CpGPT: a Foundation Model for DNA Methylation

Lucas Paulo de Lima Camillo*†
University of Cambridge, Shift Bioscience
Cambridge, 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

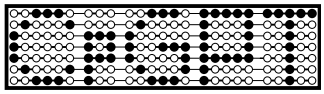
Henry E. Miller
Shift Bioscience
Cambridge, UK

Jessica A. Lasky-Su
Brigham and Women's Hospital
Boston, MA, USA

Albert T. Higgins-Chen
Yale University School of Medicine
New Haven, CT, USA

Steve Horvath
Altos Labs
Cambridge, UK

Bo Wang
University of Toronto, Vector Institute, University Health Network
Toronto, Canada



Abstract

DNA methylation is a type of epigenetic modification that plays a significant role in development, aging, and disease. Despite extensive research into the molecular mechanisms of DNA methylation, they remain poorly understood today. Foundation models are a class of machine learning model that leverage vast quantities of data to make sense of complex data types, such as genome sequences or single-cell transcriptomes. Here, we present the Cytosine-phosphate-Guanine Pretrained Transformer (CpGPT), a novel foundation model pretrained on CpGCorpus, a novel database with more than 2,000 DNA methylation datasets encompassing over 150,000 samples from diverse 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

*Main author
†Corresponding authors

contexts, CpGPT outperforms specialized models when finetuned for aging-related tasks, such as mortality risk and morbidity assessments. The model is highly adaptable and can impute beta values across different methylation platforms, tissue types, mammalian species, and even single-cell data. As a foundation model, CpGPT can be leveraged as a new tool for biological discovery in the field of epigenetics. The open-source code and model can be found at `http://github.com/lcamillo/CpGPT`.

Highlights

- CpGPT is a novel foundation model for DNA methylation analysis, pretrained on over 2,000 datasets encompassing 150,000+ 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.

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) [2]. Transformers leverage self-attention mechanisms to process sequential data more effectively, capturing long-range dependencies and complex patterns [3]. 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 [4], Geneformer [5], and Universal Cell Embeddings [6] 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 suitable for biological data, which often involve complex interactions and hierarchies. The advent of these models has also begun to influence longevity research [7, 8].

Despite significant progress in the past decade in the field of epigenetic biomarkers, many widely used models rely on relatively simple regularized linear models using Cytosine-phosphate-Guanine (CpG) DNA methylation data [9, 10, 11, 12, 13]. These predictors often do not consider the sequence context or genomic positions of CpG sites and as a result, they may overlook complex interactions and the underlying biological mechanisms driving processes like development, aging, and disease. A recent advancement involves the application of principal component decomposition before the linear model to enhance the reliability and performance of DNA methylation epigenetic proxies [14, 15]. However, few 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. The model can impute missing methylation values within a dataset, convert between different methylation platforms by reconstructing unmeasured CpG sites, and perform zero-shot reference mapping to label

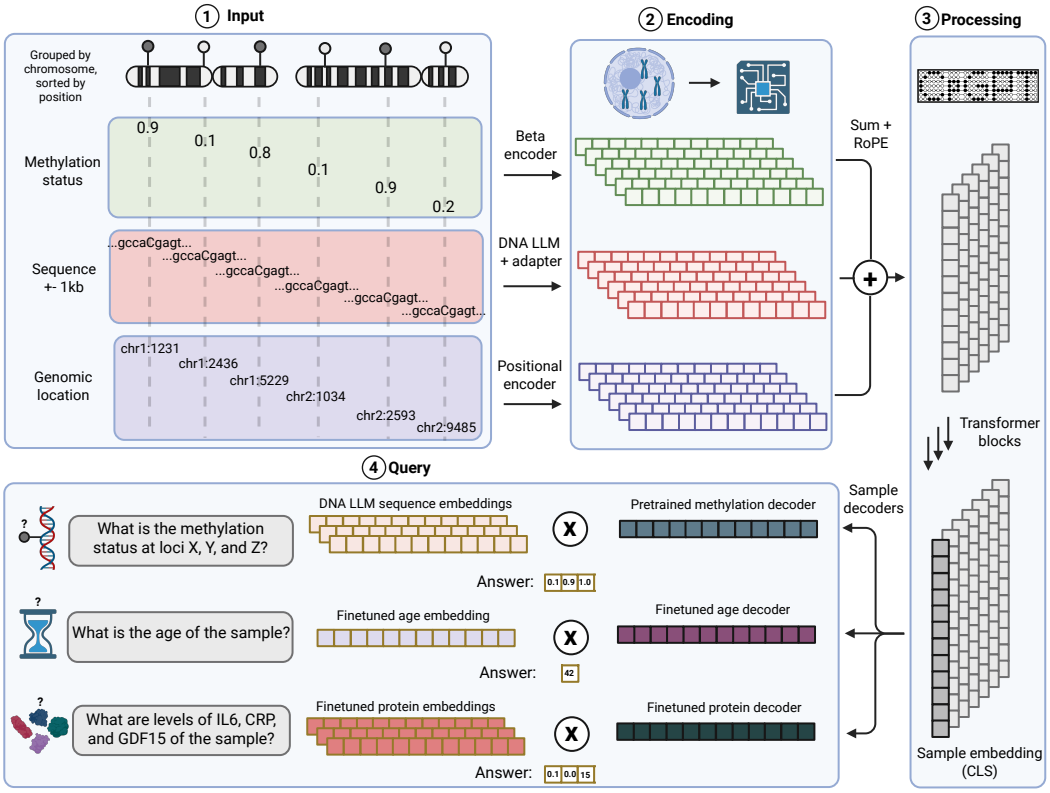


Figure 1: The CpGPT framework. The input CpG sites are grouped by chromosome and sorted. Information is encoded by the model through several neural networks, including a projection of the methylation status, a transformation of the sequence embedding from a DNA large language model (LLM), and a positional addition to incorporate genomic location. The embeddings of each information type are added followed by rotational positional encoding (RoPE). A series of transformer blocks aggregates the sample information into a CLS sample embedding. The sample embedding can be queried after transformation with a pretrained or finetuned decoder followed by a dot product of the query embedding.

samples without finetuning, all while benefiting from increasing test-time compute. Moreover, CpGPT excels when finetuned for specific tasks, achieving second place overall for chronological age prediction in the Biomarkers of Aging Challenge [18].

The CpGPT framework is highly generalizable and can be applied across various tasks, such as cancer prediction and classification, across different mammalian species, and across platforms, including single-cell data. Overall, CpGPT establishes a new standard for DNA methylation analysis and offers a versatile tool for multiple applications in the field of 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 database named *CpGCorpus* (see Methods). This database comprises over 150,000 human DNA methylation samples collected from more than 2,000 studies, encompassing a diverse array of tissue types, developmental stages, and disease conditions (Supplementary Figure 2). We preprocessed and harmonized the data to ensure consistency across different methylation array platforms, including Illumina 27k, 450k, EPIC, and EPICv2 arrays, resulting in over 1.3 billion unique tokens.

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, namely Nucleotide Transformer V2 with 500 million parameters ($\text{NTv2}_{\text{500M}}$) [19]. For global positional context, we sorted the sequence embeddings by genomic positions and grouped them by chromosomes. To prevent positional biases, we also applied stochastic shuffling per chromosome. We incorporated modifications of the original positional encoding method from Vaswani et al. [1] along with rotary positional embeddings [20] 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 1).

At the core of CpGPT is an enhanced version of the transformer architecture, known as transformer++ [21], 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 beta values and their associated uncertainty for CpG sites. The model can utilize the DNA LLM 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 multitask learning 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 (see Methods). The model is designed to handle missing data, enabling it to work effectively with incomplete methylation profiles, which are common because of varying array designs and experimental conditions. Rather than relying on a fixed vocabulary of CpG sites as current epigenetic clocks, the model's input is typically a random subset of CpG sites, varying from 5,000 to 10,000 depending on the task. This comprehensive training strategy allows CpGPT to learn rich, nuanced representations of DNA methylation patterns across diverse genomic contexts and biological conditions, allowing it to perform a series of downstream tasks (Figure 1).

We pretrained two models, a lightweight one designed to facilitate fine-tuning and a large one tailored toward zero-shot tasks. The former contains 2.5 million parameters, whereas the latter contains 101 million. They are herein referred to as CpGPT_{2M} and $\text{CpGPT}_{\text{100M}}$ respectively, with a label following the parameter size indicating the model was finetuned for that purpose (e.g. CpGPT_{2M}-Mortality).

2.2 CpGPT learns meaningful methylation and sample embeddings

Understanding how well a foundation model captures the underlying structure of biological data is essential. We assessed whether CpGPT can learn meaningful embeddings in an unsupervised manner by examining whether 1. features (individual CpG sites) and 2. samples naturally cluster according to their biological attributes — even without explicit labels. To this end, we generated low-dimensional representations of the high-dimensional embeddings using Uniform Manifold Approximation and Projection (UMAP) [22] at both the CpG site (locus) level and the sample level across a variety of datasets.

Our first experiment aimed to test whether the locus embeddings produced by CpGPT capture biologically relevant genomic annotations. Specifically, CpGPT transforms the $\text{NTv2}_{\text{500M}}$ embeddings of nucleotide sequences flanking each CpG site using an adapter network to match its internal dimension. We hypothesized that these transformed locus embeddings would better reflect functional genomic annotations than the raw DNA LLM embeddings, given the model's knowledge of epigenetics. To test this, we sampled approximately 100,000 CpG sites from the Illumina EPICv2 array and applied UMAP for visualization. We then examined the resulting clusters with chromatin state annotations from ChromHMM [23] — a framework that uses combinations of histone modifications to

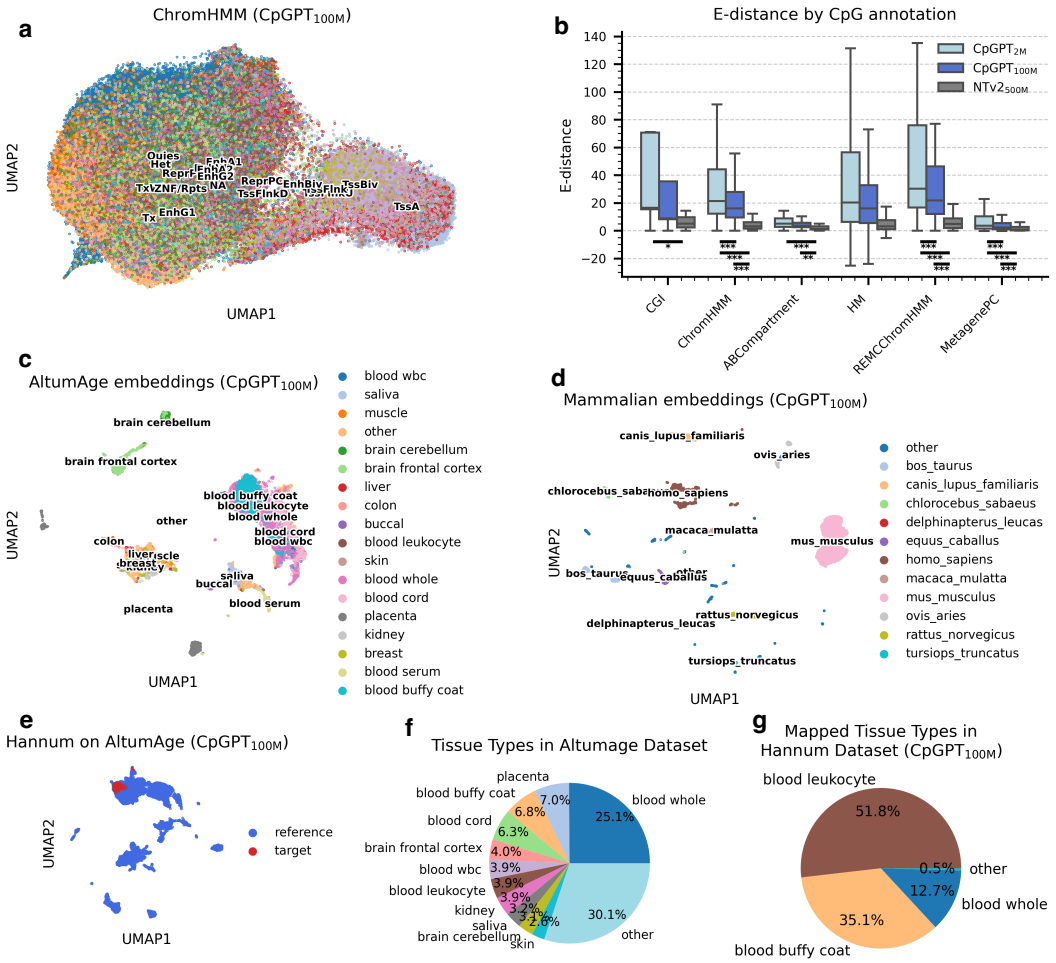


Figure 2: CpGPT embeddings capture biologically meaningful information. (a) UMAP of $\text{CpGPT}_{100\text{M}}$ locus embeddings for 100,000 randomly sampled CpG sites (Illumina EPICv2 array), colored by ChromHMM annotation. (b) Boxplot showing energy distances for six genomic annotations, comparing the genomic location embeddings of $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and the Nucleotide Transformer V2 500M. (c) UMAP of $\text{CpGPT}_{100\text{M}}$ sample embeddings for the AltumAge dataset, colored by tissue type (with "other" indicating tissue labels totaling less than 10% of the data). (d) UMAP of $\text{CpGPT}_{100\text{M}}$ sample embeddings for 55 mammalian species (GSE223748), colored by species (with "other" indicating species labels totaling less than 10% of the data). (e) UMAP of $\text{CpGPT}_{100\text{M}}$ sample embeddings comparing the AltumAge dataset (reference in blue) and the Hannum dataset (target in red). (f) Pie chart illustrating the tissue composition of the AltumAge dataset. (G) Pie chart illustrating the tissue composition of the Hannum dataset.

segment the genome into distinct states such as active transcription start sites (TssA), bivalent promoters (TssBiv; characterized by both activating and repressive marks often seen in developmentally regulated genes), enhancers (Enh), and repressed or inactive regions. In the UMAP projection of the $\text{CpGPT}_{100\text{M}}$ locus embeddings (Figure 2a), CpG sites linked to active TssA regions formed a distinct cluster, while bivalent promoters and enhancers occupied intermediate positions between active and repressed states. These results indicate that CpGPT's locus embeddings segregate CpG sites according to their chromatin state, demonstrating the model's capacity to capture complex epigenetic information solely from sequence context, particularly when compared to $\text{NTv2}_{500\text{M}}$ (Supplementary Figure 3).

