.

One of the notable strengths of CpGPT is its ability to generalize across different datasets and conditions. The successful zero-shot reference mapping of the Hannum [9] samples onto the AltumAge [16] tissue annotations demonstrates that CpGPT's sample embeddings capture sufficient biological information to enable accurate classification without additional training. CpGPT's capacity to perform zero-shot imputation and array conversion has significant practical implications. With the introduction of different methylation array platforms over time, researchers often face challenges in integrating data from various sources or utilizing epigenetic clocks developed on older platforms [39]. Our evaluation using the Hannum dataset simulated the scenario of reconstructing 450k array probes from MSA array data [27]. CpGPT demonstrated high accuracy in imputing missing beta values, enabling the application of multiple epigenetic clocks that would otherwise be incompatible with the newer array.

Interpretability remains a crucial aspect of deploying machine learning models in biomedical research. CpGPT addresses this by utilizing the attention mechanism to provide sample-specific importance scores for CpG sites. Our analysis of attention weights revealed that the model highlights genes relevant to specific tissues, such as *KCNQ1* in heart tissue [36, 37] and *GDNF* in brain tissue [38]. This level of interpretability offers valuable insights into the biological underpinnings of methylation patterns and can guide hypothesis generation and experimental validation. By contrast, traditional linear models assign static importance weights that do not account for sample-specific variations, limiting their interpretative power.

The superior performance of CpGPT in predicting mortality and chronological age, as evidenced by its top placements in the Biomarkers of Aging Challenge [18], signifies its potential impact on aging research. Additionally, in our evaluation across three training cohorts and one test cohort, CpGPT robustly predicted mortality with C-index values ranging from 0.68 to 0.82 in training cohorts and 0.82 in the test cohort, indicating strong concordance between predicted and observed survival times. AUC values were equally impressive, ranging from 0.70 to 0.87 in training cohorts and reaching 0.90 in the test cohort, showcasing high predictive accuracy across different populations. Despite a weaker association in Training Cohort 3, CpGPT effectively differentiated between individuals with high and low survival probabilities when cohorts were stratified into quartiles based on age-residualized CpGPT scores. Significant differences in survival curves were observed: in the test cohort, the most age-decelerated group had a median lifespan of 6,462 days compared to 4,625 days in the most age-accelerated group. In Training Cohort 3, restricted mean survival times [42] differed significantly between the decelerated (6,677 days) and accelerated quartiles (5,886 days). These findings demonstrate the model's ability to capture biologically meaningful variations in aging and mortality risk.

CpGPT has shown strong predictive capabilities across a variety of morbidity outcomes. The model predicted a range of diseases and functional morbidity measures in different cohorts. In Training Cohort 2, ROC analysis revealed that CpGPT could predict nine diseases at both baseline and four-year follow-up, while in the test cohort, it could predict eight diseases. Dementia and cardiovascular diseases had the highest AUCs at baseline (0.66) and follow-up (0.70) in the training cohort, highlighting the model's potential utility in neurological and cardiovascular aging research [43, 44]. A similar trend was observed in the test cohort, with baseline AUCs reaching up to 0.74 for conditions like angina and stroke, further emphasizing the applicability of the model and epigenetics in cardiovascular and neurovascular aging research [45, 46]. Functional outcomes such as the total number of conditions and difficulty with mobility showed strong associations with CpGPT scores in the training cohort, with baseline z-scores of 11.93 and 10.08, respectively. These associations remained significant at the four-year follow-up, emphasizing the model's relevance in assessing overall health and functional status—two critical metrics for measuring frailty in humans [47]. In the test cohort, functional measures like DEXA scan body fat percentage [48] and healthy eating index [48, 49] were significantly associated with CpGPT scores, suggesting the model's

applicability in metabolic aging and lifestyle-related health assessments [50]. Moreover, in both cohorts, we observed strong predictions of baseline and future cancer disease status using the model, underscoring the relationship between cancer, epigenetics, and aging [51]. Another notable finding was the association with psychosocial measures and disorders, such as depression in the test cohort and CES-D scores [52] in the training cohort, potentially indicating a link between aging, depression, and mental health deterioration [53].

These results not only validate CpGPT's predictive capabilities but also highlight its ability to generalize across different datasets and conditions, with the potential to build disease-specific risk scores using epigenetics [54]. The consistent performance in mortality and morbidity prediction across multiple cohorts underscores the robustness of the model and its potential for broad applications in biomedical research. By capturing complex epigenetic signatures associated with aging and disease states, CpGPT contributes to the development of more precise biomarkers and enhances the predictive power of epigenetic clocks. While the application of transformers to aging research has precedent [55, 56, 57], up to until now it has not been shown to achieve state-of-the-art results.

CpGPT establishes a new benchmark in DNA methylation analysis by combining deep learning with comprehensive epigenetic data. Its ability to learn meaningful, biologically relevant embeddings and perform a variety of tasks without explicit supervision highlights the potential of foundational models in epigenetics. By bridging gaps between different array platforms, enabling sample-specific interpretations, and demonstrating strong predictive capabilities, CpGPT contributes valuable tools and insights to the field of aging and epigenetics. CpGPT not only enhances our ability to predict aging-related outcomes but also opens new avenues for exploring the epigenetic mechanisms underlying human health and disease.

