



Detecting ultra-rare neuronal cell-free DNA in blood with Pleiades-Neuron

We used our epigenetic foundation model, Pleiades, to detect ultra-rare neuronal cell-free DNA and applied it to the detection of diseases of the brain.

Pooja Kathail ^{1,*}, **Saumya Gupta** ^{1,*}, **Netanel Loyfer** ¹, **Javkhlan-Ochir Ganbat** ¹,
Anjeet Jhutti¹, **Husam Babikir**¹, **Francisco M. Martín-Zamora** ¹, **Anooj Patel** ¹,
Luca Giacomoni ¹, **Yuval Dor** ², **Pouya Niki** ¹, and **Ravi Solanki** ¹

20 min read

¹Prima Mente, London, UK; and San Francisco, USA

²The Hebrew University of Jerusalem, Jerusalem, Israel

*These authors contributed equally to the work

Correspondence: pooja@primamente.com, saumya@primamente.com, and ravi@primamente.com

Our approach to modeling cell-free DNA

At Prima Mente, we believe that minimally invasive biomarkers for diseases of the brain are essential for early detection, longitudinal profiling and prognostication, treatment response determination and prediction, and development of molecular endpoints. Cell-free DNA (cfDNA) is one of many such biomarkers we study and model, and are fragments of DNA that are released by dying cells into the bloodstream¹. Chemical marks on cfDNA, known as DNA methylation, can help elucidate the cell-of-origin of each fragment²⁻⁶. This has been the fundamental basis behind using blood-based cfDNA liquid biopsies to detect changes or damage to organs or tissues that are otherwise difficult to access⁷⁻⁹. It is the

biology that underpins many of the modern-day early cancer detection, minimum residual disease monitoring, and non-invasive prenatal tests.

From a modeling perspective, cfDNA is particularly challenging. A single patient blood sample contains hundreds of millions of cfDNA fragments; a 50× coverage sample corresponds to approximately 150 billion tokens - equivalent to reading the Harry Potter series more than 50,000 times over. Often only a small fraction of fragments - 1% or below - have disease-relevant signal^{10,11}. We introduced our approach to this problem in Niki *et al.*¹², and continue to develop new methods in each area:

In this post we present Pleiades-Neuron, a new model for **per-fragment representation learning** built to detect neuron-derived cfDNA. This is a particularly challenging task because brain-derived cfDNA is present at extremely low concentrations in blood^{4,8,15,16}. Pleiades-Neuron identifies neuron-derived fragments with extremely high sensitivity and distinguishes samples containing one part-per-million neuron-derived DNA from clean samples. As a proof-of-concept to demonstrate Pleiades-Neuron's utility for brain-disease detection, we show that the model can accurately distinguish samples with potential injury to the brain - such as Neurological Intensive Care Unit patients - from healthy controls.

Key findings

- **Model scaling:** Performance improves consistently with model scale, from Pleiades 90M to 600M to 7B, without signs of plateauing at 7B parameters.
- **Data scaling:** Performance also continues to improve as we scale the size of the training corpus.
- **Distinguishing neuron-derived DNA at one part-per-million concentration:** At 50× sequencing depth, Pleiades-Neuron can separate samples containing one part-per-million (10^{-6}) neuron-derived DNA from blank samples. Hard example mining, data diversity, and a likelihood-ratio scoring method are critical components to model performance.
- **60× speedup with model distillation:** Distilling Pleiades-Neuron into substantially smaller versions retains 99.8% of performance while providing 60× faster inference, making it practical to inference large disease cohorts.
- **Mechanistic interpretability:** Pleiades-Neuron identifies both known and novel non-CpG methylation patterns as biomarkers of neuron-derived DNA, suggesting attributes of neuron DNA methylation for further study.
- **Blood-based detection of brain-related conditions:** Pleiades-Neuron predicts elevated neuron-derived cfDNA for multiple conditions involving potential injury to the brain: donors after brain death (DBD), patients in the Neurological Intensive Care Unit (Neuro-ICU), and patients with measles.

Challenges and opportunities in brain-derived cfDNA detection

Estimating the cell-type composition of cell-free DNA is a clinically important problem with numerous existing methods^{3,5,6,17–20}. Existing approaches, which compare observed methylation to reference atlases, often perform well for

abundant cell types but poorly for rare cell types - such as the brain - that may be present at less than a 1% concentration in cfDNA derived from blood plasma^{3,6}. In addition, existing approaches typically operate on cfDNA fragments from a small set of differentially methylated regions (DMRs)^{3,5,17,19,20}, and effectively discard the vast majority of cfDNA fragments from genome-wide methylation sequencing.

Brain-derived cfDNA in particular is present at extremely low concentrations in blood plasma^{4,8,15,16,21,22}. Prior work estimated circulating tumor DNA from glioblastoma patients to be in the parts-per-million range¹⁵, indicating that parts-per-million detection sensitivity may be necessary for blood-based liquid biopsies for brain pathologies. Prior literature has identified DNA methylation at non-CpG sites as a neuron-specific biomarker, which suggests a promising path for increasing detection sensitivity to the parts-per-million range²³⁻²⁸. Recent work demonstrated a detection method based upon non-CpG methylation using relatively simple heuristics²⁹.

In parallel, epigenetic foundation modeling has emerged as a promising approach for representation learning of cell-type specific DNA methylation patterns^{12,30}. Foundation model representations can capture subtle epigenetic signals from rare cell types and from genome-wide data. In prior work, we developed Pleiades, an autoregressive transformer decoder at three scales - 90M, 600M, and 7B parameters - pre-trained on 1.9T tokens of healthy human cellular and cell-free methylation and genomic data¹². We previously showed Pleiades to be state-of-the-art (SOTA) for multi-class cell-type-of-origin prediction.

Here we introduce Pleiades-Neuron, a fine-tuned version of Pleiades focused on the problem of detecting ultra-rare, neuron-derived cell-free DNA fragments.

Training and evaluation data

We first assembled DNA methylation sequencing datasets for training and evaluating Pleiades-Neuron. Our primary datasets^{3,31,32} are from samples with a known cell-type-of-origin. These contain **fragment-level labels**: each fragment can be labeled as neuron-derived or non-neuron-derived. However, for the use case of disease detection, we are interested in predicting **sample-level labels**. A sample - a set of cfDNA fragments from an individual patient - can be labeled as “disease” or “healthy”, but the cell-type origin of each individual cfDNA fragment remains unknown. To evaluate performance in this setting, we created *in silico* data mixtures as described below.

Neuron classification data: the fragment-level labels

We created a corpus of DNA methylation sequencing data containing ~630 million fragments and a total of ~120 billion tokens (Fig. 2a)^{3,31,32}. Neuron-derived fragments come from post-mortem brain tissue of eight donors (seven male, one female) aged 19–65 with no recorded neurological diagnosis. Non-neuron-derived fragments come from 177 donors aged 3–83 (where reported), primarily healthy donors, as well as 20 patients with mild cognitive impairment or Alzheimer’s disease. For robustness across methylation sequencing protocols, the corpus includes both whole-genome bisulfite sequencing (WGBS) and enzymatic methylation sequencing (EM-seq) data. However, EM-seq makes up less than 0.5% of training fragments, all in the non-neuron class - a known limitation of our corpus.

We processed every dataset from raw FASTQ files with one standardized pipeline to minimize preprocessing differences. We further processed fragments by applying an adaptive clipping algorithm to remove spurious methylation calls at fragment ends²⁹. To remove easy negative fragments, we filtered to fragments with at least 3 methylated CpA sites based upon prior work that methylated CpA is a neuron-specific marker²⁹. Finally, we created training, validation, and test splits stratified by donor. Our final training set is composed of 347M neuron-derived and 16M non-neuron-derived fragments, a total of ~71B tokens.

