Human whole epigenome modelling for clinical applications with Pleiades

Pouya Niki[1†], Christoforos Nalmpantis[1†], Javkhlan-Ochir Ganbat[1†], Donal Byrne[1†], Husam Babikir[1], Anjeet Jhutty[1], Will Rowe[1], Timing Liu[1], Netanel Loyfer[2], Sofia Toniolo[3,4], Masud Husain[3,4], Sanjay G. Manohar[3,4], Sian A. Thompson[5], Ivan Koychev[1,6,7], Henrik Zetterberg[1,8-15], Robert Sugar[1], Augustinas Malinauskas[16], Khaled Saab[17], Hannah Madan[1], Jonathan C. M. Wan[1,18,19*], Ravi Solanki[1,5*]

[1]Prima Mente, London, UK and San Francisco, USA.
[2]Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel.
[3]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK.
[4]Department of Experimental Psychology, University of Oxford, Oxford, UK.
[5]Department of Neurology, Oxford University Hospitals NHS Trust, Oxford, UK.
[6]Department of Psychiatry, University of Oxford, Oxford, UK.
[7]Department of Brain Sciences, Imperial College London, London, UK.
[8]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, The Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden.
[9]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden.
[10]Department of Pathology and Laboratory Medicine, University of Wisconsin School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA.
[11]Wisconsin Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA.
[12]Department of Neurodegenerative Disease, UCL Institute of Neurology, London, UK.
[13]UK Dementia Research Institute, UCL, London, UK.
[14]Hong Kong Center for Neurodegenerative Diseases, InnoHK, Hong Kong, China.
[15]Centre for Brain Research, Indian Institute of Science, Bangalore, India.
[16]Eternis Labs, San Francisco, USA.
[17]Independent Researcher, San Francisco, USA.
[18]University College London Hospital, London, UK.
[19]Francis Crick Institute, London, UK.

*Corresponding author(s). E-mail(s): jonathan@primamente.com; ravi@primamente.com;
†These authors contributed equally to this work.

Abstract

Gene regulation in humans extends beyond the four letter genetic code. Cytosine methylation, in particular, functions as a critical epigenetic switchboard, dynamically programming cellular identity, adapting gene expression in response to environmental cues, and underpinning the onset and progression of numerous diseases. Here we present Pleiades, a series of whole-genome epigenetic foundation models spanning three sizes: 90M, 600M, and 7B parameters. Pleiades is trained upon an extensive proprietary dataset of methylated and unmethylated human DNA sequences totalling 1.9T tokens. We introduce alignment embeddings and stacked hierarchical attention techniques to provide precise epigenetic modelling without the need for extended context lengths. Collectively, these advances enable Pleiades to perform a diverse range of downstream biological and clinical tasks, including nucleotide-level regulatory

prediction, realistic generation of cell-free DNA fragments and fragment-level cell-type-of-origin classification, within a unified and scalable computational framework. We specifically apply Pleiades to the early detection of real-world cohorts of clinical Alzheimer's disease and Parkinson's disease, achieving high-accuracy. We integrate Pleiades with leading protein biomarkers, achieving state-of-the-art results, underscoring the complementary value of epigenomic and proteomic multi-modal approaches. By advancing beyond the modelling of pure DNA sequences and relying on limited genomic regions, Pleiades establishes genome-wide epigenomic modelling as a new paradigm for clinical diagnostics, synthetic biology, and precision medicine.

1 Introduction

The creation of tools to understand human biology has been critical for the advancement of health and the prevention of disease for centuries [1–9]. Applications of artificial intelligence and advanced language modelling to the life sciences have established a new era for the discovery of biological knowledge, promising diagnostics and therapeutics across a range of complex human diseases [10–14].

Many such tools and models have focused upon the human genome, effectively capturing the underlying statistics of DNA patterns and coevolutions [11, 15–19]. However, these models do not capture the epigenome, the set of dynamic environmental and chemical changes to the genetic code critical for organismal development, cellular fate, and both the onset and progression of multiple diseases [20–25]. Of the variety of epigenetic alterations, DNA methylation is a critical class [26, 27]. Its study has led to the development of diagnostics for cancer detection and identification of novel mechanisms responsible for age-related diseases [28, 29].

The dynamic role of DNA methylation is increasingly evident in the pathophysiology of neurodegenerative conditions, including Alzheimer's disease (AD), Parkinson's disease (PD), and amyotrophic lateral sclerosis (ALS) [30–35]. These diseases are among the leading causes of morbidity and mortality in the world: first in the UK, sixth in the USA, and seventh globally [36–38]. They have historically been challenging to study due to the lack of routine tissue biopsy, the poor translation of animal models, and an insidious disease onset with pathology preceding symptoms by up to two decades [39–42]. The lack of precision diagnostic tools poses a challenge to both patient outcomes and clinical trials for these neurodegenerative conditions [43–45]. While recent advances in proteomic tests for AD (most notably plasma pTau-217) and PD (mainly cerebrospinal fluid tests detecting pathological alpha-synuclein seeds) show promise, they remain limited to a subset of patients and stages of disease, leading to real-world clinical ambiguities [46–51].

In recent years, similar challenges faced in oncology have been addressed with the study and application of cell-free DNA (cfDNA), fragments of DNA that freely circulate in a variety of biofluids such as plasma and cerebrospinal fluid (CSF) [28, 52]. Importantly, the methylation status of cfDNA is indicative of its cellular origin as well as of changes that may occur throughout the disease process [52, 53]. To date, cfDNA has been successfully utilised for the early detection of cancers, monitoring of treatment response, and the discovery of novel biology for downstream therapeutic application [52, 54, 55]

Here we introduce Pleiades, a series of biological foundation models for the human epigenome spanning three sizes: 90M, 600M, and 7B. The models are built using an autoregressive transformer decoder architecture with multi-tier hierarchical attention for set modelling. Pleiades is trained to explicitly capture the environmental contexts and changes to genetic code. We orchestrated a unique data corpus for pretraining including a high-quality human tissue-specific methylation atlas, plasma-derived cfDNA, and a graph of human genomic diversity [53, 56, 57]. We show that Pleiades achieves state-of-the-art performance at identifying human genomic annotations, compared to leading genomic foundation models [12, 15].

We further showcase that appropriately developed foundation models for biology could provide meaningful clinical utility today. We choose neurodegenerative diseases due to their global mortality, biological complexity, and need for precision tools. We apply Pleiades to

a series of tasks involving cfDNA, with a view to early detection of AD and PD [58–61]. We show that Pleiades can be used to identify the cellular origin of circulating cfDNA in human plasma, generate synthetic fragments *in silico* with high fidelity, and enrich a patient sample for certain fragments of interest.

By utilising *hierarchical attention* — a mechanism that first learns patterns in individual cfDNA reads and then progressively pools them into sample-level signals — Pleiades detects clinical neurodegenerative disease from plasma-derived cfDNA in real-world AD and PD cohorts. When combined with proteomic information, the model demonstrates detection of AD with high-performance.

Pleiades represents a foundation for novel biological discovery for neurodegenerative diseases, for diagnostic applications and future mechanistic interpretation for novel therapeutic targets.

2 Results

2.1 Pleiades

The Pleiades series scales to 7B parameters and is trained upon a unique data corpus of 1.9T tokens of methylated and unmethylated DNA sequences, using a context length of 1,024 tokens (Fig. 1).

2.1.1 Pretraining

Pleiades adopts an autoregressive transformer decoder architecture, provided at three scales — 90M, 600M, and 7B parameters.

To effectively pretrain Pleiades, we required large amounts of high-quality sequence data from a variety of human cell types and samples. While large epigenetic datasets are publicly available - notably the NIH Roadmap Epigenomics Mapping Consortium and ENCODE Consortium - heterogeneous sample quality limits their utility for foundation modelling [22, 53, 62].

To overcome this, we curated our own consolidated corpus of human methylation and genomic data:

1. **The methylation atlas of normal human cell types**: whole-genome bisulfite sequencing (WGBS) of 39 cell-type groups obtained via fluorescence-activated cell sorting from 205 tissue samples across 137 healthy donors [53].
2. **Plasma-derived cfDNA**: WGBS and enzymatic methylation sequencing (EM-Seq) of plasma cell-free DNA from 20 healthy individuals [56].
3. **Human genome diversity**: a genome graph representative of the 1000 Genomes Project [57].

Comprehensive dataset preparation and tokenisation details are provided in Section 4.1.4.

2.1.2 Alignment Embeddings

Accurate epigenomic modelling requires precise representation of genomic context, particularly to capture long-range regulatory interactions critical for gene expression and biological function [63]. To address this, we introduce Alignment Embeddings (AEs), a novel architectural component that explicitly encodes genomic coordinates directly into Pleiades' sequence representations. By embedding positional information, AEs equip Pleiades to recognise subtle yet biologically meaningful differences among genomic and methylomic sequences, and without relying upon prohibitively large transformer context windows.

For each nucleotide position within a read, we recovered its GRCh38 genomic position, using start position, strand and the CIGAR string, then decomposed that position into four integers: the chromosome, as well as the millions, thousands and ones offsets ($\leq$ 249, 999, 999). Each component passed through its own embedding table and the four vectors were concatenated to form the Alignment Embedding token. These embeddings

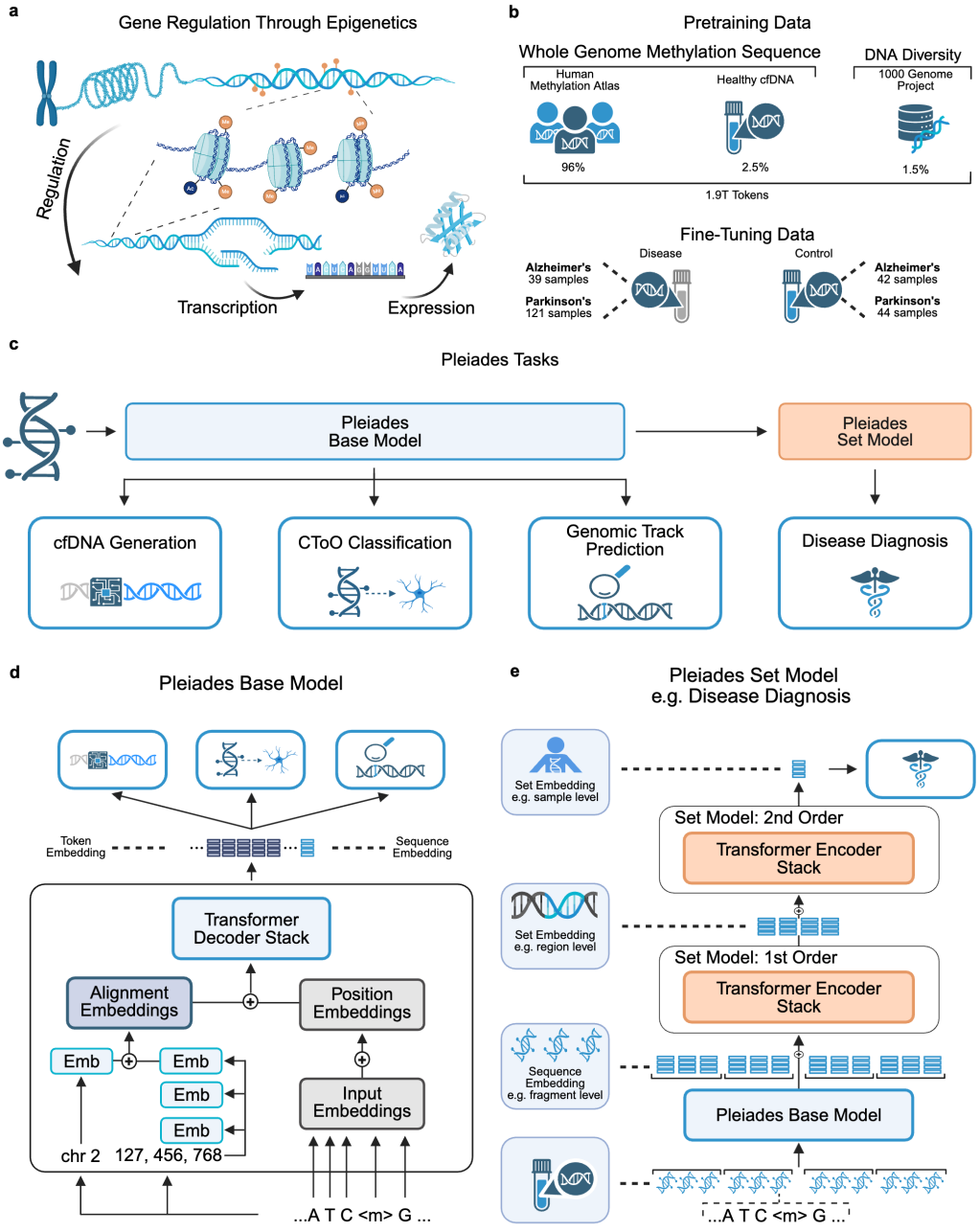

Fig. 1: **Epigenomic Foundation Modelling with Pleiades (a)** Epigenetic regulation. DNA and histone modifications modulate the accessibility, transcription, and downstream expression of DNA, without altering base sequence. Cell type–specific profiles encode high-resolution regulatory signals. **(b)** Pretraining and fine-tuning. Pleiades is pretrained on 1.9T tokens of whole-genome methylated DNA sequences and fine-tuned on disease-specific and control cfDNA. **(c)** Model and tasks. The base model supports cfDNA generation, sequence classification such as Cell-Type-of-Origin and genomic track prediction. The set model enables set level tasks such as disease diagnosis. **(d)** Base model architecture. Token-level inputs (sequence, methylation, position, alignment) are processed by a transformer decoder to produce per-token and sequence embeddings. Emb refers to embedding. **(e)** Hierarchical set model. Fragment-level embeddings are first aggregated within 1kb genomic windows to yield region-level vectors; these region vectors are subsequently pooled by a second transformer encoder to form a single sample-level embedding for downstream inference.

were trained jointly with the model, covering every base, thereby capturing long-range epigenomic structure inside a fixed-length transformer.

Unlike recent positional embedding approaches such as CpGPT [16] that encode location only for CpG dinucleotides (≈ 30 million loci in GRCh38), our AEs encode the absolute chromosome and single-base offset for every nucleotide in the human genome (≈ 3.1 billion positions; Fig. 1d). Full details of the AEs are outlined in 4.1.2.

2.1.3 Hierarchical Set Modelling

Many downstream clinical-genomics tasks require reasoning over an entire bio-sample rather than individual reads. For such tasks, we treat the collection of all cfDNA reads originating from one sample as a set of sequences. As sets can be expansive in size — on the order of 10^8–10^9 sequences — efficient aggregation is essential.

To summarise these large sets, we adopt a multi-tier *Hierarchical Attention Transformer* (HAT), inspired by Chalkidis *et al.* [64]. In our implementation we stack N HAT blocks: each block pools fixed-size groups of lower-level tokens and passes the condensed representation upward such that successively higher-order summaries are built in N steps. The top-level token yields a compact sample-level embedding that feeds directly into a classifier. This hierarchical pooling preserves long-range dependencies, allowing Pleiades to tackle large sequencing tasks without the need for long context methods or alternative base architectures. Full architectural and training details appear in Section 4.1.3.

2.2 DNA Sequence Classification Benchmarks

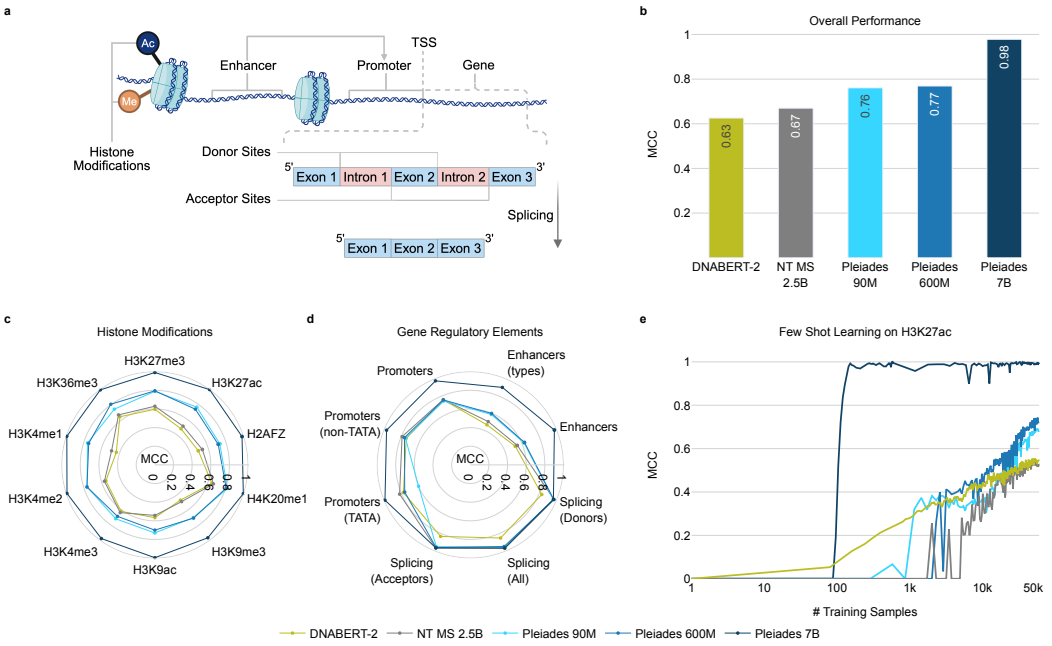

Fig. 2: **Pleiades Performance Evaluation on Unbiased Nucleotide Transformer Benchmarks** **(a)** Schematic breakdown of Nucleotide Transformer benchmark tasks. **(b)** Overall performance of all models in terms of Matthews Correlation Coefficient (MCC). **(c)** MCC performance on histone modification tasks. **(d)** MCC performance on gene regulatory elements, including Enhancers, Promoters and Splice Sites. **(e)** Few-shot learning capability of Pleiades 7B on an example NT Benchmark (H3K27ac). For Pleiades 7B, MCC of 0.9925 was achieved after training with 152 sequences.

