Learning Extremely Sparse Signals in High-Dimensional Cell-Free DNA Data Using Modern Hopfield Attention for Colorectal Cancer Detection

Anonymous Authors[1]

Abstract

Next generation sequencing-based early detection of colorectal cancer from cell-free DNA (cfDNA) is a clinically important supervised learning problem on complex molecular data with extraordinarily sparse signal. It presents an extreme-scale multiple instance learning challenge: identifying rare tumor signals from high-dimensional, multi-resolution data with millions of instances per sample and witness rates as low as $<0.0001\%$. We propose Fragment-Level Deep Learning (FLDL), an end-to-end deep learning framework utilizing Modern Hopfield Networks to perform dense associative retrieval over the massive instance space. Using held-out real-world clinical and challenging contrived test sets, we compare FLDL's performance to a state-of-the-art machine learning model and to a deep learning model without attention (max pooling). Our results demonstrate that only the attention-based FLDL model outperforms the machine learning model, in spite of a modest training set size ($n = 4,394$). FLDL also scales effectively with sample size and with number of instances per sample while offering useful biological insights due to the interpretability of its attention weights and intermediate learned representations. This work establishes a new frontier for highly scalable, attention-based deep learning in the field of clinical cfDNA diagnostics.

1. Introduction

Structured data plays a central role in modern healthcare, from tabular records and irregular clinical measurements to high-dimensional molecular assays derived from sequencing assays. Among these modalities, next-generation sequencing (NGS) of cell-free DNA (cfDNA) represents an especially challenging form of health data: each sample contains millions of DNA fragments, yielding a high-dimensional and naturally multi-resolution signal spanning nucleotide-level, fragment-level, and sample-level structure. Attention-based deep learning models, particularly transformers (Vaswani et al., 2017), excel in many domains (Berroukham et al., 2023; He et al., 2023; Latif et al., 2023; Nerella et al., 2024; Liang et al., 2024) but remain difficult to apply when inputs number in the millions due to the quadratic cost of standard self-attention, and clinical applications compound this challenge with modest labeled cohort sizes (Bertsch et al., 2023; Alva Principe et al., 2025). The detection of colorectal cancer (CRC) from cfDNA combines all of these difficulties: extreme instance counts, sparse disease-associated signal, and limited training data, making it a uniquely demanding structured health data problem.

Colorectal cancer (CRC) remains the second most common cause of cancer-related death in the US (Siegel et al., 2025). Despite the availability of multiple screening modalities, only an estimated 63% of eligible individuals aged 45 years and older were up to date with guideline-recommended screening in 2023 (Bandi et al., 2025), motivating the development of more convenient, non-invasive blood-based tests.

Cell-free DNA (cfDNA) analysis is central to non-invasive blood-based tests. cfDNA in healthy individuals consists largely of fragments of $\approx$166 base pairs released by dying cells (Snyder et al., 2016; Gao et al., 2022; Thierry, 2023). Unlike tumor tissue (where tumor DNA may exceed 50%), cancer signal in cfDNA is usually very sparse: circulating tumor DNA (ctDNA) ranges from $<0.05\%$ to 90% of cfDNA, with some "low-shedding" cancers releasing little to none (Bettegowda et al., 2014; Luo et al., 2021). This paper focuses on **DNA methylation**: the conversion of cytosine to 5-methylcytosine (mC) at CpG dinucleotides. Aberrant methylation patterns in cfDNA are a well-established non-invasive cancer biomarker (Yamaguchi et al., 2003), and NGS-based methylation blood tests for CRC and advanced precancerous lesions (APLs) have recently achieved or are approaching FDA approval (Chung et al., 2024; Shaukat et al., 2025). These tests present a formidable computational challenge: identifying a minute number of ctDNA "needles"

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

within an enormous haystack of millions of healthy cfDNA fragments.