Key evaluation metric

Neuron-derived fragments are at extremely low concentration in cfDNA, so our deployment setting is highly class-imbalanced. We therefore focused on the low false positive rate (FPR) regime, measured by a **log-scale partial AUC (pAUC)** over the 10^{-2} to 10^{-6} FPR range. A perfect classifier scores 1, and a random classifier scores only $\approx 1.1 \times 10^{-3}$ rather than the 0.5 of standard AUC. We use a log scale because on a linear scale the 10^{-6} – 10^{-5} range, critical for ultra-rare detection, would contribute almost nothing to the score. We empirically observe that pAUC tracks neuronal fragment recovery in low concentration *in silico* spike-in samples (Fig. 3a).

In silico spike-in data: the sample-level labels

To evaluate neuron-derived cfDNA detection accuracy at extremely low concentrations, we created a set of *in silico* data mixtures where neuron-derived fragments are “spiked in” to a set of background non-neuron-derived fragments at a known, fixed concentration, as low as one part-per-million (10^{-6}). We considered two different *in silico* data mixtures, each with distinct advantages and drawbacks.

The **in-distribution data mixture** (Fig. 2b) is made up of a pseudo-plasma background set; labeled non-neuron cell type data are mixed together in proportions typically found in plasma cfDNA³ to mimic a 50× coverage sample. Its advantage is a background set free of contamination from brain fragments; its drawback is that it lacks characteristics of real cfDNA, such as the distinctive fragment length distribution^{1,33–36}. Also, we note that megakaryocytes make up a large proportion of typical blood plasma cfDNA but are excluded from our *in silico* mixture due to data availability. Finally, even though the donors are held out from training and validation, both the background and neuron-derived fragments come from studies used to train Pleiades-Neuron, hence we refer to this data mixture as in-distribution.

We also evaluated an **out-of-distribution (OOD) data mixture** (Fig. 2c). In this case, the background set comes from real plasma cfDNA from healthy individuals that were sequenced in-house. This background may itself contain some neuron-derived fragments, making it a more realistic proxy for disease detection, where the goal is to detect an elevated level of neuron-derived cfDNA relative to a healthy baseline. To assess robustness across assays, we considered background sets from both WGBS and EM-seq. Finally, we spiked in neuron-derived fragments from an entirely held-out WGBS study³⁷ to ensure the data mixture is out-of-distribution relative to the training corpus.

Key evaluation metric

We report a **sample-level AUC** between “blank” samples consisting of only background fragments and “spike-in” samples that also contain neuron-derived fragments at a known, fixed concentration (e.g. 10^{-6}). We report the sample-level AUC on the three data mixtures described above: (1) in-distribution, (2) out-of-distribution (WGBS), and (3) out-of-distribution (EM-seq).

Results

Pleiades-Neuron exhibits benefits from both model and data scaling

To assess the utility of model capacity, we trained simple MLP probes on frozen representations from Pleiades 90M, 600M and 7B, measuring how much neuron-specific signal the pre-trained representations already capture. All probes outperform the baseline heuristic²⁹, and, importantly, **performance improves with model scale without signs of plateauing** (Fig. 3a, b). pAUC correlates strongly with the in-distribution spike-in sample-level AUC at 10^{-6} concentration ($R^2 = 0.976$; Fig. 3a), supporting its use as our primary classification metric.

Along sequencing depths from 10x to 85x, larger models keep their lead and the gap widens with depth (Fig. 3b); all models perform best at 85x, highlighting the importance of high-coverage samples for low-concentration neuronal fragment detection. Based on these data, we moved forward with Pleiades 7B as the base model for all subsequent experiments.

We next assessed the utility of data scale by performing a probing sweep with frozen Pleiades 7B representations using up to 10M training examples. **We observe consistent improvement in model performance with dataset size** (Fig. 3c), motivating us to scale up subsequent experiments to the full >350M training set.

Pleiades-Neuron detects ultra-rare neuron-derived fragments

Pleiades-Neuron, our final model, distinguishes neuron-derived DNA from background at concentrations as low as one part-per-million. It has three major components:

- Pleiades-Neuron is initialized with weights from the Pleiades 7B base model, an autoregressive transformer decoder pre-trained on 1.9T tokens of human methylation and genomic data (Fig. 4a).
- Pleiades-Neuron is then LoRA fine-tuned³⁸ to classify neuron-derived versus non-neuron-derived DNA fragments (Fig. 4b). Ablation studies reveal that hard example mining and data diversity across sequencing assays are critical components to model performance (Table 1).
- Finally, we developed a likelihood-ratio scoring method to derive a single sample-level score from a set of fragments (Fig. 4c). Across all model variants, this scoring method outperforms using a fixed score threshold (Fig. 6).

We compared Pleiades-Neuron against three baselines: a heuristic based on methylated CpA counts²⁹, MethylBERT³⁰ (a recent Transformer-based CpG methylation model), and a goM transformer decoder trained from scratch. Pleiades-Neuron outperforms all three at 10^{-6} concentration, with a sample-level AUC of 0.933 in-distribution (Table 1). **Its gain over the baseline heuristic comes mainly from sensitivity: at a fixed specificity it captures 4.4 times as many true positives, and at the optimal threshold from a calibration set the improvement comes almost entirely from sensitivity rather than specificity** (Fig. 5a).

We evaluated Pleiades-Neuron in multiple scenarios meant to represent realistic forms of data distribution shift. First, on the out-of-distribution evaluation sets (real plasma cfDNA background, neuron-derived fragments from a held-out study; see above), performance drops relative to in-distribution, but Pleiades-Neuron still outperforms every baseline (Table 1). **Furthermore, we see strong performance on the OOD dataset with EM-seq background fragments, even though EM-seq data makes up less than 0.5% of our training corpus.** The gap between the WGBS and EM-seq OOD datasets may reflect several factors. Because the spiked-in neuron-derived fragments are always WGBS, the task may be easier on an EM-seq background. Differences in background plasma quality and sources of experimental noise such as methylation conversion efficiency also contribute. We expect that expanding the training corpus to reflect the heterogeneity of methylation sequencing data will further improve generalization.

We also considered the impact of sequencing depth on performance. Pleiades-Neuron outperforms the baseline heuristic at all depths considered (10x to 100x) and steadily improves at higher sequencing depths on both the in- and out-of-distribution evaluations (Fig. 5b–d). **On the in-distribution set, it surpasses a 0.9 spike-in AUC at 10^{-6} concentration for sequencing depths 40x and above** (Fig. 5b).

Model hyperparameters			Spike-in sample-level AUC at 10^{-6} concentration		
Model name	Data corpus	Hard mining	In-distribution (50x)	OOD – WGBS (50x)	OOD – EM-seq (50x)
CpA heuristic	-	-	0.573	0.523	0.531
MethylBERT	No EM-seq	Pos & neg (2%)	0.567	0.526	0.528
goM from scratch	Full	Neg only (2%)	0.872	0.591	0.729
Pleiades-Neuron	Full	Neg only (2%)	0.933	0.625	0.786
Pleiades-Neuron distilled goM	-	-	0.933	0.618	0.769
Pleiades-Neuron distilled 12M	-	-	0.931	0.623	0.776
without hard mining	Full	None	0.923	0.613	0.761
with positive and negative mining	Full	Pos & neg (2%)	0.937	0.615	0.768
without EM-seq	No EM-seq	Pos & neg (2%)	0.937	0.615	0.714

Table 1. Performance comparison across Pleiades-Neuron model variants and baselines.

We then ablated the Pleiades-Neuron training recipe to see what each part contributes. As expected, including EM-seq data in the training corpus improves performance on the EM-seq OOD dataset (Table 1). We reasoned that hard positive and hard negative mining may not be equally useful: not every neuron-derived fragment carries markers of its origin, so

hard positive mining may concentrate learning on fragments without informative sequence features. Indeed, hard negative mining alone improves performance, especially OOD, while mining both classes has little to no effect (Table 1).