To systematically evaluate the capabilities of the Pleiades architecture and epigenetic-focussed pretraining, we benchmarked the model on established genomic classification tasks originally introduced by Nucleotide Transformer [15]. These benchmarks include the ability of a model to identify a variety of genomic features, such as promoters, enhancers,

splice sites, and histone modification sites, and are commonly used to assess the predictive performance of DNA sequence models [12, 15] (Fig. 2a).

During our analysis, we identified a significant positional bias within the original Nucleotide Transformer dataset, where negative sequences consistently started from genomic positions divisible by $1,000$ (Supplementary Fig. S1). This bias could inadvertently provide any model with an unintended positional signal to easily distinguish between positive and negative samples, thereby inflating the previously reported performances. To address this, we introduced a randomised jitter in the range $[-500, 499]$ to the start positions of negative sequences, effectively removing the positional bias from our findings (Supplementary Fig. S1). We refer to this as the Unbiased Nucleotide Transformer Benchmark. A comprehensive explanation of this adjustment is detailed in Supplementary Section S1.

We benchmarked our three Pleiades models — 90M, 600M, and 7B — against two popular baseline genomic foundation models, the largest Nucleotide Transformer model (Multi-species, 2.5B parameters; NT MS 2.5B) [15] and DNA-BERT2 [12]. Each model was fine-tuned for exactly five epochs on our Unbiased Nucleotide Transformer Benchmark. To alleviate potential distribution shift effects between predominantly methylomic pretraining data for Pleiades (98.5% of total pretraining data) and this purely genomic fine-tuning dataset, small Pleiades models (90M, 600M) were fine-tuned for one epoch on the DNA sub-portion of the pretraining dataset prior to evaluation (1.5% of total pretraining data). Pleiades 7B did not undergo any specific DNA fine-tuning.

Pleiades consistently outperformed baseline models on these tasks, as measured by Matthews correlation coefficient (MCC) (Fig. 2b). Specifically, Pleiades 7B achieves the highest MCC scores in 15 of 18 tasks, and is on par with baselines on the remainder with an overall macro-average MCC of 0.98. The smaller 90M and 600M models exceed baseline performance in 12 out of 18 tasks as well as on macro-average with overall MCC of 0.76 and 0.77, respectively. This is in contrast to the results for DNABERT-2 (MCC 0.63) and NT MS 2.5B (MCC 0.67). Full results are detailed in Table S1.

Notably, on histone modification prediction tasks, even small Pleiades models demonstrated significantly higher performance compared with DNA-only models (Fig. 2c). Pleiades 90M consistently outperformed the best baseline model, NT MS 2.5B, despite having 27 times fewer parameters. This is not entirely unexpected, as DNA methylation is dynamically connected to histone modification in the human genome [65, 66].

We further investigate the Pleiades series by examining the few-shot learning capabilities on these tasks. Fig. 2e shows the performance on the representative histone modification task for H3K27ac for all models under study, trained for two epochs on the entire dataset. Pleiades 7B achieves near-perfect MCC (0.9925) after training on only 152 examples. No other model exhibits few-shot learning capabilities and instead achieve lower overall performance, even after training on the full dataset of approximately $30,000$ samples for 2 epochs.

2.3 Epigenomic Sequence Generation

We next explored the generative capabilities of Pleiades. We applied this specifically to cfDNA, to assess the feasibility of generation of *in silico* biological data (Fig. 3a).

We frame the task on cfDNA bio-samples, held out during pretraining, each sequenced to a depth of $20 - 50x$. In each sample, we designated 10% of fragments to a seed set, from which prompts were created for generation. The remaining fragments in the sample (90%) were assigned to the ground truth set. Each prompt consisted of five full fragments in the same 1kb region within the seed set and the first three nucleotides of a fragment in the ground truth set. This target fragment was non-overlapping with the prompt and from the same 1kb region. We focused on 68 repeat-masked high-coverage 1kb regions. A full list of these regions can be found in Table S2.

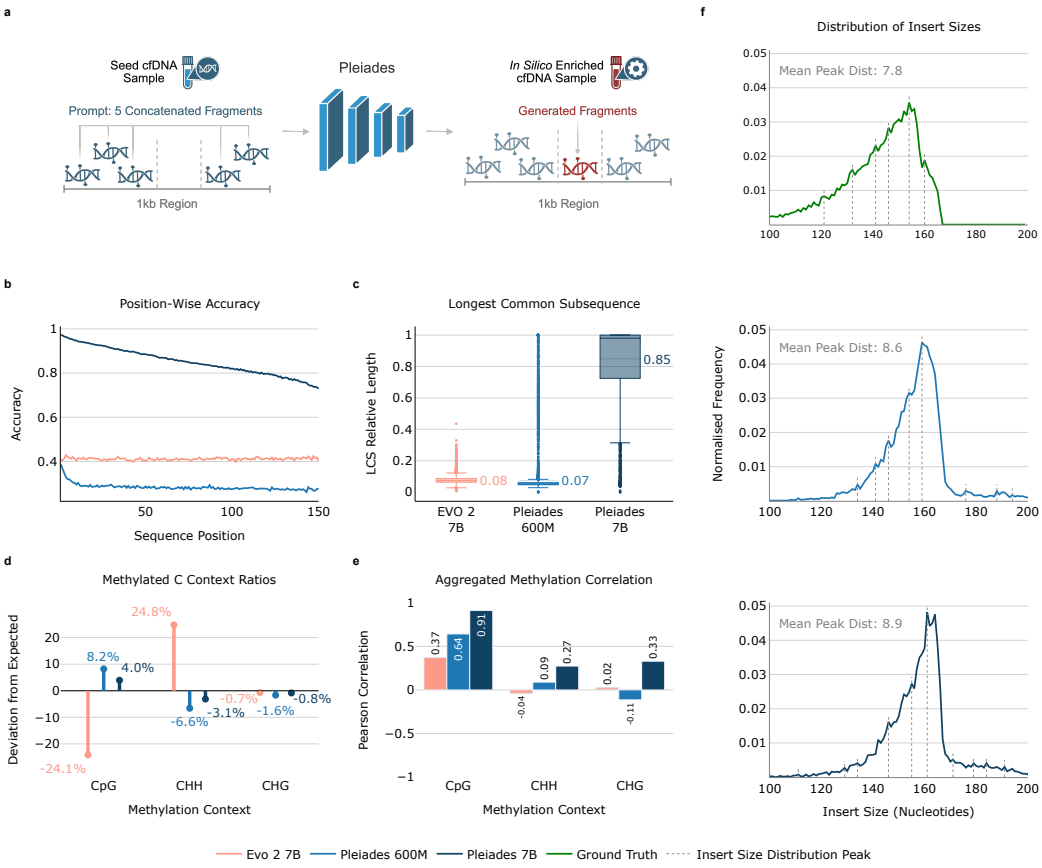

Fig. 3: ***In silico* cfDNA Generation with Pleiades (a)** Task schematic: five observed cfDNA fragments prompted the model to generate a sixth, non-overlapping fragment. **(b)** Position-wise nucleotide accuracy of the generated fragment. **(c)** Longest common *subsequence* (LCS) length between generated and true fragments, expressed relative to the ground-truth length. **(d)** Cytosine-context distribution of methylated sites. **(e)** Pearson correlation of 1kb-binned methylation ratios between generated and true samples; higher values indicate better concordance. **(f)** Insert size distribution of generated fragments for Pleiades 600M and 7B. Evo 2 7B was omitted because its decoding does not terminate.

We established three classes of evaluation metrics to assess generation quality at three complementary resolutions:

(i) *Nucleotide fidelity*: Per-position accuracy and longest-common-subsequence (LCS) length relative to ground truth fragment length (Fig. 3b,c).
(ii) *Methylome concordance*: Pearson correlation of aggregated methylation ratios (CpG/CHG/CHH) across 1kb bins and analysis of methylation with respect to cytosine context (Fig. 3d,e).
(iii) *Fragmentomics*: Insert-length distribution and insert size distribution peak periodicity (Fig. 3f).

For baseline comparisons, we chose Evo 2 7B, a DNA-only language model trained upon genomic sequences across multiple species [19]. This allowed for a direct assessment of the utility of DNA-only models without specialised training or data inputs to capture complex epigenetic signatures.

Assessing nucleotide accuracy, Pleiades 7B achieved 97% accuracy of the first nucleotide and 73% at base 150, for a mean of 83% (Fig. 3b). Pleiades 600M started at 40% and declined to 19%, for a mean of 25%. In comparison, Evo 2 7B remained largely static across the trace (mean 42%). We observed that parameter scaling substantially improved single-nucleotide level accuracy.

LCS analysis indicated similar results (Fig. 3c). Pleiades 7B reproduced *on average* 85% of ground-truth fragments contiguously ($\approx$ 125 nucleotides of a 150 nucleotide window). The model achieved a *median* relative LCS length of 0.98. This indicates that using the 10% seed set alone, 50% of generated fragments matched at least 98% of the original sequence. In contrast, Pleiades 600M and Evo 2 7B had a relative LCS length of only 0.07 and 0.08 ($\approx$ 11–12 nt) on average, respectively.

We next assessed the concordance of methylation across cytosine context (Fig. 3d,e). Methylation can occur across three genomic contexts: CpG, CHG, and CHH. While the vast majority of methylation in the human genome occurs at CpG sites, non-CpG methylation is critical especially for brain biology and brain disease [30, 67–69]. In our ground truth set, approximately 91% of all methylation occurs at CpG sites, with only 7% at CHH and 2% at CHG sites. Methylation–context ratio analysis shows that Evo 2 7B under-represented CpG contexts by 24.1% while over-generating CHH by 24.8%, indicating a bias towards non-canonical methylation sites. In contrast, Pleiades 600M reduced these errors to +8.2% (CpG) and −6.6% (CHH) and Pleiades 7B further to +4.0% (CpG) and −3.1% (CHH). All three models maintained CHG deviations below 1.6% (Figure 3d).

We then compared aggregated methylation correlation between predicted and true across context (Figure 3e). Pleiades 7B attains a Pearson correlation of 0.91 with ground-truth CpG methylation, in contrast to 0.64 for Pleiades 600M and 0.37 for Evo 2 7B. Correlations for non-CpG methylation are lower, with the best performance seen with Pleiades 7B in each of CHH (0.27 vs. 0.09 Pleiades 600M and −0.04 Evo 2 7B) and CHG (0.33 vs. −0.11 600M and 0.02 Evo 2 7B). As expected, larger scale and methylation-focussed pretraining together reduce context-distribution biases and improve locus-specific methylation recall.

In silico generated fragments were plotted by insert size (Fig. 3f). The distribution reflects the canonical nucleosome-related architecture with a modest right-shift in their length distribution. Modal length increased from 154nt in the ground-truth library to 163nt in the generated set ($\Delta = +9$ nt; +5.8%). Rotational phasing of cfDNA, quantified by the spacing between successive mini-peaks in the insert size spectrum, was virtually preserved (8.9 ± 0.3nt *in silico* vs. 7.8 ± 0.2nt empirically). Interestingly, these higher-order chromatin signatures arise *de novo*; the model was never *directly* supplied with fragment-length targets or nucleosome annotations, underscoring that base-resolution sequence pretraining alone suffices to learn nucleosome organisation.

Collectively, these results show that scaling from 600M to 7B parameters raises mean per-nucleotide accuracy from 25% to 83%, increases relative LCS from 0.07 to 0.85, sharpens CpG methylation correlation from 0.64 to 0.91, and preserves nucleosome-driven fragment lengths. In contrast, the DNA-only baseline despite matching the 7B parameter budget, unsurprisingly underperforms across all metrics.

2.4 Cell Type-of-Origin (CToO)

Cell Type-of-Origin (CToO) refers to the specific cell type from which a circulating cfDNA fragment originates. As plasma-derived cfDNA inherently comprises fragments from diverse cell types, accurate determination of the CToO of cfDNA fragments is necessary for precise evaluation of a specific tissue's physiological state and facilitates early diagnosis, disease monitoring, and targeted therapeutic interventions [53, 70].

Existing cell type deconvolution methods typically produce aggregate, sample-level estimates [53, 56]. Here, we propose the Cell Type-of-Origin task, an approach that leverages methylation patterns of individual cfDNA fragments (100–300 bp) to determine their originating cell type on a per fragment level (Fig. 4a).

We compare overall averaged F1 score of our fragment-level CToO classifier across progressively larger panels of Differentially Methylated Regions (DMRs) [71]: (i) the 25 marker regions published the tissue methylation atlas [53], as well as the (ii) the top 100 and (iii) the top 1000 regions discovered from our training cohort (Fig. 4b). Per cell type results are detailed in Fig. S3, while the DMR calling method is described in Section 4.4. The shift in the UMAP embedding of fragment representations after fine-tuning (Fig. S2) further illustrates how the model leverages these DMRs to separate cell types in latent space.

We observe that applying stricter filtering criteria to the DMR sets enhances classification accuracy. This rigorous filtering is especially beneficial for downstream applications that require precision within narrowly defined genomic regions.

In contrast, employing larger DMR sets enables broader genomic coverage, allowing identification of a larger number of fragments originating from any cell type of interest. Across all three panel sizes, Pleiades $90M$ and $600M$ achieve higher macro F1 than the tuned random forest baseline (e.g. Top-25 panel: 0.67 vs 0.46; Fig. 4b). As the panel expands to Top 100 and Top 1k regions, this performance gap widens, highlighting the Pleiades' abilities to capitalise on broader marker sets.

We then assessed performance on an unseen and independent out-of-distribution (OOD) dataset containing 6 different cell types [72] (Fig. 4b). Fig. S4 illustrates the classification F1 scores on this OOD dataset across all cell types and models. The model demonstrates comparable or improved performance across several cell types when evaluated on previously unseen data.

2.4.1 Cell Type Enrichment

Having demonstrated the capability of Pleiades to accurately predict fragment-level cell type origins, we next explore the model's potential for cell type enrichment. Enrichment involves selectively retaining fragments classified by the model as originating from a specific cell type, thereby enhancing the representation of a cell-of-interest within naturally mixed cfDNA samples. To evaluate enrichment, we utilised the OOD dataset used earlier [72] to generate an *in silico* admixture for downstream manipulation (Fig. 4c). We applied predictions from the Pleiades 600M model to select DNA fragments with the highest relative probabilities assigned to each target cell type. Due to computational constraint, we did not fine-tune Pleiades 7B.

Results revealed large increases in the target cell type proportions within the top 1K DMR regions (Fig. 4c). Notably, even rare cell types exhibited substantial enrichment: neuronal fractions increased from 7.37% to 58.5%, a 7.9-fold enhancement and hepatocytes from 4% to 73.3%, an 18.3-fold increase.

The enrichment process balanced precision, indicated by high post-enrichment cell type proportions, and recall, reflecting the retention of relevant fragments (Fig. S5). Furthermore, broader DMR regions enhance genomic coverage, allowing retrieval of more diverse fragments.

Additionally, we validated this enrichment strategy on a real-world cfDNA sample with unknown cell type composition. Employing the UXM deconvolution tool to estimate proportions before and after enrichment (Fig. 4d), we observe substantial improvements across all evaluated cell types, further confirming the effectiveness of Pleiades for cell type enrichment.

2.4.2 Cell Type Deconvolution

To assess the effectiveness of our fragment-level approach, we benchmarked Pleiades against two leading deconvolution methods, UXM [53] and CelFiE [56], which infer cell type ratios based on global aggregated features from cfDNA samples.