4 Methods

4.1 Data curation

4.1.1 CpGCorpus

To train an effective foundation model capable of generalizing across diverse biological contexts, we curated a comprehensive DNA methylation dataset named *CpGCorpus*. This dataset aggregates publicly available DNA methylation data from the Gene Expression Omnibus (GEO), encompassing a total of 1,502 studies and 106,795 human samples. These samples were measured using various Illumina methylation array platforms, including the 27k, 450k, EPIC, EPIC+, and EPICv2 arrays, providing a wide spectrum of coverage across the human methylome. The collected data represent a rich diversity of tissue types, developmental stages, disease conditions, and demographic backgrounds. This diversity is crucial for training a foundation model that can capture the complex patterns and variations in DNA methylation across different biological states.

To ensure consistency and quality across the dataset, we performed the following:

- **Processing of raw data:** For datasets where raw IDAT files were available, we utilized the R package *SeSAMe* [58] for data processing. *SeSAMe* offers advanced normalization and preprocessing algorithms tailored for Illumina methylation arrays, including background correction, dye bias correction, detection p-value computation, and beta value estimation.
- **Handling of processed data:** For datasets lacking raw IDAT files, we employed the normalized beta value matrices provided by the original studies. These beta values were assumed to be preprocessed according to the methodologies described in their respective publications.
- **Quality control:** We implemented quality control measures to identify and exclude poor-quality samples and probes (*SeSAMe* argument `prep=``QCDPB"`). This included filtering out probes with detection p-values above a threshold, probes targeting single-nucleotide polymorphisms (SNPs), and those known to cross-hybridize.

For data that was already processed, we excluded datasets from which beta values were outside of the expected 0 to 1 range.

- **Probe harmonization:** To address differences in probe sets across array platforms, we mapped probes to a common set of CpG sites based on their genomic coordinates (*SeSAMe* argument `collapseToPfx=TRUE`). This allowed us to integrate data from different arrays and ensured that the model learned from a consistent set of features.

The final *CpGCorpus* dataset provides a harmonized and high-quality resource for model training.

Data splitting To assess the performance and generalization capabilities of CpGPT, we partitioned *CpGCorpus* into training, validation, and test sets:

- **Training set:** Consists of 100,965 samples from 1,443 studies (median of 25 samples per study). This set was used for model training.
- **Validation set:** Includes 489 samples from 10 studies (median of 46 samples per study). This set was used for hyperparameter tuning and model selection.
- **Test set:** Contains 5,341 samples from 49 studies (median of 24 samples per study). This set was held out during training and validation and used exclusively for the final model evaluation.

We ensured that there was no overlap of samples or studies between the splits to prevent data leakage. A detailed list of GEO Series (GSE) entries included in each split is provided in **Supplementary Table 1**.

4.2 Model architecture

The CpGPT model is designed to integrate sequence, positional, and epigenetic information to effectively learn the complex relationships inherent in DNA methylation data. Below, we detail each component of the model architecture and the associated preprocessing steps.

4.2.1 Input representation and preprocessing

The model input comprises DNA sequence embeddings, methylation beta values, and positional information for each CpG site. Specifically:

- **DNA sequence embeddings** ($\mathbf{E}$): For each CpG site, we extracted a nucleotide sequence of length L centered on the target cytosine. These sequences were embedded into numerical representations using a pretrained DNA language model.
- **Methylation beta values** (β): The methylation level at each CpG site, represented as a beta value ranging from 0 (unmethylated) to 1 (fully methylated).
- **Positional information** ($\mathbf{p}$): The genomic coordinates and chromosome indices for each CpG site.

To optimize the utilization of genomic context and mitigate potential biases, we applied the following preprocessing steps:

1. **Intra-chromosomal sorting:** Within each chromosome, CpG sites were sorted in ascending order based on their genomic coordinates. This preserves local genomic context and facilitates the modeling of spatial dependencies between neighboring CpG sites.
2. **Chromosome grouping:** CpG sites were grouped by chromosome to maintain proximate loci together.
3. **Stochastic chromosome shuffling:** The order of chromosomes was randomly shuffled for each input batch during training. This prevents the model from developing positional biases tied to specific chromosome orders and promotes generalization across the genome.

This approach allows the model to capture both local (within chromosome) and global (across chromosomes) genomic contexts.

4.2.2 Sequence encoding

To encode the local genomic context of each CpG site, we employed a pretrained DNA language model to generate sequence embeddings:

- **Sequence Eextraction:** For each CpG site, a nucleotide sequence of length L centered on the CpG site was extracted.
- **Embedding generation:** These sequences were input into a DNA LLM [19, 20, 21], yielding embeddings $\mathbf{E} \in \mathbb{R}^{b \times l \times d_{\text{dna}}}$, where b is the batch size, l is the sequence length, and d_{dna} is the dimension of the DNA embeddings.
- **Embedding transformation:** A trainable multi-layer perceptron (MLP) f_{seq} projected these embeddings into the model's internal embedding space:

$$\mathbf{S} = f_{\text{seq}}(\mathbf{E}), \tag{1}$$

  where $\mathbf{S} \in \mathbb{R}^{b \times l \times d_{\text{emb}}}$ and d_{emb} is the model embedding dimension.

This process captures the local sequence context, enabling the model to learn patterns associated with sequence motifs and methylation patterns.

