Appendix

A. Abbreviation List of Cancer Types

We use the following abbreviations for the selected cancer types in this paper:

1) BRCA: Breast invasive carcinoma;
2) COADREAD: Colorectal adenocarcinoma;
3) HNSC: Head and neck squamous cell carcinoma;
4) LUAD: Lung adenocarcinoma;
5) LUSC: Lung squamous cell carcinoma;
6) KIPAN: Pan-kidney cancer;
7) STES: Stomach and esophageal carcinoma;
8) THCA: Thyroid carcinoma;
9) BLCA: Bladder urothelial carcinoma;
10) PAAD: Pancreatic adenocarcinoma;
11) SKCM: Skin cutaneous melanoma;
12) TGCT: Testicular germ cell tumor.

B. Phenotypic Features Definitions

We extract the following variables from LinkedOmics [15] as phenotypic features for our constructed dataset (original column names from the raw CSV file are shown in parentheses):

1) Biological Sex (`sex`): The biological sex of the patient.
2) Radiation Therapy (`radiation_therapy`): A binary indicator denoting whether the patient received radiation therapy during the recorded treatment period.
3) Pathologic Overall Stage (`pathologic_stage`): The overall pathological stage based on the TNM system, which summarizes tumor size, lymph node involvement, and metastasis status. A value of "*is*" corresponds to Pathologic Stage 0 (carcinoma in situ), indicating that abnormal cells are present but have not yet invaded surrounding tissues.
4) Pathologic T Stage (`pathology_T_stage`): The "T" component of TNM staging, describing the size and local invasion of the primary tumor.
5) Pathologic N Stage (`pathology_N_stage`): The "N" component of TNM staging, indicating the extent of regional lymph node involvement.
6) Pathologic M Stage (`pathology_M_stage`): The "M" component of TNM staging, representing the presence or absence of distant metastases.

We also incorporate several additional variables from the raw phenotypic data in our analysis:

1) `years_to_birth`: Represents the chronological age of the patient at the time of the clinical visit, used as a proxy for biological age.
2) `cancer_type`: Identifies the specific cancer cohort to which the patient belongs.
3) `overall_survival`: Denotes the total survival time (in days) from diagnosis or clinical visit to either death or last follow-up.
4) `status`: Indicates the patient's vital status at the last follow-up, with 1 representing deceased and 0 representing alive.

C. Details about Feature Pre-processing

1) DNA Methylation Feature Pre-processing: All models in this study are trained on DNA methylation data obtained from the Illumina HM450K array. The original feature identifiers include both gene annotations and CpG site IDs (e.g., `GENE_cg12345678`). To standardize the input features, we remove the gene annotations and retain only the CpG site IDs. We also perform deduplication to eliminate redundant features. These preprocessing steps are applied consistently across all pipelines to ensure fair comparison.

2) Model Specific Pre-processing: For the baseline *epigenetic clock models* (*e.g.*, Horvath, Hannum, PhenoAge, and YingCausAge), we adopt their implementations from the open-source package `biolearn` [39]. Our methylation dataset, provided by LinkedOmics, is reported in a centered format (Beta value $-$ 0.5) according to the platform; therefore, we uniformly shift the DNAm values by $+0.5$ to meet the input requirement of `biolearn`, which expects standard Beta values in the range [0,1].

For the baseline *phenotypic clock models* (*e.g.*, Linear Regression, Random Forest, XGBoost, LightGBM, and tabNN), we adopt their implementations from the open-source package `AutoGluon` [17]. Similar to Equation (5) in EPICAGE, each categorical feature is mapped into monotonically increasing integers via ordinal encoding.

D. Implementation Details about Our EPICAGE

1) DNAm Feature Pre-processing: Beyond the preprocessing steps detailed in Appendix C.1, we impute missing methylation values using the column-wise mean of the training data. Additionally, we apply Z-score normalization using the mean and standard deviation computed exclusively from the training set.

2) Second Feature Selection Step for Epigenetic Clock: BorutaSHAP: All steps of feature selection are conducted within the training set to avoid information leakage. We employ BorutaSHAP on $\mathcal{S}_0$ to obtain the final subset $\mathcal{S}$ with the following hyperparameters.

Hyperparameter	**Value**
model type	LightGBM Regressor
n estimators	100
max depth	7
learning rate	0.05
min gain to split	1e-4
min data in leaf	5
subsample	0.8
colsample bytree	0.8
importance measurement	SHAP
SHAP arguments	pvalue=0.05
max feature count	≤ 492

3) Dimension Reduction of DNAm Data for Skip-connection: Principal Component Analysis (PCA) is performed on the training set $\mathbf{X}^{\text{train}}$, and the top $r = 400$ components are retained. These same $r = 400$ components are applied consistently during inference.