Each experiment involved random sampling of 500,000 fragments, with known ground-truth cell type ratios and repeated 5 times with different sampling seeds. All methods require predefined marker regions, restricting the total number of fragments for evaluation. Performance is quantified using Jensen-Shannon divergence [73]. For UXM and Pleiades, we used the published Top 25 official marker set to achieve full comparability [53].

Pleiades achieved competitive overall performance compared to both UXM and CelFie (Fig. 4e). Pleiades performed best in 3 of 5 categories and was tied with UXM in 1. UXM performed best in the category of 4 cell type mixtures. CelFie did not achieve superior performance in any of the mixtures (Fig. 4e).

For these tasks, high-performance was achieved with the small Pleiades models. At this scale, fragment-level CToO yielded robust, generalisable cfDNA cell-type calls.

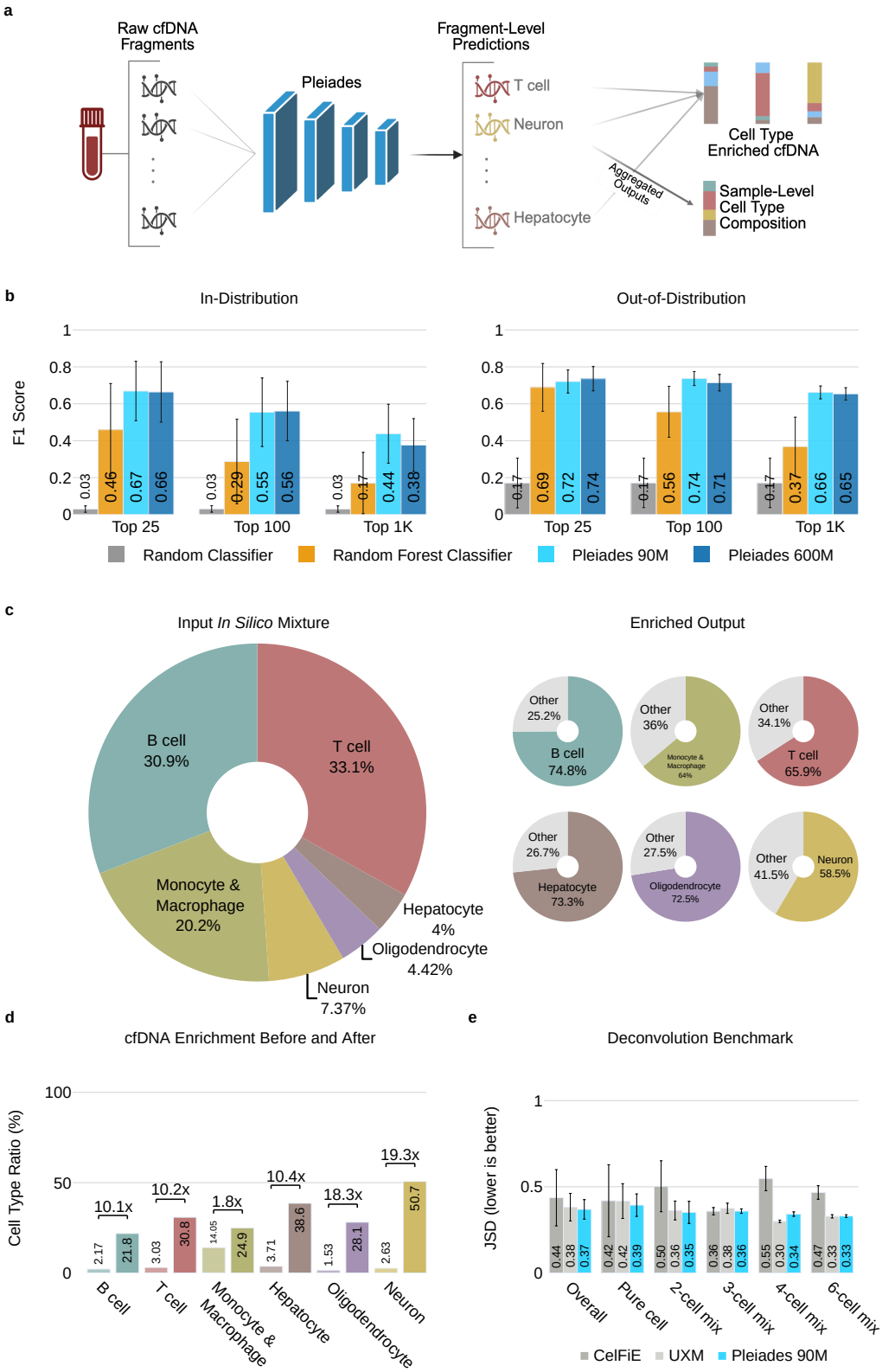

Fig. 4: **Cell Type-of-Origin**. CToO performance and evaluation over in-distribution, out-of distribution (OOD) as well as cfDNA. **(a)** An overview of CToO and downstream tasks. **(b)** Macro F1 scores on in-distribution and external out-of distribution dataset. In-distribution contains 39 cell types and out-of distribution(OOD) dataset contains 6 cell types. Bars represent means; error lines denote standard deviation. **(c)** OOD *in silico* composition before and after cell type enrichment. Enrichment was done with fine-tuned Pleiades 600M model over Top 1K cell type markers. **(d)** Cell type enrichment over actual cell-free DNA samples. Cell type proportion was estimated using UXM deconvolution tool. **(e)** Deconvolution benchmark against 2 well known deconvolution tools. Bars present the mean Jensen-Shannon divergence score between true cell type ratio and estimated ratios; error bars indicate the standard deviation.

2.5 Applications of Pleiades to Neurodegenerative Diseases

With a deep understanding of the human epigenome and cfDNA, we hypothesised Pleiades could be applied to the detection of early-stage neurodegenerative diseases. As a modality, cfDNA promises a minimally-invasive alternative to traditional diagnostic and prognostic methods such as cerebrospinal fluid (CSF) analysis and amyloid-PET scans [74–76].

A cohort of 81 age and sex-matched patients were curated from the Cognitive Disorders Clinic at Oxford University Hospitals, Oxford, UK. Patients were identified as having either mild AD dementia or mild cognitive impairment (MCI), defined under the ATN criteria [77]. Patients additionally underwent comprehensive cognitive evaluation and neuroimaging for completeness; full details are described in 4.5.1. To take advantage of Pleiades global genomic understanding, we utilise a whole-genome approach rather than targeted sequencing methods as previously reported [60]. cfDNA was extracted from plasma and prepared into libraries for EM-Seq using short-read sequencing at depths > 30x.

For modelling and evaluation, clinical diagnosis was framed as an epigenomic *set* problem. Each cfDNA fragment was processed by the pretrained Pleiades base model with a trailing `[CLS]` token: the final-layer `[CLS]` embedding represents that fragment. Embeddings from the same sample are concatenated and passed to a HAT block (Fig. 1e) that builds three nested representations: (i) individual fragments, (ii) genomic regions (aggregating fragments), and (iii) the complete cfDNA sample (aggregating regions). This hierarchy mirrors biological organisation and enables region-level attribution alongside sample-level prediction.

Diagnostic performance of Pleiades for early clinical AD demonstrated high-performance (Fig. 5b). We utilised a nested cross-validation approach as described in Section 4.5.2. Pleiades 7B achieved AuROC scores for classification of AD vs control across specific cell-type identities as follows: 0.81 for microglia, 0.82 for neuron, 0.82 for B cell, and 0.80 for T cell, on average. Smaller models, as expected, exhibited less accurate and more variable performance; Pleiades 90M achieved average AuROC scores ranging between 0.60 and 0.64 and Pleiades 600M between 0.64 and 0.71. Notably, we observed significant scaling improvements with the 7B model in both mean AuROC scores and consistency of results (Fig. 5). By combining predictions across all available cell types using average pooling, we achieved an AuROC of 0.89, suggesting the presence of complementary signals across cell types.

Further, when we trained Pleiades on experimental, non-human quality control DNA introduced within each cfDNA sample during sequencing (pUC19, Lambda DNA), all regions were rejected by our marker selection t-test with no regions deemed to contain any signal relating to clinical disease status (Fig. S6).

In recent years, protein biomarkers have emerged as promising tools in the diagnosis of Alzheimer's disease due to their strong biological relevance and established clinical significance [49, 78, 79]. In this cohort of patients, individual proteomic biomarkers achieved a range of AuROC scores: 0.46 (Aβ40), 0.61 (Aβ42), 0.66 (Aβ42/40), 0.79 (NFL), 0.76 (GFAP), 0.78 (pTau-181), and 0.90 (pTau-217) (Fig.5b). Comparing the protein biomarkers to Pleiades, we observe improved performance ($p < 0.05$) of Pleiades relative to markers Aβ40, Aβ42, $\frac{\text{A}\beta 42}{\text{A}\beta 40}$, and NFL. The performance of Pleiades 7B showed a trend toward outperformance compared to protein-related biomarkers GFAP and pTau-181. The performance of Pleiades 7B was non-significantly different to pTau-217 (Wilcoxon test details on this comparison can be found in Table S4). Next, we combine pTau-217 with predictions from Pleiades 7B across all individual cell type marker regions. This integration results in significant improvements for all cell types, stabilising diagnosis and notably enhancing accuracy by approximately $2-5$ percentage points. This leads to a peak diagnostic performance of 0.97, underscoring the substantial diagnostic advantage gained by integrating epigenetic and proteomic biomarkers in one model. The best performing multi-omic combination of all cell type marker regions and pTau-217, outperforms both proteomic and cfDNA approaches (Wilcoxon p-value of 0.06 and 0.03 respectively).

To assess the transferable capabilities of Pleiades across conditions, we apply the model to the early detection of PD. 160 human plasma samples from patients with PD and age- and sex-matched controls were procured from a commercial supplier. Samples were processed in-house identically to AD samples. Microglial region AuROC scores for Pleiades

7B models are on average 0.82, and neuronal 0.83 across 5 outer folds (Fig. 5c). The average pooled version of these models achieves AuROC of 0.84.

All disease diagnosis experiments used a nested 5-fold cross-validation, with outer folds giving error bars (Fig. 5b,c). A detailed example of the methodology given an outer fold is described in Fig. S6.

Overall, multi-cell type ensembles using Pleiades matched or exceeded state-of-the-art proteomic markers. Multi-modal combination with pTau-217 demonstrated state-of-the-art results (AuROC 0.97).

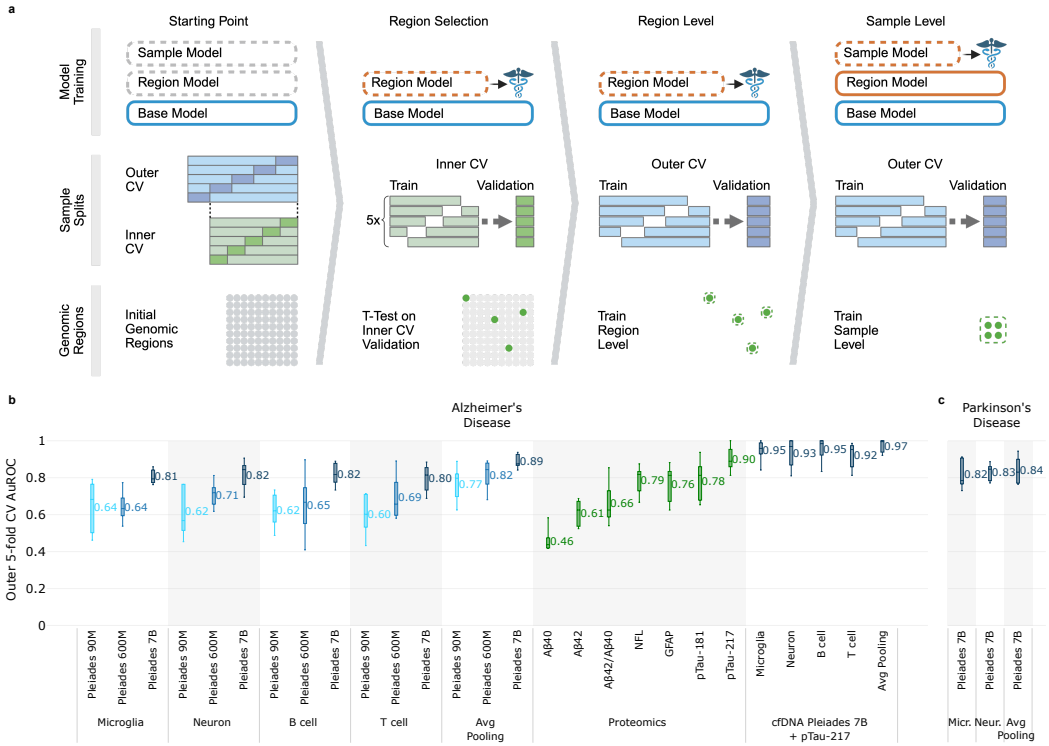

Fig. 5: **Clinical Neurodegenerative Disease Diagnosis with Pleiades (a)** Schematic representation of the diagnosis process using hierarchical Pleiades models. Process begins with broad genomic regions. The entire sample set is divided into nested 5-fold cross validations. Inner folds are used to train region level models and perform a statistical test to select high-performing regions on the inner validation sets. Then the outer fold train sets are used to train region- and sample-level models and report final performance. **(b)** Clinical AD diagnosis performance, reported using all Pleiades models on cfDNA samples, 7 proteomic markers, and combinations of Pleiades 7B with pTau-217 protein marker. **(c)** Clinical PD diagnosis with Pleiades 7B.

.

3 Discussion

Our work demonstrates that modelling both DNA and methylation unlocks capabilities beyond DNA-only language models. Methylation represents a critical feature set of the epigenome, the dynamic set of modifications of DNA that extensively influence cellular identity, function, and change throughout age and disease [26, 27, 29]. The Pleiades Series of models were created to capture these changes and showcase the utility epigenetic pretraining across technical, biological, and clinical applications.

Pleiades was trained upon a unique corpus of human DNA and methylation sequences totalling 1.9T tokens. Importantly, this corpus includes a comprehensive atlas of the human methylome, spanning 39 cell-type groups [53]. Pleiades spans three parameter sizes - 90M, 600M and 7B parameters - and utilises alignment embeddings to build a latent understanding of the human genomic space.

We assessed Pleiades using the popular Nucleotide Transformer benchmark, which we modified to remove a bias that may influence evaluation. We showed that all Pleiades model sizes exhibit state-of-the-art macro performance on this benchmark. Small Pleiades models outperform DNA-only models with many-fold higher in parameter count, and Pleiades 7B achieves MCC of 0.98. It is likely that incorporation of high-quality methylation data during pretraining will improve the performance and capabilities of genomic language models in many downstream evaluations, perhaps beyond those presented in this work. Uniquely, Pleiades 7B additionally demonstrated few-shot learning capabilities on this benchmark.

In our work, we are interested in the biology of the brain in age and disease. Historically, this has been prohibited by the lack of ground-truth data and the presence of complex clinical phenotypes [38–40]. In oncology and other complex diseases, applications of epigenetics have enabled the discovery of both novel diagnostics and therapeutic targets [52, 54, 55]. With recent research identifying a key role for epigenetic disruption early in neurological pathology [30], we hypothesised a role for epigenetics to dissect the complexity of neurological disease.

We therefore explored a role for Pleiades as a downstream discovery tool for neurodegenerative conditions. We first showcased a variety of applications for the model series on cfDNA, starting with their generative capabilities. Pleiades 7B generated *in silico* fragments of cfDNA with high accuracy across methylation context and fragment size. This indicates the largest model in the series has developed a good latent understanding of the statistical properties of human methylation and cfDNA. Downstream evaluation of *in silico* generated fragments in clinical tasks will be necessary to confirm utility, which, if demonstrated, may bolster small clinical datasets, calibrate diagnostic assays, and/or create controlled settings for studying cell-specific epigenetic changes.

Pleiades was additionally able to identify the cellular origin of plasma-derived cfDNA, matching state-of-the-art deconvolution methods. One step further, Pleiades was able to enrich a sample for fragments derived from a particular cell-of-interest. Future work will characterise the potential clinnical utility of cell-type based enrichment.

We then demonstrated Pleiades' ability in biomarker discovery, given a starting set of DMRs. Our results support the high-performance of both cfDNA and proteomic approaches independently. We suspect a combination of foundation modelling for both to be particularly advantageous, for improved AuROCs and further stratification or subtyping of disease. It is possible that the underlying mechanisms for both proteomic and epigenetic signatures identified are via different mechanisms, e.g. neuronal loss and microglia-induced inflammation for cfDNA and protein oligomerisation/aggregation for proteomic markers. Downstream work would benefit from interrogation of the subtypes of methylation and regions that are driving model decisions.