In order to quantitatively validate our visual observations, we computed the energy distance (E-distance) between clusters of embeddings. E-distance is a metric that measures how separate two groups of data points are [24]. We compared embeddings from $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and the raw $\text{NTv2}_{500\text{M}}$ model across multiple genomic annotation frameworks, including CpG Islands (CGI), ChromHMM, ABCompartment, REMCChromHMM, and MetagenePC. Overall, the energy distances were significantly higher for both versions of CpGPT compared to $\text{NTv2}_{500\text{M}}$ (e.g. for ChromHMM annotations, Mann-Whitney U test $p < 0.0005$ for both CpGPT models), indicating that the adapter network enriches the biological information present in the sequence embeddings. These findings confirm that CpGPT produces more distinctly clustered embeddings than those generated by the underlying DNA LLM alone.

Next, we evaluated whether the sample-level embeddings learned by CpGPT encapsulate meaningful biological variation. For this purpose, we extracted sample embeddings from several publicly available datasets: (i) a multi-tissue DNA methylation dataset previously used to develop the AltumAge clock [16], (ii) a Reduced Representation Bisulfite Sequencing (RRBS) atlas (GSE233417) [25] encompassing diverse human cell types, (iii) a pan-mammalian dataset (GSE223748) spanning dozens of species, and (iv) a brain single-cell methylation dataset generated using the recent sciMETv3 method (GSE273592) [26]. UMAP was then used to reduce these high-dimensional embeddings to two dimensions. In the AltumAge dataset (Figure 2c), distinct clusters corresponding to tissues such as placenta, brain, and blood emerged, indicating that CpGPT captures tissue-specific methylation signatures, in contrast to the UMAP of the original beta values (Supplementary Figure 4). Similarly, in the RRBS atlas (Supplementary Figure 5), samples mostly grouped by cell type — with separations observed among divergent lineages (e.g., testis, white blood cells, and pituitary gland). In the pan-mammalian dataset (Figure 2d, Supplementary Figure 6), the sample embeddings clustered by species, with species phylogenetically closer to Homo sapiens (for example, Macaca mulata and Chlorocebus sabaeus) positioned in closer proximity. In the single-cell dataset, the finetuned ($\text{CpGPT}_{100\text{M}}$-sciMETv3) sample embeddings appear much more biologically plausible compared to the UMAP derived from the raw methylation values (Supplementary Figures 7). These observations demonstrate that the global methylation patterns encoded by CpGPT's sample embeddings accurately reflect cellular identity, tissue type, and evolutionary relationships — even in the absence of explicit labels.

An important application of robust sample embeddings is the ability to perform zero-shot reference mapping, where labels from a well-annotated reference dataset can be transferred to an unlabeled dataset without additional training. To evaluate this capability, we projected the sample embeddings of the Hannum dataset (GSE40279) [10] onto those from the AltumAge dataset, which contains comprehensive tissue-specific annotations (Figure 2e). Our objective was to determine whether the blood-derived samples from the Hannum cohort could be accurately classified into the tissue categories defined by AltumAge. The mapping results were compelling: out of 656 Hannum samples, 340 were classified as "blood leukocyte," 230 as "blood buffy coat," 83 as "blood whole," and 3 as "other" (Figure **??**). These results indicate that CpGPT's sample embeddings capture subtle epigenetic differences among cell types, allowing accurate tissue classification without the need for further training.

Finally, to further demonstrate the practical utility of CpGPT's sample embeddings, we applied zero-shot reference mapping to a dataset capturing the reprogramming of fibroblasts to pluripotency (GSE54848) [27]. Using the comprehensive CpGCorpus as the reference,

we mapped each sample from the reprogramming time course to its most similar entry in the Gene Expression Omnibus. Notably, the three replicates of day 0 fibroblasts (prior to OSKM induction) mapped to GSM868011, GSM868013, and GSM868015, corresponding to "somatic primary HDFs." By day 28, the replicates mapped to GSM867971, GSM867983, and GSM867975, now annotated as "induced pluripotency stem cell." Day 15 was the last time point where the mapped embeddings retained fibroblast descriptions, in agreement with previous research that identifies this period as the transition point when fibroblasts lose their cellular identity [28]. This example demonstrates that CpGPT's sample embeddings are not only biologically informative but also robust enough to capture dynamic changes in cell state during reprogramming.

2.3 CpGPT faithfully reconstructs the methylome

DNA methylation data often contain missing values or unmeasured CpG sites because of platform differences and variations in experimental design. An ideal foundation model should be able to leverage its learned knowledge to fill in these gaps and infer the methylation status of CpG sites that were never directly observed. This ability, often referred to as zero-shot imputation, is particularly valuable for applications like biomarker discovery, epigenetic clock predictions, and data integration across different methylation arrays. Here, we systematically evaluated CpGPT's capacity to perform zero-shot imputation and array conversion in several scenarios, including reconstructing sites omitted during pretraining, converting between Illumina arrays, mapping methylation patterns across mammalian species, and even handling single-cell methylation data.

In the first experiment, our objective was to assess the generality of CpGPT's imputation when it encounters CpG sites that were deliberately excluded from its pretraining. Specifically, we removed 1% of the probes (i.e., CpG sites) during model training and then tested how well CpGPT could infer their methylation levels in the CpGCorpus test set (Figure 3a). In the simplest case, the model receives beta values from CpG sites it had encountered during pretraining ("seen, seen") and predicts the methylation of target loci also within the pretraining universe. Both CpGPT_{2M} and $\text{CpGPT}_{\text{100M}}$ performed markedly better than the mean baseline (baseline mean absolute error (MAE) = 0.3295; CpGPT_{2M} MAE = 0.06036; $\text{CpGPT}_{\text{100M}}$ MAE = 0.03831). We then tested the more challenging "seen, unseen" case, in which the model predicts completely new loci that were never observed in training. Despite an increase in error, $\text{CpGPT}_{\text{100M}}$ still outperformed the baseline by a substantial margin (baseline MAE = 0.3297; CpGPT_{2M} MAE = 0.1515; $\text{CpGPT}_{\text{100M}}$ MAE = 0.0764). Even in the most stringent "unseen, unseen" setup, where both the input and predicted sites were not in the pretraining universe, CpGPT retained a substantial edge over the baseline (baseline MAE = 0.3376; CpGPT_{2M} MAE = 0.1470; $\text{CpGPT}_{\text{100M}}$ MAE = 0.0781). These outcomes underscore the model's ability to generalize to novel genomic regions purely from previously learned representations, indicating robust zero-shot performance.

We next explored a more practical scenario of reconstructing methylation sites for an emerging array platform, the Illumina MSA [29], for which no paired MSA and conventional array data exist. We simulated this situation using the Hannum dataset (GSE40279) [10], originally measured on the Illumina 450k array. To mimic partial MSA data, we kept only the 113,585 probes shared between 450k and MSA, then used CpGPT to impute the remaining 450k probes needed for various epigenetic clocks (Figure 3b). Even without direct MSA-450k paired training data, $\text{CpGPT}_{\text{100M}}$ produced an MAE of 0.03144, and a fine-tuned version ($\text{CpGPT}_{\text{100M}}$-Hannum) achieved an MAE of 0.02447, both notably outperforming the mean baseline (0.3027). As shown by the near-perfect identity line fit for an example sample (Figure 3c), the model accurately recovered large segments of the 450k methylome from a subset of MSA-like inputs. Unlike tissue-specific methods such as *mLiftOver* [30], CpGPT's multi-tissue training in *CpGCorpus* enables tissue-agnostic imputation, making it more broadly applicable.

Because epigenetic clocks often rely on sets of probes that may be missing from a given platform, we then tested the downstream impact of CpGPT-based array conversion on several clocks, including *encen40* [31], *intrinclock* [32], and *stemtoc* [33]. We computed these clocks on the Hannum dataset using fully measured 450k data (considered the ground truth) and compared them to predictions derived from CpGPT-imputed values with a subset

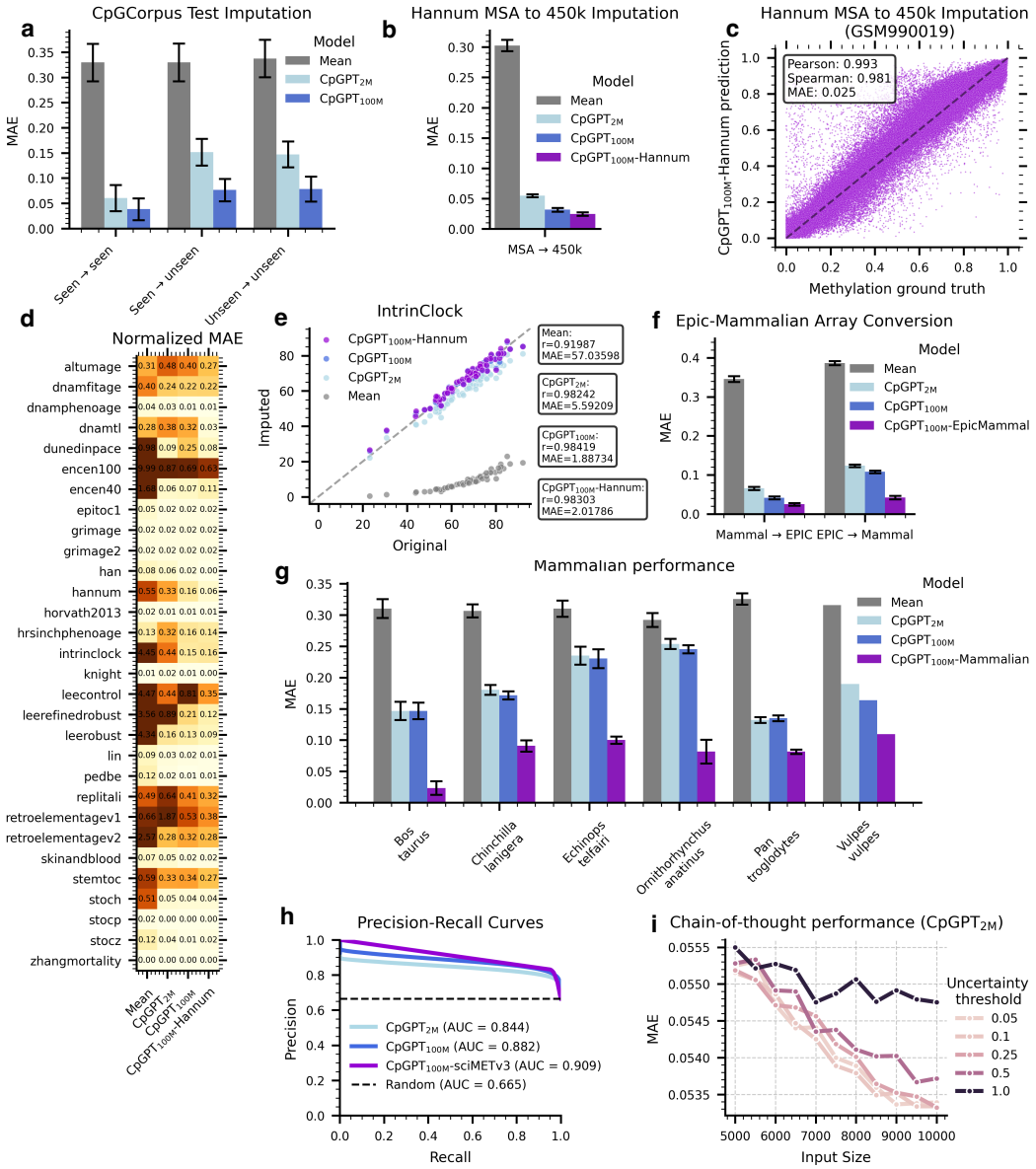


Figure 3: CpGPT performs zero-shot imputation across various conditions. (a) Barplot of average mean absolute error (MAE) per sample for $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and a mean-baseline model, tested on the CpGCorpus dataset. The 5,000 input CpG sites and the 5,000 queried CpG sites are either from the pretraining vocabulary ("seen") or not ("unseen"). Error bars represent the standard deviation of the sample-level MAE. (b) Barplot of the average MAE per sample in the Hannum dataset for mean-baseline, $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and finetuned $\text{CpGPT}_{100\text{M}}$-Hannum models. The query includes all CpG sites in the 450k array not present in the MSA array. Error bars represent the standard deviation of the sample-level MAE. (c) Scatterplot comparing the ground-truth beta values versus $\text{CpGPT}_{100\text{M}}$-Hannum predictions for sample GSM990019, showing the conversion from MSA probes to remaining 450k probes. (d) Heatmap of the MAE between epigenetic clock calculations using only MSA-overlapping probes (zeroed-out missing probes) or CpGPT-imputed probes ($\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, finetuned $\text{CpGPT}_{100\text{M}}$-Hannum), compared against the full 450k values (Hannum dataset). (e) Scatterplot comparing intrinclock predictions using only MSA-overlapping probes (missing probes either zeroed out or imputed with $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, or finetuned $\text{CpGPT}_{100\text{M}}$-Hannum) to the ground-truth 450k array (Hannum dataset). (f) Barplot of the average MAE per sample for mean-baseline, $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and finetuned $\text{CpGPT}_{100\text{M}}$-EpicMammal when converting between paired Mammalian40k and EPIC arrays. Error bars represent the standard deviation. (G) Barplot of the average MAE per sample for a mean-baseline, $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, or finetuned $\text{CpGPT}_{100\text{M}}$-Mammalian model. Results are shown for a random subset of 5,000 CpG sites in six unseen mammalian species. Error bars represent the standard deviation. (h) Precision-recall curves for random baseline, $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and finetuned $\text{CpGPT}_{100\text{M}}$-sciMETv3, showing imputation performance on single-cell methylation (sciMETv3 dataset). (i) Lineplot of the MAE for MSA-to-450k conversion (Hannum dataset) using $\text{CpGPT}_{2\text{M}}$ with chain-of-thought inference. Different uncertainty quantiles are used to filter generated CpG methylation values.

of the MSA probes as input (Figure 3d; Supplementary Figure 8). The clock estimates from CpGPT-imputed data closely matched those from the fully measured set. For instance, the correlation of the *intrinclock* with chronological age improved from 0.920 (mean baseline imputation) to 0.983 when relying on $\text{CpGPT}_{100\text{M}}$-Hannum predictions (Figure 3e. These results demonstrate that CpGPT's array conversion can restore missing probes in a manner that preserves or even enhances the performance of modern epigenetic clocks.