Traditional approaches to ctDNA detection largely rely on biological prior knowledge and summarize methylation measurements through aggregation at the per-CpG (β) or per-fragment (α) level (Li et al., 2018). While effective, these methods depend on manual definitions rather than learning optimal representations directly from raw data. Recently proposed multi-step deep learning approaches for methylation sequencing data (Jeong et al., 2025; Niki et al., 2025; Deng et al., 2023) demonstrate the value of fragment-level representation learning. However, these methods are typically not trained end-to-end for final patient-level classification. We hypothesize that robust detection of early-stage cancer in this low-signal regime benefits from a) the ability to identify and upweight specific informative fragments, and b) representations jointly optimized with the classification objective to capture subtle task-specific signatures.

To enable direct multi-resolution learning (nucleotide and CpG, fragment and sample level), we propose the **Fragment-Level Deep Learning (FLDL) model**, a method that formulates early cancer detection as a multiple instance learning (MIL) problem (Carbonneau et al., 2018; Ilse et al., 2018). FLDL incorporates a specialized attention module based on Modern Hopfield Networks (MHN) (Ramsauer et al., 2020), which have been previously shown to be successful for immune repertoire classification (Widrich et al., 2020). where they effectively aggregate signals from "bags" of hundreds of thousands of immune receptor sequences to predict disease status. By adapting this mechanism to ctDNA detection, our method captures cancer signals from a multi-modal representation of cfDNA fragments, solving a needle-in-a-haystack challenge of extreme scale.

In this work, we show that FLDL for blood-based detection of CRC a) operates on billions of nucleotides from millions of cfDNA fragments per subject, b) directionally outperforms a state-of-the-art machine learning baseline on contrived and real-world clinical test sets, c) is capable of implicit denoising of large and sparse input spaces, d) supports biological interpretability via attention weights and generalizability assessment via learned sample embeddings, and e) scales consistently with training data volume. These results establish FLDL as an effective end-to-end deep learning model for clinically impactful cancer screening from complex high-dimensional health data.

2. Fragment-Level Deep Learning model

We base our FLDL model on the attention mechanism of continuous MHNs (Ramsauer et al., 2020), a generalization of Hopfield Networks (Hopfield, 1982; 1984), that provides a trainable associative memory for deep learning architectures (Ramsauer et al., 2020; Hu et al., 2023). MHNs have been applied to reinforcement and contrastive learning (Widrich et al., 2021; Fürst et al., 2022), tabular data (Schäfl et al., 2022), immune repertoire classification (Widrich et al., 2020; Al Hajj et al., 2024), and chemical reaction prediction (Seidl et al., 2021). Due to their exponential storage capacity, they excel in low signal-to-noise problems, as demonstrated by the DeepRC model in Widrich et al. (2020). Our FLDL extends DeepRC to the prediction of CRC from cfDNA.

Problem formulation. We follow the formulation of the MIL problem in Widrich et al. (2020), where a bag $\mathcal{X} = \{s_1, \ldots, s_N\}$ of N instances constitutes a sample. Assuming binary classification, each such instance s_i is associated with an inaccessible label $y_i \in \{0, 1\}$. Only a sample-level label $y = \max_i y_i$ is observed per bag. In order to correctly classify a positive sample, the instances that are responsible for the label y have to be identified (Foulds & Frank, 2010). Correct classification of a positive sample therefore requires identifying the instances responsible for the positive label (Foulds & Frank, 2010).

Each cfDNA fragment constitutes an instance and each patient sample a bag, with N typically ranging from 1 to 10 million and witness rates (fraction of instances indicating a positive label) as low as $< 0.0001\%$ in low shedding cases (compared to $N \approx 300{,}000$ and 0.01% in DeepRC).

FLDL model. To address these challenges, we design FLDL as an end-to-end trainable deep learning architecture (Fig. 1). A dedicated sub-network $\phi(\cdot)$ (*Fragment Embedding*) maps each fragment s_i independently to a fixed-size vector $\boldsymbol{h}_i = \phi(s_i) \in \mathbb{R}^m$. FLDL then aggregates the embedded fragments into a sample representation using a MHN (*Sample Embedding*) and predicts CRC status using a fully connected *Output Network*. Each fragment s_i is represented by five modalities: a) sequence and CpG methylation, b) methylation statistics, c) DNA foundation model embedding, d) genomic position, e) strand. Each modality is projected to a latent embedding by a dedicated sub-network. The individual modality embeddings are concatenated and further processed to form a unified fragment representation $\boldsymbol{h}_i \in \mathbb{R}^m$. See also Appendix Section B.1 and Fig. A2.