A final component of Pleiades-Neuron is a likelihood-ratio scoring method (Fig. 4c). Intuitively, it carries more information than a fixed threshold, which reduces model scores to binary predictions, and is better calibrated than raw model outputs (Fig. 6a). Both methods use the same calibration set of held-out positive and negative fragments: the fixed-threshold approach selects a score threshold from it, and likelihood-ratio scoring estimates how often each score occurs among neuronal versus background calibration fragments. In both cases we then sum across fragments to obtain a sample-level score reflecting evidence of neuron-derived DNA. **Across all model variants, likelihood-ratio scoring outperforms the fixed threshold approach** (Fig. 6b).

Lastly, we distilled Pleiades-Neuron to enable inference on disease cohorts, which can contain hundreds to thousands of 50× coverage samples, approximately 150B tokens per patient. Distillation trains a smaller student model to reproduce the predictions and internal representations of a larger teacher, retaining as much of the teacher’s performance as possible at a fraction of the compute requirement during inference. We trained student models with a loss that is a weighted sum of three terms: a classification loss on the ground truth labels, a logit-matching loss against the teacher’s output, and a representation-alignment loss on the trunk activations. **Student models distilled to 90M and 12M both perform comparably to the full 7B Pleiades-Neuron model, enabling a ~60× speedup while preserving 99.8% of performance** (Table 1).

Pleiades-Neuron learns known and novel non-CpG methylation patterns that are markers of neuron-derived DNA

We interpreted the Pleiades-Neuron model to identify the key features driving its predictions. Prior literature identifies non-CpG, and especially CpA, methylation as a neuron-specific marker^{27,28}, so we expected it to be a key feature. Nonetheless, we took an unbiased approach and simply asked which sequence patterns drive model predictions. As expected, methylated CpA sites have the largest causal impact on Pleiades-Neuron’s predictions, relative to their share of sequence positions (Fig. 7a). Methylated CpT and CpC sites also have a disproportionate impact on predictions. **An unbiased insertion analysis over all di-, tri- and tetranucleotide contexts identified mCAC, mCAG, and mCAT as the strongest trinucleotide neuron markers and mCACC as the strongest tetranucleotide marker, in agreement with prior literature^{27,39}** (Fig. 7b). Interestingly, examining tetranucleotide contexts showed substantial variability within mCAC and mCAG that to our knowledge has not yet been previously reported, suggesting that more complex motifs or sequence patterns may be markers of neuron-derived DNA.

Finally, we previously reported that Pleiades learns fragment length features as a marker of Alzheimer’s disease⁴⁰. By contrast, the performance of Pleiades-Neuron in fragment-length stratified quartiles closely matches the overall performance, indicating that it is not learning length-based features (Fig. 7c). Furthermore, a simple predictor based on fragment length alone has no capability to predict neuron-derived DNA (pAUC < 0.0009). This matches our expectation: the model was trained only upon genomic DNA from healthy individuals, where fragment length carries no biologically meaningful variability.

Pleiades-Neuron enables brain disease detection from blood-based cell-free DNA

The ultimate goal of Pleiades-Neuron is the detection, stratification, and profiling of diseases of the brain. As a proof-of-concept, we evaluated it on clinical cfDNA samples. We use cerebrospinal fluid (CSF) derived cfDNA as an easy first test case because CSF carries far more neuron-derived cfDNA than plasma. As expected, both Pleiades-Neuron and the baseline heuristic easily differentiate CSF samples from healthy blood plasma samples (Fig. 8a, b: right-most box; Fig. 8c: bottom bars). The fold-change in predicted neuronal fraction between CSF and plasma is significantly higher for Pleiades-Neuron compared to the baseline heuristic (1827× vs. 121×).

We then consider the more challenging test of plasma-derived cfDNA samples from conditions involving potential injury to the brain: donors after brain death (DBD), patients in the Neuro-ICU, and patients with measles. Pleiades-Neuron predicts elevated neuron-derived cfDNA for all three conditions relative to healthy controls (DBD AUC = 1.0, Neuro-ICU AUC = 0.98, Measles AUC = 0.83), consistent with our expectation that individuals with potential injury to the brain may have elevated neuron-derived cfDNA in blood plasma (Fig. 8a, c). In samples derived from patients with measles, we see a broad range of scores, consistent with prior literature that encephalitis - inflammation of the brain - occurs in a small subset of measles cases⁴¹. To check that our results reflect brain-specific signal rather than general disease markers, we also scored non-healthy controls with non-neurological conditions; Pleiades-Neuron does not predict elevated neuron-derived cfDNA for these (AUC = 0.51 against healthy). **Importantly, Pleiades-Neuron separates brain-related conditions from controls better than the baseline heuristic (Fig. 8c), indicating that the sensitivity gains established above carry over to real-world detection of neurological disease.**

What we learned and what comes next

In this post, we showcase how we have tackled the problem of detecting ultra-rare neuron-derived cfDNA in blood plasma with Pleiades-Neuron, aimed at achieving a clinically relevant detection sensitivity of one part-per-million¹⁵.

Along the way, we demonstrate scaling laws on several fronts. Larger parameter Pleiades models consistently outperform smaller models, with our 7B parameter foundation model performing best. Excitingly, we note that performance did not plateau at 7B parameters, indicating potentially significant room for future improvement from model scaling. We also see consistent data scaling: larger and more diverse datasets lead to model improvement. This has inspired significant efforts to expand atlas-scale multi-omic datasets with greater cell-type resolution for our Pleiades 2 generation of models. Finally, we observe that performance strongly depends upon evaluation data quality. This last point is an important one, as no model can recover optimal performance from poor or shallow data. Higher sequencing depth consistently yielded the best performance, and we predict that as models improve and the cost of sequencing continues to fall, there will be material clinical benefit from deeper sequencing. However, continued experimental optimizations will also be necessary. Recently discovered systematic conversion errors in enzymatic methods limit the resolution of cell-type specific methylation information that can be captured⁴², prompting efforts from ourselves and others to develop improved protocols to mitigate such effects. Our results indicate that the accurate detection of neurological disease from plasma cfDNA will likely require both powerful models and deeply sequenced cohorts of patients with optimized experimental methods.

Overall, Pleiades-Neuron has 4.4× the detection sensitivity of the baseline heuristic. Excitingly, we see preliminary evidence that this improvement translates to real-world disease detection for multiple brain-related conditions. However, there is still more to be done. Here, we have showcased our work in per-fragment representation learning (Fig. 1). We are actively expanding our prediction framework to other cell types of the brain, such as microglia and astrocytes, as well as extending our neuronal framework to subtypes. Further, we are pairing it with powerful set modeling approaches, enabling more comprehensive disease detection across brain pathologies such as neurodegenerative conditions as well as diseases involving neuronal injury such as acute trauma and chronic traumatic encephalopathy. Finally, with our own clinical studies and that of our partners’, we are profiling patients longitudinally to help unravel the molecular complexity of these progressive and heterogenous diseases. The advances made here can be applied to neurological diseases, and we hope beyond diseases of the brain as well.

Author contributions

P.K., S.G., and N.L. designed the study, prepared the data corpus, trained and evaluated models, analyzed results, and drafted the manuscript. J.-O.G., P.K., N.L., and F.M.M.-Z. built the methylation sequencing data processing pipeline. A.J. and H.B. contributed to experimental data generation. S.G., L.G., and P.N. contributed to machine learning engineering and infrastructure. Y.D. contributed to clinical sample acquisition and data generation. P.N. and R.S. managed the project. R.S. conceived the original idea for this work. All authors contributed to the interpretation of the results and critically reviewed and revised the manuscript.

Acknowledgements

We thank Jonathan Wan, Hayley Holt, Qiurui (Rachel) Zeng, and Kerrigan Blake for their kind feedback on the text. We thank Shabbir Khan for enabling this work through cluster infrastructure. We are grateful to Hannah Madan for her mentorship and support, as well as to the rest of the Prima Mente team for their continued support. We thank the Nebius teams for their continued support of our work through their active support, backend infrastructure, and compute for model training and inference.