We next aimed to convert data between the Illumina EPIC and Horvath mammalian arrays [34]. Using paired blood samples, we assessed two directions: reconstructing EPIC from the mammalian array and vice versa. As illustrated in Figure 3f, for conversion to EPIC, $\text{CpGPT}_{100\text{M}}$-EpicMammal yielded a MAE of 0.02485 versus a mean baseline of 0.3459, whereas for conversion to the mammalian array, it achieved 0.04206 versus a baseline of 0.3869. Despite the lower fidelity of the epigenetic clock predictions from the reconstructed data (Supplementary Figure **??**, these findings show that CpGPT is capable of bridging even larger design gaps between array platforms, thereby unlocking cross-platform analyses such as applying mammalian-focused epigenetic clocks to EPIC data.

We further tested CpGPT's ability to generalize across species by using data from 55 mammalian taxa (GSE223748) [35, 36]. From these, 43 species comprised the training set, 6 formed a validation set, and 6 served as an unseen test set. Within each species, half of the CpG sites were masked, and CpGPT was tasked to impute those sites (Figure 3g). Remarkably, for a species it had never seen during training, Pan troglodytes (chimpanzee), $\text{CpGPT}_{100\text{M}}$-Mammalian achieved an MAE of 0.08143, compared to a mean baseline of 0.3257. Although performance tended to be stronger for species phylogenetically closer to humans, these results reveal that a model solely trained on human arrays can still capture salient epigenetic patterns in evolutionarily related mammals. This cross-species flexibility opens the door to unified frameworks for translational studies and cross-mammalian biomarker research.

Next, we tested if CpGPT could handle single-cell methylation data, which are typically sparse and largely binary. We used the sciMETv3 dataset (GSE273592) [26], focusing on precision–recall (PR) and receiver-operator (ROC) curves to evaluate its predictions (Figure 3h; Supplementary Figure 11). Even though CpGPT was pretrained on bulk methylation data, it outperformed the random baseline for single-cell measurements. For instance, $\text{CpGPT}_{100\text{M}}$ had a PR AUC of 0.882 versus a random baseline of 0.665, and a fine-tuned variant ($\text{CpGPT}_{100\text{M}}$-sciMETv3) attained 0.909. Similar trends were observed in the AUROC analyses ($\text{CpGPT}_{100\text{M}}$ AUROC = 0.8373 vs. random = 0.5000). These results indicate that CpGPT's learned representations can extend to the unique challenges of single-cell methylation.

Finally, we explored whether CpGPT could benefit from increased test-time compute. Traditional LLMs have vastly benefited from tokens that are generated prior to a final answer [37]. We translated the natural language processing approach to our biological foundation model in the following way. First, a given number of CpG sites are given to the model (context window), followed by the prediction of methylation values for a random subset of genomic locations with available DNA LLM embeddings from pretraining. For the top n sites with the lowest predicted uncertainty values, those generated outputs are appended to the original input. After a certain number of thinking steps, the methylation values for the queried genomic loci are predicted. With the same setup as for the array conversion for the Hannum dataset, we initially started with a subset of 5,000 CpG sites from the MSA array to reconstruct the remaining probes in the 450k array using $\text{CpGPT}_{2\text{M}}$ (Figure 3i). The MAE decreases by about 4% (0.0553 vs. 0.0532) after an additional 10 thinking steps of size 500 given an uncertainty threshold of 5th, 10th and 25th percentiles. We also investigated how larger step sizes of 1000 and 2000 and finetuning would affect the chain-of-thought performance curve. Overall, decreasing the uncertainty was beneficial whereas step size had minimal impact. Overall, the improved performance seems to deteriorate after a certain point, in line with traditional LLM research [38]. This would be analogous to the forgetting of the initial context given too many thinking steps. Overall, given the flexibility of the CpGPT framework, state-of-the-art reasoning techniques can be applied even without any explicit chain-of-thought training.

Finally, we investigated whether CpGPT could benefit from increased test-time compute, inspired by chain-of-thought approaches in natural language processing [37]. In this procedure, we first provided the model with a "context window" of CpG sites and generated a random subset of methylation predictions from the pretrained DNA LLM embeddings. We then appended the top n least uncertain predictions back into the input, allowing the model to refine its estimates in subsequent iterations. To evaluate this strategy, we used the Hannum dataset's array-conversion task as a test bed (Figure 3i). Starting with 5,000 MSA CpG sites and $\text{CpGPT}_{2\text{M}}$, the mean absolute error (MAE) decreased from approximately 0.0553 to 0.0532 after 10 additional inference steps of size 500 under uncertainty thresholds at the 5th, 10th, and 25th percentiles. We also examined how larger step sizes (1,000 or 2,000) and fine-tuning influenced the chain-of-thought performance curve, finding that lowering uncertainty improved accuracy while step size exerted minimal effect (Supplementary Figure **??**). However, beyond a certain number of iterations, performance gains deteriorated—paralleling traditional chain-of-thought "forgetting" phenomena [38]. Taken together, these results indicate that even without explicit chain-of-thought training, CpGPT's flexible architecture can incorporate state-of-the-art reasoning methods to enhance imputation accuracy at test time.

2.4 CpGPT displays state-of-the-art performance when finetuned

Foundation models often gain substantial performance improvements when finetuned on specific tasks, because the pretraining process captures broad, generalizable representations that can be readily adapted to particular datasets. We hypothesized that this principle would extend to CpGPT, enabling the model to excel in a variety of downstream prediction challenges once it had been exposed to the unique nuances of each problem. To test this, we employed our lighter version, $\text{CpGPT}_{2\text{M}}$, and examined its performance on (i) regression tasks of varying complexity, (ii) multiclass regression, and (iii) binary classification.

In the first experiment, we assessed whether $\text{CpGPT}_{2\text{M}}$ could predict chronological age from multi-tissue data with greater accuracy than conventional methods. We partitioned the AltumAge dataset [16] into training, validation, and testing sets, and compared $\text{CpGPT}_{2\text{M}}$-Age to an ElasticNet model and a multilayer perceptron (MLP). In terms of mean squared error (MSE), $\text{CpGPT}_{2\text{M}}$-Age achieved 38.12 years-squared, outperforming both ElasticNet (270.4) and MLP (86.56). Moreover, the Pearson correlation coefficient (r) rose to 0.9847 with $\text{CpGPT}_{2\text{M}}$-Age compared to 0.9116 and 0.9679 for ElasticNet and MLP, respectively (Figure 4a). These results suggest that incorporating sequence and epigenetic context through CpGPT yields a more nuanced representation of methylation data, ultimately enhancing the accuracy of age estimation.

We then investigated a more demanding multiclass regression problem involving proteomic measurements. Here, our goal was to predict the levels of 322 proteins that were quantified alongside methylation data for the Biomarkers of Aging Consortium (GSE246337) [39]. This task is especially challenging because proteomic assays can be prone to measurement variability. Likely due to this potential noise, $\text{CpGPT}_{2\text{M}}$-Proteins attained a median Spearman correlation coefficient of 0.2100 across all proteins, but still surpassing both ElasticNet (0.1180) and MLP (0.1131) (Figure 4b and Supplementary Figure 12). These findings underscore CpGPT's performance when predicting hundreds of variables with a single forward pass.

We next applied CpGPT to a binary classification task aimed at distinguishing between normal and adjacent cancerous tissues, using a dataset from multiple studies [16]. When comparing precision–recall and receiver operating characteristic curves (Figure 4c and Supplementary Figure 13), $\text{CpGPT}_{2\text{M}}$-Cancer achieved an area under the ROC curve (AUROC) of 0.9540 and an area under the precision–recall curve (AUPRC) of 0.9680, closely matching the MLP (AUROC 0.9520, AUPRC 0.9720) and outperforming logistic regression (AUROC 0.9360, AUPRC 0.9650). Thus, CpGPT not only excels at regression tasks but also maintains robust classification performance.

We further examined how CpGPT would handle data beyond human samples by finetuning on a mammalian dataset to predict relative age, adult bodyweight, and maximum lifespan. Overall, the model exhibited lower absolute error (MAE) compared to ElasticNet and MLP

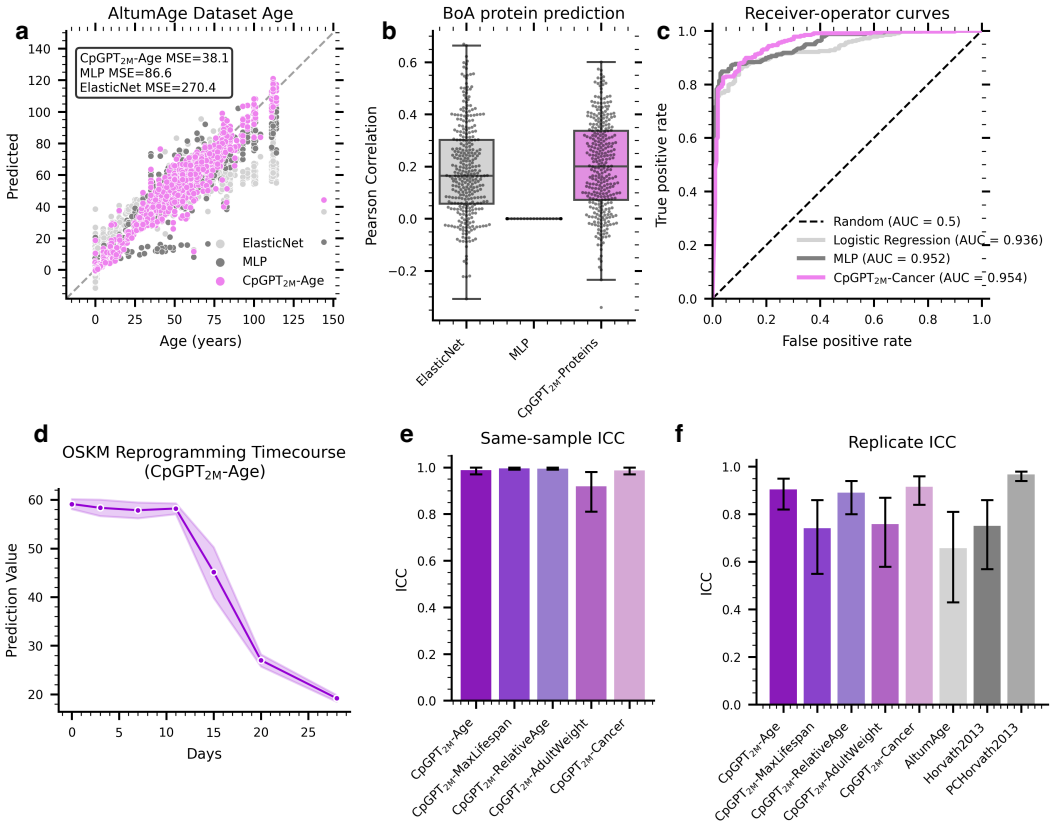


Figure 4: CpGPT outperforms neural networks and linear models, while showing reproducible results. (a) Scatterplot comparing ground-truth age to predictions on the AltumAge dataset using a linear model (ElasticNet), a multi-layer perceptron (MLP), and $\text{CpGPT}_{2\text{M}}$-Age. (b) Boxplot of the Spearman correlation between ground-truth and predicted blood protein levels for 322 proteins in the Biomarkers of Aging dataset (GSE246337), comparing logistic regression, MLP, and $\text{CpGPT}_{2\text{M}}$-Proteins. (c) Receiver operating characteristic (ROC) curves of cancer prediction on the cancer dataset for a random baseline, logistic regression, MLP, and $\text{CpGPT}_{2\text{M}}$-Cancer. (d) Lineplot (with 95% confidence intervals) of $\text{CpGPT}_{2\text{M}}$-Age predictions over an OSKM reprogramming time course using random subsets of CpG sites. (e) Barplot of the intraclass correlation coefficient (ICC) for various finetuned CpGPT models on the OSKM reprogramming dataset, computed over 10 random CpG subset inputs. Error bars show the 95% confidence intervals. (f) Barplot of the ICC for various CpGPT models and three epigenetic clocks on a reliability dataset with technical replicates, all using a fixed input subset of 10,000 CpG sites.

baselines (Supplementary Figure **??**), suggesting that the pretrained embeddings encode relationships that generalize across different mammalian species. Although the performance was not as strong as in the age prediction or cancer classification tasks, these results highlight the flexibility of the framework.

To evaluate the consistency of CpGPT's predictions when given different subsets of methylation data, we analyzed the OSKM dataset (GSE54848) with 10 distinct random seeds to select different subsets of 10,000 probes from the Illumina 450k array. We specifically tracked how $\text{CpGPT}_{2\text{M}}$-Age behaves throughout a reprogramming time course, alongside other finetuned models (Supplementary Figure 17). As shown in Figure 4d, the predicted chronological age declined over time, reflecting the dedifferentiation process, and only the adult bodyweight predictor exhibited a slight upward drift. To quantify stability, we computed the intraclass correlation coefficient (ICC) across predictions from different subsets of CpG sites for each finetuned model (Figure 4e). Most predictors had ICC values above 0.99, and the lowest ($\text{CpGPT}_{2\text{M}}$-AverageAdultWeight) reached 0.9193. These high ICCs il-

lustrate that CpGPT remains consistent even when the input methylation sites are changed, affirming its reliability in scenarios where coverage or probe availability may vary.