To aggregate these fragment representations into a sample-level embedding, we employ a Hopfield Pooling layer following DeepRC. As shown in Fig. 1, we learn a set of K cancer-indicative fragment prototypes (*state patterns* or *queries*) $\boldsymbol{Q} \in \mathbb{R}^{K \times d}$ in a high-dimensional association space. A fully-connected self-normalizing neural network (Klambauer et al., 2017), $\mathrm{NN} : \mathbb{R}^m \to \mathbb{R}^d$, is applied row-wise to project the N embedded fragments $\boldsymbol{Y} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N] \in \mathbb{R}^{N \times m}$ (*stored patterns*) to the same high-dimensional association space $\mathbb{R}^{N \times d}$ (yielding *keys*). The softmax-normalized dot-product similarity (attention values) $\boldsymbol{A} = \mathrm{softmax}\left(\beta\, \boldsymbol{Q}\, \mathrm{NN}(\boldsymbol{Y})^T\right) \in \mathbb{R}^{K \times N}$

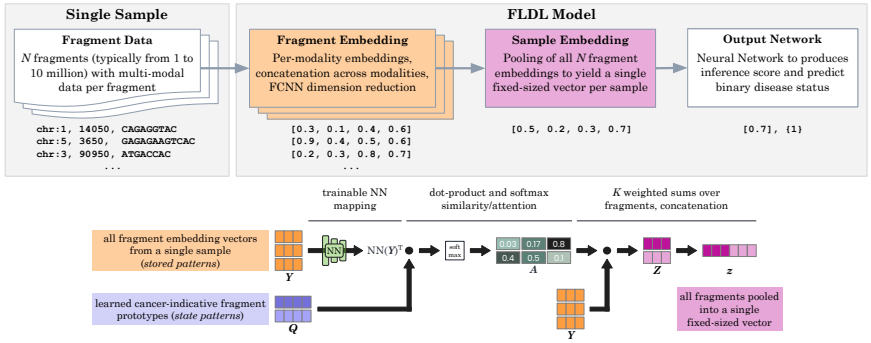


Figure 1. Overview of the FLDL architecture. **Top**: The end-to-end pipeline embeds multi-modal data from millions of cfDNA fragments and aggregates them into a sample-level representation. **Bottom**: Hopfield Pooling uses an attention matrix ($\boldsymbol{A}$) to aggregate fragment embeddings ($\boldsymbol{Y}$) into a sample representation ($\boldsymbol{Z}$) based on similarity to learned cancer-indicative prototypes ($\boldsymbol{Q}$).

between these queries and the keys are then used to aggregate the embedded fragments $\boldsymbol{Y}$ into a sample embedding $\boldsymbol{Z} = \text{softmax}\left(\beta\, \boldsymbol{Q}\, \text{NN}(\boldsymbol{Y})^T\right)\ \boldsymbol{Y} \in \mathbb{R}^{K\times m}$, where β controls the attention distribution sharpness. Finally, $\boldsymbol{Z}$ is flattened into a vector $\boldsymbol{z} \in \mathbb{R}^{mK}$ and passed to an MLP to predict CRC probability and other biologically relevant auxiliary tasks. See Appendix Section B.2 for FLDL training.

3. Experimental setup and results

We evaluate three FLDL configurations, each differing by the subset of the targeted capture panel used: ER-FLDL uses the full capture panel to test performance in a larger, noisier input space, C-FLDL uses the CRC panel, a section of the full panel most relevant for CRC, and PD-FLDL uses biologically informed pre-filtering on the CRC panel. We further compare against MaxPool variants that replace Hopfield pooling with max-activation-based fragment aggregation, and against a state-of-the-art machine learning baseline (ML baseline) (Shaukat et al., 2025) developed on the CRC panel. See Appendix Section D for details.