Limitations

Despite these advances, several constraints remain:

1. **Limited training data.** Pretraining relied solely upon methylated DNA from healthy human tissues and cfDNA. The absence of other epigenetic marks and non-human genomes curtails multi-omic and cross-species generalisation. As shown in Section 2.5, even simple multimodal methods significantly boost diagnostic accuracy, implying that broader multi-omic data at both pretraining and fine-tuning stages could yield further gains. Recently published work indicated strong potential for the inclusion of large, multi-parametric proteomic information for downstream diagnostics and prognostics for neurological conditions [42, 80, 81]. In addition, our work has focussed on whole-genome sequencing of cfDNA (30-50x). While this allows for a global view of a patient biosample, it does increase the risk of over-fitting. Additionally, drop in whole-genome sequencing costs does require consideration of much higher depths of sampling (i.e. >1,000x). Target approaches such as bespoke methylation panels may also be beneficial.
2. **Cohort size and diversity.** The clinical cohorts assessed for downstream diagnosis are modest and demographically narrow. Larger studies with broader demographic

coverage and independent replication sets will be needed to validate Pleiades-based diagnostics for routine clinical use.

3. **Interpretability gap.** The marker regions within the genome, automatically found by our biomarker discovery method, are a positive step towards explainable AI within biology. While these regions can reveal *where* the model finds signal, the model does not yet annotate *what* exactly the extracted features are, *why* they drive model decisions, and *what* biological mechanisms they hint at. Bridging this gap will require more rigorous interpretability techniques.

Future Work

Multi-modal information will improve the capabilities of Pleiades for both biomarker discovery and future mechanism identification for downstream targets. Expanding Pleiades to additional epigenomic modalities (ATAC-seq, ChIP-seq, etc.) alongside the integration of transcriptomic and proteomic data may improve diagnostic and prognostic capabilities. Incorporation of high-quality single-cell brain atlases of ageing and degeneration could yield a multi-omic brain foundation model [69]. Incorporating set-level pretraining objectives alongside the autoregressive loss, together with architectural advances for representing very large sets, may produce more expressive sample-level embeddings and boost diagnostic and generative performance.

In clinical practice, Pleiades may represent a future of multi-modal early detection of brain disease. Alone, Pleiades applied to cfDNA approaches the accuracy of proteomic measures such as pTau-217, with combinations yielding the highest AuROC. With the scale of data present within a single cfDNA (and in the future multi-omic) sample, our approach may showcase meaningful utility toward staging and sub-classification of neurodegenerative diseases. To rigorously confirm utility and expand both biological and clinical scope, models should be expanded on multi-centre, longitudinal cohorts that span diverse ancestries, age ranges, and diseases.

Building upon results observed when scaling to 7B parameters, we also plan to investigate larger foundation models and memory-efficient transformer variants that accommodate substantially longer context windows — potentially on the order of megabases — so that Pleiades can tackle genome-scale set problems end-to-end. Finally, we seek to develop a dedicated interpretability pipeline aiming to turn model attributions into mechanistic insights and clinically actionable biomarkers. The incorporation of reinforcement learning will likely be necessary for further improvements in Pleiades' capabilities.

Multi-modal foundation modelling for biology offers promise to enable precision medicine and unlock novel insights for complex diseases. Pleiades establishes the first step and displays the effectiveness of jointly modelling DNA and methylation in a unified, general-purpose foundation model. Our work lays the groundwork for accelerated biomarker discovery, and deeper, interpretable mechanistic insights into the genomic regulation of brain ageing and disease. Near-term, effective staging and subtyping of disease would improve clinical outcomes and could enable a molecular classification of neurological disease. Longer-term, wider expansions of modality and improvements in training architecture and methodology would encourage the creation of a brain foundation model for discovery of novel mechanisms and targets for precision therapies.

4 Methods

4.1 Pleiades Pretraining

4.1.1 Architecture and Pretraining Procedure

The Pleiades base model is an auto-regressive language model based on the generative pretrained transformer architecture [82]. It uses character-level tokenisation that represents each nucleotide with a capital letter (`A, C, T, G`) and methylation as (`<m>`). For the activation function of the Multi Layer Perceptron (MLP) layers, we used squared ReLU [83] for three primary reasons: (a) it is comparatively performant on several language-related tasks [84, 85]; (b) it can produce sparse representations; and (c) it has been empirically

shown to be the most efficient activation function for sparse LLMs [86]. Rotary positional embeddings (RoPE) [87] were utilised to represent the relative position of input tokens.

The models are optimised using the AdamW algorithm [88] with a learning rate of 10^{-4} and a cosine annealing learning schedule. In order to speed up training and reduce memory usage, we used bfloat16 mixed precision [89]. Full hyper parameters are specified in Table 1

For pretraining, we used the standard cross-entropy objective function. Given a sequence $x = (x_1, \ldots, x_T)$, the autoregressive language model is trained to minimise the negative log-likelihood:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}; \theta)$$

where $P(x_t \mid x_{<t}; \theta)$ is the model's predicted probability of token x_t given the previous context $x_{<t}$.

Table 1: Pleiades Architecture Details

	Epigenomic Sequence Model			Epigenomic Set Model	
Property	$90M$	$600M$	$7B$	1'st Order	2'nd Order
Layers	12	32	42	4	2
Model Dimension	768	1280	4096	768[1]	768
FFN Dimension	768	1280	4096	768	768
Attention Heads	12	20	32	12	12
Peak Learning Rate	10^{-4}	10^{-4}	10^{-5}	10^{-6}	10^{-6}
Warmup Steps	200	200	2000	0	0
Vocabulary Size	598	598	598	—	—
Max Context	1024	1024	1024	1024	1024

[1]For larger 600M and 7B models, a simple linear projection layer was used to convert the higher dimensionality base model embeddings to 768 dimensions for the set transformers.

The input to the model during pretraining consisted of sequences from various sources including the human genome, specific cells, and plasma. The variability of the type of sequences was introduced to the model with special tokens at the beginning and the end with `<dna>,<mdna>` and `<cfdna>` for pure nucleotide, sequences that include methylation, and cfDNA fragments respectively. If the cell type of origin is known, a `<cell_type>` token is used followed by the name of the cell type and the closing token `</cell_type>`. Table 2 shows how these three different modalities would look like as sequences of tokens.

Table 2: Examples of DNA Representations

Property	**Value**
DNA Sequence	`<dna>ATGCGTACGTTAGCTAGCTAGCTAGCGTACGTTAGCTAGCTAGCT</dna>`
Methylated DNA	`<mdna>ATGCGTAC<m>GTTAGCTAGCTAGC<m>GTACGTT</mdna>` `<cell_type>neuron</cell_type>`
cfDNA	`<cfdna>ATGCGTACGTTAGCTAGCTAGCTAGCGTACGTTAGCTAGCTAGCT</cfdna>`

4.1.2 Alignment Embeddings

Alignment Embeddings (AEs) encode genomic location by explicitly defining the chromosome number and precise position for each nucleotide, computed with reference to the Concise Idiosyncratic Gapped Alignment Report (CIGAR) strings against the GRCh38 reference genome. Given that the human genome comprises over 3 billion base pairs, positional information can span extensive ranges, exemplified by the largest contiguous segment (chr1) containing approximately 249 million base pairs [90]. To be equally sensitive to changes in genomic position in the largest to the smallest scale, we segmented the position within the

chromosome into three distinct parts: millions, thousands and ones. Consequently, each nucleotide was represented by a set of four tensors corresponding to the chromosome and segmented positional values:

$$[P_{\text{chr}}, P_m, P_k, P_u]$$

with constraints

$$0 < P_{\text{chr}} < 26, \quad 0 < P_m < 250, \quad 0 < P_k < 1000, \quad 0 < P_u < 1000$$

Note that we encode chromosome X, Y, and the mitochondrial chromosome as 23, 24 and 25 respectively. These components are individually embedded via learned embeddings as follows:

$$P_i = \text{LearnedEmbedding}(i), \quad i \in \{\text{chr}, m, k, u\}$$

Formally, the complete positional embedding matrix $P \in \mathbb{R}^{L \times d}$ is defined such that each row $P[i, :]$ encodes positional information for the nucleotide at position i:

$$P_i = \text{LearnedEmbedding}(i)$$

4.1.3 Set-level modelling with a multi-tier Hierarchical Attention Transformer

Many clinical and biological applications rely on *sample-level* information drawn from sets that contain $10^8 - 10^9$ cfDNA fragments. To obtain a compact representation of such large sets we place a **Hierarchical Attention Transformer (HAT)** [64] on top of the frozen sequence-level Pleiades model (Fig. 1e).

Architecture.

Each fragment is first processed by the base decoder; the corresponding contextual `[CLS]` embedding forms the input to the HAT. A single HAT block attends over this collection of embeddings and emits a *set token* that summarises the group. We *stack* N such blocks, so that the output token of tier k serves as one element in the input set of tier $k{+}1$. This recursive design yields successively higher-order representations — fragment → region → sample — while keeping memory requirements fixed.

Data flow.

(i) **Tier 0**: sequence decoder produces one `[CLS]` vector per fragment.
(ii) **Tier 1**: a HAT encoder attends over all fragment vectors within a 1kb window (default) and outputs a region vector.
(iii) **Tier** $2 \ldots N$: region vectors are concatenated and passed through additional HAT encoders; the final `[CLS]` token forms the input to the task-specific head.

Advantages.

The multi-tier scheme (a) preserves information at multiple genomic scales, (b) accommodates arbitrary set sizes, and (c) distinguishes fragments from different loci without enlarging the transformer context window. All hyper-parameters are listed in 1.

4.1.4 Pretraining Datasets

The pretraining data corpus is a compilation of four different sources: (1) the DNA methylation atlas of normal human cell types [53]; (2) WGBS of cfDNA from healthy individuals [56]; (3) EMSeq of cfDNA from healthy individuals; (4) and a genomic graph of the 1000G dataset [57], a human WGS dataset that captures high haplotypic variation. The corpus includes both single-end and paired-end sequencing reads. For single-end reads and haplotype reference reads, each read is treated as an individual fragment. For paired-end reads, we merged overlapping read pairs into a single fragment to increase the effective sequence context and reduce redundancy.

The DNA methylation atlas is a comprehensive WGBS dataset of 39 human cell types from 205 healthy tissue samples. It provides high-resolution methylation maps at the fragment level rather than just individual CpG sites. Loyfer et al. (2023) [53] identify over 1000 cell type specific Differentially Methylated Regions (DMRs), focusing on uniquely unmethylated/methylated regions, which usually reside in enhancers and contain binding sites for tissue-specific transcriptional regulators. These DMRs are consistent between individuals, reflecting the cell lineage and cell type specific programmes. To leverage this dataset for pretraining, we prioritise CpG-rich regions by including all CpG islands, shores, and shelves, while downsampling open sea regions to 5%.

We utilise cfDNA datasets prepared with experimental methods to capture methylation information. Specifically, we include 10 healthy samples from Caggiano et al. (2021)'s [56] study generated via WGBS alongside 10 healthy samples from our clinical samples generated via EMSeq. In both cases, the data format is FASTQ and is processed using a modified MethylSeq pipeline [91] with Bismark [92] as the aligner. After we generated the downstream BAM files, we followed the same pre-processing and filtering steps as the DNA methylation atlas [53].

Following a similar process to that in Dalla-Torre et al. (2024) [15], we combined the human reference genome (GRCh38) with phased haplotype variant data from the 1000 Genomes Project (release version 20220422) to construct chromosome-specific variation graphs [57]. The construction process used VG (version 1.62.0) [93], employing the "construct" command to integrate sequence and variation data. Each variation graph was indexed using the 1000G preset, ensuring the retention of haplotype-specific pathways. To analyse localised genomic variation, a sliding window approach was applied to each variation graph. A window size of 1 million base pairs was defined and the window moved incrementally along reference positions. For each window, 40 haplotype paths were randomly sampled from the preserved haplotype pathways. This subsampling was conducted to represent diverse genomic contexts within a manageable computational framework. Simulated sequencing reads were generated from the subsampled haplotype paths for each window. Read simulation was performed to achieve approximately 20x coverage per haplotype. Each read was assigned a random start position within the window and read lengths were restricted to a range of 500 to 1000 nucleotides. These parameters were selected to mimic the characteristics of real-world sequencing while maintaining high coverage for downstream analysis.

4.1.5 Computational Resources

We pretrained Pleiades 7B for approximately 10 days on 256 H100 GPUs (32 nodes $\times$ 8). For the Unbiased Nucleotide Transformer benchmark, we fine-tuned between $\sim$4 minutes and $\sim$1 hour (depending on the task) on a single H200 GPU. Pleiades 600M was trained for $\sim$18h on 8 H200 GPUs (1 node) for Top 25 regions, $\sim$11h on 64 H200 GPUs (8 nodes) for Top 100 regions and $\sim$22h on 64 H200 GPUs (8 nodes) for Top 1000 regions. Finally, for our disease diagnosis task we performed $\sim$17h of reconstruction-loss training of Pleiades $7B$ on 128 H100 GPUs (16 nodes) followed by disease diagnosis fine-tuning for $\sim$8h per cell-type region on 32 H200 GPUs (4 nodes).

4.2 Nucleotide Transformer Benchmarks

In order to test the performance of Pleiades and the baseline models on the Nucleotide Transformer benchmarks, we first had to create a newly randomised version of the Nucleotide Transformer benchmarks to counter the effects of the strong bias in positions of negative sequences in the dataset. Supplementary Section S1 and Fig. S1 explain the problem within the official Nucleotide Transformer benchmarks and our proposed solution in detail.

The next step was to fine-tune all models for exactly five epochs on the revised Unbiased Nucleotide Transformer Benchmark train set and measure the MCC. The reported values are the test set MCC at the end of training.

For Pleiades, a `[CLS]` token was appended to the end of the input sequence and its embeddings were directly fed to a classification head. This classification head is a simple

2−layer MLP with a ReLU non-linearity in between. The first layer has the same input and output dimension which equals the base model's hidden dimension size. The second layer projects down to as many dimensions as the task has labels, 2 for binary classification tasks and 3 for the tasks that have 3 labels, *Splicing (All)* and *Enhancers (Types)*. Table 3 shows the exact hyperparameters used for fine-tuning Pleiades models on NT tasks. All Pleiades models were fine-tuned on H200 GPUs.

Table 3: Pleiades Hyperparameters for Unbiased Nucleotide Transformer Benchmarks

Property	90M	600M	7B	7B High LR[1]
Peak Learning Rate	10^{-4}	10^{-4}	10^{-6}	6×10^{-5}
Global Batch Size	48	288	288	96
Unfrozen Layers	*All*	*All*	*All*	Last 2 Layers

[1]This setting was only used for the two tasks *Splicing (All)* and *Promoter (TATA)*.

Pleiades 90M and 600M were fine-tuned for an epoch on the DNA-only portion of our pretraining dataset, to bring their representations closer to pure DNA before fine-tuning for the benchmark tasks. This was not performed for Pleiades 7B.

Both baseline models were fine-tuned using their published code and with default hyperparameters. DNABERT-2 was fine-tuned using the fine-tune script in this official github page with learning rate 10^{-4} on V100 GPUs with the maximum batch size that would fit per device. NT 2.5B MS was fine-tuned using LORA and with a learning rate 5×10^{-4} on H200 GPUs.

4.3 Epigenomic Sequence Generation

To evaluate model performance on cfDNA generation, we started from real world biosamples, held out during pretraining and aimed to reconstruct them *in silico*. Two biosamples from the test set of the pretraining data were used for this analysis [56], both of which were processed using WGBS for library preparation and sequenced to a depth of $30-50x$. In each sample, we designated 10% of fragments to a seed set, from which prompts were created for generation. The remaining fragments in the sample (90%) were assigned to the ground truth set. Each prompt consisted of five full fragments in the same 1kb region within the seed set and the cfDNA start special token `<cfdna>`, followed by the first three nucleotides of a fragment in the ground truth set. This target fragment was non-overlapping with the prompt and from the same 1kb region. We focused on 68 repeat-masked high-coverage 1kb regions. A full list of these regions can be found in Table S2. During generation we applied a top k sampling method where $k = 2$ with a temperature $T = 0.7$. Decoding terminated when either

- The cfDNA end special token `</cfdna>` was generated.
- Maximum token limit was reached, which is set to $c - length(p)$, where c is model context (1024) and p is the prompt.

4.4 CToO

4.4.1 Contrastive Training Procedure and Loss

Within a marker region for any cell type, fragments from the target cell type are vastly outnumbered by fragments from other cell types. This severe class-imbalance destabilises training and can cause the model to overfit the dominant class. To counteract this problem, we adopted a contrastive learning strategy. The method clusters the scarce positive examples while repelling the overwhelming pool of negatives by (a) sampling a tractable subset of negative reads, (b) augmenting the positives, and (c) re-weighting the loss. During training, the network directly compared reads from different cell types, encouraging sequences from the same cell type to co-locate in representation space and pushing sequences from different types apart. We implemented this behaviour with a contrastive loss that combines

mean-based loss and hard negative mining [94, 95]. In order to maximise model generalisability during each epoch of training, random negative examples were re-sampled for each anchor.