4.2.3 Dual positional encoding strategy

Capturing positional information at multiple genomic scales is essential for modeling DNA methylation patterns. We employed a dual positional encoding strategy:

Absolute positional encoding We adapted the sinusoidal positional encoding scheme from Vaswani et al. [1] to suit genomic distances:

$$\mathbf{t} = \frac{\mathbf{p}}{p_{\max}}, \tag{2}$$

$$\mathbf{w} = 2\pi \frac{\mathbf{p}}{p_{\max}}, \tag{3}$$

$$f_i = 10^{-4} + i\left(\frac{d_{\text{emb}}}{2} - 1 - 10^{-4}\right) \Big/ \left(\frac{d_{\text{emb}}}{2} - 1\right), \quad i = 0, 1, \dots, \frac{d_{\text{emb}}}{2} - 1, \tag{4}$$

$$\text{PE}(\mathbf{p}) = \left[\mathbf{t}, \mathbf{t}, \cos(f_0\mathbf{w}), \sin(f_0\mathbf{w}), \dots, \cos(f_{d_{\text{emb}}/2-1}\mathbf{w}), \sin(f_{d_{\text{emb}}/2-1}\mathbf{w})\right], \tag{5}$$

where $p_{\max}$ approximates the length of the largest human chromosome (i.e., 3×10^8 base pairs). This encoding $\text{PE}(\mathbf{p}) \in \mathbb{R}^{b \times l \times d_{\text{emb}}}$ captures global positional information, which is needed for distinguishing between CpG sites with similar sequences but located in different genomic regions.

Relative positional encoding To capture local positional relationships, we applied Rotary Positional Embeddings (RoPE) [22]:

$$\mathbf{L} = \text{RoPE}(\mathbf{S} + \text{PE}(\mathbf{p})), \tag{6}$$

where $\mathbf{L} \in \mathbb{R}^{b \times l \times d_{\text{emb}}}$ is the locus embedding. RoPE allows the model to incorporate relative positional information efficiently, enhancing its ability to model interactions between nearby CpG sites.

4.2.4 Methylation state encoding

The methylation beta values were embedded to capture the epigenetic state:

- **Beta Value Embedding:** A trainable MLP f_β transformed the beta values into embeddings:
$$\mathbf{B} = f_\beta(\beta), \tag{7}$$
where $\mathbf{B} \in \mathbb{R}^{b \times l \times d_{\text{emb}}}$.
- **CpG Site Embedding:** The final embedding for each CpG site combined the locus embedding and the beta value embedding:
$$\mathbf{C} = \frac{1}{\sqrt{2}}(\mathbf{L} + \mathbf{B}). \tag{8}$$
This normalization ensures that the combined embedding maintains a consistent scale.

4.2.5 Transformer++ architecture

The core of CpGPT is based on the Transformer++ architecture [23], which enhances the original Transformer model [1] with:

1. **SwiGLU activation function:** Swish-Gated Linear Unit (SwiGLU) replaces the standard ReLU activation, providing smoother gradients and improved performance.
2. **RMSNorm pre-normalization:** Root Mean Square Layer Normalization stabilizes training in deep networks.
3. **Bias-free linear layers:** Removing bias terms reduces overfitting and parameter count.

The model consists of N layers of Transformer++ blocks. Each layer computes:

$$\hat{\mathbf{H}}_{i-1} = \text{RMSNorm}(\mathbf{H}_{i-1}), \tag{9}$$
$$\mathbf{A}_i = \mathbf{H}_{i-1} + \text{MultiHeadAttention}(\hat{\mathbf{H}}_{i-1}, \hat{\mathbf{H}}_{i-1}, \hat{\mathbf{H}}_{i-1}), \tag{10}$$
$$\tilde{\mathbf{A}}_i = \text{RMSNorm}(\mathbf{A}_i), \tag{11}$$
$$\mathbf{H}_i = \mathbf{A}_i + \text{SwiGLU}(\text{MLP}(\tilde{\mathbf{A}}_i)), \tag{12}$$

where $\mathbf{H}_0 = [\mathbf{c}_{\text{cls}}; \mathbf{C}]$, and $\mathbf{c}_{\text{cls}} \in \mathbb{R}^{1 \times d_{\text{emb}}}$ is a learnable classification token that aggregates information across the sequence.

4.2.6 Decoders

We designed specialized decoders to extract meaningful outputs from the model:

Beta value decoder Predicts methylation beta values for CpG sites:

$$\hat{\beta} = \sigma\left(\left\langle g_\beta(\mathbf{S}), \mathbf{h}_{\text{cls}} \right\rangle\right), \tag{13}$$

where g_β is a projection function, $\mathbf{h}_{\text{cls}}$ is the final hidden state of the classification token, $\langle \cdot, \cdot \rangle$ denotes the inner product, and σ is the sigmoid activation function.

Uncertainty decoder Estimates the uncertainty of the beta value predictions:

$$\hat{\sigma}\beta = \sigma\left(\langle g_{\text{unc}}(\mathbf{S}), \mathbf{h}_{\text{cls}} \rangle\right), \tag{14}$$

where g_{unc} is a projection function.

Condition decoder Used during fine-tuning for specific downstream tasks:

$$\hat{\mathbf{y}} = \langle g_{\text{cond}}(\mathbf{T}), \mathbf{h}_{\text{cls}} \rangle, \tag{15}$$

where $\mathbf{T}$ is a set of learnable condition tokens, and g_{cond} projects them into the embedding space.

Table 1: Comparison of CpGPT Model Versions

Parameter	**Large**	**Small**
Total Parameters	65M	2.5M
d_embedding	512	128
d_hidden	512	128
n_attention_heads	16	8
n_layers	16	8
dropout	0.01	0.01
n_mlp_blocks	3	2
dna_context_len	2001	2001
max_length	10000	5000

4.3 Model parameters

The task of compressing the entire methylome into an embedding is challenging and requires a large model capacity. However, when finetuning for a specific task using only a subset of CpG sites, not as much flexibility is required. Hence, we used a large model for all zero-shot tasks and a smaller one for fine-tuning (Table 1).