Citation

```
@misc{primamente2026pleiadesneuron,
  title      = {Detecting ultra-rare neuronal cell-free DNA in blood with Pleiades-Neu},
  author     = {Kathail, Pooja and Gupta, Saumya and Loyfer, Netanel and Ganbat, Javkhlán-Ochir and Jhutti, Anjeet and Babikir, Husam and Martín-Zamora, Francisco M. and Patel, Anooj and Giacomoni, Luca and Dor, Yuval and Niki, Pouya and Solanki, Ravi},
  year      = {2026},
  month     = {Oct},
```

```
url          = {https://www.primamente.com/blog/pleiades-neuron/},  
organization = {Prima Mente}  
}
```

References

1. Lo, Y. M. D., Han, D. S. C., Jiang, P. & Chiu, R. W. K. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, eaaw3616 (2021). <https://doi.org/10.1126/science.aaw3616>.
2. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine. *Lancet* **392**, 777–786 (2018). [https://doi.org/10.1016/S0140-6736\(18\)31268-6](https://doi.org/10.1016/S0140-6736(18)31268-6).
3. Loyfer, N. et al. A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023). <https://doi.org/10.1038/s41586-022-05580-6>.
4. Lehmann-Werman, R. et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl Acad. Sci. USA* **113**, E1826–E1834 (2016). <https://doi.org/10.1073/pnas.1519286113>.
5. Sun, K. et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl Acad. Sci. USA* **112**, E5503–E5512 (2015). <https://doi.org/10.1073/pnas.1508736112>.
6. Moss, J. et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018). <https://doi.org/10.1038/s41467-018-07466-6>.
7. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020). <https://doi.org/10.1016/j.annonc.2020.02.011>.
8. Lubotzky, A. et al. Liquid biopsy reveals collateral tissue damage in cancer. *JCI Insight* **7**, e153559 (2022). <https://doi.org/10.1172/jci.insight.153559>.
9. Peretz, A. et al. The DNA methylome of human vascular endothelium and its use in liquid biopsies. *Med* **4**, 263–281.e4 (2023). <https://doi.org/10.1016/j.medj.2023.03.006>.
10. Diehl, F. et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl Acad. Sci. USA* **102**, 16368–16373 (2005). <https://doi.org/10.1073/pnas.0507904102>.
11. Wan, J. C. M. et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017). <https://doi.org/10.1038/nrc.2017.7>.
12. Niki, P. et al. Human whole-epigenome modelling for clinical applications with Pleiades. Preprint at <https://doi.org/10.1101/2025.07.16.665231> (2025).

13. Health and Care Research Wales. Groundbreaking research study into earlier diagnosis of dementia opens in Wales. <https://healthandcareresearchwales.org/about/news/groundbreaking-research-study-earlier-diagnosis-dementia-opens-wales> (2025).
14. BBC News. Dementia diagnosis revolutionised by AI and new blood tests. <https://www.bbc.com/news/articles/c62e5n8jlldo> (2025).
15. Iorgulescu, J. B. *et al.* Impact of higher cell-free DNA yields on liquid biopsy testing in glioblastoma patients. *Clin. Chem.* **71**, 215–225 (2025). <https://doi.org/10.1093/clinchem/hvae178>.
16. Lubotzky, A. *et al.* Elevated brain-derived cell-free DNA among patients with first psychotic episode – a proof-of-concept study. *eLife* **11**, e76391 (2022). <https://doi.org/10.7554/eLife.76391>.
17. Caggiano, C. *et al.* Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. *Nat. Commun.* **12**, 2717 (2021). <https://doi.org/10.1038/s41467-021-22901-x>.
18. Li, S. *et al.* Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. *Proc. Natl Acad. Sci. USA* **120**, e2305236120 (2023). <https://doi.org/10.1073/pnas.2305236120>.
19. Li, W. *et al.* CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res.* **46**, e89 (2018). <https://doi.org/10.1093/nar/gky423>.
20. Guo, S. *et al.* Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **49**, 635–642 (2017). <https://doi.org/10.1038/ng.3805>.
21. Gaitsch, H., Franklin, R. J. M. & Reich, D. S. Cell-free DNA-based liquid biopsies in neurology. *Brain* **146**, 1758–1774 (2023). <https://doi.org/10.1093/brain/awac438>.
22. Southwood, D., Singh, S. & Chatterton, Z. Brain-derived cell-free DNA. *Neural Regen. Res.* **17**, 2213–2214 (2022). <https://doi.org/10.4103/1673-5374.335794>.
23. Guo, J. U. *et al.* Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* **17**, 215–222 (2014). <https://doi.org/10.1038/nn.3607>.
24. Price, A. J. *et al.* Divergent neuronal DNA methylation patterns across human cortical development reveal critical periods and a unique role of CpH methylation. *Genome Biol.* **20**, 196 (2019). <https://doi.org/10.1186/s13059-019-1805-1>.
25. Brethouwer, T., de Mendoza, A. & Bogdanovic, O. Non-CG DNA methylation in animal genomes. *Nat. Genet.* **57**, 2395–2407 (2025). <https://doi.org/10.1038/s41588-025-02303-1>.
26. He, Y. & Ecker, J. R. Non-CG methylation in the human genome. *Annu. Rev. Genomics Hum. Genet.* **16**, 55–77 (2015). <https://doi.org/10.1146/annurev-genom-090413-025437>.
27. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013). <https://doi.org/10.1126/science.1237905>.

28. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017). <https://doi.org/10.1126/science.aan3351>.
29. Loyfer, N. et al. Ultra-sensitive detection of brain DNA using non-CG methylation patterns. Manuscript in preparation (2026).
30. Jeong, Y. et al. MethylBERT enables read-level DNA methylation pattern identification and tumour deconvolution using a Transformer-based model. *Nat. Commun.* **16**, 788 (2025). <https://doi.org/10.1038/s41467-025-55920-z>.
31. Cavalier, A. N. et al. Epigenetic dysregulation of transposable elements in cognitive impairment and Alzheimer's disease. *GeroScience* **47**, 6731–6748 (2025). <https://doi.org/10.1007/s11357-025-01765-9>.
32. Masson, H. O. et al. XIST expression and hypermethylation of the X chromosome in males with systemic lupus erythematosus. *Front. Immunol.* **17**, 1743606 (2026). <https://doi.org/10.3389/fimmu.2026.1743606>.
33. Noë, M. et al. DNA methylation and gene expression as determinants of genome-wide cell-free DNA fragmentation. *Nat. Commun.* **15**, 6690 (2024). <https://doi.org/10.1038/s41467-024-50850-8>.
34. Lo, Y. M. D. et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010). <https://doi.org/10.1126/scitranslmed.3001720>.
35. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016). <https://doi.org/10.1016/j.cell.2015.11.050>.
36. Mouliere, F. et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **10**, eaat4921 (2018). <https://doi.org/10.1126/scitranslmed.aat4921>.
37. 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 Biol.* **21**, 153 (2020). <https://doi.org/10.1186/s13059-020-02059-3>.
38. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. Preprint at <https://doi.org/10.48550/arXiv.2106.09685> (2021).
39. Kozlenkov, A. et al. Substantial DNA methylation differences between two major neuronal subtypes in human brain. *Nucleic Acids Res.* **44**, 2593–2612 (2016). <https://doi.org/10.1093/nar/gkv1304>.
40. Wang, N. et al. Using Interpretability to Identify a Novel Class of Biomarkers for Alzheimer's Detection. Goodfire & Prima Mente <https://www.goodfire.com/research/interpretability-for-alzheimers-detection> (2026).
41. Perry, R. T. & Halsey, N. A. The clinical significance of measles: a review. *J. Infect. Dis.* **189** (Suppl. 1), S4–S16 (2004). <https://doi.org/10.1086/377712>.
42. Loyfer, N. et al. Systematic errors in enzymatic conversion limit cell-free DNA methylation specificity. Preprint at bioRxiv <https://doi.org/10.64898/2026.03.24.713040> (2026).