4) Handling Extra Categorical Variable in External Dataset: The external clinical dataset contains one additional categorical variable compared to the internal dataset. To ensure compatibility, all categorical variables were converted to the `category` type and processed using TabPFN's internal one-hot encoder with `handle_unknown="ignore"` [14]. Under this scheme, any unseen categories present in the external dataset are mapped to an all-zero vector in the one-hot representation, effectively treating them as "unknown." As a result, predictions for these cases rely solely on the remaining known features, allowing models trained on the internal dataset to perform inference on the external dataset without additional modification.

E. Biological Validation of Selected CpG Sites

1) Supporting Evidence from Published CpG-Trait Associations: In EWAS Atlas based analysis, we find that 50 out of the 57 CpG sites have been previously reported to be associated with cancer or aging. Table VII summarizes these CpG sites grouped by their reported trait. For presentation purposes, we use the broad "TraitType" column from EWAS analysis results (*i.e.*, "TraitType" value equal to "cancer") to denote CpG sites that are statistically associated with *cancer*. And we use the fine-grained "Trait" column (*i.e.*, "Trait" value contains "aging" or "chronological age") from EWAS results to denote CpG sites that are associated with *aging*.

TABLE VII: Summary of CpG sites by trait.

Trait	CpG site
Cancer	cg01341751, cg01586506, cg03181248, cg05129081, cg05454501, cg07575466, cg11418477, cg14965220, cg22153181, cg26885220
Aging	cg04836038, cg05289022, cg05304393, cg05404236, cg07553761, cg08928145, cg10687131, cg11705975, cg12451153, cg12934382, cg13221458, cg16832267, cg21159778, cg22736354, cg24922090, cg26792755
Cancer & Aging	cg00292135, cg00590036, cg00884093, cg04875128, cg04940570, cg05207048, cg06268694, cg06458239, cg06784991, cg06933824, cg07755735, cg12920180, cg13790603, cg14780466, cg15618978, cg16015712, cg16295725, cg18795809, cg19078576, cg20809087, cg23091758, cg23606718, cg24466241, cg25352836

2) Functional Enrichment of Genes Near Selected CpGs: From our KEGG pathway enrichment analysis, we identify significant enrichment in pathways related to aging and cancer, including the Longevity Regulating Pathway and MicroRNAs in Cancer. Table VIII provides details of these enriched pathways, including the genes located near the selected CpG sites and the corresponding p-values.

TABLE VIII: KEGG pathway analysis results.

KEGG Pathway	Genes	p-val
Longevity regulating pathway	SOD2, IRS2	0.002
MicroRNAs in cancer	TP63, IRS2	0.016

F. Age Acceleration Distribution Among Cancer Types

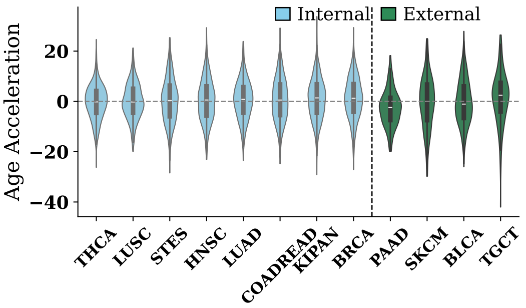

Fig. 3: Age acceleration distribution among cancer types.

Figure 3 illustrates age acceleration distributions. For the internal dataset, medians across most cancer types were close to zero, suggesting minimal bias. However, external data exhibit heterogeneous biases: TGCT showed positive median age acceleration (predicted ages higher than true), whereas PAAD and BLCA showed negative medians (predicted ages lower than true).

G. More Error Analysis

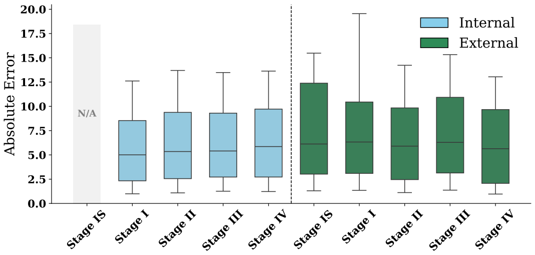

Fig. 4: Absolute error of biological age prediction by stage.

We further stratify errors by pathological stage (Figure 4). For the internal dataset, median absolute errors are relatively stable across Stage I to Stage IV (approximately 5–6 years), showing minimal variation between stages. In contrast, the external dataset exhibits higher and more variable errors, with Stage I showing the largest median error (about 6.3 years) and Stage III also displaying elevated error. Rare ambiguous stages (e.g., "I/IINOS") are excluded due to unclear definition and only one available sample.