Mean-based loss provided stable overall training by considering an average of negative examples. Hard negative mining helped with fine-grained discrimination, especially when certain cell types share similar methylation patterns. The combination enabled the model to learn both broad distinctions and subtle differences between cell types, improving its ability to generalise.

In addition, the model performed cell type classification, where it processed methylation sequences through Pleiades base model to generate numerical embeddings and predict the corresponding cell type via a classification head. This was supervised by a classification loss, which relied upon pooled sequence representations.

To optimise both learning strategies, the model was trained with a combined loss, where each component was weighted by a multiplier (m_{class} and m_{contr}) to balance their contributions.

By simultaneously learning from both tasks, the model built a deep understanding of cell type specific epigenetic features. Contrastive learning enhanced its ability to distinguish patterns, while classification provided direct supervision for accurate cell type prediction. This dual approach ensured robust representations, even in cases where some cell types were more common than others.

$$\mathcal{L}_{\text{contrastive}} = (1-\lambda)\mathcal{L}_{\text{mean}} + \lambda\mathcal{L}_{\text{hard}}$$

where

$$\mathcal{L}_{\text{mean}} = \frac{\max(0, d_{\text{pos}} - d_{\text{neg}} + \text{margin})}{\text{temperature}},$$
$$\mathcal{L}_{\text{hard}} = \frac{\max(0, d_{\text{pos}} - d_{\text{hardest\_neg}} + \text{margin})}{\text{temperature}}.$$

where

$$d_{\text{pos}} = \frac{\sum_{p_i \in S_+} \delta(p_i, a)}{|S_+|},$$
$$d_{\text{neg}} = \frac{\sum_{n_j \in S_-} \delta(n_j, a)}{|S_-|},$$
$$d_{\text{hardest\_neg}} = \min_{n_j \in S_-} \delta(n_j, a)$$

where a is the vector representation of the anchor sample and S_+ and S_- are the set of all positive and negative example's vector representations, respectively.

The distance metric δ was defined as follows:

$$\delta(v, w) = 1 - \frac{v}{\|v\|_2} . \frac{w}{\|w\|_2}$$

where v and w are input vectors and $\|v\|_2$ and $\|w\|_2$ represents l_2 norms of these vectors. The distance metric is $1 -$ cosine similarity.

The hyper-parameters used are:

$$\lambda = 0.3, \quad \text{margin} = 0.1, \quad \text{temperature} = 0.1.$$

$$\mathcal{L}_{\text{classification}} = \text{CrossEntropy}(\text{softmax}(\mathbf{y}_{\text{pred}}), y)$$

Here,

$$x = [\text{mean}(h); \max(h)] \quad \text{(concatenated pooling)},$$
$$\mathbf{y}_{\text{pred}} = W_2\sigma(W_1 x + b_1) + b_2,$$
$$y = \text{one-hot encoded cell type labels}.$$

$$\mathcal{L}_{\text{total}} = m_{\text{class}}\mathcal{L}_{\text{classification}} + m_{\text{contr}}\mathcal{L}_{\text{contrastive}}.$$

We also experimented with a regularisation term added to the contrastive loss to prevent representation collapse and increase training stability.

$$\mathcal{L}_{\text{contrastive\_with\_regularization}} = \mathcal{L}_{\text{contrastive}} + \lambda \sum_{i \neq j} \left(\tilde{\mathbf{E}}^{\top} \tilde{\mathbf{E}} \right)^2_{ij}$$

where $\tilde{\mathbf{E}} \in \mathbb{R}^{N \times d}$ is the matrix of row-wise normalised embeddings, and λ is the multiplier (regularisation coefficient) controlling the strength of the diversity penalty.

4.4.2 Cell-Type Differentially Methylated Regions

To fine-tune the model for the CToO task, we curated a specialised dataset derived from the DNA methylation atlas [53]. These samples were randomly split into train and test sets. DMRs were identified for each cell type in a 1-vs-all manner to distinguish between each cell type and all-the-rest. DMRs were called exclusively from the training set to prevent information leakage into the test set, using the open-source software wgbstools [96]. The genome segmentation and marker finding were done using `segment` and `find_markers` commands with each marker region containing ≥ 2 CpGs, marker size from 50 to 2000bp, delta mean difference ≤ 0.4, significance p-value ≤ 0.01 and employing Bayesian pseudocounts of 15 for both C and T counts. From the resulting DMRs, we selected the top 100 and top $1,000$ markers per cell type, yielding $3,801$ and $34,134$ regions in total respectively. We also used the official UXM top 25 marker set for reference [53].

For DNA sequences intersecting each DMR region, the ones belonging to the target cell type of the DMR were annotated as positives with the rest annotated as negatives. Negative reads have broadly similar methylation signals, so their labels were replaced with a generic negative label, for example `not_neuron` if the DMR target is `neuron`. We used contrastive learning to force the model to separate its representations of reads for DMR targets and other cell types within each DMR region. In order to do that, each positive read within a DMR region were used as an anchor, while a constant number of other positive examples (3) are randomly selected to match with each anchor alongside a constant number of negative examples (36) from other cell types in the same region. This random sampling was done per training epoch to maximise model generalisability.

Final classification results were calculated on the total set of reads that intersected with the DMR regions. For performance comparison, we evaluated three DMR sets: the top 25 DMR set reported by [53], and Top 100 and Top 1000 DMR sets. To assess generalisation, we tested on an out-of-distribution dataset incorporating sequence data [72], using only reads intersecting the predefined DMR sets.

4.4.3 Out-of-Distribution data for evaluation

For evaluating CToO and deconvolution tools, we obtained external data (Do et al., 2020) [72] and only considered healthy samples. This data set consists of 478 FACS-processed samples from various cell types. Of the total samples, 96 failed quality control criteria: 92 had low sequencing coverage and 4 showed evidence of failed bisulfite conversion. We grouped related cell types into major groups defined by the methylation atlas [53] - six groups were selected. Data was processed similarly to our cfDNA data, but with an underlying aligner such as Bismark [92] due to its native support for per-base DNA methylation calling.

From the entire cohort, we selected the following cell types for our mixture experiments: B cell, monocyte & macrophage cell, T cell, liver hepatocyte, oligodendrocyte, and neuronal cells. Using these 6 cell type samples, we created 14 different cell type mixtures ranging from six pure cell type, three types of 2 cell type mix, three types of 3 cell type mix, one type of 4 cell type mix and a 6 cell type mix. Each mix had an equal proportion of cell types. The complete breakdown of the mixtures is shown in Table S3.

For a fair comparison, we used the same marker regions for both Pleiades and UXM with the corresponding cell atlas. In contrast, CelFiE required a distinct cell-atlas format and marker-region caller, and relied upon raw coverage instead of methylation percentages. Therefore we built a custom atlas for CelFiE to improve comparability of the results.

We fine-tuned Pleiades 90M over official regions from the Top 25 markers from UXM [53] and performed classification over fragments. For each sample we aggregated the predictions of all fragments and calculated the cell type composition to get a deconvolution result. The same Top 25 markers from 2.4.2 were used to run the deconvolution tool of UXM with the parameter `--rlen 2`, which is the minimum number of CpG sites in each fragment.

For CelFiE, we used scripts provided by Caggiano et al. (2021) [56] to identify DMRs for each of the 39 cell type groups, selecting the top 100 CpG sites per marker as recommended. We then estimated cell type proportions by running the authors' deconvolution script with its default parameters and excluding any unknown cell types.

4.5 Neurodegenerative Disease Diagnosis

4.5.1 Dataset

We create a proprietary AD cfDNA dataset. Patients with AD dementia and elderly healthy controls were recruited from the Cognitive Disorders Clinic at the John Radcliffe Hospital in Oxford, UK, or open day events. AD dementia patients were defined as having AD clinical syndrome according to the 2018 ATN criteria [77]. Patients with AD had a progressive, multidomain, largely amnestic cognitive impairment and underwent MRI and FDG-PET imaging, the results of which were in keeping with a clinical diagnosis of AD (temporo-parietal atrophy and hypometabolism). ATN status was reviewed after plasma biomarker analysis to ensure AD dementia patients had an ATN profile compatible with their diagnosis. Elderly healthy controls were greater than 50 years old, had no psychiatric or neurological illness and were not on regular psychoactive drugs. They also underwent brain MRI imaging, and only participants with a normal MRI scan, reviewed by two independent senior neurologists, were included in the study. Participants underwent in-person blood collection and face-to-face standard cognitive testing, the Addenbrooke's Cognitive Examination-III (ACE-III), at the time of the visit. ACE scores lower than 88/100 were considered abnormal, and all healthy controls scoring below that threshold were excluded from this study. However, patients with AD were not recruited based on a fixed threshold on standard cognitive testing but rather took part in the study according to the criteria outlined above.

Blood was collected in six ethylenediaminetetraacetic acid (EDTA) tubes (10 mL each), and centrifuged (1800 g, room temperature, 10 minutes). The EDTA tubes were filled completely and gently inverted after collection to avoid coagulation. After centrifugation, plasma from all six tubes were transferred into one 50-mL polypropylene tube, mixed, aliquoted into 0.5 mL polypropylene tubes (Fluid X, Tri-coded Tube, Azenta Life Sciences), and stored at 4°C, until (less than 8 hours) it was transferred into a −80°C freezer. The time between blood collection and centrifugation was less than 30 minutes. Transfer time between 4°C and −80°C storage was < 20 minutes, and the samples were kept refrigerated during transport. All cryovials were anonymised, and the unique cryovial code was logged into a secure database, linked to the participant's anonymous code and visit number.

PD samples were procured from a commercial supplier (AMSBIO). A cohort of sex- and age-matched samples were selected. Individuals with PD were defined as having the clincial features of bradykinesia with additionally rigidity and/or tremor. Healthy controls selected had no psychiatric or neurological illness and were not on regular psychoactive drugs. Samples were collected in EDTA tubes and processed similarly to AD samples.

cfDNA was extracted from 0.5-1.0ml plasma using the QIAamp Circulating Nucleic Acid Kit. Sequencing libraries were generated with the NEBNext Enzymatic Methyl-seq and sequenced on Illumina short read sequencer for depths between 30-50x. Data was processed using a modified MethylSeq pipeline [91] with BWA-Meth [97] as the aligner. Per-read methylation calls were processed using a modified MethylDackel tool [98]; non-CpG methylation calling feature was added. Processed BAM files were converted into fragments as described in 4.1.4.

For protein measurements, samples were shipped on dry ice to the Biomarker Factory/Fluid Biomarker Laboratory, UK Dementia Research Institute at University College London (UCL), London. The Dementia Research Institute (DRI) laboratory staff carried out the analyses. Plasma Aβ40, Aβ42, GFAP, and NfL were measured by single-molecule

array (Simoa) technology using the Neurology 4-plexE assay on an HD-X analyzer (Quanterix), according to manufacturer's instructions. Plasma p-Tau181 and p-Tau217 were also measured by Simoa using the pTau-181 Advantage and ALZpath assays on an HD-X analyzer (Quanterix). Samples were analysed in one round of experiments using one batch of reagents with intra-assay coefficients of variation below 10% and the analysts blinded to clinical data.

Since using the entire human genome space would be computationally prohibitive, we utilised cell type specific DMRs to extract regions for microglia, neuron, B cells, and T cells. We utilised the top 100 DMRs as described in Section 4.4.2 for all but microglia, because they are not included in the methylation atlas. For the latter, we used the 59 DMRs proposed by Tian et al. (2023) [69].

4.5.2 Marker Discovery Methodology

We addressed the challenges of discovering markers for neurodegenerative diseases, such as the vast exploration space, low signal-to-noise ratio, and signal heterogeneity by employing Pleiades to search the epigenome, amplify weak signals, and capture cell-type-specific patterns. We curated data based upon biological relevance, targeting specific cell-type DMRs.

Starting from 59 DMRs for microglia and 100 DMRs each for neurons, B cells and T cells, Pleiades was utilised to further filter them to a concise set of high-confidence candidates, using nested 5-fold cross validation. The inner CV folds were used to train the region level HAT models and select the best marker regions, while outer folds were used for final fine-tuning and performance measurement. The overall workflow is summarised in Table 4.

To account for low signal-to-noise ratio and high variability, we selected regions with AuROC$>$ 0.6 for Alzheimer's disease and AuROC$>$ 0.65 for Parkinson's disease as well as significance of $p - value < 0.01$ for at least four of five inner folds across the last 50 epochs of training. Each inner CV fold was trained for 100 epochs, and validation AuROC scores from the last 50 epochs were examined. We applied a Student's t-test to assess statistical significance above the threshold. Regions passing the t-test in at least four of the five inner CV folds were sorted by the number of passing folds and average AuROC score in descending order. The top four were selected for further training phases to fit within GPU memory constraints, particularly for larger models.

Pleiades hierarchical architecture leverages the natural biological structure of cfDNA fragmentomics, progressing from the fragment level to the sample level. It builds on a base sequence-level model by introducing a region-level model. This region-level model is fine-tuned on the final, refined set of regions, after the inner cross validation step. Finally, a sample-level model is trained to predict the Alzheimer's disease condition. This three-tiered hierarchy — fragment $\rightarrow$ region $\rightarrow$ sample — allows the model to learn structured, set-based features in a computationally efficient manner. Importantly, it avoids the need for long-context models, which are constrained by the quadratic complexity of standard Transformer architectures.

4.5.3 Learning Biological Hierarchical Representations

Fragment Level Representation Learning: We fine-tuned the base Pleiades model to enhance representations of cfDNA fragments using a reconstruction head. Although the original pretraining dataset included cfDNA fragments, their proportion was small relative to other data, and fine-tuning improved diagnostic performance. Table 5 details the reconstruction head architecture. This lightweight transformer decoder reconstructs the input sequence conditioned on `[CLS]` token sequence embeddings. It functions as an autoencoder, incorporating global context via cross-attention into token-level reconstruction. With two layers, it remains computationally efficient compared to the base model. Dropout was applied to attention and feedforward layers to mitigate over-fitting. To promote diverse and informative representations, we combined a diversity loss with the reconstruction loss, penalising similarity among sequence embeddings. This dual objective captures input content and structural properties in a compact embedding space.

Table 4: Simplified Three-Step Workflow for Diagnostic Marker Discovery

	Marker Filtering	**Region-Level Training**	**Sample-Level Training**
Goal	Find regions that carry useful signals	Learn patterns from regions using weak labels	Make final predictions for each sample
Supervision	Weak supervision	Weak supervision	Full supervision
Model Type	1^{st} order HAT	1^{st} order HAT	2^{nd} order HAT
Approach	Use cross-validation to select the most informative regions based on performance and statistics	Train models on region data using labels from whole samples; represent each region with embeddings	Train a final model using true sample labels and region-level outputs
Input Data	All genomic regions	Filtered regions from previous step	Same filtered regions
Output	Selected genomic regions as biomarkers	Region-level predictions or embeddings	Final diagnosis for each sample

Table 5: Reconstruction Head Configuration

Property	**Value**
Architecture	Transformer Decoder
Input Dimension	d (same as base model)
Attention Heads	h (same as base model)
Layers	2
Dropout Rate (Attention)	0.1
Dropout Rate (Feedforward)	0.1
Cross-Attention	Enabled
Peak Learning Rate	10^{-6}

We defined the reconstruction head's input and output mathematically as follows. The input sequence is given by:

$$X \in \mathbb{R}^{B \times L \times d}$$

where B is the batch size, L is the sequence length, and d is the embedding dimension. Sequence embeddings are computed by extracting the `[CLS]` token representation from the input and used to form the context:

$$E_{CLS} \in \mathbb{R}^{B \times 1 \times d}$$

The decoder receives both the input sequence and the context embeddings, producing token-level output logits:

$$Y = \text{Decoder}(X, E_{CLS}), \quad Y \in \mathbb{R}^{B \times L \times V}$$

where V is the vocabulary size.

Training was supervised via a combination of two loss functions. The first is the standard reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \text{CrossEntropyLoss}(Y, \text{labels})$$

which encourages the model to accurately reconstruct the input sequence. The second is a correlation-based diversity loss:

$$\mathcal{L}_{\text{div}} = \lambda \sum_{i \neq j} \left(\langle \hat{e}_i, \hat{e}_j \rangle\right)^2$$

where $\hat{e}_i = \frac{e_i}{\|e_i\|_2}$ is the normalised `[CLS]` embedding for sample i, and λ is a small scaling factor, in our case 10^{-6}. This loss penalises cosine similarity between distinct embeddings, encouraging orthogonality and diversity across the batch. The total loss is the sum of both components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{div}}$$

This setup enabled the model to learn compact, informative embeddings that serve downstream tasks while remaining efficient and robust.