Figures

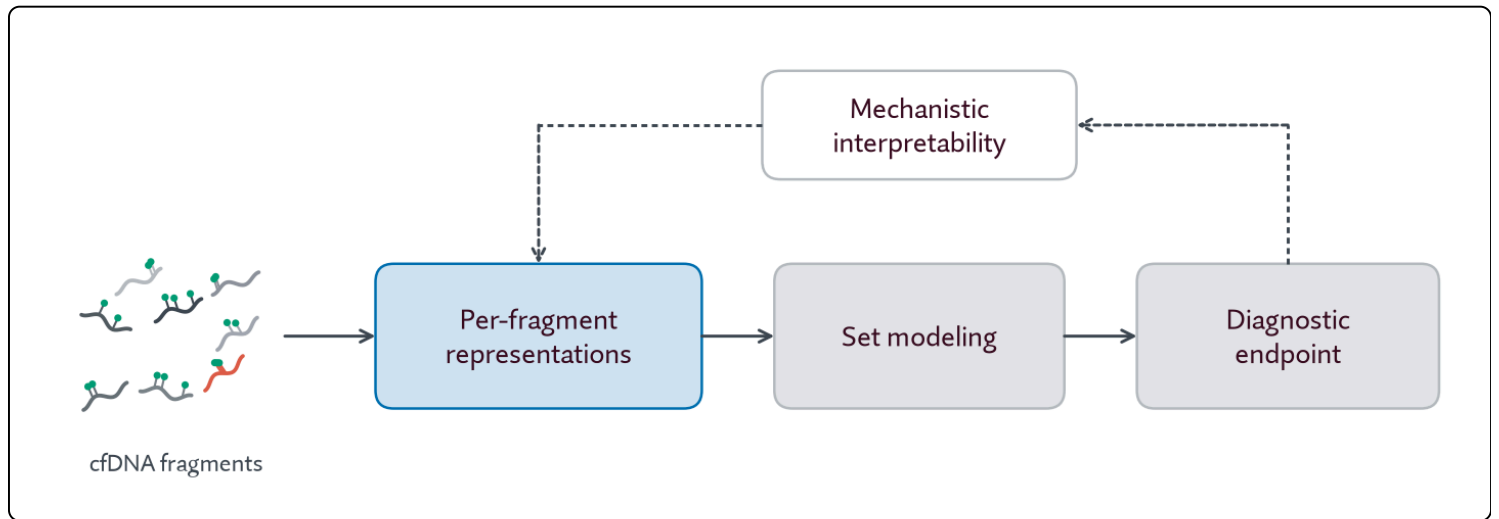


Figure 1. Our approach to modeling cell-free DNA.

Per-fragment representation learning: sequence-based epigenetic foundation models are powerful tools for learning compact representations of cell-type specific DNA methylation patterns. Our first model, Pleiades, demonstrated that these representations can be used for a wide range of biologically relevant tasks. Pleiades-Neuron is our newest model in this category.

Set modeling: aggregating sparse disease-relevant signal from a large set of cfDNA fragments requires specialized architectures. In prior work, we developed a multi-tier Hierarchical Attention Transformer for this purpose¹².

Diagnostic endpoint: disease phenotypes are much more high-dimensional than binary clinical labels. Through both global partnerships with studies of aging and age-related diseases as well as our clinical studies, such as SANDBOX^{13,14}, we are collecting rich clinical information that can be used to create better prediction endpoints.

Mechanistic interpretability: interpreting model predictions allows us to gain mechanistic insight into disease biology, and use learned insights to build better models.

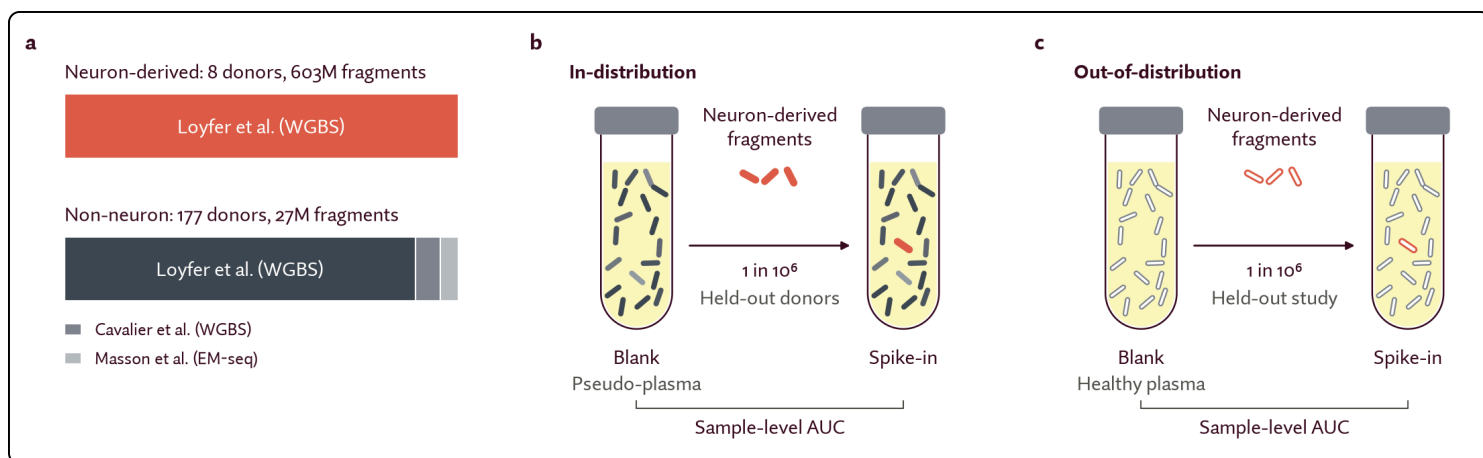


Figure 2. Pleiades-Neuron training and evaluation data.

a, The Pleiades-Neuron data corpus: 603M neuron-derived and 27M non-neuron-derived fragments (~120B tokens) from three studies. **b**, The in-distribution *in silico* spike-in evaluation set. The pseudo-plasma background mixes non-neuron cell types in the proportions typical of normal plasma³. Neuron-derived fragments from held-out donors are spiked in at a known, fixed concentration. **c**, The out-of-distribution set. The background is real plasma cfDNA from healthy individuals, sequenced in-house. The spiked-in neuron-derived fragments come from an entirely held-out study.

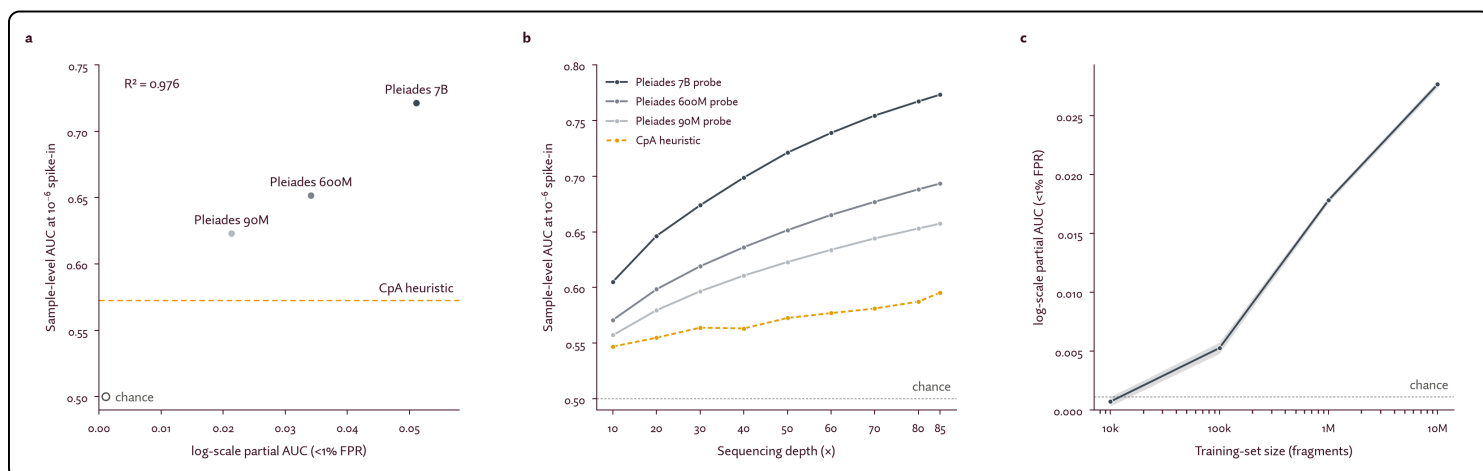


Figure 3. Neuron-derived cfDNA detection performance improves with both model and data scale.