Finally, we benchmarked CpGPT's stability against replicate samples in a blood methylation dataset (GSE55763) [40]. We compared CpGPT predictions with three well-known multi-tissue epigenetic clocks (AltumAge [16], Horvath2013 [9], and PCHorvath2013 [14]), fixing the input to a randomly selected subset of 10,000 CpG sites (Figure 4f). Although PCHorvath2013 exhibited a higher ICC (0.9773) than CpGPT_{2M}-Age (0.9039), the dynamic range of PCHorvath2013's predictions was notably small — only about 0.7 years — whereas CpGPT_{2M}-Age spanned approximately 12 years (Supplementary Figure 18). As a result, CpGPT achieved a lower mean squared error (86.42) relative to these clocks (AltumAge MSE of 362.88, Horvath2013 MSE of 393.19, and PCHorvath2013 MSE of 136.50), indicating that the model not only preserves reproducibility but also captures a wider range of variation in biological age estimates from a small subset of the methylome. This combination of accuracy, robustness, and generalizability underscores the advantages of finetuning CpGPT for specific tasks, paving the way for more precise and flexible analyses of methylation data

2.5 CpGPT strongly predicts mortality and morbidity

To showcase the downstream utility of finetuned CpGPT predictions, we developed a mortality predictor named CpGPTGrimAge3 based on the CpGPT_{2M}-Proteins model. Specifically, we adopted GrimAge's methodology by first predicting protein proxies from methylation data and then applying a Cox proportional hazards model to predict mortality [11, 12] (see Methods). Trained and validated on the Framingham Heart Study (FHS) cohort, and tested on the Health and Retirement Study (HRS) cohort, we evaluated this model's ability to predict mortality through association analyses with time to mortality in both the training (N = 3,658, deaths n = 319) and test cohorts (N = 3,941, n = 443), adjusting Cox models for age and assessing predictor coefficient z-scores (Figures **??**).

In the training cohort, CpGPTGrimAge3 achieved a z-score of 14.67, outperforming six other well-known clocks, including GrimAge2 (14.078), SystemsAge (13.762), GrimAge (13.028), PCGrimAge (12.705), PCPhenoAge (10.665), and DunedinPACE (11.028). In the HRS test cohort, CpGPTGrimAge3 was the second-best with a z-score of 14.854, trailing only SystemsAge (15.106), which was trained with HRS data, potentially biasing its superior performance.

Further analysis involved stratifying participants into quartiles based on age-residualized CpGPTGrimAge3 risk scores and comparing survival curves between the most age-decelerated (lowest quartile) and age-accelerated (highest quartile) groups (Figures **??**). In both the training and test cohorts, survival curves differed significantly between these groups ($p < 0.0001$), demonstrating CpGPTGrimAge3's capability to stratify individuals based on biological aging profiles.

Additionally, we explored CpGPTGrimAge3's correlations with 13 prevalent disease categories using baseline and follow-up data from HRS. This included conditions like hypertension, diabetes, cancer, and more diseases (Figure **??**). The model showed the strongest or near-strongest scaled z-score associations across almost all disease endpoints at baseline and over time. Notably, GrimAge2 protein proxies and CpGPT_{2M}-Proteins protein proxies, forming features of CpGPTGrimAge3, were highly linked with disease risks, particularly for cancer (Supplementary Figures **??**).

Overall, these findings affirm that CpGPTGrimAge3 predicts a range of diseases and functional morbidity outcomes across diverse cohorts, surpassing other epigenetic clocks for all-cause mortality. Its associations with neurodegenerative diseases, cardiovascular conditions, and physical function measures emphasize its potential in identifying individuals at risk for significant health decline, enhancing its utility in both clinical and research contexts.

3 Discussion

In this study, we introduce CpGPT, a foundation model specifically designed for DNA methylation analysis. By leveraging the comprehensive dataset *CpGCorpus* comprising

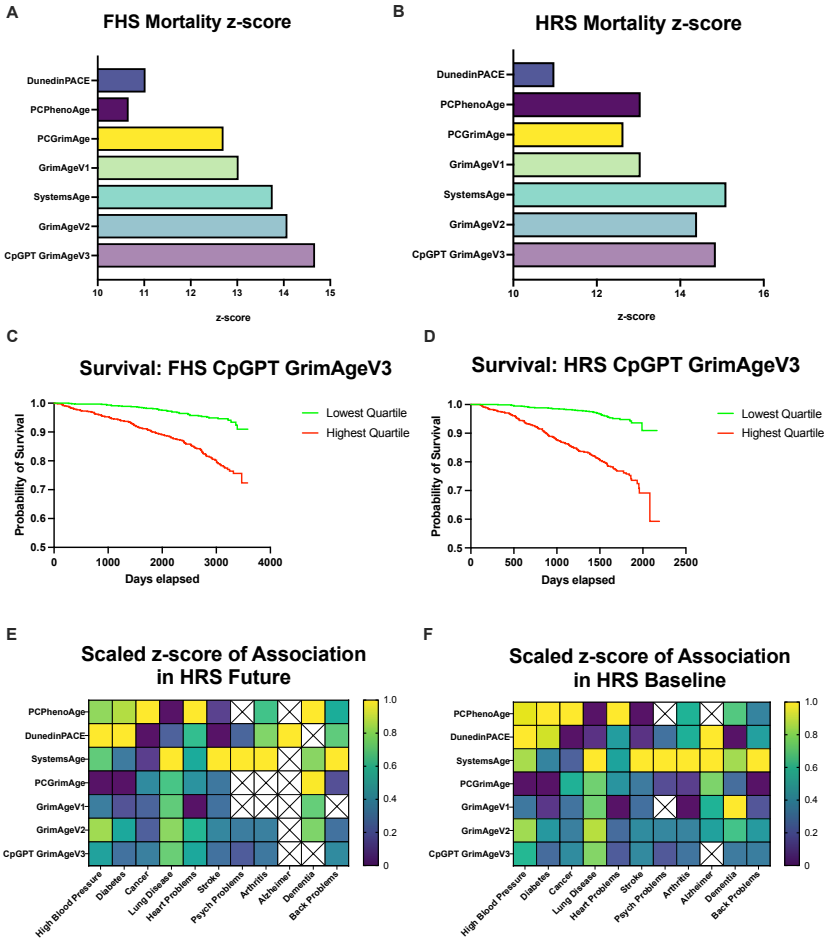


Figure 5: Predictive performance of CpGPT for mortality across FHS and HRS cohorts. (a-b) Z-scores for mortality prediction, representing the strength of association between CpGPTGrimAge3 and mortality (and six other well-known clocks) in training and test cohorts, with higher scores reflecting stronger associations. (c-d) Kaplan-Meier survival curves comparing the most age-decelerated (lowest quartile) and most age-accelerated (highest quartile) groups. 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. (e-f) Six commonly used clocks and CpGPTGrimAge3's associations with 13 common disease categories using baseline and future follow-up data from the HRS cohort. Disease categories included hypertension, diabetes, cancer, lung disease, heart problems, stroke, psychiatric disorders, arthritis, Alzheimer's disease, dementia, and back pain.

over 150,000 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 [41, 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 [42, 43]. 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 [10] 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 [41]. Our evaluation using the Hannum dataset simulated the scenario of reconstructing 450k array probes from MSA array data [29]. 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.

The superior performance of CpGPT in predicting mortality and chronological age, as evidenced by its top placement in the Biomarkers of Aging Challenge [18, 44], signifies its potential impact on aging research. Additionally, in our evaluation, CpGPTGrimAge3 demonstrated the highest z-score when adjusted by chronological age in the HRS cohort out of any clock that did not use it as part of its training data. Moreover, we observed strong predictions of baseline and future disease status using the model and its input protein proxies, particularly for cancer. These findings demonstrate the model's ability to capture biologically meaningful variations in aging and mortality risk.

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 [45]. 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 [46, 47, 48], 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 foundation models in 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 database named *CpGCorpus*. This database aggregates publicly available DNA methylation data from the Gene Expression Omnibus (GEO) [49], encompassing a total of 2,042 studies, 155,151 human samples, and 1,331,566,900 unique tokens (number of samples times the number of CpG sites). 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* [50] 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 splits To assess the performance and generalization capabilities of CpGPT, we partitioned *CpGCorpus* into training, validation, and test sets:

- **Training set:** Consists of 150,719 samples from 1,986 studies (median 30 samples per dataset). This set was used for model training. One percent of the available genomic locations were masked during pretraining.
- **Validation set:** Includes 174 samples from 10 studies (median 13 samples per dataset). This set was used for hyperparameter tuning and model selection.
- **Test set:** Contains 4,258 samples from 46 studies (median of 48 samples per study). This set was held during training and validation and used exclusively for the final evaluation of the model.

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**.

4.1.2 AltumAge

The AltumAge dataset was obtained from a public Google Drive repository [16]. This curated dataset includes age-labeled DNA methylation data from multiple studies, some of which make part of the training, validation or testing of *CpGCorpus*. We used a study-level splitting strategy, assigning 70% of the unique studies to training, 10% to validation, and 20% to testing.

4.1.3 Cancer

The cancer dataset was downloaded from a public Google Drive repository [16]. It contains DNA methylation data from normal and adjacent cancerous tissue for multiple tissue types from several studies. Any beta values outside the 0–1 range were set to `NaN`. A study-level 50%-10%-40% train-val-test split was applied to create training, validation, and test sets.

4.1.4 EPIC-Mammalian comparison

The EPIC-Mammalian dataset contains paired measurements from the EPIC and Horvath mammalian arrays for human blood samples. A 50%-10%-40% train-val-test split was used.

4.1.5 Biomarkers of Aging Consortium

We accessed the Biomarkers of Aging Consortium (GSE246337) dataset via the *biolearn* Python package. Paired DNA methylation and protein measurements were taken from the same individuals, and proteins with duplicate measurements were averaged. For CpG site selection, we chose the superset of CpG sites used by GrimAge, GrimAge2, DunedinPACE, DNAmPhenoAge, and HRSInchPhenoAge [**lu2019**, **lu2022**, 14, 13, 51]. The resulting dataset was divided into training (70%), validation (10%), and test (20%) splits.

4.1.6 Mammalian

The Mammalian dataset, from GSE223748, captures DNA methylation data for over 100 mammalian species. However, we limited ourselves to the 55 species with the appropriate genomic mapping of the mammalian probes [52]. For data splits, human (*homo_sapiens*) and mouse (*mus_musculus*) data were retained in the training set; remaining species were split into training (80%), validation (10%), and test (10%). Two percent of the probes for training were reserved for out-of-distribution testing.

4.1.7 Mortality

We utilized the Framingham Heart Study (FHS) and the Health and Retirement Study (HRS) cohorts to develop and validate CpGPTGrimAge3, with harmonized methylation data and clinical information. We excluded beta values outside the 0–1 range.

4.1.8 OSKM

The OSKM dataset (GSE54848) includes 450k DNA methylation data from a reprogramming timecourse experiment of fibroblasts with cell sorting over a period of 28 days.

4.1.9 Hannum

The Hannum dataset (GSE40279) contains blood DNA methylation data from the 450k array. The dataset was split with 80%-10%-10% for training, validation, and testing.

4.1.10 RRBS atlas

The Human RRBS Atlas (GSE233417) captures methylation profiles using Reduced Representation Bisulfite Sequencing (RRBS). Original hg19 coordinates were converted to hg38, retaining only CpGs that correctly mapped to the updated reference and that overlapped the genomic loci from pretraining. We used a 50%-10%-40% train-val-test split. This ensured a balanced representation of tissue types across all subsets and facilitated tissue-specific analyses.

4.1.11 sciMETv3

The sciMETv3 dataset (GSE273592) comprises single-cell methylation profiles generated with both Illumina and Ultima sequencers. We used the Illumina dataset. Only cells with at least 1,000 measured beta values from the genomic loci from pretraining were kept. The dataset was subsequently split into training (50%), validation (10%), and test (40%) sets.

4.1.12 Reliability

The reliability dataset (GSE55763) was used to calculate intraclass correlation coefficients. The dataset comprises 36 whole blood samples with two technical replicates.

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 using the annotation from Ensembl version 112 [53]. 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 extraction:** 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 [54, 19, 55], 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

We employed a dual positional encoding strategy that allows the model to uniquely identify each CpG location and understand CpG positions relative to each other:

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 in base pairs 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) [20]:

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

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 [21], 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:

$$\hat{\mathbf{H}}_{i-1} = \mathrm{RMSNorm}(\mathbf{H}_{i-1}), \tag{9}$$

$$\mathbf{A}_i = \mathbf{H}_{i-1} + \mathrm{MultiHeadAttention}(\hat{\mathbf{H}}_{i-1}, \hat{\mathbf{H}}_{i-1}, \hat{\mathbf{H}}_{i-1}), \tag{10}$$

$$\tilde{\mathbf{A}}_i = \mathrm{RMSNorm}(\mathbf{A}_i), \tag{11}$$

$$\mathbf{H}_i = \mathbf{A}_i + \mathrm{SwiGLU}(\mathrm{MLP}(\tilde{\mathbf{A}}_i)), \tag{12}$$

where $\mathbf{H}_0 = [\mathbf{c}_{\mathrm{cls}}; \mathbf{C}]$, and $\mathbf{c}_{\mathrm{cls}} \in \mathbb{R}^{1 \times d_{\mathrm{emb}}}$ 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:

M value decoder Predicts methylation M values for CpG sites:

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

where g_M is a projection function, $\mathbf{h}_{\mathrm{cls}}$ is the final hidden state of the classification token, and $\langle \cdot, \cdot \rangle$ denotes the inner product. We opted for M values for the prediction as they have better statistical properties compared to beta values and there is no need for a sigmoid after the last hidden layer, which can cause gradient issues.