4.4 Pretraining procedure

4.4.1 Training loop

We trained CpGPT using a multi-task learning approach. For each training batch, we performed the following steps:

1. **Data preparation:** A batch of samples $\mathcal{B} = \beta, \mathbf{E}, \mathbf{c}, \mathbf{p}, \mathbf{O}$ was prepared, where $\mathbf{O}$ contains optional observed variables for downstream tasks.
2. **Missing value masking:** We created a mask $\mathbf{M}_{\text{na}}$ to handle missing beta values due to array differences or quality control filtering.
3. **Input-target split:** The beta values were split into input and target sets:

$$\beta_{\text{input}} = \beta[:, : l/2], \tag{16}$$

$$\beta_{\text{target}} = \beta[:, l/2 :], \tag{17}$$

   encouraging the model to reconstruct missing data.
4. **Encoding steps:**
   - Sequence encoding to obtain $\mathbf{S}$.
   - Positional encoding to obtain $\mathbf{L}$.
   - Methylation state encoding to obtain $\mathbf{B}$ and $\mathbf{C}$.
5. **Transformer++ processing:** The combined embeddings were processed through the Transformer++ layers to obtain $\mathbf{H}_N$.
6. **Prediction:**
   - Beta value prediction $\hat{\beta}$.
   - Uncertainty estimation $\hat{\sigma}_\beta$.
   - Condition prediction $\hat{\mathbf{y}}$ (if applicable).

4.4.2 Loss functions

The total loss $\mathcal{L}$ is a weighted sum of several component losses:

$$\mathcal{L} = \sum_i w_i \mathcal{L}_i \tag{18}$$

where w_i are the weights for each loss component. The main loss components are:

Beta losses A mean absolute error loss is used both for the predicted beta values and the estimated error of the prediction itself. In addition, a Wasserstein distance loss, also known as Earth mover's distance, is used to improve the distribution of the predicted beta values when compared to the real distribution.

$$\mathcal{L}_{\text{mae}} = \frac{1}{|\mathbf{M}_{\text{valid}}|} \sum_{i,j \in \mathbf{M}_{\text{valid}}} |\hat{\beta}_{ij} - X_{ij}| \tag{19}$$

$$\mathcal{L}_{\text{mae_unc}} = \frac{1}{|\mathbf{M}_{\text{valid}}|} \sum_{i,j \in \mathbf{M}_{\text{valid}}} \left| |\hat{\beta}_{ij} - X_{ij}| - \hat{\sigma}_{\beta,ij} \right| \tag{20}$$

$$\mathcal{L}_{\text{wd}} = \mathbb{E}[|\text{sort}(\hat{\beta}) - \text{sort}(\mathbf{X})|] \tag{21}$$

where $\mathbf{M}_{\text{valid}} = \{\text{non-NA indices in } \mathbf{X}_{\text{target}}\}$

Sample embedding losses To facilitate meaningful representations of the samples, the Kullback-Leibler divergence is incorporated to constrain the embedding space to be normally distributed with mean zero and variance one. Moreover, a contrastive loss first used by scGPT [3] is employed to separate dissimilar samples and bring together similar ones.

$$\mathcal{L}_{\text{kld_sample}} = \text{KL}(\mathcal{N}(\mu_{\mathbf{h}}, \sigma^2_{\mathbf{h}}) || \mathcal{N}(0, \mathbf{I})) \tag{22}$$

$$\mathcal{L}_{\text{contrastive}} = \mathbb{E}[1 - (\max(0, \cos(\mathbf{h}_{\text{sample},i}, \mathbf{h}_{\text{sample},j}) - \tau))^2] \tag{23}$$

where τ is a threshold parameter, and $\mu_{\mathbf{h}}$, $\sigma^2_{\mathbf{h}}$ are the mean and variance of $\mathbf{h}_{\text{sample}}$ across the batch dimension.

Condition prediction loss (if enabled for finetuning) The condition prediction loss $\mathcal{L}_{\text{cond}}$ is task-specific and depends on the nature of the predicted conditions. Let $\hat{\mathbf{y}} \in \mathbb{R}^{b \times k}$ be the predicted conditions and $\mathbf{O} \in \mathbb{R}^{b \times k}$ be the observed variables, where b is the batch size and k is the number of conditions or features. The general form of the condition prediction loss is:

$$\mathcal{L}_{\text{cond}} = \mathcal{L}_{\text{task}}(\hat{\mathbf{y}}, \mathbf{O}) \tag{24}$$

The specific form of $\mathcal{L}_{\text{task}}$ depends on the prediction task:

1. For regression tasks:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{bk} \sum_{i=1}^{b} \sum_{j=1}^{k} |\hat{y}_{ij} - O_{ij}| \tag{25}$$

$$\mathcal{L}_{\text{MSE}} = \frac{1}{bk} \sum_{i=1}^{b} \sum_{j=1}^{k} (\hat{y}_{ij} - O_{ij})^2 \tag{26}$$

2. For binary classification tasks:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{bk} \sum_{i=1}^{b} \sum_{j=1}^{k} [O_{ij} \log(\sigma(\hat{y}_{ij})) + (1 - O_{ij}) \log(1 - \sigma(\hat{y}_{ij}))] \tag{27}$$

   where σ is the sigmoid function.