Predictive performance, real-world clinical sample test set. One primary indicator of the model's efficacy is its predictive performance on the real-world clinical sample test set. For blood-based CRC screening, a specificity at or near 90% is considered to be the most clinically appropriate (Chung et al., 2024; Shaukat et al., 2025). As a result, the most relevant performance metric is not the area under the receiver operating characteristic curve (AUROC), but rather, the sensitivity at a fixed specificity of 90%. We report the achieved APL and CRC sensitivities at a specificity of 90%, along with Wilson's method 95% confidence intervals (CI), in Table 1. For each method, a classification threshold is selected to achieve the desired specificity among the negative samples in the real-world clinical sample set. As shown in Table 1, PD-FLDL outperforms all competing methods, achieving sensitivities of 30.2% for APLs and 89.6% for CRCs. This improvement over the ML Baseline is more pronounced in the challenging APL group: a gain of 3.1 points over the ML Baseline vs. a gain of 1.4 points for CRCs. Further, PD-FLDL outperforms C-FLDL and ER-FLDL, which suggests that leveraging explicit biological priors to denoise background signals before FLDL training is an effective mechanism for improving model detection sensitivity, at least at the current training set sample size.

Table 1. Real-world clinical sample test set sensitivity at 90% specificity, with Wilson's method 95% CIs, for all models.

	APL	CRC
ML Baseline	27.1 (22.8, 31.9)	88.2 (82.9, 92.1)
PD-FLDL	**30.2 (25.7, 35.0)**	**89.6 (84.5, 93.2)**
C-FLDL	28.9 (24.5, 33.7)	89.1 (83.9, 92.8)
ER-FLDL	28.4 (24.0, 33.2)	88.2 (82.9, 92.1)
PD-MaxPool	19.9 (16.1, 24.3)	79.2 (73.0, 84.3)
C-MaxPool	17.3 (13.8, 21.5)	76.3 (69.9, 81.8)

It is noteworthy that the C-FLDL model, which operates on the CRC panel but does not use an additional negative sample set for biologically informed filtering, achieves a sensitivity of 28.9% for APL and 89.1% for CRC, also outperforming the ML Baseline. Additionally, even when the input space is significantly expanded to include more task-irrelevant genomic regions (ER-FLDL), the FLDL architecture performs competitively to the ML Baseline model. In contrast, models employing max pooling (PD-MaxPool, C-MaxPool) show significantly degraded performance, with sensitivities dropping below 20% for APLs and below 80% for CRCs. This evidence supports the hypothesis that a simple max pooling aggregation is insufficient for identifying sparse ctDNA signals; instead, the ability to dynamically attend to rare informative fragments (here via MHN) contributes to high-sensitivity cancer detection.

Predictive performance, challenging contrived test set. Performance on the challenging contrived test set (Appendix Section E.1 and Appendix Fig. A4) recapitulates the trends observed in the real-world clinical sample set. PD-FLDL

achieves a positive call rate of 84.2%, substantially outperforming the ML Baseline (70.2%). PD-FLDL again outperforms C-FLDL, which in turn outperforms ER-FLDL. C-FLDL (74.3%) again exceeds the ML Baseline, but in contrast to the real-world clinical sample set, here ER-FLDL (66.9%) underperforms relative to the ML Baseline model. Finally, both PD-MaxPool and C-MaxPool struggle in this task and fail to achieve a competitive detection rate when presented with cases with a challenging ctDNA level. These results further validate the importance of the MHN attention mechanism for low-signal regimes and demonstrate that while biological prior denoising yields improved FLDL performance (at the current training set sample size), the architecture is capable of effective implicit denoising.