Uncertainty decoder Estimates the uncertainty of the M value predictions:

$$\hat{\sigma}_M = \langle g_{\mathrm{unc}}(\mathbf{S}), \mathbf{h}_{\mathrm{cls}} \rangle, \tag{14}$$

where g_{unc} is a projection function.

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

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

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

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).

Table 1: Comparison of CpGPT Model Versions

Parameter	$\text{CpGPT}_{100\text{M}}$	$\text{CpGPT}_{2\text{M}}$
Total Parameters	101M	2.5M
d_embedding	512	128
d_hidden	512	128
n_attention_heads	16	8
n_layers	32	8
dropout	0.01	0.01
n_mlp_blocks	3	2
dna_context_len	2001	2001
max_length	10000	5000

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:**
   - M value prediction $\hat{M}$.
   - Uncertainty estimation $\hat{\sigma}_M$.
   - 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:

Methylation losses A mean absolute error loss is used both for the predicted M 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 M values when compared to the real distribution.

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

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

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

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 [4] 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.

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 Chain-of-thought inference

To enhance prediction accuracy during inference, we implemented a novel chain-of-thought (CoT) approach for CpGPT. This iterative refinement procedure allows the model to progressively improve its methylation predictions by incorporating its own confident estimates into subsequent iterations.

For a set of n thinking steps with a fixed step size k, the procedure operates as follows:

1. **Initialization:** Begin with an initial batch $\mathcal{B}_0 = \{\beta_0, \mathbf{E}_0, \mathbf{c}_0, \mathbf{p}_0\}$ containing a small set of known CpG sites.
2. **Prediction generation:** For each thinking step $i \in \{1, 2, \dots, n\}$:
   - Calculate the candidate pool size $s_i = \frac{k}{q}$, where $q \in (0, 1)$ is the predetermined uncertainty quantile threshold.
   - Sample a random subset of s_i genomic locations $\mathcal{L}_i = \{l_1, l_2, \dots, l_{s_i}\}$ with available DNA LLM embeddings.
   - Extract DNA embeddings $\mathbf{E}_{\mathcal{L}_i}$ for these locations.
   - Generate methylation predictions and uncertainties:
   $$\hat{\beta}_i, \hat{\sigma}_{\beta,i} = f_{\text{CpGPT}}(\mathcal{B}_{i-1}, \mathbf{E}_{\mathcal{L}_i}) \tag{28}$$
3. **Confidence-based filtering:** Compute the uncertainty threshold that selects approximately k predictions:
   $$\mathcal{C}i = \{j \mid \hat{\sigma}_{\beta,\mathcal{L}i,j} \leq Q_{\hat{\sigma}_{\beta,\mathcal{L}_i}}(q)\} \tag{29}$$
   where $Q_{\hat{\sigma}_{\beta,\mathcal{L}_i}}(q)$ represents the q-th quantile of the uncertainty distribution $\hat{\sigma}_{\beta,\mathcal{L}i}$. The size of $\mathcal{C}_i$ will be approximately k sites, ensuring a consistent step size across iterations.
4. **Batch augmentation:** Update the batch with confident predictions for the next iteration:
   $$\mathcal{B}_i = \mathcal{B}_{i-1} \cup \{\hat{\beta}_{i,\mathcal{C}_i}, \mathbf{E}_{\mathcal{L}_i,\mathcal{C}_i}, \mathbf{c}_{\mathcal{L}_i,\mathcal{C}_i}, \mathbf{p}_{\mathcal{L}_i,\mathcal{C}_i}\} \tag{30}$$
5. **Sorting:** Reorder the augmented batch according to the chosen sorting strategy to maintain genomic context:
   $$\mathcal{B}_i = \text{Sort}(\mathcal{B}_i) \tag{31}$$
   where the sorting strategy preserves the local genomic context as described in the input representation section.
6. **Final prediction:** After n thinking steps, a final forward pass is made using the target locations to generate the final predictions:
   $$\hat{\beta}_{\text{final}}, \hat{\sigma}_{\beta,\text{final}} = f_{\text{CpGPT}}(\mathcal{B}_n, \mathbf{E}_{\text{target}}) \tag{32}$$

This approach leverages the model's uncertainty estimates to selectively incorporate its most confident predictions into subsequent iterations. Crucially, it maintains a fixed step size k across iterations while adapting the initial candidate pool size $s_i = \frac{k}{q}$ to ensure that after filtering, approximately k sites are added at each step. The uncertainty quantile threshold q determines the strictness of the filtering process — smaller values (e.g., 0.05) result in more selective filtering and require larger initial candidate pools, while larger values (e.g., 0.25) are less selective but require sampling fewer initial candidates.

Our experimental results show that this iterative refinement process leads to progressive improvements in prediction accuracy, with diminishing returns and even detrimental effects after the the chain-of-thought is approximately the same size as the original input — a phenomenon similar to the "forgetting" effect observed in traditional chain-of-thought reasoning [38].

4.6 Mortality predictor

To develop CpGPTGrimAge3, we started with the FHS dataset for both training and validation. We calculated all GrimAge2 and $\text{CpGPT}_{2\text{M}}$-Proteins proxies within the dataset. For feature selection, a 5-fold cross-validation with a greedy search was utilized. Beginning with age plus a single proxy from all available proxies, proxies were added one by one until the validation z-score, corrected by age, no longer improved. This process resulted in the selection of seven GrimAge2 proxies and 23 $\text{CpGPT}_{2\text{M}}$-Proteins proxies. These proxies, along with age, compose the input features. Finally, for the final prediction in unit of time, the Cox risk score is converted back to a year scale to align with the mean and standard deviation of the training cohort.

4.7 Statistics

4.7.1 Energy distance

For the comparison between the energy distance (E-distances) between the different models, a two-sided Mann Whitney U test was done to get the p values. For the energy distance calculation, the python package *scperturb* was used.

4.7.2 Intraclass correlation

Intraclass correlation coefficient was calculated with the *pengouin* python package.

4.7.3 UMAP

To compute the UMAP, we initially reduce the data using the first 50 principal components from a PCA. Subsequently, the UMAP is determined using 30 neighbors via the *umap* python package.

4.7.4 ElasticNet and MLP baselines

4.7.5 Mortality association prediction

We conducted association analyses using time-to-mortality data across one training cohorts and one test cohor. 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).

4.7.6 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 CpGPTGrimAge3 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. Statistical significance of the differences between survival curves was assessed using the log-rank test, implemented in Prism 9.

4.7.7 Morbidity association analysis

To evaluate the association of CpGPTGrimAge3 scores with morbidity outcomes, we performed ROC analyses for 13 disease outcomes 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). All statistical analyses were performed using R (version 4.4.0).

5 Code and data availability

CpGCorpus is publicly available for access in the S3 bucket `s3://cpgpt-lucascamillo-public/data/cpgcorpus/raw`.

The code and the tutorials are available in our GitHub repository under `http://github.com/lcamillo/CpGPT`. CpGPTGrimAge3 is available on pyaging `http://github.com/rsinghlab/pyaging`.

6 Acknowledgements

We would like to thank the Zhou lab for their comprehensive, publicly available annotation of the Illumina methylation arrays [52]. 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

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] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

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

[4] 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.

[5] 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.

[6] 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.

[7] 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).

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[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] 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. "Nucleotide Transformer: building and evaluating robust foundation models for human genomics". In: *Nature Methods* (2024), pp. 1–11.

[20] 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.

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

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

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

[24] Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. "scPerturb: harmonized single-cell perturbation data". In: *Nature Methods* 21.3 (2024), pp. 531–540.

[25] 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.

[26] Ruth V Nichols, Lauren E Rylaarsdam, Brendan L O'Connell, Zohar Shipony, Nika Iremadze, Sonia N Acharya, and Andrew C Adey. "Atlas-scale single-cell DNA methylation profiling with sciMETv3". In: *Cell Genomics* 5.1 (2025).

[27] Mari Ohnuki, Koji Tanabe, Kenta Sutou, Ito Teramoto, Yuka Sawamura, Megumi Narita, Michiko Nakamura, Yumie Tokunaga, Masahiro Nakamura, Akira Watanabe, et al. "Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential". In: *Proceedings of the National Academy of Sciences* 111.34 (2014), pp. 12426–12431.

[28] Diljeet Gill, Aled Parry, Fátima Santos, Hanneke Okkenhaug, Christopher D Todd, Irene Hernando-Herraez, Thomas M Stubbs, Inês Milagre, and Wolf Reik. "Multi-omic rejuvenation of human cells by maturation phase transient reprogramming". In: *Elife* 11 (2022), e71624.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] Adriana Arneson, Amin Haghani, Michael J Thompson, Matteo Pellegrini, Soo Bin Kwon, Ha Vu, Emily Maciejewski, Mingjia Yao, Caesar Z Li, Ake T Lu, et al. "A mammalian methylation array for profiling methylation levels at conserved sequences". In: *Nature communications* 13.1 (2022), p. 783.

[35] 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.

[36] 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.

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[38] Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. "When More is Less: Understanding Chain-of-Thought Length in LLMs". In: *arXiv preprint arXiv:2502.07266* (2025).

[39] Mahdi Moqri, Jesse R Poganik, Chiara Herzog, Kejun Ying, Qingwen Chen, Mehrnoosh Emamifar, Alexander Tyshkovskiy, Eames W Alec, Jure Mur, Benyamin Matei-Dediu, et al. "Integrative epigenetics and transcriptomics identify aging genes in human blood". In: *bioRxiv* (2024), pp. 2024–05.

[40] Benjamin Lehne, Alexander W Drong, Marie Loh, Weihua Zhang, William R Scott, Sian-Tsung Tan, Uzma Afzal, James Scott, Marjo-Riitta Jarvelin, Paul Elliott, et al. "A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies". In: *Genome biology* 16 (2015), pp. 1–12.

[41] 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.

[42] 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.

[43] 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).

[44] Kejun Ying, Seth Paulson, Julian Reinhard, Lucas Paulo de Lima Camillo, Jakob Träuble, Stefan Jokiel, Dane Gobel, Chiara Herzog, Jesse R Poganik, Mahdi Moqri, et al. "An Open Competition for Biomarkers of Aging". In: *bioRxiv* (2024).

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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, et al. "NCBI GEO: archive for functional genomics data sets—update". In: *Nucleic acids research* 41.D1 (2012), pp. D991–D995.

[50] 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.

[51] Morgan E. Levine, Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, Andrea A. Baccarelli, James D. Stewart, Yun Li, Eric A. Whitsel, James G. Wilson, Alex P. Reiner1, Abraham Aviv1, Kurt Lohman, Yongmei Liu, Luigi Ferrucci, and Steve Horvath. "An epigenetic biomarker of aging for lifespan and healthspan". In: *Aging* 10.4 (Apr. 2018), pp. 573–591. ISSN: 19454589. DOI: 10.18632/aging.101414.

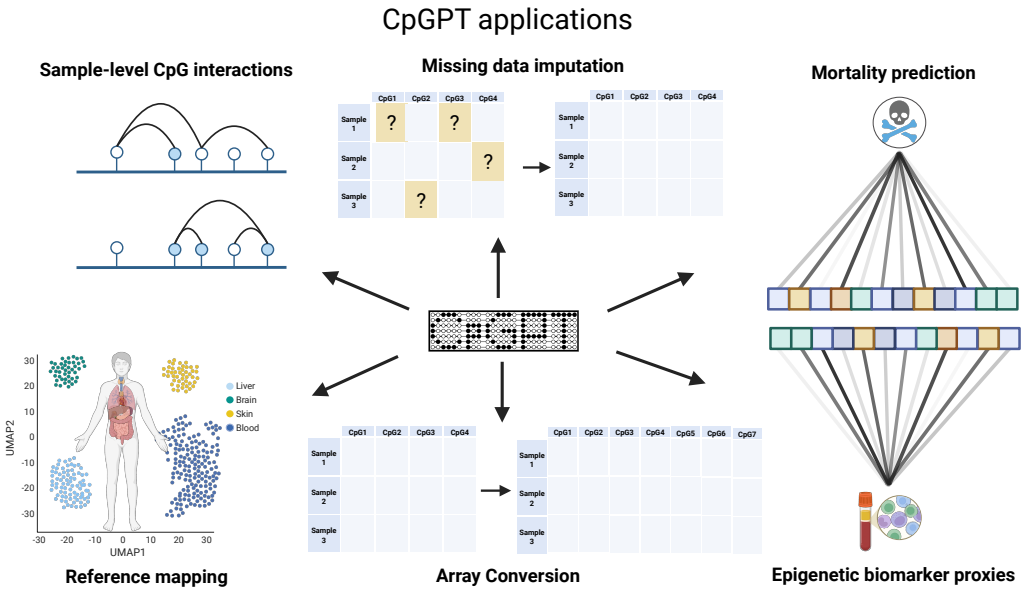
[52] Wanding Zhou, Peter W Laird, and Hui Shen. "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes". In: *Nucleic acids research* 45.4 (2017), e22–e22.

[53] Sarah C Dyer, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Vianey Paola Barrera-Enriquez, Arne Becker, Ruth Bennett, Martin Beracochea, Andrew Berry, et al. "Ensembl 2025". In: *Nucleic acids research* 53.D1 (2025), pp. D948–D957.

[54] 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).

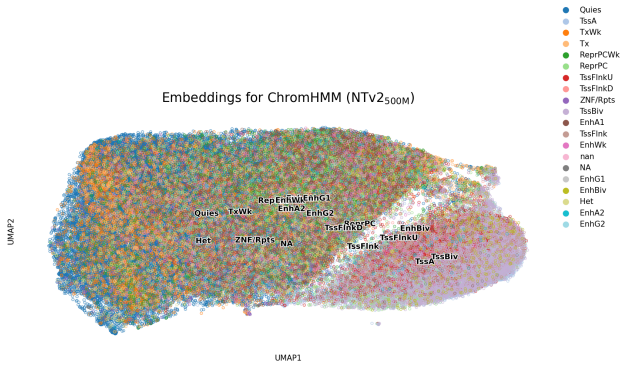
[55] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. "Dnabert-2: Efficient foundation model and benchmark for multi-species genome". In: *arXiv preprint arXiv:2306.15006* (2023).