3. For survival analysis tasks, we use the Cox Proportional Hazards (CPH) loss:

$$\mathcal{L}_{\text{CPH}} = -\frac{1}{\sum_{i=1}^{b} e_i} \sum_{i:e_i=1} \left(\hat{y}_i - \log \left(\sum_{j:t_j \geq t_i} \exp(\hat{y}_j) \right) \right) \tag{28}$$

   where $\hat{y}_i$ is the predicted risk score for the i-th sample, t_i is the observed time, and e_i is the event indicator (1 if the event occurred, 0 if censored).

4.4.3 Software and hardware

CpGPT was developed and pretrained using *python* version 3.10.15 and *torch* version 2.4.1. A list of other dependencies will be available on our GitHub after publication.

The standard CpGPT model was trained on an NVIDIA H100 GPU for approximately 10 days.

4.5 Finetuning procedure

CpG sites associated with mortality were selected using the following two criteria. First, CpGs were required to have an intra-class correlation coefficient (ICC) exceeding 0.75, ensuring high reproducibility across samples. Second, we applied an absolute z-score threshold, selecting CpGs with a z-score greater than 4, based on their association with mortality adjusted for age and sex in the training cohort 1. These criteria were established to prioritize robust, reproducible CpG markers with strong statistical associations to mortality outcomes. Beyond these CpG sites used in GrimAge2 [11] and DunedinPACE [12] were also used.

The small CpGPT model was then trained with the aforementioned modified Cox proportional hazard loss.

4.6 Mortality association prediction

To evaluate the predictive power of the CpGPT fintuned for mortality, we conducted association analyses using time-to-mortality data across three training cohorts and one test cohort. The training cohorts consisted of Cohort 1 (n = 3,935, n_d = 319), Cohort 2 (n = 3,941, n_d = 443), and Cohort 3 (n = 2,107, n_d = 563), where n represents the total number of individuals and n_d denotes the number of death events observed during the follow-up period. The test cohort included 828 individuals (n_d = 333). We performed Cox proportional hazards regression models and receiver operating characteristic (ROC) analyses, adjusting for chronological age in each cohort. Three key metrics were used to evaluate performance: the concordance index (C-index) from the Cox model, z-scores of the predictor coefficients in the Cox model, and the area under the ROC curve (AUC) for mortality prediction. The Cox proportional hazards analyses were conducted using the `coxph()` function from the *survival* package (version 3.6.4) in R (version 4.4.0), while ROC analyses were performed using the `roc()` function from the *pROC* package (version 1.18.5).

4.7 Survival analysis

To assess the model's ability to differentiate between individuals with high and low survival probabilities, we divided each cohort into quartiles based on age-residualized CpGPT scores. The age-residualized scores were obtained by fitting a linear model of CpGPT scores against chronological age using the `lm()` function in R, and extracting the residuals using the `resid()` function. Survival curves for the most age-decelerated quartile (lowest residuals) and the most age-accelerated quartile (highest residuals) were compared using Kaplan-Meier analysis, performed using GraphPad Prism 9 software. Median survival times were calculated for the test cohort and Training Cohort 2. In Training Cohort 3, due to an insufficient number of events within the follow-up period, we calculated the restricted mean survival time (RMST) using the `survfit()` function from the *survival* package (version 3.6.4) in R. Statistical significance of the differences between survival curves was assessed using the log-rank test, implemented in Prism 9.

4.8 Morbidity association analysis

To evaluate the association of CpGPT scores with morbidity outcomes, we performed ROC analyses for nine disease outcomes in Training Cohort 2 at both baseline and four-year follow-up. The diseases analyzed included Alzheimer's disease, arthritis, cancer, dementia, diabetes, cardiovascular disease (CVD), high blood pressure, lung disorders, and stroke. ROC analyses, adjusted for age, were conducted using the `roc()` function from the *pROC* package (version 1.18.5) in R (version 4.4.0).

In addition to disease outcomes, we assessed associations between CpGPT scores and functional outcomes in Training Cohort 2. The functional measures included body mass index (BMI), Center for Epidemiologic Studies Depression Scale (CESD) score, cognitive function, total number of chronic conditions, difficulty with activities of daily living (ADLs), and difficulty with mobility. Linear regression models adjusted for age were applied to estimate z-scores for these outcomes at baseline and at the four-year follow-up, using the `lm()` function in R.

In the test cohort, we conducted ROC analyses for eight conditions—hypertension, myocardial infarction (MI), angina, stroke, diabetes, cancer, arthritis, and depression—at baseline and at a two-year follow-up, following the same methodology as in Training Cohort 2. Functional outcomes in the test cohort were also evaluated, including cognitive function, walk index (a measure of mobility), grip strength, glucose level category, healthy eating index, BMI, and dual-energy X-ray absorptiometry (DEXA) scan body fat percentage. Associations were assessed using linear regression models adjusted for age, similar to those applied in Training Cohort 2.

All statistical analyses were performed using R (version 4.4.0), and statistical significance was determined at a threshold of $p < 0.05$.

5 Code and data availability

The code will be released upon publication and the finetuned models will be available on the Python package *pyaging* [29] and the R package *methylCYPHER* [59]. If you would like to get early access to CpGPT, please contact L.P.D.L.C. at lucas_camillo@alumni.brown.edu.

6 Acknowledgements

We would like to thank the Zhou lab for their comprehensive, publicly available annotation of the Illumina methylation arrays [35]. We would also like to thank Alexander Mathiasen for the fruitful conversations. Moreover, part of this work was supported by the Gruber Science Fellowship at Yale University (R.S.).