a, Probes on frozen Pleiades 90M, 600M and 7B representations, scored by log-scale partial AUC (pAUC) and by sample-level AUC at 10^{-6} spike-in. All sizes beat the baseline heuristic, and larger models score higher on both metrics. **b**, The same ordering holds at every sequencing depth from 10x to 85x. All models improve with depth, and the gap widens at higher depth. **c**, A probing sweep with frozen Pleiades 7B representations over dataset size, up to 10M training examples. Performance improves consistently with dataset size.

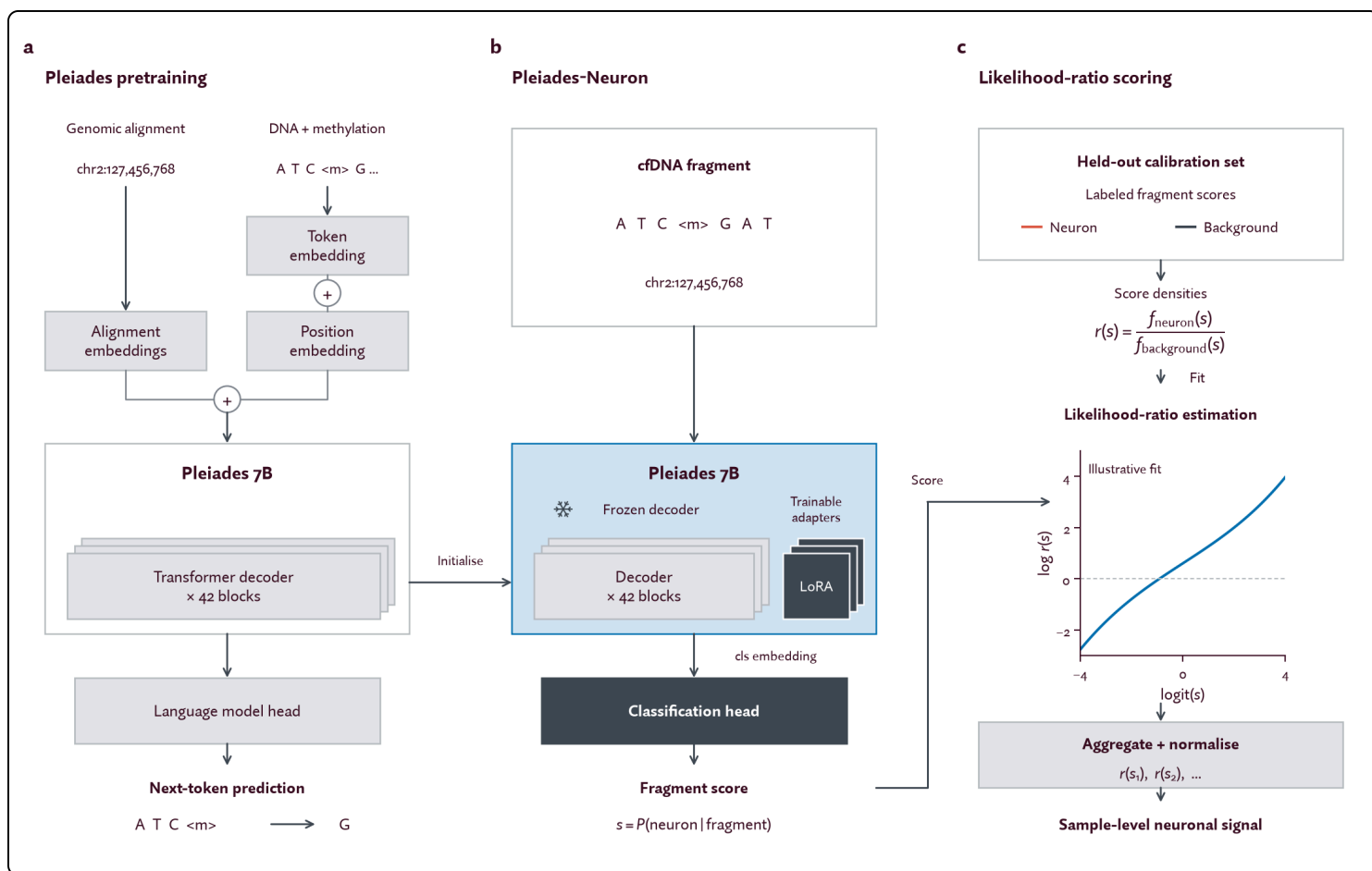


Figure 4. Pleiades-Neuron model architecture and likelihood-ratio scoring.

a, Pleiades-Neuron starts from the Pleiades 7B base model¹², an autoregressive transformer decoder pre-trained on 1.9T tokens of human methylation and genomic data. **b**, It is LoRA fine-tuned to classify neuron-derived versus non-neuron-derived fragments. **c**, Likelihood-ratio scoring turns a set of fragment scores into one sample-level score.