Scaling behavior: training dataset size. In order to understand the scale of data needed for effective training for the CRC detection task and to identify the onset of potential performance plateaus, we explore the scaling behavior of the FLDL architecture when trained on different subsampled datasets (20% - 100% of the full training dataset at 20% increments as described in Appendix Section E.2). As shown in Appendix Fig. A5, at all but the smallest subsampling fraction, the ensemble outperforms its 5 individual members, suggesting that ensembling smooths the noisy performance of individual models. We also observe a clear improvement in model performance as the training dataset size increases, with no plateau even near the full training dataset size. The increasing performance trend indicates that the FLDL model, as currently parameterized, is likely to continue to benefit from even more training samples, and that there may be a potential for further performance improvements by scaling the complexity of the model alongside increased training data volume in the future.

Latent space analysis and biological interpretability. For FLDL interpretability at the fragment level, we analyze the model's attention weights and evaluate genomic localization of high attention fragments. This is described in detail in Appendix Section E.3 and Appendix Fig. A6 including a comparison with the ML Baseline that demonstrates substantial agreement as well as some complementarity in the prioritized genomic loci between the two models.

At the sample level, embeddings after MHN aggregation from C-FLDL are used to visualize sample distributions and qualitatively assess the model's generalizability to unseen test data. In Fig. 2 we project both training and test set embeddings into a two-dimensional space using UMAP (McInnes et al., 2018) to inspect the latent structure for potential distribution shifts and biological relevance. The UMAP visualization reveals a coherent, biologically relevant continuum: we observe a smooth gradient transitioning from no-ctDNA negative clinical samples (blue) to challenging low-ctDNA clinical blends (orange), to high-ctDNA clinical blends (purple) and real-world clinical APL and CRC samples (red). Furthermore, the artificial no-methylation samples, which lack all CpG methylation and are hence biologically implausible, form a distinct cluster that is well separated from real-world clinical samples and clinical blends. Together, these results demonstrate that the FLDL model learns a robust, biologically meaningful representation that effectively distinguishes among biologically plausible samples based on their ctDNA content, and clearly separates biologically implausible inputs.

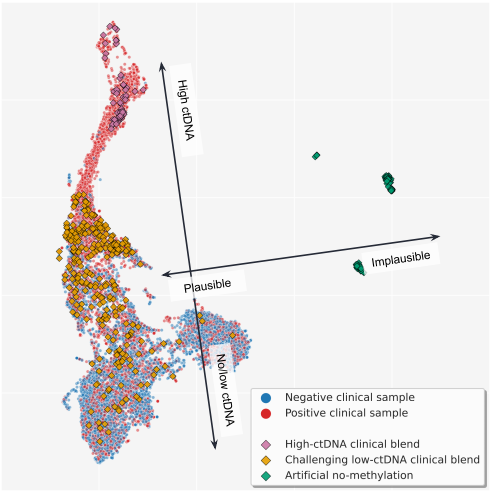


Figure 2. UMAP projection of learned FLDL sample embeddings from the C-FLDL model, applied to real-world clinical, contrived blend, and artificial no-methylation samples.

4. Conclusion

In this work, we present Fragment-Level Deep Learning (FLDL), an end-to-end multiple instance learning architecture designed to address the extreme needle-in-a-haystack challenge of early cancer detection from cfDNA in blood. Leveraging Modern Hopfield Networks to attend to rare ctDNA signals amidst millions of uninformative background fragments, FLDL outperforms a state-of-the-art machine learning approach and a max pooling deep learning alternative, particularly at identifying low-signal APLs and challenging contrived samples. Beyond predictive performance, FLDL also provides interpretable and clinically relevant representations. In addition, the finding that FLDL can implicitly denoise large genomic search spaces suggests suitability for future multi-cancer detection tasks using a capture panel designed for more than one cancer. Finally, scaling behavior analysis reveals that the model's predictive performance improves consistently with increasing training data volume, without yet reaching a plateau, positioning FLDL well for larger future datasets. These results show the effectiveness of specialized attention-based deep learning for a clinically important supervised learning problem on high-dimensional and multi-resolution data.