Supplementary Figures

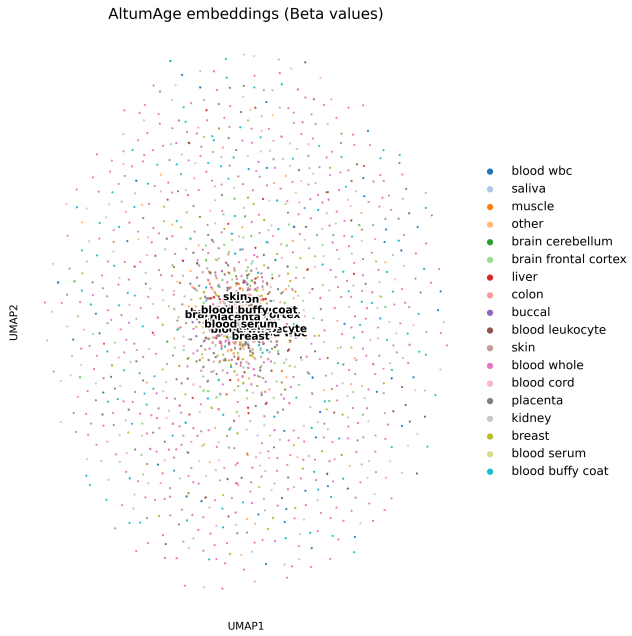


Supplementary Figure 1: Schematic showing some CpGPT applicatios. CpGPT can perform a series of downstream tasks, including zero-shot imputation, array conversion, reference mapping, and age and mortality predictions.

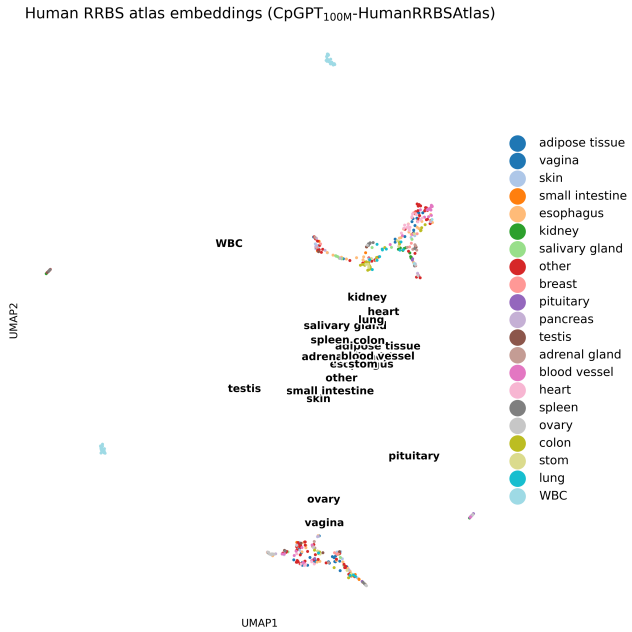
Supplementary Figure 2: UMAP of the CpGPT$_{100M}$ embeddings of the training set of CpGCorpus colored by dataset.



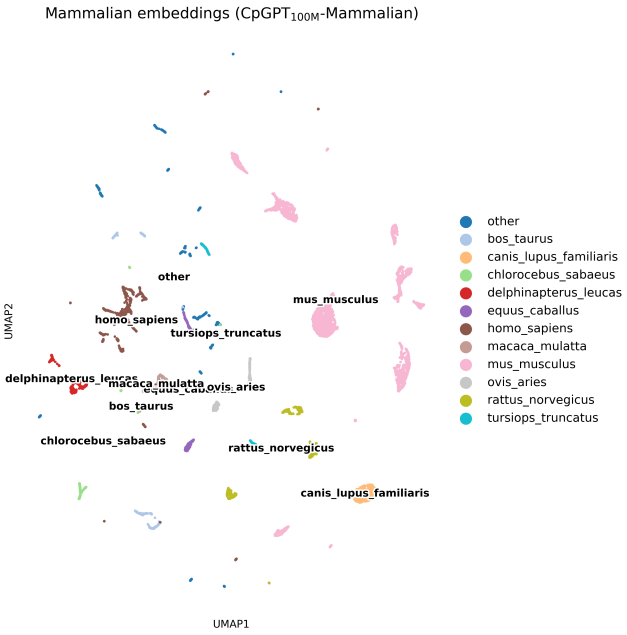
Supplementary Figure 3: UMAP of the locus embeddings from NTv2$_{500M}$ colored by ChromHMM genomic annotation



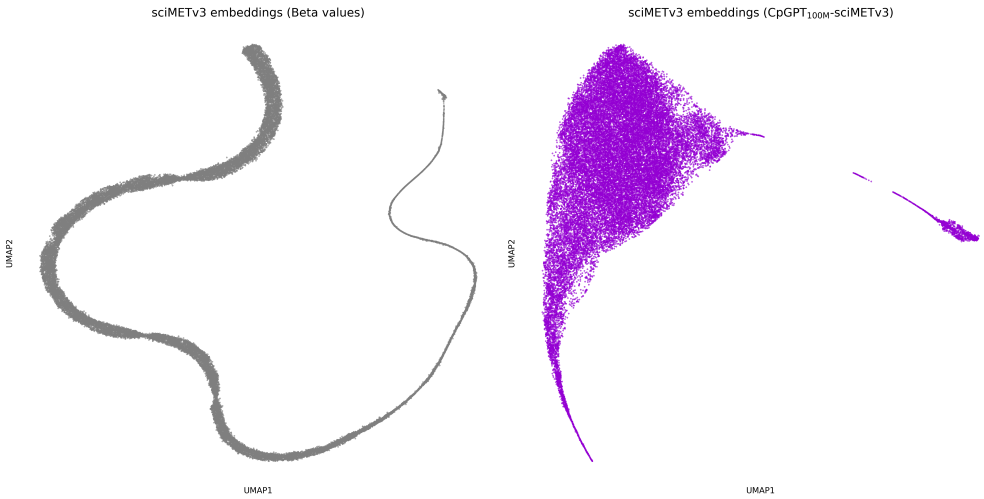
Supplementary Figure 4: UMAP of the beta values for the AltumAge dataset colored by tissue type.



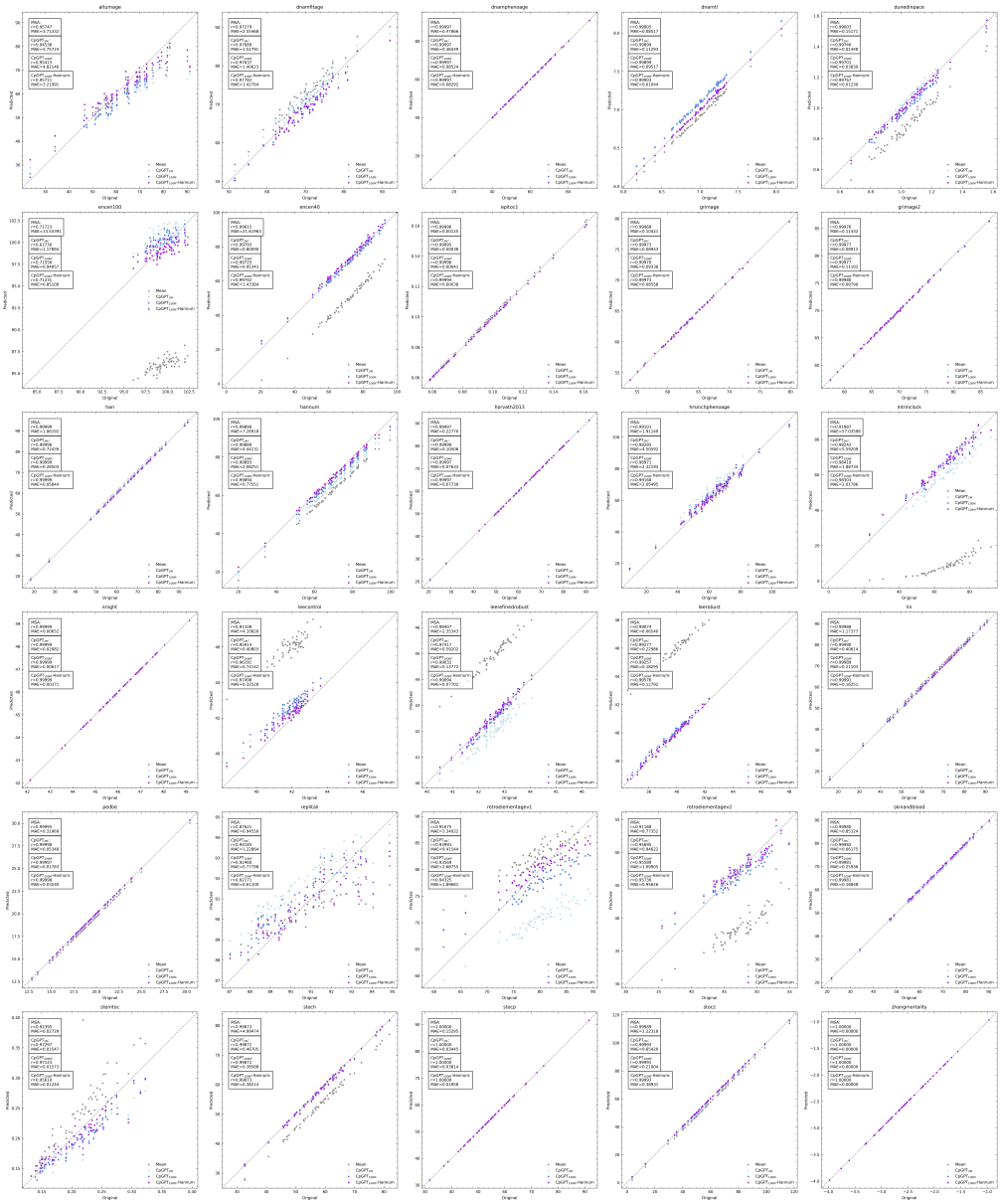
Supplementary Figure 5: UMAP of the $\text{CpGPT}_{100\text{M}}$-HumanRRBSAtlas samples embeddings of the human RRBS dataset colored by tissue type.



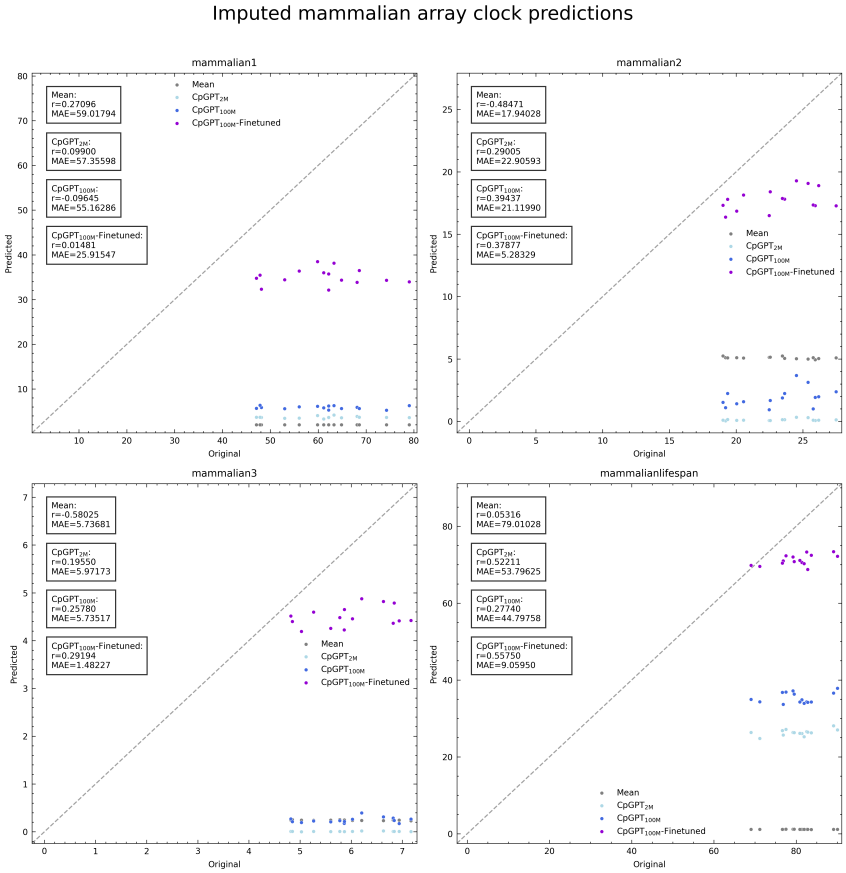
Supplementary Figure 6: UMAP of the $\text{CpGPT}_{100\text{M}}$-Mammalian samples embeddings of the mammalian dataset colored by species.