Region Level Representation Learning By keeping the base model weights frozen, we trained a hierarchical transformer with a classification head on the task of predicting whether an individual has the target disease. The input to the model is a batch of fragments grouped in genomic regions. The base model outputs the embedding representations of these fragments, which are grouped accordingly to form a region. This collection of fragment embeddings is the input to the second transformer, which is trained in a weakly supervised way; the disease label is assigned to each of the regions of the sample's DNA. The model is guided by a classification head that learns to predict these labels.

The input to the base model is $X \in \mathbb{R}^{B\times L\times d}$ and the sequence embeddings are described by $E_{CLS} \in \mathbb{R}^{B\times d}$

The architecture groups sequences by their region IDs. Let

- $\mathbf{R} \in \mathbb{N}^B$ denote the region ID assigned to each of the B sequences,
- $\mathbf{R}_{\text{unique}} = \text{unique}(\mathbf{R})$ be the set of unique region IDs,
- $N = |\mathbf{R}_{\text{unique}}|$ be the total number of distinct regions.

For each region $r \in \mathbf{R}_{\text{unique}}$, define the set of embeddings associated with that region as:

$$\mathbf{H}_r = \{\mathbf{e}_i \mid \mathbf{R}[i] = r\}$$

Here, $\mathbf{H}_r \in \mathbb{R}^{n_r\times d}$, where n_r is the number of sequences belonging to region r, and d is the embedding dimension.

Each region-specific embedding set $\mathbf{H}_r$ is then passed through an encoder:

$$\mathbf{Z}_r = \text{Encoder}_r(\mathbf{H}_r)$$

which is used by a classification head f_r with the following loss function:

$$\mathcal{L}_{\text{r}} = \text{CrossEntropyLoss}(\mathbf{Y}_r, labels)$$

where $\mathbf{Y}_r = f_r(\mathbf{Z}_r)$. To encourage diversity among attention heads by penalizing their similarity, we define the loss $L_{attndiv}$ as:

$$\mathcal{L}_{\text{attndiv}} = \frac{\lambda}{L}\sum_{l=1}^{L}\frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}\left\|A^{(l,b,i)}(A^{(l,b,j)})^{\top}\right\|_F$$

where:

- λ: scaling factor,
- $A^{(l,b,h)} \in \mathbb{R}^{S\times S}$: normalised attention map,
- $\mathcal{P}$: all head pairs $(i,j), i<j$,
- $\|\cdot\|_F$: Frobenius norm,
- L, B, H, S: number of layers, batches, heads, and sequence length.

The total loss becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{r}} + \mathcal{L}_{\text{attndiv}}$$

Sample Level Representation Learning: The architecture can be extended hierarchically by reusing the base and the region level models and adding a third sample-level model. The first two models were then frozen and the third received its inputs by grouping region-level embeddings into individual samples. Let:

- $\mathbf{S} \in \mathbb{N}^N$ denote the sample ID assigned to each of the N regions,
- $\mathbf{S}_{\text{unique}} = \text{unique}(\mathbf{S})$ be the set of unique sample IDs,
- $M = |\mathbf{S}_{\text{unique}}|$ be the total number of distinct samples.

For each sample $s \in \mathbf{S}_{\text{unique}}$, we collected the embeddings of all regions belonging to that sample:

$$\mathbf{H}_s = \{\mathbf{z}_{ri} \mid \mathbf{S}[i] = s\}$$

Here, $\mathbf{H}_s \in \mathbb{R}^{n_s \times d}$, where n_s is the number of regions in sample s and $r \in \mathbf{R}_{\text{unique}}$.

Each sample-level set of region embeddings $\mathbf{H}_s$ was then passed through a transformer encoder:

$$\mathbf{Z}_s = \text{Encoder}_s(\mathbf{H}_s)$$

and the sample representation $\mathbf{z}_s \in \mathbb{R}^d$ was passed to a classification head f_s.

The model was trained by assigning the sample's condition label (e.g., AD vs. non-AD) to the collection of all regions. The classification loss was defined as:

$$\mathcal{L}_{\text{s}} = \text{CrossEntropyLoss}(\mathbf{Y}_s, \text{labels})$$

where $\mathbf{Y}_s = f_s(\mathbf{Z}_s)$.

The overall loss combined attention diversity regularization, and sample-level classification:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{s}} + \mathcal{L}_{\text{attndiv}}$$

4.5.4 Evaluation

The performance of Pleiades models and other biomarkers is measured using AuROC metric on the outer 5-fold cross validation of both our disease datasets. All results were reported as mean of the 5 outer folds in Fig. 5 and Fig. S6.

5 Data Availability

This study makes use of several publicly available datasets: the DNA methylation atlas of normal human cell types [53], cfDNA WGBS from healthy individuals [56], and the 1000 Genomes Project reference dataset [57]. Additional healthy cfDNA EMSeq data and all Alzheimer's disease cfDNA data generated in this study are not publicly available due to participant privacy and ethical restrictions. Requests for access to these datasets may be considered by the corresponding author on reasonable request, subject to appropriate data sharing agreements and ethical approval. Processed code and pipelines are available from the authors upon request.

6 Code Availability

The code and model weights used in this study are not publicly available, but collaborations are welcome. Interested researchers may contact the corresponding author.

7 Acknowledgements

We thank Vivek Natarajan, Eric Nguyen, Yuval Dor, and Taya Reed for their support and contributions to this work. We are also grateful to the wider Prima Mente team, including Marie Schildt, Devin Gilliam, and Hayley Holt. This project was supported in great part by resources and services provided by NVIDIA DGX Lepton, Nebius, Siam AI, Eternis Labs, Google Cloud Platform, and Amazon Web Services.

8 Declarations

Husam Babikir, Donal Byrne, Javkhlan-Ochir Ganbat, Anjeet Jhutty, Timing Liu, Hannah Madan, Christoforos Nalmpantis, Pouya Niki, Will Rowe, Ravi Solanki, Robert Sugar and Jonathan C. M. Wan are shareholders of Prima Mente.

Henrik Zetterberg (HK) is a Wallenberg Scholar and a Distinguished Professor at the Swedish Research Council supported by grants from the Swedish Research Council (#2023-00356, #2022-01018 and #2019-02397), the European Union's Horizon Europe research and innovation programme under grant agreement No 101053962, and Swedish State Support for Clinical Research (#ALFGBG-71320). The UK DRI Biomarker Factory is funded by the National Institute for Health and Care Research University College London Hospitals Biomedical Research Centre, the UK Dementia Research Institute at UCL (UKDRI-1003), and the Weston Family Foundation. HZ has served at scientific advisory boards and/or as a consultant for Abbvie, Acumen, Alector, Alzinova, ALZpath, Amylyx, Annexon, Apellis, Artery Therapeutics, AZTherapies, Cognito Therapeutics, CogRx, Denali, Eisai, Enigma, LabCorp, Merck Sharp & Dohme, Merry Life, Nervgen, Novo Nordisk, Optoceutics, Passage Bio, Pinteon Therapeutics, Prothena, Quanterix, Red Abbey Labs, reMYND, Roche, Samumed, ScandiBio Therapeutics AB, Siemens Healthineers, Triplet Therapeutics, and Wave, has given lectures sponsored by Alzecure, BioArctic, Biogen, Cellectricon, Fujirebio, LabCorp, Lilly, Novo Nordisk, Oy Medix Biochemica AB, Roche, and WebMD, is a co-founder of Brain Biomarker Solutions in Gothenburg AB (BBS), which is a part of the GU Ventures Incubator Program, and is a shareholder of Prima Mente and MicThera (outside submitted work).

Ivan Koychev (IK) has received honoraria for advisory board roles from J&J and Novo Nordisk and non-promotional speaker fees from Eisai, is in receipt of an investigator-initiated grant from Novo Nordisk to explore the effects of a GLP-1 receptor agonist in preclinical dementia and is a medical advisor (stock options and/or retainer fees) to the following health technology companies: Five Lives, Oxford Brain Diagnostics, Leaf AI, Paloma Health and Prima Mente.

Sofia Toniolo (ST), Masud Husain (MH), Sian Thompson (SiT) are funded by the Wellcome Trust. Sanjay G. Manohar (SGM) is funded by a Medical Research Council (MRC) Clinician Scientist Fellowship and National Institute of Health and Care Research (NIHR) Oxford Biomedical Research Centre (BRC) and NIHR Oxford Health BRC. MH has received speaker and advisory board honoraria from Lilly, Otsuka, and Sumitomo.

Khaled Saab is a shareholder in Alphabet, Inc.

Netanel Loyfer, ST, SiT, and SGM declare no conflict of interest.

Two patents have been filed encompassing this work.

References

[1] Gram, H. Ueber die isolirte färbung der schizomyceten in schnitt- und trockenpräparaten von dr. gram, kopenhagen. — fortschritte der medicin 1884 no. 6. ref. dr. becker. *DMW - Deutsche Medizinische Wochenschrift* **10**, 234–235 (1884). URL https://doi.org/10.1055/s-0029-1209285.

[2] Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage t4. *Nature* **227**, 680–685 (1970). URL https://doi.org/10.1038/227680a0.

[3] Julius, M. H., Masuda, T. & Herzenberg, L. A. Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. *Proceedings of the National Academy of Sciences* **69**, 1934–1938 (1972). URL https://doi.org/10.1073/pnas.69.7.1934.

[4] Sanger, F., Nicklen, S. & Coulson, A. R. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977). URL https://doi.org/10.1073/pnas.74.12.5463.

[5] Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994). URL https://doi.org/10.1126/science.8303295.

[6] Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006). URL https://doi.org/10.1016/j.cell.2006.07.024.

[7] Tang, F. *et al.* mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009). URL https://doi.org/10.1038/nmeth.1315.

[8] Jinek, M. *et al.* A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012). URL https://doi.org/10.1126/science.1225829.

[9] Kadoshima, T. *et al.* Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human es cell-derived neocortex. *Proceedings of the National Academy of Sciences* **110**, 20284–20289 (2013). URL https://doi.org/10.1073/pnas.1315710110.

[10] Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021). URL https://doi.org/10.1038/s41586-021-03819-2.

[11] Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

[12] Zhou, Z. *et al. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes* (2023).

[13] Žiga Avsec *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**, 1196–1203 (2021). URL https://doi.org/10.1038/s41592-021-01252-x.

[14] Cui, H. *et al.* scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods* 1–11 (2024).

[15] Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **22**, 287–297 (2024). URL http://dx.doi.org/10.1038/s41592-024-02523-z.

[16] de Lima Camillo, L. P. *et al.* Cpgpt: a foundation model for dna methylation. *bioRxiv* 2024–10 (2024).

[17] Ying *et al.* Methylgpt: a foundation model for the dna methylome. *bioRxiv* (2024). URL https://www.biorxiv.org/content/early/2024/11/04/2024.10.30.621013.

[18] Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with evo. *Science* **386**, eado9336 (2024).

[19] Brixi *et al.* Genome modeling and design across all domains of life with evo 2. *bioRxiv* (2025). URL https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918.

[20] Holliday, R. & Pugh, J. E. Dna modification mechanisms and gene activity during development. *Science* **187**, 226–232 (1975). URL https://doi.org/10.1126/science.187.4173.226.

[21] Ferguson-Smith, A. C., Cattanach, B. M., Barton, S. C., Beechey, C. V. & Surani, M. A. Embryological and molecular investigations of parental imprinting on mouse chromosome 7. *Nature* **351**, 667–670 (1991). URL https://doi.org/10.1038/351667a0.

[22] Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

[23] Wang, A. *et al.* Epigenetic priming of enhancers predicts developmental competence of hesc-derived endodermal lineage intermediates. *Cell Stem Cell* **16**, 386–399 (2015). URL https://doi.org/10.1016/j.stem.2015.02.013.

[24] Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357** (2017). URL https://doi.org/10.1126/science.aal2380.

[25] Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015). URL https://doi.org/10.1038/nature13835.

[26] Vanyushin, B. F., Tkacheva, S. G. & Belozersky, A. N. Rare bases in animal dna. *Nature* **225**, 948–949 (1970). URL https://doi.org/10.1038/225948a0.

[27] Schübeler, D. Function and information content of dna methylation. *Nature* **517**, 321–326 (2015). URL https://doi.org/10.1038/nature14192.

[28] Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour dna. *Nature Reviews Cancer* **17**, 223–238 (2017). URL https://doi.org/10.1038/nrc.2017.7.

[29] Dai, W. *et al.* Epigenetics-targeted drugs: current paradigms and future challenges. *Signal Transduction and Targeted Therapy* **9** (2024). URL https://doi.org/10.1038/s41392-024-02039-0.

[30] Shireby, G. *et al.* Dna methylation signatures of alzheimer's disease neuropathology in the cortex are primarily driven by variation in non-neuronal cell-types. *Nature Communications* **13** (2022). URL https://doi.org/10.1038/s41467-022-33394-7.

[31] Sun, Y. *et al.* Identification of candidate dna methylation biomarkers related to alzheimer's disease risk by integrating genome and blood methylome data. *Translational Psychiatry* **13** (2023). URL https://doi.org/10.1038/s41398-023-02695-w.

[32] Nativio, R. *et al.* An integrated multi-omics approach identifies epigenetic alterations associated with alzheimer's disease. *Nature Genetics* **52**, 1024–1035 (2020). URL https://doi.org/10.1038/s41588-020-0696-0.

[33] Tsalenchuk, M., Gentleman, S. M. & Marzi, S. J. Linking environmental risk factors with epigenetic mechanisms in parkinson's disease. *npj Parkinson's Disease* **9** (2023). URL https://doi.org/10.1038/s41531-023-00568-z.

[34] Kochmanski, J., Kuhn, N. C. & Bernstein, A. I. Parkinson's disease-associated, sex-specific changes in dna methylation at park7 (dj-1), slc17a6 (vglut2), ptprn2 (ia-2β), and nr4a2 (nurr1) in cortical neurons. *npj Parkinson's Disease* **8** (2022). URL https://doi.org/10.1038/s41531-022-00355-2.

[35] Nabais, M. F. *et al.* Significant out-of-sample classification from methylation profile scoring for amyotrophic lateral sclerosis. *npj Genomic Medicine* **5** (2020). URL https://doi.org/10.1038/s41525-020-0118-3.

[36] Li, X. *et al.* Global, regional, and national burden of alzheimer's disease and other dementias, 1990-2019. *Frontiers in Aging Neuroscience* **14** (2022). URL https://doi.org/10.3389/fnagi.2022.937486.

[37] Farida B. Ahmad, R. N. A., Jodi A. Cisewski. Leading causes of death in the us, 2019-2023. *JAMA* **332**, 957 (2024). URL https://doi.org/10.1001/jama.2024.15563.

[38] Livingston, G. *et al.* Dementia prevention, intervention, and care: 2024 report of the lancet standing commission. *The Lancet* **404**, 572–628 (2024). URL https://doi.org/10.1016/s0140-6736(24)01296-0.

[39] Nichols & et al., E. Global, regional, and national burden of alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology* **18**, 88–106 (2019). URL https://doi.org/10.1016/S1474-4422(18)30403-4.

[40] Granzotto, A., Vissel, B. & Sensi, S. L. Lost in translation: Inconvenient truths on the utility of mouse models in alzheimer's disease research. *eLife* **13** (2024). URL https://doi.org/10.7554/elife.90633.

[41] Schindler, S. E. *et al.* Predicting symptom onset in sporadic alzheimer disease with amyloid pet. *Neurology* **97** (2021). URL https://doi.org/10.1212/wnl.0000000000012775.

[42] Jia, J. *et al.* Biomarker changes during 20 years preceding alzheimer's disease. *New England Journal of Medicine* **390**, 712–722 (2024).

[43] Zhang, J. *et al.* Recent advances in alzheimer's disease: mechanisms, clinical trials and new drug development strategies. *Signal Transduction and Targeted Therapy* **9** (2024). URL https://doi.org/10.1038/s41392-024-01911-3.

[44] Boxer, A. L. & Sperling, R. Accelerating alzheimer's therapeutic development: The past and future of clinical trials. *Cell* **186**, 4757–4772 (2023). URL https://doi.org/10.1016/j.cell.2023.09.023.

[45] DeKosky, S. T. & Marek, K. Looking backward to move forward: Early detection of neurodegenerative disorders. *Science* **302**, 830–834 (2003). URL https://doi.org/10.1126/science.1090349.

[46] Grande, G. *et al.* Blood-based biomarkers of alzheimer's disease and incident dementia in the community. *Nature Medicine* **31**, 2027–2035 (2025). URL https://doi.org/10.1038/s41591-025-03605-x.

[47] Barthélemy, N. R. *et al.* Highly accurate blood test for alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. *Nature Medicine* **30**, 1085–1095 (2024). URL https://doi.org/10.1038/s41591-024-02869-z.