Impact Statement

Impact on deep learning and on diagnostics. By successfully addressing the extreme-scale multiple instance learning challenge, the FLDL architecture may encourage evaluation of Modern Hopfield Networks and attention mechanisms in a broader class of biological problems with extremely low signal-to-noise ratios. The architecture's ability to implicitly denoise large genomic search spaces suggests it could be effectively repurposed for cfDNA-based detection of other cancers or diseases. The interpretability of the attention mechanism allows for the identification of high-attention genomic regions, potentially aiding in the discovery or better understanding of disease biomarkers.

Broader impact on society. Colorectal cancer remains a critical public health challenge and is the second most common cause of cancer-related death in the US despite being preventable through screening (Siegel et al., 2025). While existing screening methods like colonoscopy and stool-based tests are available, adherence rates remain suboptimal (Bandi et al., 2025). Blood-based tests offer a non-invasive alternative that can significantly increase screening adherence. A blood test with better sensitivity will be useful in this context, particularly for harder-to-detect cases with low tumor fraction, which is the stage at which the disease is most treatable. The FLDL model improves upon the current state-of-the-art machine learning model for this application, and it demonstrates the potential for further performance gains through training data scaling, pointing towards even greater cancer detection sensitivity in the future.

Data considerations for bias. Confounding factors such as age, sex, race/ethnicity, and comorbidities could inadvertently be used by the model for classification if they correlate with the target label in the training data. Our current training set consists of 4,394 samples, a size that is substantial for the clinical diagnostics space but small compared to image or large language model training datasets. Techniques for controlling confounding of this type (e.g., dynamic minibatch balancing) may be less effective in smaller training datasets. Further, there may be unknown confounders. We are encouraged by FLDL's performance on the fully independent real-world clinical test set. However, continued monitoring and validation in additional diverse, representative datasets will be essential for confirming generalizability. It will also be important for these expanded datasets to increase sample counts for currently underrepresented subpopulations.

Privacy, data availability. The clinical data used in this study cannot be shared or distributed because they contain sensitive patient information and high-resolution genomic sequences that are subject to strict privacy regulations. Written informed consent was obtained from each subject who met eligibility criteria and contributed biosample to this study. All personally identifiable information (PII) in the source data was fully redacted prior to use in the model development process, to protect patient privacy and adhere strictly to all relevant regulations and guidelines. The de-identification process was approved and monitored by our Data Governance Committee.

References

Al Hajj, G. S., Hubin, A., Kanduri, C., Pavlovic, M., Rand, K. D., Widrich, M., Solberg, A. S., Greiff, V., Pensar, J., Klambauer, G., and Sandve, G. K. Incorporating probabilistic domain knowledge into deep multiple instance learning. In *Forty-first International Conference on Machine Learning*, 2024.

Alva Principe, R., Chiarini, N., and Viviani, M. Long document classification in the transformer era: A survey on challenges, advances, and open issues. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2):e70019, 2025.

Bandi, P., Star, J., Mazzitelli, N., Nargis, N., Islami, F., Siegel, R. L., Yabroff, K. R., and Jemal, A. Prevalence and review of major modifiable cancer risk factors, hpv vaccination, and cancer screenings in the united states: 2025 update. *Cancer Epidemiology, Biomarkers & Prevention*, 34(6):836–849, 2025.

Berroukham, A., Housni, K., and Lahraichi, M. Vision transformers: a review of architecture, applications, and future directions. In *2023 7th IEEE congress on information science and Technology (CiSt)*, pp. 205–210. IEEE, 2023.

Bertsch, A., Alon, U., Neubig, G., and Gormley, M. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36:35522–35543, 2023.

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., et al. Detection of circulating tumor dna in early-and late-stage human malignancies. *Science translational medicine*, 6(224):224ra24–224ra24, 2014.

Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern recognition*, 77:329–353, 2018.

Chung, D. C., Gray, D. M., Singh, H., Issaka, R. B., Raymond, V. M., Eagle, C., Hu, S., Chudova, D. I., Talasaz, A., Greenson, J. K., et al. A cell-free dna blood-based test for colorectal cancer screening. *New England Journal of Medicine*, 390(11):973–983, 2024.