7 Conflicts of interest

The methodology described in this manuscript is the subject of a pending patent application where L.P.D.L.C. is named as the sole inventor. L.P.D.L.C is the primary owner, and R.S. is the secondary owner. L.P.D.L.C. is the Head of Machine Learning at Shift Bioscience. R.S. has received consulting fees from TruDiagnostic, LongevityTech.fund and Cambrian BioPharma. A.H.C. has received consulting fees from TruDiagnostic and FOXO Biosciences. S.H. works for Altos Labs Limited UK and is a founder and consultant of the nonprofit Epigenetic Clock Development Foundation. B.W. serves as a scientific advisor to Shift Bioscience, Vevo Therapeutics, Deep Genomics. B.W. receives consulting fees from Arsenal Bioscience, Viecure Inc.

References

[1] A Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017).

[2] Jacob Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[3] 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". In: *Nature Methods* (2024), pp. 1–11.

[4] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. "Transfer learning enables predictions in network biology". In: *Nature* 618.7965 (2023), pp. 616–624.

[5] Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. "Universal cell embeddings: A foundation model for cell biology". In: *bioRxiv* (2023), pp. 2023–11.

[6] Barbara Steurer, Quentin Vanhaelen, and Alex Zhavoronkov. "Multimodal transformers and their applications in drug target discovery for aging and age-related diseases". In: *The Journals of Gerontology: Series A* 79.9 (2024).

[7] Alex Zhavoronkov, Polina Mamoshina, Quentin Vanhaelen, Morten Scheibye-Knudsen, Alexey Moskalev, and Alex Aliper. "Artificial intelligence for aging and longevity research: Recent advances and perspectives". In: *Ageing research reviews* 49 (2019), pp. 49–66.

[8] Steve Horvath. "DNA methylation age of human tissues and cell types". In: *Genome biology* 14 (2013), pp. 1–20.

[9] Gregory Hannum, Justin Guinney, Ling Zhao, LI Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. "Genome-wide methylation profiles reveal quantitative views of human aging rates". In: *Molecular cell* 49.2 (2013), pp. 359–367.

[10] Ake T Lu, Austin Quach, James G Wilson, Alex P Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, Andrea A Baccarelli, Yun Li, James D Stewart, et al. "DNA methylation GrimAge strongly predicts lifespan and healthspan". In: *Aging (albany NY)* 11.2 (2019), p. 303.

[11] Ake T Lu, Alexandra M Binder, Joshua Zhang, Qi Yan, Alex P Reiner, Simon R Cox, Janie Corley, Sarah E Harris, Pei-Lun Kuo, Ann Z Moore, et al. "DNA methylation GrimAge version 2". In: *Aging (Albany NY)* 14.23 (2022), p. 9484.

[12] Daniel W Belsky, Avshalom Caspi, David L Corcoran, Karen Sugden, Richie Poulton, Louise Arseneault, Andrea Baccarelli, Kartik Chamarti, Xu Gao, Eilis Hannon, et al. "DunedinPACE, a DNA methylation biomarker of the pace of aging". In: *Elife* 11 (2022), e73420.

[13] Albert T Higgins-Chen, Kyra L Thrush, Yunzhang Wang, Christopher J Minteer, Pei-Lun Kuo, Meng Wang, Peter Niimi, Gabriel Sturm, Jue Lin, Ann Zenobia Moore, et al. "A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking". In: *Nature aging* 2.7 (2022), pp. 644–661.

[14] Sheng Fong, Kamil Pabis, Djakim Latumalea, Nomuundari Dugersuren, Maximilian Unfried, Nicholas Tolwinski, Brian Kennedy, and Jan Gruber. "Principal component-based clinical aging clocks identify signatures of healthy aging and targets for clinical intervention". In: *Nature Aging* (2024), pp. 1–16.

[15] Lucas Paulo de Lima Camillo, Muhammad Haider Asif, Steve Horvath, Erica Larschan, and Ritambhara Singh. "Histone mark age of human tissues and cells". In: *bioRxiv* (2023), pp. 2023–08.

[16] Lucas Paulo de Lima Camillo, Louis R Lapierre, and Ritambhara Singh. "A pan-tissue DNA-methylation epigenetic clock based on deep learning". In: *npj Aging* 8.1 (2022), p. 4.

[17] Fedor Galkin, Polina Mamoshina, Kirill Kochetov, Denis Sidorenko, and Alex Zhavoronkov. "DeepMAge: a methylation aging clock developed with deep learning". In: *Aging and disease* 12.5 (2021), p. 1252.

[18] Sage Bionetworks. *Biomarkers of Aging Challenge 2024*. https://www.synapse.org/Synapse:syn52966292/wiki/624696. Accessed: 19/10/2024. 2024.

[19] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. "Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution". In: *Advances in neural information processing systems* 36 (2024).

[20] 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, et al. "The nucleotide transformer: Building and evaluating robust foundation models for human genomics". In: *BioRxiv* (2023), pp. 2023–01.

[21] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. "Dnabert-2: Efficient foundation model and benchmark for multi-species genome". In: *arXiv preprint arXiv:2306.15006* (2023).

[22] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. "Roformer: Enhanced transformer with rotary position embedding". In: *Neurocomputing* 568 (2024), p. 127063.

[23] Albert Gu and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces". In: *arXiv preprint arXiv:2312.00752* (2023).

[24] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[25] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature methods* 9.3 (2012), pp. 215–216.