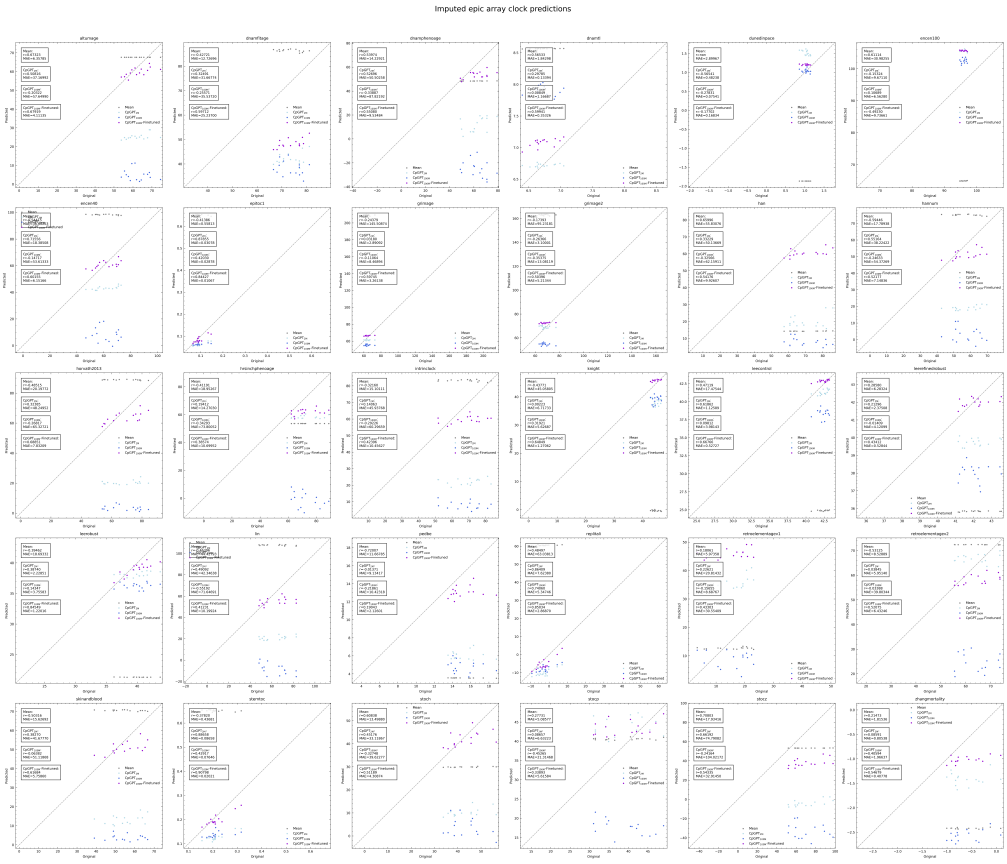
Supplementary Figure 7: UMAP of the beta values (left) and $\text{CpGPT}_{100\text{M}}$-sciMETv3 (right) samples embeddings of the single-cell sciMETv3 dataset.



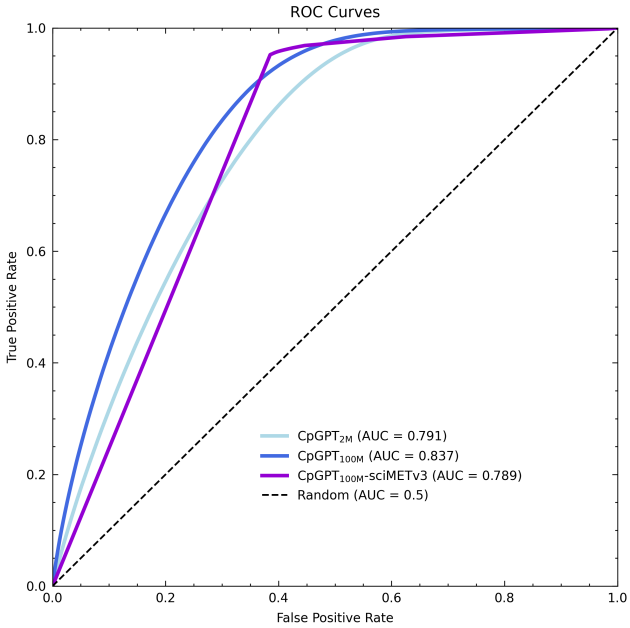
Supplementary Figure 8: Scatterplot of 30 epigenetic clocks predicted based on different imputation methods for the 450k probes missing from the MSA array in the Hannum dataset with a mean baseline, $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and $\text{CpGPT}_{100\text{M}}$-Hannum.



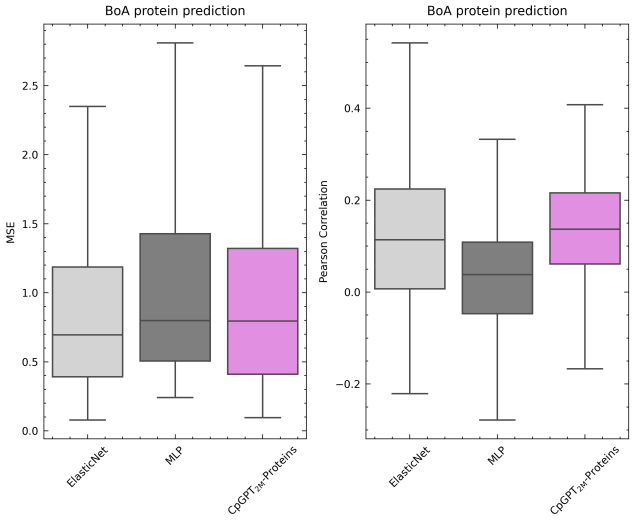
Supplementary Figure 9: Scatterplot of four epigenetic clocks applied to the reconstructed mammalian array from EPIC data with a mean baseline, $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and $\text{CpGPT}_{100\text{M}}$-EpicMammal.



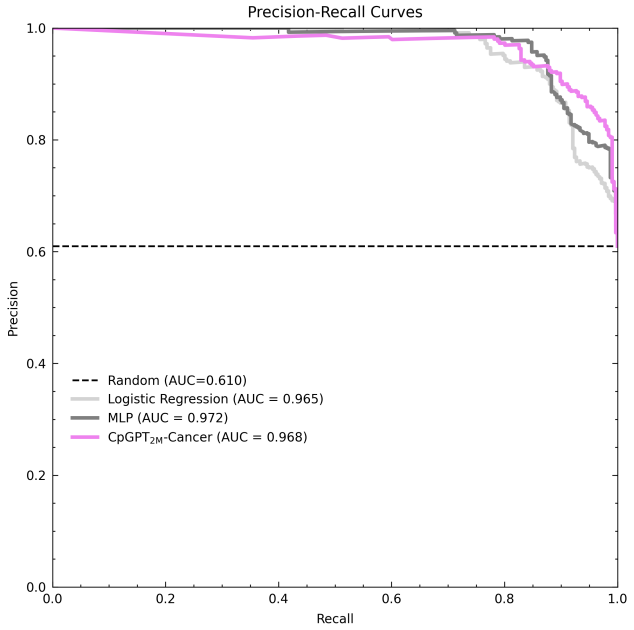
Supplementary Figure 10: Scatterplot of 30 epigenetic clocks applied to the reconstructed EPIC array from mammalian array data with a mean baseline, $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and $\text{CpGPT}_{100\text{M}}$-EpicMammal.



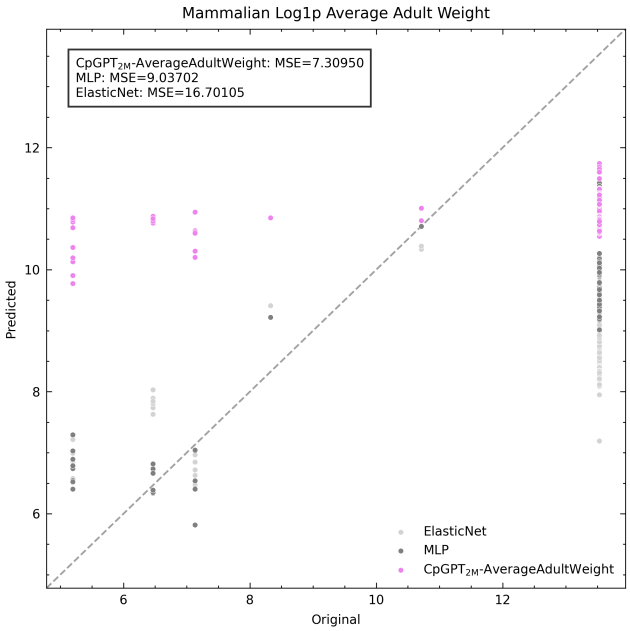
Supplementary Figure 11: Receiver-operator curve of the reconstruction of the single-cell sciMETv3 dataset with a random baseline, $\text{CpGPT}_{2\text{M}}$, $\text{CpGPT}_{100\text{M}}$, and $\text{CpGPT}_{100\text{M}}$-sciMETv3.



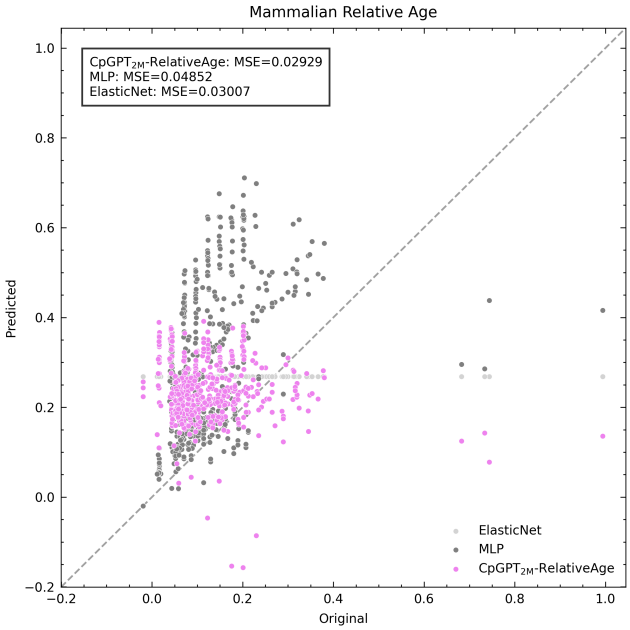
Supplementary Figure 12: Boxplot of the mean squared error (MSE) and Pearson correlation coefficients for the prediction of 322 proteins from blood methylation of the Biomarkers of Aging Consortium dataset comparing ElasticNet, MLP, and CpGPT_{2M}-Proteins.



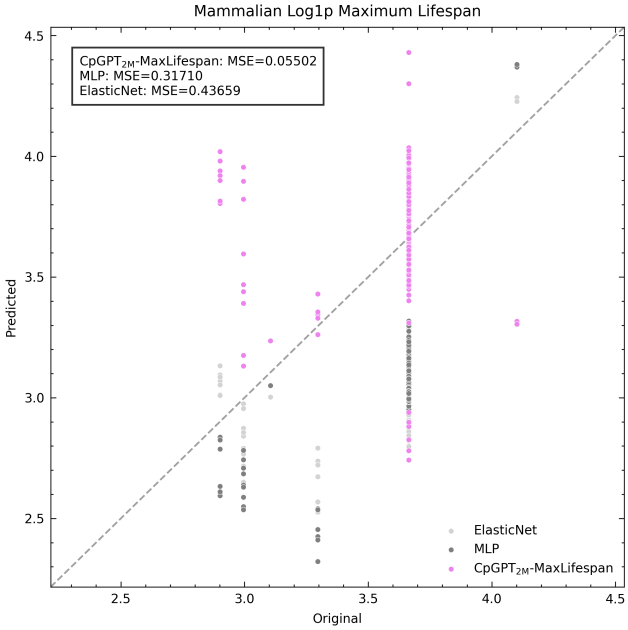
Supplementary Figure 13: Precision-recall curve for the cancer prediction in the AltumAge dataset comparing logistic regression, MLP, and CpGPT_{2M}-Cancer



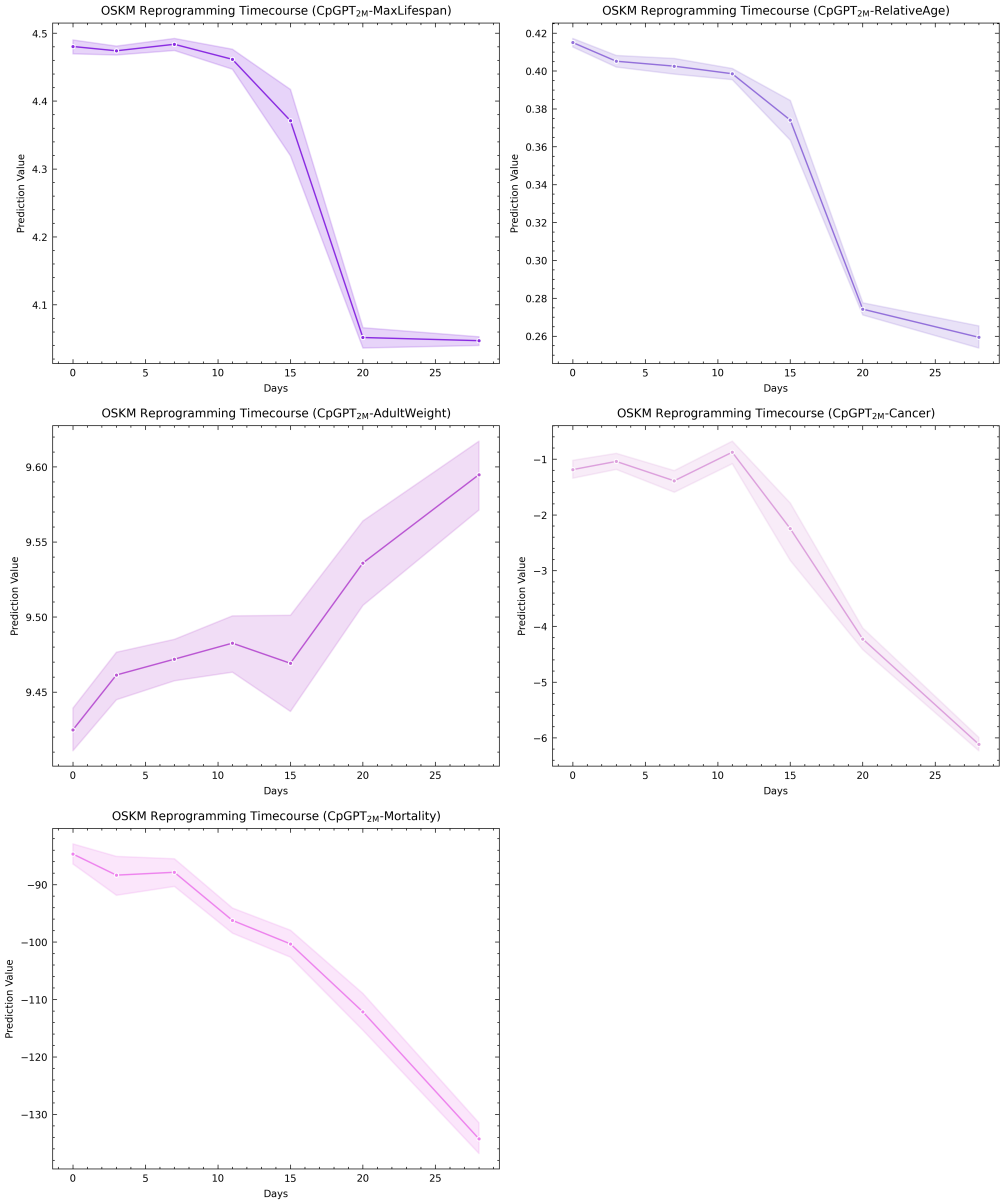
Supplementary Figure 14: Scatterplot for the prediction of the log1p average adult weight comparing ElasticNet, MLP, and CpGPT$_{2M}$-AverageAdultWeight.