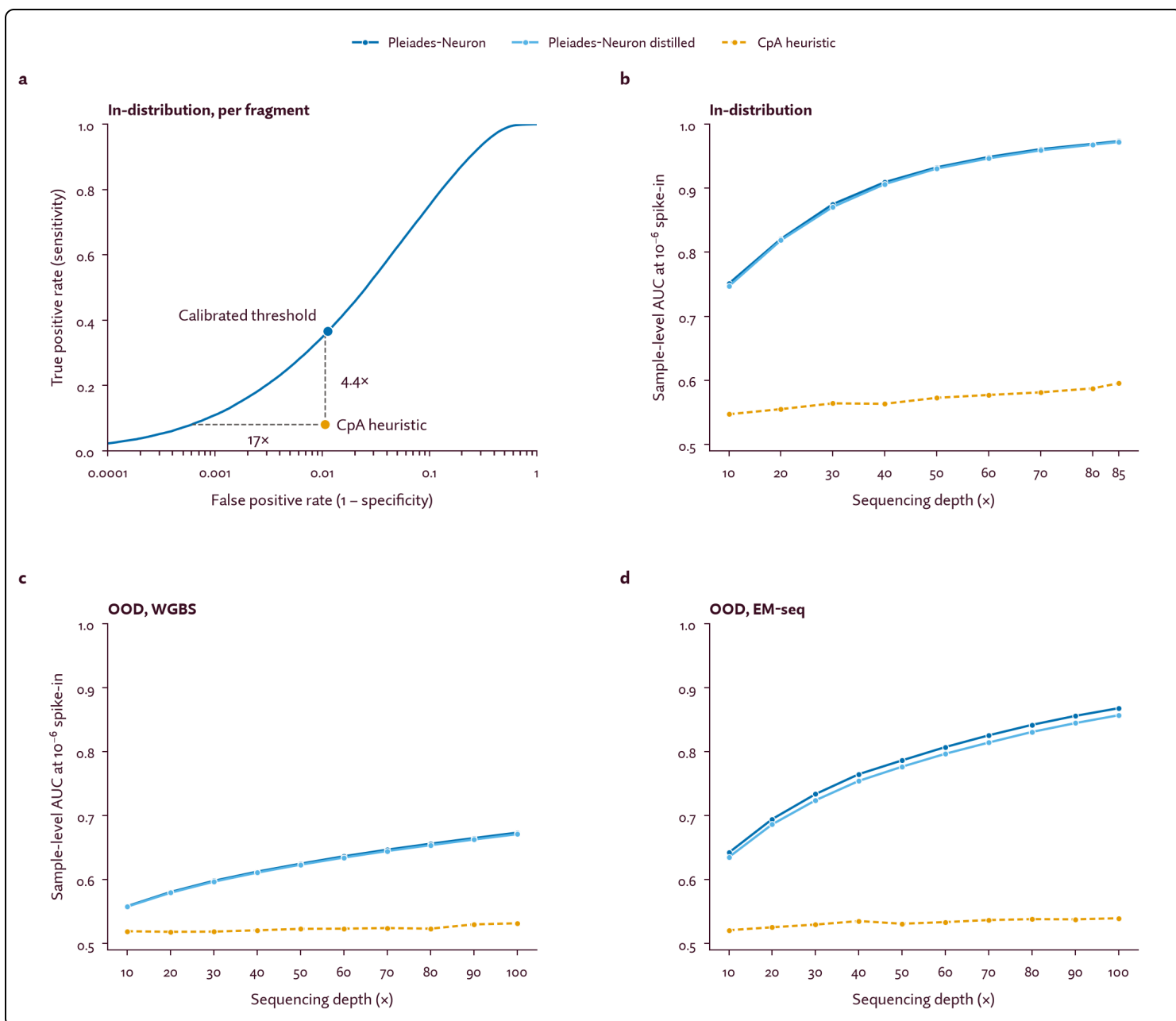


Figure 5. Pleiades-Neuron has increased sensitivity for detecting low concentration neuron-derived fragments, especially at high sequencing depth.

a, Receiver operating characteristic (ROC) curves for Pleiades-Neuron and the CpA heuristic. At a fixed specificity, Pleiades-Neuron captures 4.4x as many true positives. At a fixed sensitivity, it admits 17x fewer false positives. At the threshold calibrated on a held-out set, the gain is almost entirely in sensitivity. **b,** In-distribution sample-level AUC for 10^{-6} neuron-derived DNA versus background-only samples, against sequencing depth. Both Pleiades-Neuron models and the baseline improve with depth, and the gap widens at higher depth. **c, d,** The same evaluation on the WGBS and EM-seq out-of-distribution datasets.

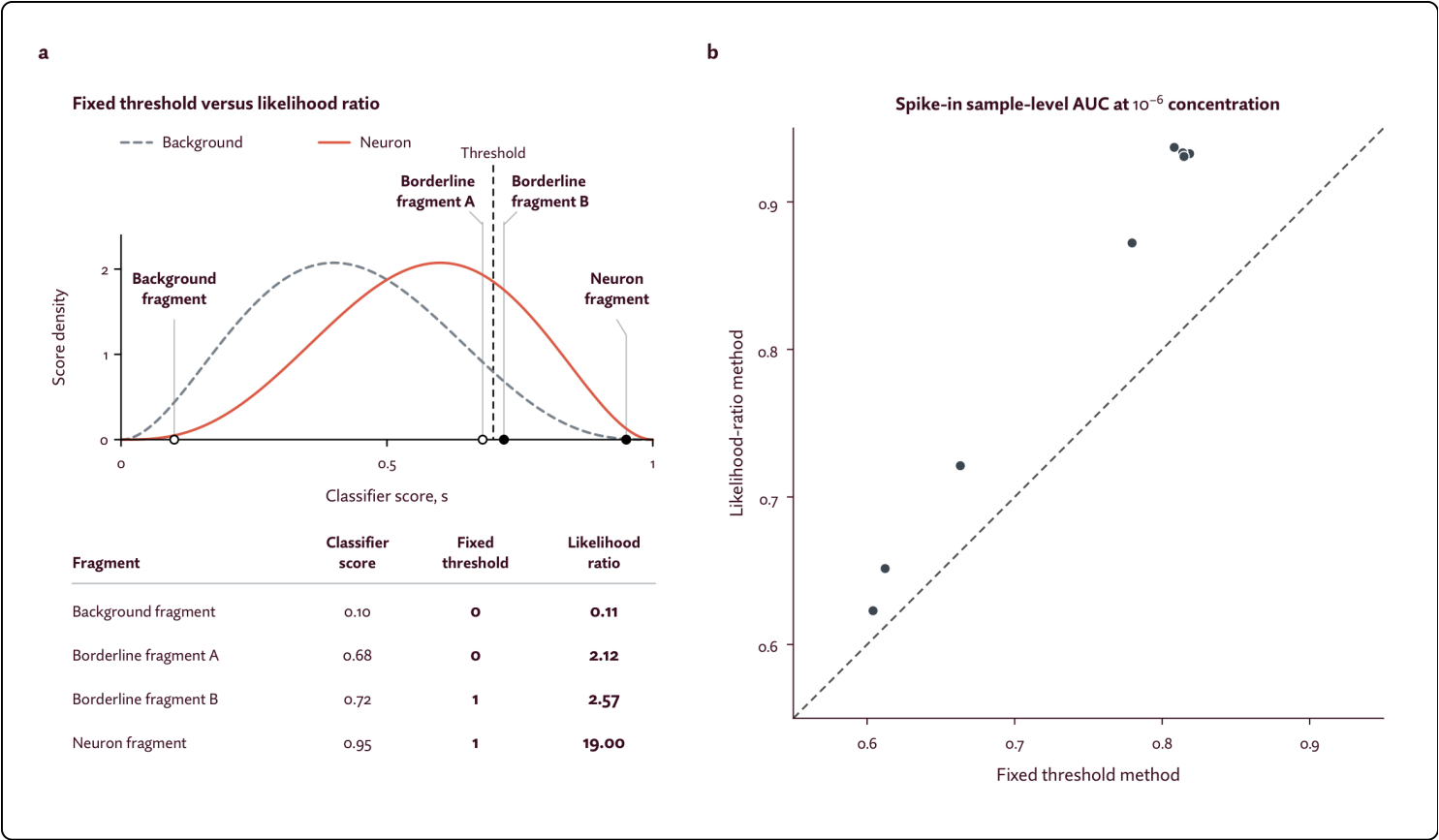


Figure 6. Likelihood-ratio scoring method outperforms a fixed threshold.

a, The two scoring methods. A fixed threshold gives a borderline fragment the same binary score as one strongly predicted towards a class. The likelihood-ratio gives nearby borderline fragments similar scores. **b**, The two methods across model variants on the in-distribution *in silico* spike-in evaluation at 10^{-6} concentration. The likelihood-ratio outperforms the fixed threshold for every model.