[26] Shuo Li, Weihua Zeng, Xiaohui Ni, Qiao Liu, Wenyuan Li, Mary L Stackpole, Yonggang Zhou, Arjan Gower, Kostyantyn Krysan, Preeti Ahuja, et al. "Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring". In: *Proceedings of the National Academy of Sciences* 120.28 (2023), e2305236120.

[27] David C Goldberg, Cameron Cloud, Sol Moe Lee, Bret Barnes, Steven Gruber, Elliot Kim, Anita Pottekat, Max Westphal, Luana McAuliffe, Elisa Majournie, et al. "MSA: scalable DNA methylation screening BeadChip for high-throughput trait association studies". In: *bioRxiv* (2024), pp. 2024–05.

[28] Brian H Chen and Wanding Zhou. "mLiftOver: harmonizing data across Infinium DNA methylation platforms". In: *Bioinformatics* 40.7 (2024), btae423.

[29] Lucas Paulo de Lima Camillo. "pyaging: a Python-based compendium of GPU-optimized aging clocks". In: *Bioinformatics* (2024), btae200.

[30] Eric Dec, James Clement, Kaiyang Cheng, George M Church, Michael B Fossel, David H Rehkopf, Luis Rosero-Bixby, Michael S Kobor, David TS Lin, Ake T Lu, et al. "Centenarian clocks: epigenetic clocks for validating claims of exceptional longevity". In: *Geroscience* 45.3 (2023), pp. 1817–1835.

[31] Alan Tomusiak, Ariel Floro, Ritesh Tiwari, Rebeccah Riley, Hiroyuki Matsui, Nicolas Andrews, Herbert G Kasler, and Eric Verdin. "Development of an epigenetic clock resistant to changes in immune cell composition". In: *Communications Biology* 7.1 (2024), p. 934.

[32] Tianyu Zhu, Huige Tong, Zhaozhen Du, Stephan Beck, and Andrew E Teschendorff. "An improved epigenetic counter to track mitotic age in normal and precancerous tissues". In: *Nature Communications* 15.1 (2024), p. 4211.

[33] Ake T Lu, Zhe Fei, Amin Haghani, Todd R Robeck, JA Zoller, CZ Li, R Lowe, Q Yan, J Zhang, H Vu, et al. "Universal DNA methylation age across mammalian tissues". In: *Nature aging* 3.9 (2023), pp. 1144–1166.

[34] Amin Haghani, Caesar Z Li, Todd R Robeck, Joshua Zhang, Ake T Lu, Julia Ablaeva, Victoria A Acosta-Rodríguez, Danielle M Adams, Abdulaziz N Alagaili, Javier Almunia, et al. "DNA methylation networks underlying mammalian traits". In: *Science* 381.6658 (2023), eabq5693.

[35] Wanding Zhou, Peter W Laird, and Hui Shen. "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes". In: *Nucleic acids research* 45.4 (2017), e22–e22.

[36] Jacques Barhanin, Florian Lesage, Eric Guillemare, Michel Fink, Michel Lazdunski, and Georges Romey. "KvLQT1 and IsK (minK) proteins associate to form the I Ks cardiac potassium current". In: *Nature* 384.6604 (1996), pp. 78–80.

[37] Q Wang, M_E Curran, I Splawski, TC Burn, JM Millholland, TJ VanRaay, J Shen, KW Timothy, GM Vincent, T De Jager, et al. "Positional cloning of a novel potassium channel gene: KVLQT1 mutations cause cardiac arrhythmias". In: *Nature genetics* 12.1 (1996), pp. 17–23.

[38] Leu-Fen H Lin, Daniel H Doherty, Jack D Lile, Susan Bektesh, and Frank Collins. "GDNF: a glial cell line-derived neurotrophic factor for midbrain dopaminergic neurons". In: *Science* 260.5111 (1993), pp. 1130–1132.

[39] Christopher G Bell, Robert Lowe, Peter D Adams, Andrea A Baccarelli, Stephan Beck, Jordana T Bell, Brock C Christensen, Vadim N Gladyshev, Bastiaan T Heijmans, Steve Horvath, et al. "DNA methylation aging clocks: challenges and recommendations". In: *Genome biology* 20 (2019), pp. 1–24.

[40] Raghav Sehgal, Yaroslav Markov, Chenxi Qin, Margarita Meer, Courtney Hadley, Aladdin H Shadyab, Ramon Casanova, JoAnn E Manson, Parveen Bhatti, Eileen M Crimmins, et al. "Systems Age: A single blood methylation test to quantify aging heterogeneity across 11 physiological systems". In: *bioRxiv* (2023), pp. 2023–07.

[41] Qingwen Chen, Varun B Dwaraka, Natàlia Carreras-Gallo, Kevin Mendez, Yulu Chen, Sofina Begum, Priyadarshini Kachroo, Nicole Prince, Hannah Went, Tavis Mendez, et al. "OMICmAge: An integrative multi-omics approach to quantify biological age with electronic medical records". In: *bioRxiv* (2023).

[42] Kyunghwa Han and Inkyung Jung. "Restricted mean survival time for survival analysis: a quick guide for clinical researchers". In: *Korean Journal of Radiology* 23.5 (2022), p. 495.

[43] Aoshuang Zhou, Zimu Wu, Aung Zaw Zaw Phyo, Daniel Torres, Swarna Vishwanath, and Joanne Ryan. "Epigenetic aging as a biomarker of dementia and related outcomes: a systematic review". In: *Epigenomics* 14.18 (2022), pp. 1125–1138.