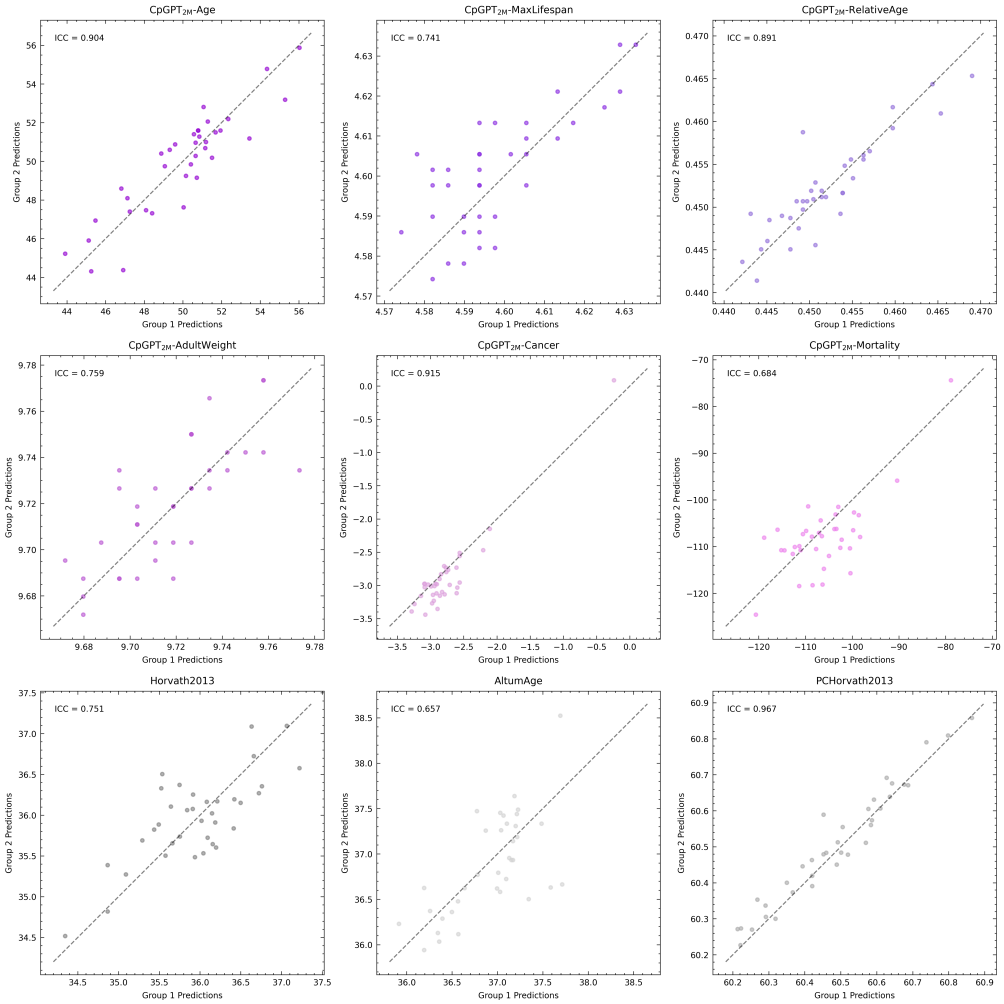
Supplementary Figure 15: Scatterplot for the prediction of the relative age comparing ElasticNet, MLP, and CpGPT$_{2M}$-RelativeAge.



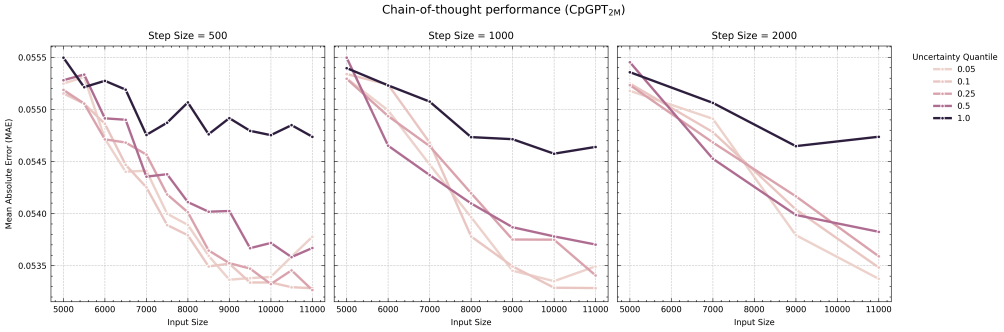
Supplementary Figure 16: Scatterplot for the prediction of the log1p maximum lifespan comparing ElasticNet, MLP, and CpGPT$_{2M}$-MaxLifespan.



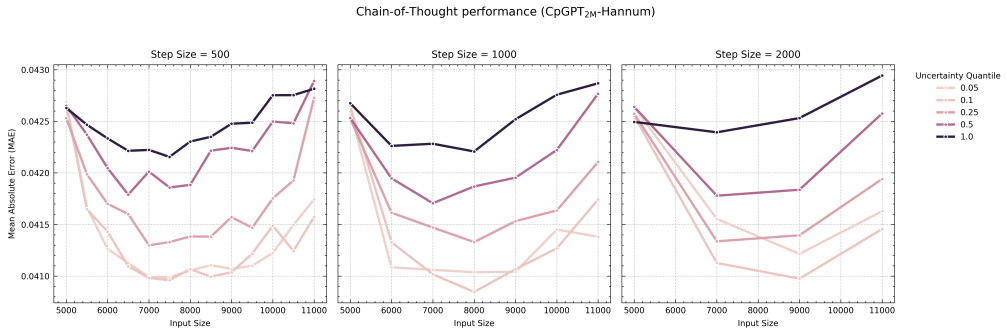
Supplementary Figure 17: Lineplots showing the 95% confidence interval of five CpGPT models in the OSKM dataset with 10 different random input subsets of CpGs.



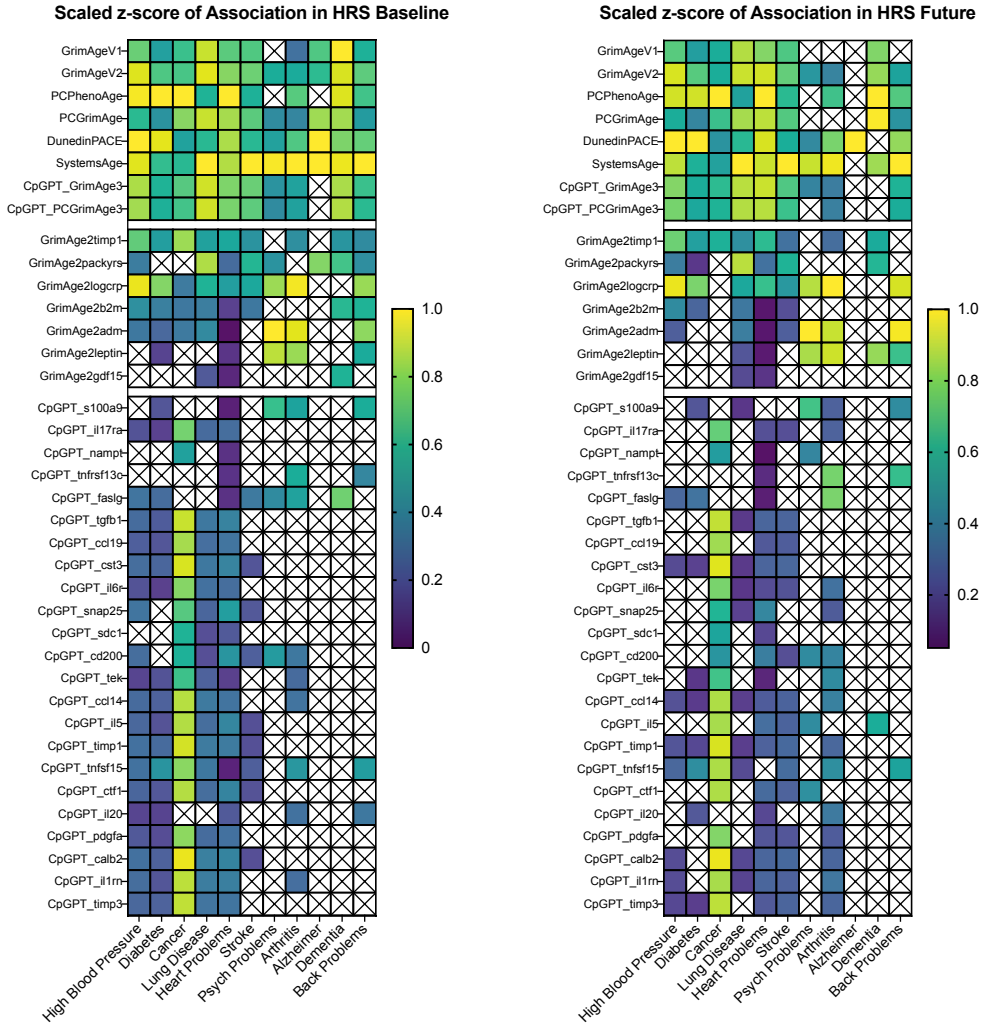
Supplementary Figure 18: Scatterplot showing the predictions for each technical replicate in the reliability dataset with different CpGPT models and three epigenetic clocks.



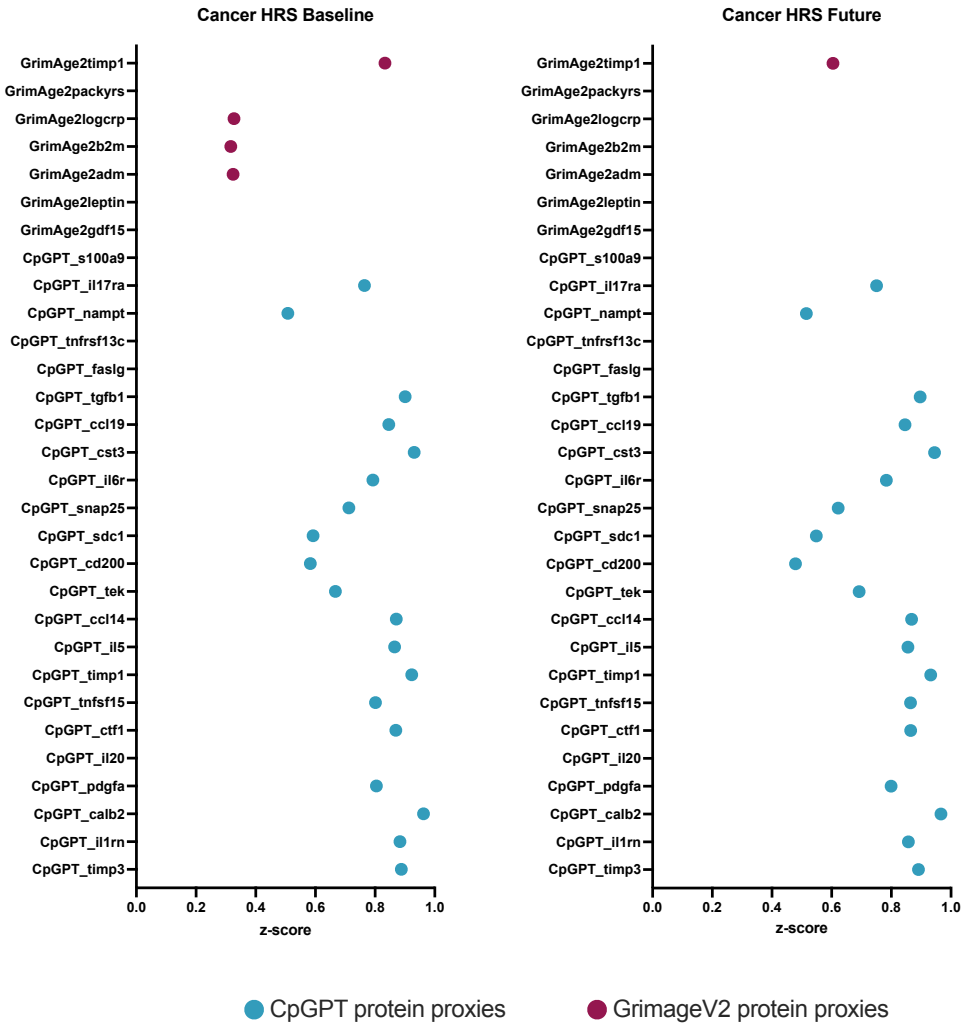
Supplementary Figure 19: Lineplot showing the chain-of-thought performance of $\text{CpGPT}_{2\text{M}}$ in reconstructing non-MSA 450k probes in the Hannum dataset based on different step sizes and uncertainty thresholds.



Supplementary Figure 20: Lineplot showing the chain-of-thought performance of $\text{CpGPT}_{2\text{M}}$-Hannum in reconstructing non-MSA 450k probes in the Hannum dataset based on different step sizes and uncertainty thresholds.



Supplementary Figure 21: Heatmap showing the scaled z-scores of the association of several well-known epigenetic clocks, GrimAge2 protein proxies and $\text{CpGPT}_{2\text{M}}$-Proteins proxies with baseline and future risk of diseases in the HRS cohort.



Supplementary Figure 22: Scatterplot showing the scaled z-scores of the association of several GrimAge2 protein proxies and $\text{CpGPT}_{2\text{M}}$-Proteins proxies with baseline and future cancer risk in the HRS cohort.