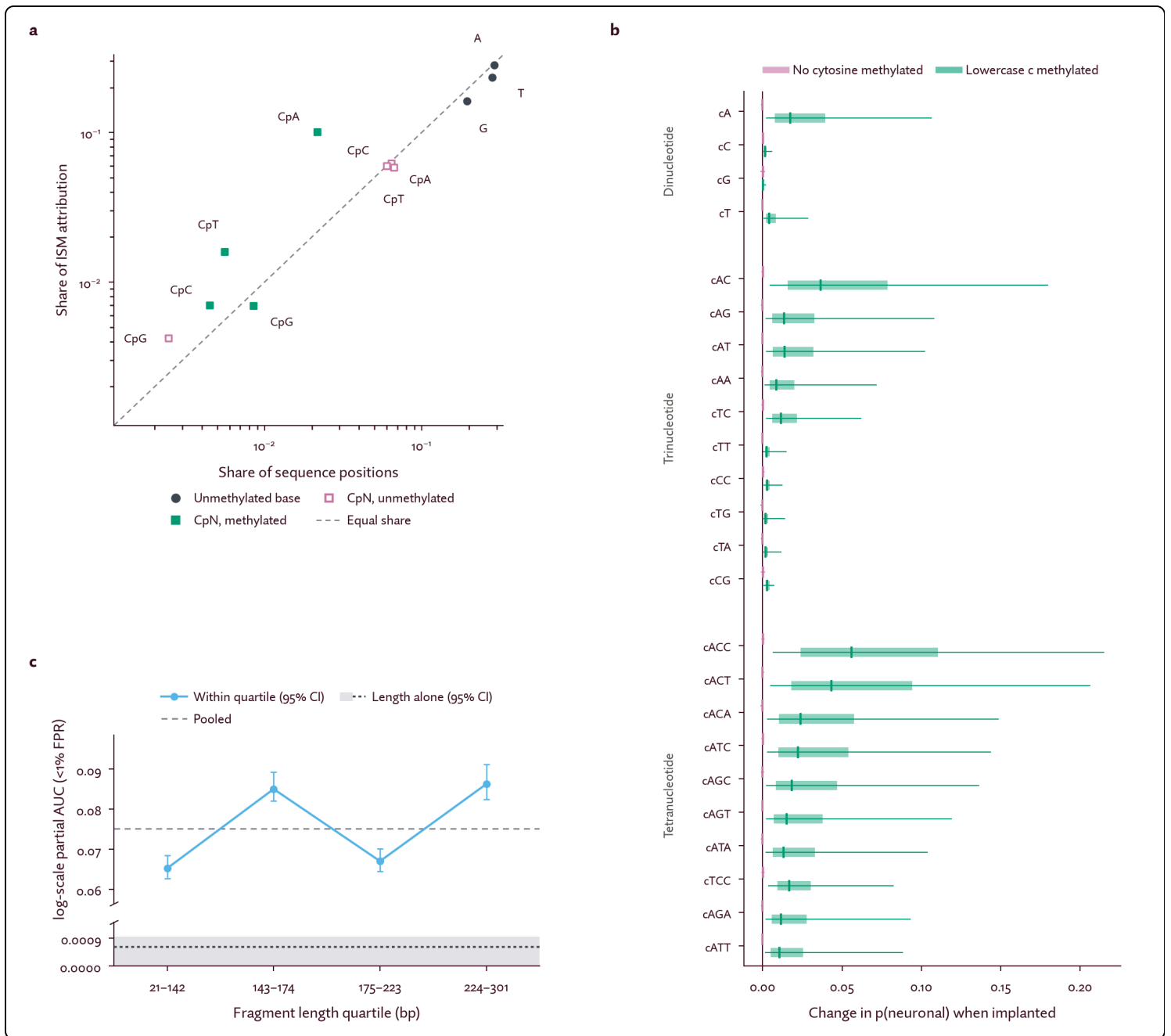


Figure 7. Interpretability analysis reveals key biomarkers of neuron-derived DNA learned by Pleiades-Neuron.

a, Share of *in silico* mutagenesis (ISM) attribution for each nucleotide and for the methylated and unmethylated form of each CpN site. Each is plotted against its share of sequence positions. Methylated CpA and CpT sites have the largest causal impact. **b**, *In silico* insertion of every cytosine di-, tri- and tetranucleotide context. mCA, mCAC and mCACC raise $p(\text{neuronal})$ the most (labeled cA, cAC, cACC; lower-case c marks the methylated cytosine). **c**, Neuronal classification performance stratified by fragment length quartile indicates that length is not a predictive feature. Dashed gray, pooled performance across all fragments; dotted dark gray, performance from length features alone.

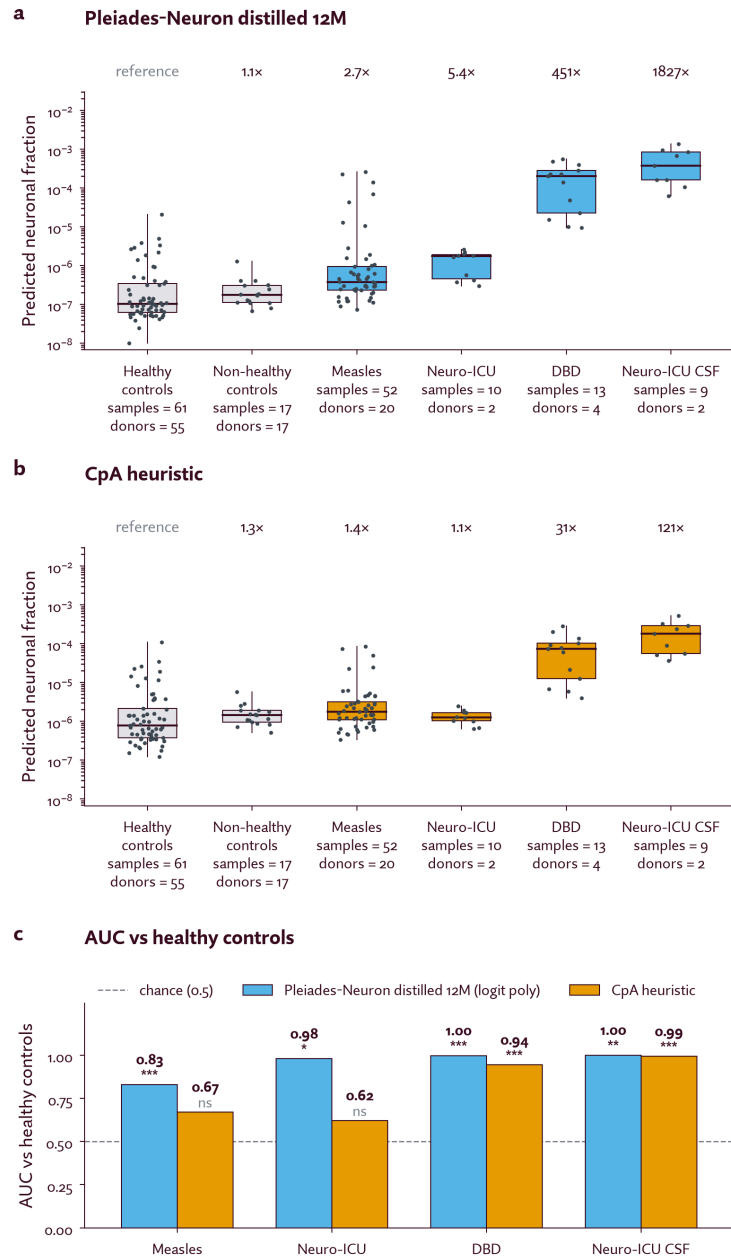


Figure 8. Pleiades-Neuron predicts elevated neuron-derived cfDNA for multiple brain-related conditions.

a, Neuronal cfDNA fraction predicted by Pleiades-Neuron (distilled 12M) for healthy controls, non-healthy controls (patients with a non-neurological condition), patients with conditions involving potential injury to the brain - measles, Neurological Intensive Care Unit (Neuro-ICU) patients, and donors after brain death (DBD) - and cerebrospinal fluid (CSF) samples (known to have elevated neuron-derived cfDNA relative to plasma). Fold-change is reported relative to the healthy controls, using a linear mixed-effects model with random donor intercept to account for dependence between samples from the same donor. **b**, The same samples and fold-change metric for the CpA heuristic. **c**, Sample-level AUC for brain-related conditions versus healthy controls using Pleiades-Neuron (distilled 12M) and the CpA heuristic. Pleiades-Neuron outperforms the baseline for all brain-related conditions. P-values are from a one-sided t-test with number of donor-based degrees of freedom and Benjamini-Hochberg correction (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).