[44] Madeleine Carbonneau, Yi Li, Brenton Prescott, Chunyu Liu, Tianxiao Huan, Roby Joehanes, Joanne M Murabito, Nancy L Heard-Costa, Vanessa Xanthakis, Daniel Levy, et al. "Epigenetic Age Mediates the Association of Life's Essential 8 With Cardiovascular Disease and Mortality". In: *Journal of the American Heart Association* (2024), e032743.

[45] Francine Grodstein, Bernardo Lemos, Lei Yu, Hans-Ulrich Klein, Artemis Iatrou, Aron S Buchman, Gemma L Shireby, Jonathan Mill, Julie A Schneider, Philip L De Jager, et al. "The association of epigenetic clocks in brain tissue with brain pathologies and common aging phenotypes". In: *Neurobiology of disease* 157 (2021), p. 105428.

[46] Carolina Soriano-Tárraga, Eva Giralt-Steinhauer, Marina Mola-Caminal, Angel Ois, Ana Rodríguez-Campello, Elisa Cuadrado-Godia, Israel Fernández-Cadenas, Natalia Cullell, Jaume Roquer, and Jordi Jiménez-Conde. "Biological age is a predictor of mortality in ischemic stroke". In: *Scientific reports* 8.1 (2018), p. 4148.

[47] Samuel D Searle, Arnold Mitnitski, Evelyne A Gahbauer, Thomas M Gill, and Kenneth Rockwood. "A standard procedure for creating a frailty index". In: *BMC geriatrics* 8 (2008), pp. 1–10.

[48] John A Shepherd, Bennett K Ng, Markus J Sommer, and Steven B Heymsfield. "Body composition by DXA". In: *Bone* 104 (2017), pp. 101–105.

[49] EILEEN T KENNEDY, James Ohls, Steven Carlson, and Kathryn Fleming. "The healthy eating index: design and applications". In: *Journal of the American dietetic association* 95.10 (1995), pp. 1103–1108.

[50] Tiina Föhr, Arne Hendrix, Anna Kankaanpää, Eija K Laakkonen, Urho Kujala, Kirsi H Pietiläinen, Terho Lehtimäki, Mika Kähönen, Olli Raitakari, Xiaoling Wang, et al. "Metabolic syndrome and epigenetic aging: a twin study". In: *International Journal of Obesity* (2024), pp. 1–10.

[51] Christopher J Minteer, Kyra Thrush, John Gonzalez, Peter Niimi, Mariya Rozenblit, Joel Rozowsky, Jason Liu, Mor Frank, Thomas McCabe, Raghav Sehgal, et al. "More than bad luck: cancer and aging are linked to replication-driven changes to the epigenome". In: *Science Advances* 9.29 (2023), eadf4163.

[52] Andy P Siddaway, Alex M Wood, and Peter J Taylor. "The Center for Epidemiologic Studies-Depression (CES-D) scale measures a continuum from well-being to depression: Testing two key predictions of positive clinical psychology". In: *Journal of Affective Disorders* 213 (2017), pp. 180–186.

[53] Breno S Diniz, Johanna Seitz-Holland, Raghav Sehgal, Jessica Kasamoto, Albert T Higgins-Chen, and Eric Lenze. "Geroscience-centric perspective for geriatric psychiatry: integrating aging biology with geriatric mental health research". In: *The American Journal of Geriatric Psychiatry* 32.1 (2024), pp. 1–16.

[54] Nora Pashayan, Daniel Reisel, and Martin Widschwendter. "Integration of Genetic and Epigenetic Markers for Risk Stratification: Opportunities ad Challenges". In: *Personalized medicine* 13.2 (2016), pp. 93–95.

[55] Anatoly Urban, Denis Sidorenko, Diana Zagirova, Ekaterina Kozlova, Aleksandr Kalashnikov, Stefan Pushkov, Vladimir Naumov, Viktoria Sarkisova, Geoffrey Ho Duen Leung, Hoi Wing Leung, et al. "Precious1GPT: multimodal transformer-based transfer learning for aging clock development and feature importance analysis for aging and age-related disease target discovery". In: *Aging (Albany NY)* 15.11 (2023), p. 4649.

[56] Denis Sidorenko, Stefan Pushkov, Akhmed Sakip, Geoffrey Ho Duen Leung, Sarah Wing Yan Lok, Anatoly Urban, Diana Zagirova, Alexander Veviorskiy, Nina Tihonova, Aleksandr Kalashnikov, et al. "Precious2GPT: the combination of multiomics pretrained transformer and conditional diffusion for artificial multi-omics multi-species multi-tissue sample generation". In: *npj Aging* 10.1 (2024), p. 37.

[57] Fedor Galkin, Vladimir Naumov, Stefan Pushkov, Denis Sidorenko, Anatoly Urban, Diana Zagirova, Khadija M Alawi, Alex Aliper, Ruslan Gumerov, Aleksand Kalashnikov, et al. "Precious3GPT: Multimodal Multi-Species Multi-Omics Multi-Tissue Transformer for Aging Research and Drug Discovery". In: *bioRxiv* (2024), pp. 2024–07.

[58] Wanding Zhou, Timothy J Triche Jr, Peter W Laird, and Hui Shen. "SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions". In: *Nucleic acids research* 46.20 (2018), e123–e123.

[59] Kyra L Thrush, Albert T Higgins-Chen, Zuyun Liu, and Morgan E Levine. "R methylCIPHER: a methylation clock investigational package for hypothesis-driven evaluation & research". In: *biorxiv* (2022), pp. 2022–07.