[48] Therriault, J. *et al.* Diagnosis of alzheimer's disease using plasma biomarkers adjusted to clinical probability. *Nature Aging* **4**, 1529–1537 (2024). URL https://doi.org/10.1038/s43587-024-00731-y.

[49] Palmqvist, S. *et al.* Plasma phospho-tau217 for alzheimer's disease diagnosis in primary and secondary care using a fully automated platform. *Nature Medicine* **31**, 2036–2043 (2025). URL https://doi.org/10.1038/s41591-025-03622-w.

[50] della Monica, C. *et al.* P-tau217 and other blood biomarkers of dementia: variation with time of day. *Translational Psychiatry* **14** (2024). URL https://doi.org/10.1038/s41398-024-03084-7.

[51] Thomas Coysh, S. M. The future of seed amplification assays and clinical trials. *Frontiers in Aging Neuroscience* **14** (2022). URL https://doi.org/10.3389/fnagi.2022.872629.

[52] Wan, J. C. M., Sasieni, P. & Rosenfeld, N. Promises and pitfalls of multi-cancer early detection using liquid biopsy tests. *Nature Reviews Clinical Oncology* (2025). URL https://doi.org/10.1038/s41571-025-01033-x.

[53] Loyfer, N. *et al.* A dna methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).

[54] Baca, S. C. *et al.* Liquid biopsy epigenomic profiling for cancer subtyping. *Nature Medicine* **29**, 2737–2741 (2023). URL https://doi.org/10.1038/s41591-023-02605-z.

[55] Alexandra Bartolomucci, M. N. *et al.* Circulating tumor dna to monitor treatment response in solid tumors and advance precision oncology. *npj Precision Oncology* **9** (2025). URL https://doi.org/10.1038/s41698-025-00876-y.

[56] Caggiano, C. *et al.* Comprehensive cell type decomposition of circulating cell-free dna with celfie. *Nature communications* **12**, 2717 (2021).

[57] Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).

[58] Bao, H. *et al.* Early detection of multiple cancer types using multidimensional cell-free dna fragmentomics. *Nature Medicine* (2025). URL https://doi.org/10.1038/s41591-025-03735-2.

[59] Bruhm, D. C. *et al.* Genomic and fragmentomic landscapes of cell-free dna for early cancer detection. *Nature Reviews Cancer* **25**, 341–358 (2025). URL https://doi.org/10.1038/s41568-025-00795-x.

[60] Pollard, C., Aston, K., Emery, B. R., Hill, J. & Jenkins, T. Detection of neuron-derived cfdna in blood plasma: a new diagnostic approach for neurodegenerative conditions. *Frontiers in Neurology* **14**, 1272960 (2023).

[61] Khemka, S., Sehar, U., Manna, P. R., Kshirsagar, S. & Reddy, P. H. Cell-free dna as peripheral biomarker of alzheimer's disease. *Aging and disease* **16**, 787–803 (2025).

[62] Consortium, T. E. P. *et al.* Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). URL https://doi.org/10.1038/s41586-020-2493-4.

[63] Chang, H. Y. Anatomic demarcation of cells: genes to patterns. *Science* **326**, 1206–1207 (2009).

[64] Chalkidis, I., Dai, X., Fergadiotis, M., Malakasiotis, P. & Elliott, D. An exploration of hierarchical attention transformers for efficient long document classification (2022). URL https://arxiv.org/abs/2210.05529. arXiv:2210.05529.

[65] Cyrus Martin, Y. Z. The diverse functions of histone lysine methylation. *Nature Reviews Molecular Cell Biology* **6**, 838–849 (2005). URL https://doi.org/10.1038/nrm1761.

[66] Howard Cedar, Y. B. Linking dna methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics* **10**, 295–304 (2009). URL https://doi.org/10.1038/nrg2540.

[67] Jeong, H. *et al.* Evolution of dna methylation in the human brain. *Nature Communications* **12**, 2021 (2021). URL https://doi.org/10.1038/s41467-021-21917-7.

[68] Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341** (2013). URL https://doi.org/10.1126/science.1237905.

[69] Tian, W. *et al.* Single-cell dna methylation and 3d genome architecture in the human brain. *Science* **382**, eadf5357 (2023).

[70] Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology* **31**, 745–759 (2020). Publisher: Elsevier.

[71] Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* **13**, R83 (2012).

[72] Do, C. *et al.* Allele-specific dna methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory snps. *Genome biology* **21**, 1–39 (2020).

[73] Sun, T. *et al.* Systematic evaluation of cell type deconvolution methods for plasma cell-free dna. *bioRxiv* 2024–03 (2024).

[74] McKhann, G. M. *et al.* The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia* **7**, 263–269 (2011). URL https://pubmed.ncbi.nlm.nih.gov/21514250/.

[75] Schindler, S. E. & Atri, A. The role of cerebrospinal fluid and other biomarker modalities in the alzheimer's disease diagnostic revolution. *Nature Aging* **3**, 460–462 (2023). URL https://doi.org/10.1038/s43587-023-00400-6.

[76] Hansson, O., Blennow, K., Zetterberg, H. & Dage, J. Blood biomarkers for alzheimer's disease in clinical practice and trials. *Nature Aging* **3**, 506–519 (2023). URL https://doi.org/10.1038/s43587-023-00403-3.

[77] Jack, C. R. *et al.* Nia-aa research framework: Toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia* **14**, 535–562 (2018). URL https://doi.org/10.1016/j.jalz.2018.02.018.

[78] Lewczuk, P., Lukaszewicz-Zajac, M., Mroczko, P. & Kornhuber, J. Clinical significance of fluid biomarkers in Alzheimer's disease. *Pharmacological Reports* **72**, 528–542 (2020). URL https://doi.org/10.1007/s43440-020-00107-0.

[79] Palmqvist, S. *et al.* Discriminative accuracy of plasma phospho-tau217 for Alzheimer disease vs other neurodegenerative disorders. *JAMA* **324**, 772–781 (2020). URL https://doi.org/10.1001/jama.2020.12134.

[80] Imam, F. *et al.* The global neurodegeneration proteomics consortium: biomarker and drug target discovery for common neurodegenerative diseases and aging. *Nature Medicine* (2025). URL https://doi.org/10.1038/s41591-025-03834-0.

[81] Oh, H. S.-H. *et al.* Plasma proteomics links brain and immune system aging with healthspan and longevity. *Nature Medicine* (2025). URL https://doi.org/10.1038/s41591-025-03798-1.

[82] Brown, T. B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[83] So, D. *et al.* Searching for efficient transformers for language modeling. *Advances in neural information processing systems* **34**, 6010–6022 (2021).

[84] Alayrac, J.-B. *et al.* Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* **35**, 23716–23736 (2022).

[85] Adler, B. *et al.* Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704* (2024).

[86] Zhang, Z. *et al.* Relu ^{2} wins: Discovering efficient activation functions for sparse llms. *arXiv preprint arXiv:2402.03804* (2024).

[87] Su, J. *et al.* Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).

[88] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (2019).

[89] Micikevicius, P. *et al. Mixed precision training* (2018).

[90] Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022). URL https://doi.org/10.1126/science.abj6987.

[91] Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology* **38**, 276–278 (2020).

[92] Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics* **27**, 1571–1572 (2011).

[93] Garrison, E., Sirén, J., Novak, A. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology* **36**, 875–879 (2018).

[94] Robinson, J. D., Chuang, C.-Y., Sra, S. & Jegelka, S. *Contrastive learning with hard negative samples* (2020).

[95] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. *A simple framework for contrastive learning of visual representations*, 1597–1607 (PMLR, 2020).

[96] Loyfer, N., Rosenski, J. & Kaplan, T. wgbstools: A computational suite for dna methylation sequencing data representation, visualization, and analysis. *bioRxiv* 2024–05 (2024).

[97] Pedersen, B. S., Eyring, K., De, S., Yang, I. V. & Schwartz, D. A. Fast and accurate alignment of long bisulfite-seq reads. *bioRxiv* (2014). URL https://arxiv.org/abs/1401.1129.

[98] Ryan, D. P. Methyldackel: A (mostly) universal methylation extractor for bs-seq experiments. https://github.com/dpryan79/MethylDackel (2021).

Supplementary Materials

S1: Nucleotide Transformer Benchmarks

The official Nucleotide Transformer Benchmarks [15] consist of 18 classification tasks, where the sequences with the positive label are sampled from genomic regions with the special characteristics such as promoter and enhancer, and the negative sequences are sampled from the rest of the genome. Negative sequences in the official benchmark are not fully random, with the genomic positions of the random sequences starting only from genomic loci divisible by 1000. Fig. S1a depicts this bias in genomic start positions. Sequences with label 0, always start from genomic loci that have modulo 1000 with respect to 1000, but sequences with label 1 and 2 have a uniformly random distribution with respect to start position modulo 1000.

This systematic offset introduces a positional bias that inadvertently reduces the diversity of the negative set. Because Pleiades explicitly encodes genomic coordinates, it achieves near-perfect scores on the original NT benchmarks (see Fig. S1c)

In order to fix this bias, we add a random jitter in the range of $[-500, 499)$ to the start positions of negative sequences (Fig. S1b). All three labels have an indistinguishable distribution with respect to the start position modulo 1000. We called this new dataset the *Unbiased* Nucleotide Transformer Benchmark.

Fig. S1d shows that our results after fine-tuning the baseline models DNABERT-2 and NT MS 2.5B for five epochs on our Unbiased NT Benchmarks.

Table S1: Performance (AuROC) of all models on unbiased Nucleotide-Transformer benchmark tasks (values rounded to four decimal places).

Task	DNABERT-2	NT MS 2.5 B	Pleiades 90 M	Pleiades 600 M	Pleiades 7 B
H2AFZ	0.4903	0.5358	0.7325	0.7147	**0.9837**
H3K27ac	0.4778	0.5141	0.7658	0.7390	**0.9988**
H3K27me3	0.5965	0.6294	0.7893	0.7948	**0.9912**
H3K36me3	0.6387	0.6636	0.7391	0.8043	**0.9953**
H3K4me1	0.4353	0.4886	0.7602	0.7466	**0.9933**
H3K4me2	0.5523	0.5695	0.7651	0.7661	**0.9915**
H3K4me3	0.6171	0.6328	0.7148	0.6827	**0.9768**
H3K9ac	0.5662	0.5423	0.7307	0.7001	**0.9960**
H3K9me3	0.4675	0.4876	0.6996	0.7063	**0.9670**
H4K20me1	0.6434	0.6608	0.8192	0.8051	**0.9982**
Enhancers	0.5231	0.5461	0.6230	0.6297	**0.9770**
Enhancers (types)	0.4658	0.5031	0.5860	0.5977	**0.9000**
Promoters	0.7546	0.7581	0.7416	0.7540	**0.9759**
Promoters (non-TATA)	0.7440	0.7921	0.7530	0.7643	**0.9725**
Promoters (TATA)	0.7767	0.8211	0.6002	0.7616	**0.9908**
Splicing (Acceptors)	0.8308	**0.9657**	0.9509	0.9593	0.9644
Splicing (All)	0.8496	**0.9693**	0.9597	0.9481	0.9659
Splicing (Donors)	0.8309	**0.9736**	0.9652	0.9694	0.9660

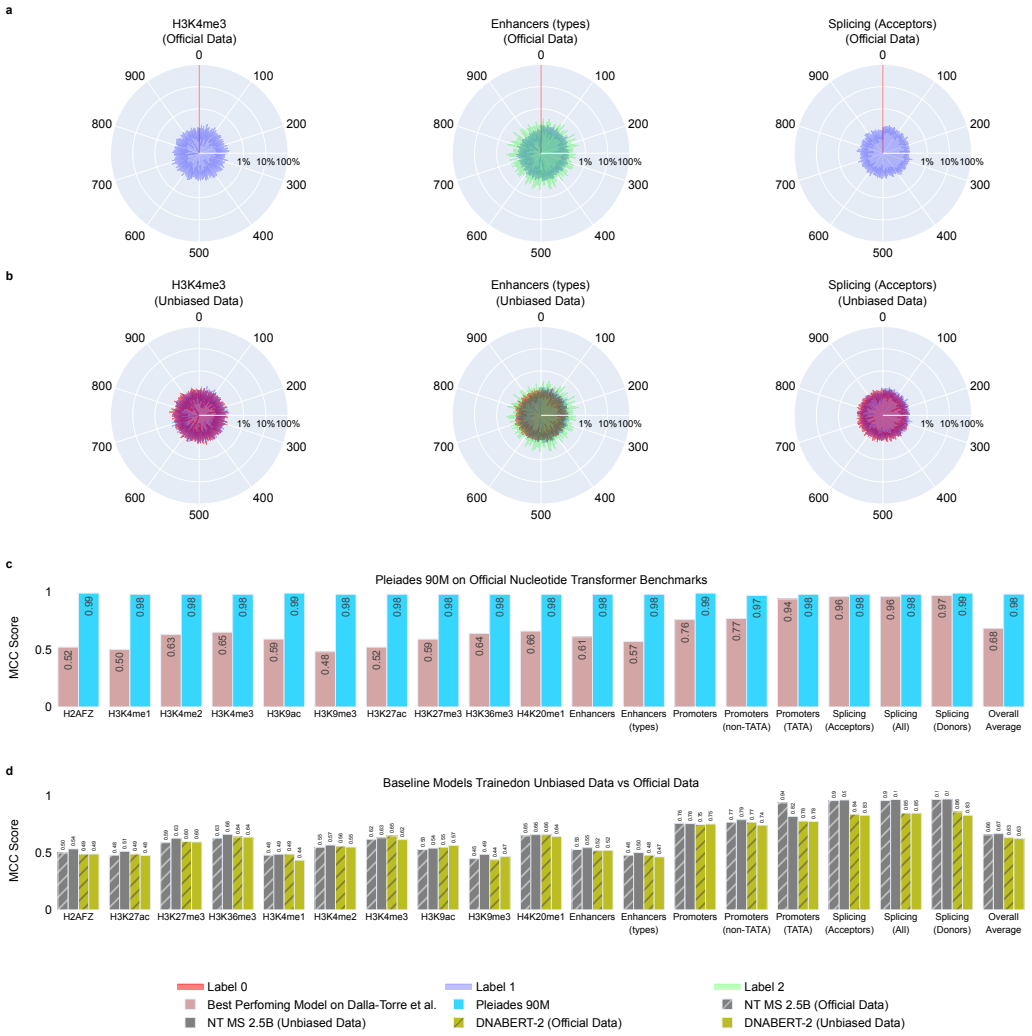

Fig. S1: **NT Benchmarks Positional Bias, Unbiased Dataset and Performance Comparisons (a)** Polar plot showing distribution of start position modulo 1000 of Official NT Benchmarks, separated by label. **(b)** Unbiased dataset where the bias is removed. **(c)** Comparison of MCC for Pleiades 90M vs Best Performing Baseline Models from Nucleotide Transformer [15] on official NT Benchmarks Data. **(d)** MCC Comparison for Baseline Models on NT-Benchmarks Official Data vs our Unbiased Data.

S2: cfDNA Generation Task Regions

Supplementary Table S2 lists all the regions used in the analysis in section 2.3. The stable nucleosome was determined using NucPosDB[1]. To identify these regions, we selected 80 high-coverage 1kb intervals from our cfDNA test-set biosamples that do not overlap any repetitive elements (as defined by the UCSC Genome Browser's RepeatMasker track).

Table S2: 1kb Regions Evaluated for cfDNA Generation

Chromosome	Region Start	Region End	Stable Chromatin
chr1	16,617,000	16,618,000	No
chr1	107,143,000	107,144,000	Yes

Continued on next page...

[1] https://generegulation.org/stable-nucleosomes/

Table S2 (continued)

Chromosome	Region Start	Region End	Stable Nucleosome
chr2	89,768,000	89,769,000	No
chr2	174,338,000	174,339,000	Yes
chr2	192,197,000	192,198,000	Yes
chr3	107,527,000	107,528,000	No
chr3	113,945,000	113,946,000	Yes
chr4	105,149,000	105,150,000	No
chr4	183,101,000	183,102,000	Yes
chr5	120,466,000	120,467,000	No
chr5	141,363,000	141,364,000	Yes
chr5	146,379,000	146,380,000	Yes
chr6	1,514,000	1,515,000	No
chr6	10,885,000	10,886,000	No
chr6	53,648,000	53,649,000	No
chr6	56,543,000	56,544,000	No
chr6	70,960,000	70,961,000	No
chr6	79,079,000	79,080,000	No
chr6	84,770,000	84,771,000	No
chr6	96,833,000	96,834,000	Yes
chr6	98,838,000	98,839,000	No
chr7	28,177,000	28,178,000	Yes
chr7	28,850,000	28,851,000	No
chr7	29,196,000	29,197,000	Yes
chr8	30,386,000	30,387,000	No
chr8	64,584,000	64,585,000	No
chr8	66,173,000	66,174,000	No
chr8	66,922,000	66,923,000	Yes
chr8	80,488,000	80,489,000	No
chr8	123,271,000	123,272,000	No
chr8	124,369,000	124,370,000	No
chr9	968,000	969,000	Yes
chr9	86,942,000	86,943,000	Yes
chr9	89,003,000	89,004,000	No
chr10	22,314,000	22,315,000	No
chr10	133,526,000	133,527,000	Yes
chr11	59,062,000	59,063,000	Yes
chr11	125,168,000	125,169,000	No
chr12	67,271,000	67,272,000	Yes

Continued on next page...

Table S2 (continued)

Chromosome	Region Start	Region End	Stable Nucleosome
chr12	79,687,000	79,688,000	No
chr13	46,384,000	46,385,000	Yes
chr13	99,976,000	99,977,000	No
chr14	19,344,000	19,345,000	No
chr14	61,697,000	61,698,000	Yes
chr15	20,360,000	20,361,000	No
chr15	35,118,000	35,119,000	No
chr15	52,567,000	52,568,000	No
chr15	66,295,000	66,296,000	Yes
chr15	67,523,000	67,524,000	No
chr15	81,004,000	81,005,000	No
chr15	96,335,000	96,336,000	No
chr16	46,390,000	46,391,000	No
chr16	79,596,000	79,597,000	Yes
chr17	44,828,000	44,829,000	Yes
chr17	70,166,000	70,167,000	No
chr18	3,497,000	3,498,000	No
chr18	3,502,000	3,503,000	No
chr18	31,103,000	31,104,000	No
chr18	57,356,000	57,357,000	Yes
chr18	70,204,000	70,205,000	No
chr18	80,146,000	80,147,000	No
chr19	266,000	267,000	No
chr20	34,214,000	34,215,000	No
chr20	59,943,000	59,944,000	Yes
chr21	10,625,000	10,626,000	No
chr21	32,414,000	32,415,000	Yes
chr22	11,624,000	11,625,000	No
chr22	25,567,000	25,568,000	Yes

S3: Cell Type-of-Origin

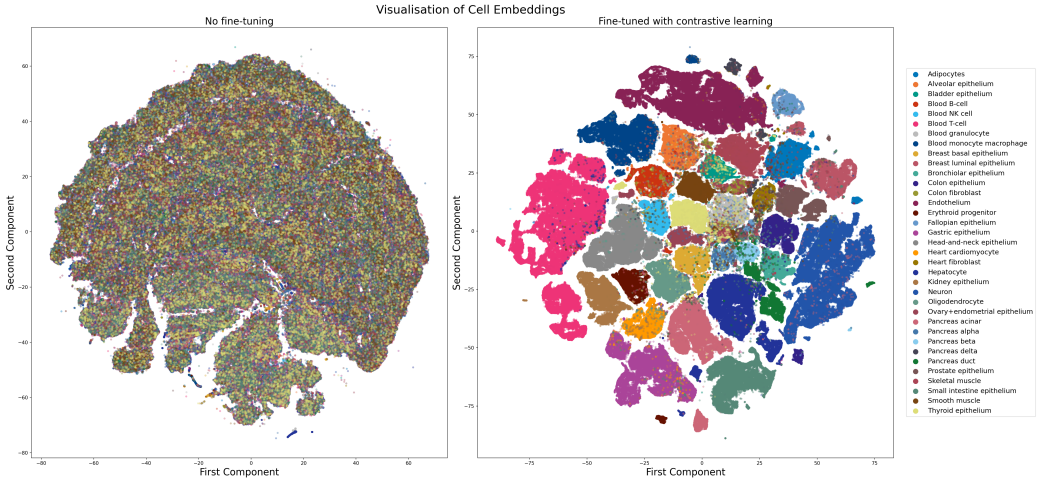

Fig. S2: Pleiades 90M Embeddings reduced to 2 dimensions using UMAP and colour-coded by cell type. The left is the pretrained model's representations with no discernible separation between cell types. The right image shows the model representations following fine-tuning.

Mixture Name	B cell	T cell	Monocyte & Macrophage	Neuron	Oligodendrocyte	Hepatocyte
1-Cell Mix	1					
1-Cell Mix		1				
1-Cell Mix			1			
1-Cell Mix				1		
1-Cell Mix					1	
1-Cell Mix						1
2-Cell Mix	0.5	0.5				
2-Cell Mix			0.5	0.5		
2-Cell Mix		0.5				0.5
3-Cell Mix		0.33		0.34	0.33	
3-Cell Mix				0.34	0.33	0.33
3-Cell Mix	0.33			0.34		0.33
4-Cell Mix	0.25	0.25	0.25	0.25		
6-Cell Mix	0.17	0.17	0.17	0.16	0.16	0.17

Table S3: Cell-Type mixture proportions used for deconvolution benchmarking. Blank entries indicate zero proportion.

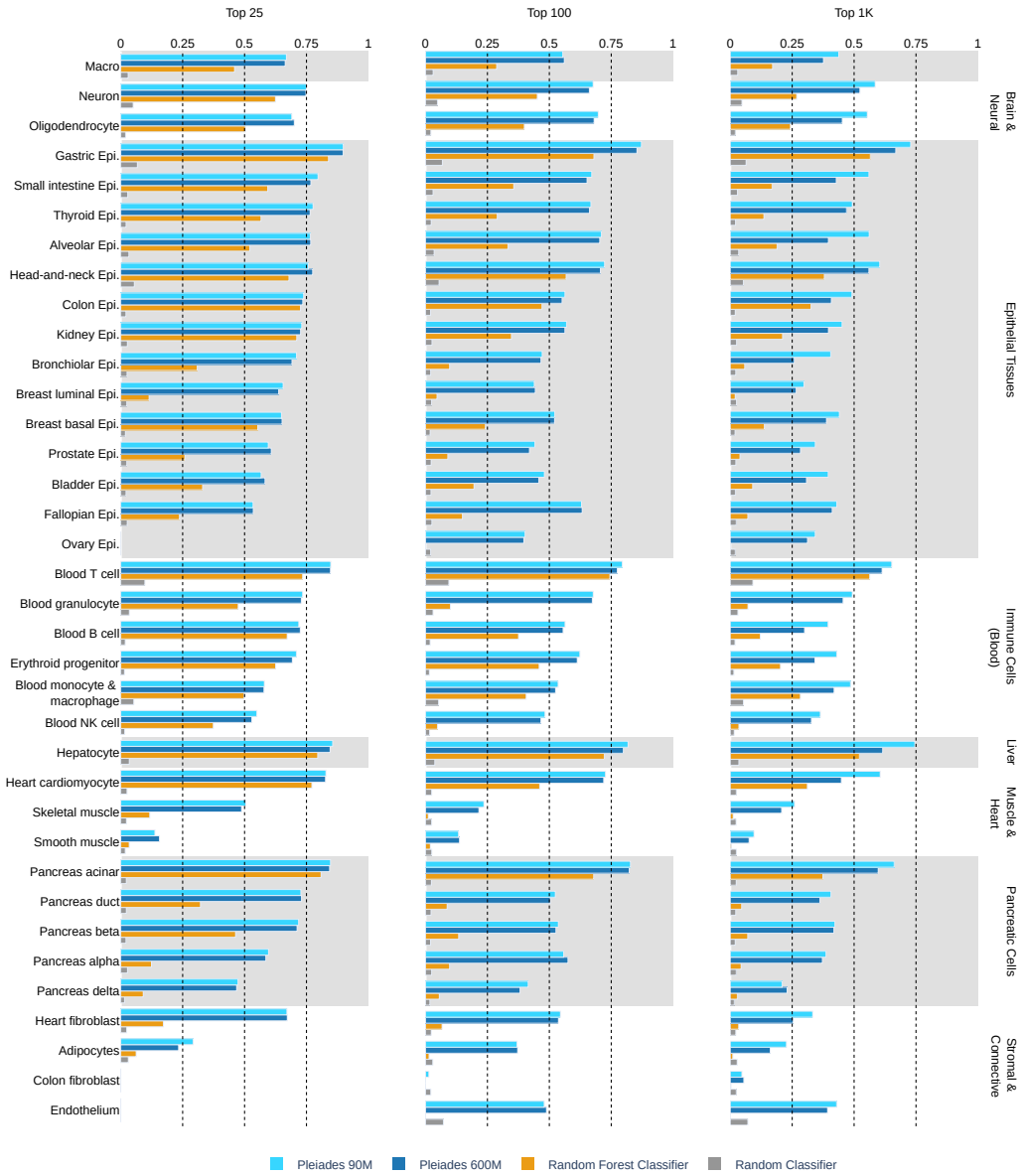

Fig. S3: CToO F1 scores over official Top 25, Top 100 and Top 1000 markers called from training data.

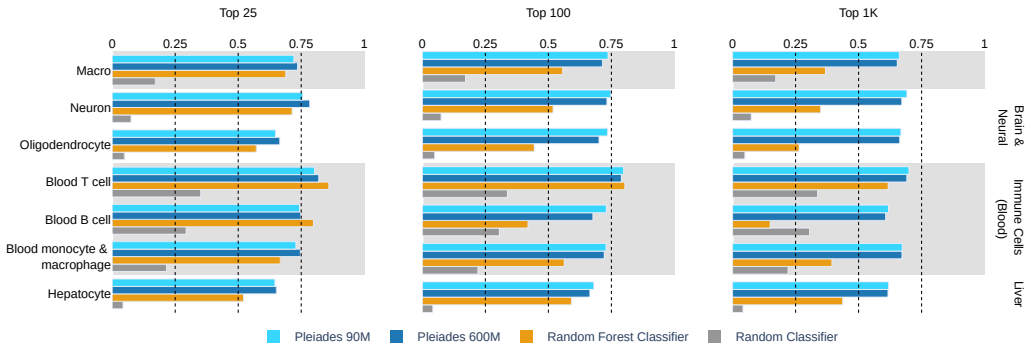

Fig. S4: Out-of-distribution CToO F1 scores over official Top 25, Top 100 and Top 1000 markers called from training data.

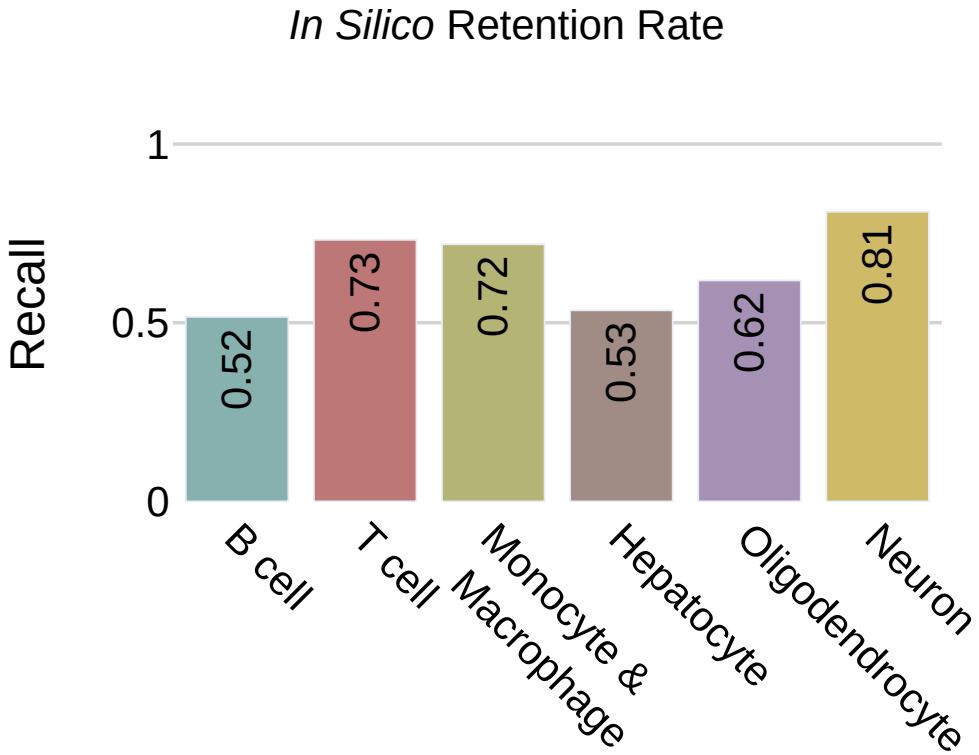

Fig. S5: Out-of-distribution CToO Recall over Top 1000 markers called from training data.

S4: Neurodegenerative Disease Diagnosis

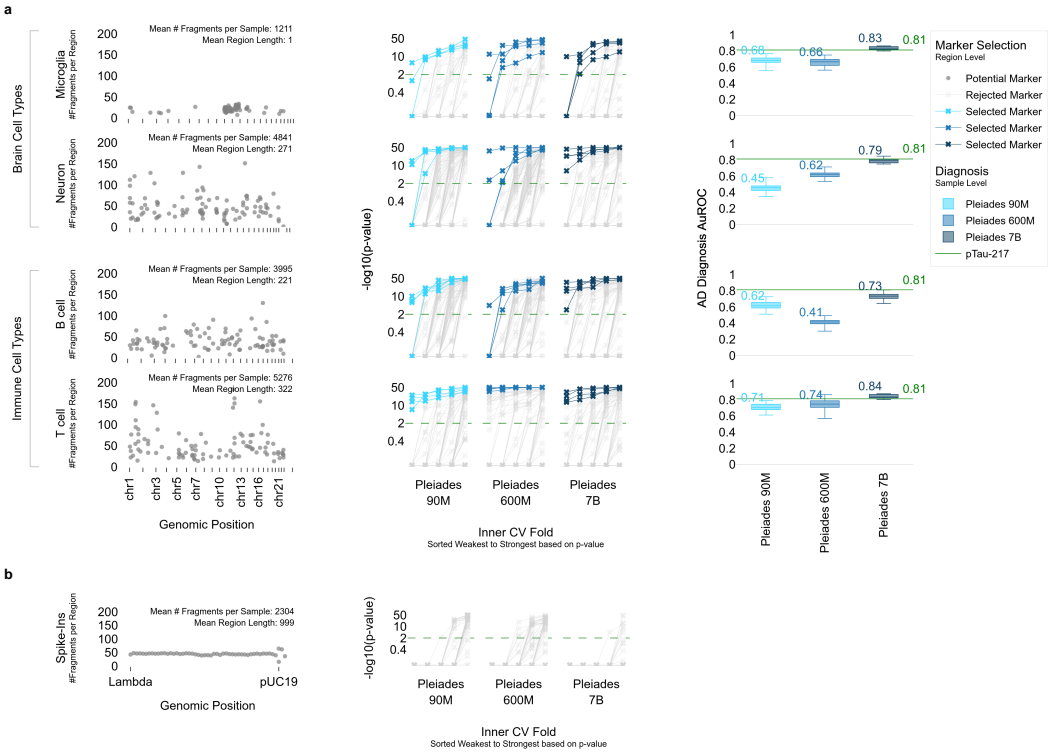

Fig. S6: **AD Marker Discovery Process with Pleiades (a)** AD marker discovery pipeline on an example fold of the outer five-fold split. The train set is divided into five inner folds. Starting from broad genomic regions across four cell type marker region sets, we select the regions with AuROC > 0.6 with p-value < 0.01 for at least four out of five inner folds. If more than four regions pass this test, we sort by number of folds passing the test and average AuROC, both descending, and select only top four. If any regions are selected, we train the region-level and sample-level models on those selected regions only. Final sample level performance in this example outer fold is shown along with the starting regions and the T-test outcomes for region selection. **(b)** Same method applied to pUC19 and Lambda DNA spike-ins; no signal was detected and the model rejected all regions.

.

Proteomic Marker	Test Direction	Wilcoxon Statistic	p−value
AB40 (pg/ml)	Greater	15.0	0.0312
AB42 (pg/ml)	Greater	15.0	0.0312
$\frac{\text{AB42 (pg/ml)}}{\text{AB40 (pg/ml)}}$	Greater	15.0	0.0312
NFL (pg/ml)	Greater	15.0	0.0312
GFAP (pg/ml)	Greater	14.0	0.0625
pTau-181 (pg/ml)	Greater	13.0	0.0938
pTau-217 (pg/ml)	Greater	4.0	0.6875
pTau-217 (pg/ml)	Less	4.0	0.4375

Table S4: Comparison of Pleiades 7B AuROC (average pooled on all cell type markers) vs. Proteomic Markers using one-sided Wilcoxon Test. Test direction column represents the direction used during testing, with *Greater* meaning Pleiades AuROC > proteomic AuROC and *Less* meaning Pleiades AuROC < proteomic AuROC. For all markers other than pTau-217, only *Greater* direction was used due to the large distance in means in favour of Pleiades 7B.