Deng, Z., Ji, Y., Han, B., Tan, Z., Ren, Y., Gao, J., Chen, N., Ma, C., Zhang, Y., Yao, Y., et al. Early detection of hepatocellular carcinoma via no end-repair enzymatic methylation sequencing of cell-free dna and pre-trained neural network. *Genome Medicine*, 15(1):93, 2023.

Foulds, J. and Frank, E. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1): 1–25, 2010.

Fürst, A., Rumetshofer, E., Lehner, J., Tran, V. T., Tang, F., Ramsauer, H., Kreil, D., Kopp, M., Klambauer, G., Bitto, A., et al. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems*, 35:20450–20468, 2022.

Gao, Q., Zeng, Q., Wang, Z., Li, C., Xu, Y., Cui, P., Zhu, X., Lu, H., Wang, G., Cai, S., et al. Circulating cell-free dna for cancer early detection. *The Innovation*, 3(4), 2022.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., and Shen, D. Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78, 2023.

Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10): 3088–3092, 1984. doi: 10.1073/pnas.81.10.3088.

Hu, J. Y.-C., Yang, D., Wu, D., Xu, C., Chen, B.-Y., and Liu, H. On sparse modern hopfield model. *Advances in neural information processing systems*, 36:27594–27608, 2023.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Jeong, Y., Gerhäuser, C., Sauter, G., Schlomm, T., Rohr, K., and Lutsik, P. Methylbert enables read-level dna methylation pattern identification and tumour deconvolution using a transformer-based model. *Nature Communications*, 16(1):788, 2025.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pp. 971–980, 2017.

Latif, S., Zaidi, A., Cuayahuitl, H., Shamshad, F., Shoukat, M., and Qadir, J. Transformers in speech processing: A survey. *arXiv preprint arXiv:2303.11607*, 2023.

Li, W., Li, Q., Kang, S., Same, M., Zhou, Y., Sun, C., Liu, C.-C., Matsuoka, L., Sher, L., Wong, W. H., et al. Cancerdetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free dna methylation sequencing data. *Nucleic Acids Research*, 46(15):e89, 2018.

Liang, J. T., Yang, C., and Myers, B. A. A large-scale survey on the usability of ai programming assistants: Successes and challenges. In *Proceedings of the 46th IEEE/ACM international conference on software engineering*, pp. 1–13, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luo, H., Wei, W., Ye, Z., Zheng, J., and Xu, R. Liquid biopsy of methylation biomarkers in cell-free dna. *Trends in Molecular Medicine*, 27(5):482–500, 2021.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., et al. Transformers and large language models in healthcare: A review. *Artificial intelligence in medicine*, 154:102900, 2024.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36: 43177–43201, 2023.

Niki, P., Nalmpantis, C., Ganbat, J.-O., Byrne, D., Babikir, H., Jhutty, A., Rowe, W., Liu, T., Loyfer, N., Toniolo, S., et al. Human whole epigenome modelling for clinical applications with pleiades. *bioRxiv*, pp. 2025–07, 2025.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. *ArXiv*, 2008.02217, 2020.

Schäfl, B., Gruber, L., Bitto-Nemling, A., and Hochreiter, S. Hopular: Modern hopfield networks for tabular data. *arXiv preprint arXiv:2206.00664*, 2022.

Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., Hochreiter, S., and Klambauer, G. Modern hopfield networks for few- and zero-shot reaction prediction. *ArXiv*, 2104.03279, 2021.

Shaukat, A., Burke, C. A., Chan, A. T., Grady, W. M., Gupta, S., Katona, B. W., Ladabaum, U., Liang, P. S., Liu, J. J., Putcha, G., et al. Clinical validation of a circulating tumor dna–based blood test to screen for colorectal cancer. *JAMA*, 334(1):56–63, 2025.

Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., and Jemal, A. Cancer statistics, 2025. *Ca*, 75(1):10, 2025.

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., and Shendure, J. Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, 164(1):57–68, 2016.

Thierry, A. R. Circulating DNA fragmentomics and cancer screening. *Cell Genomics*, 3(1):100242, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

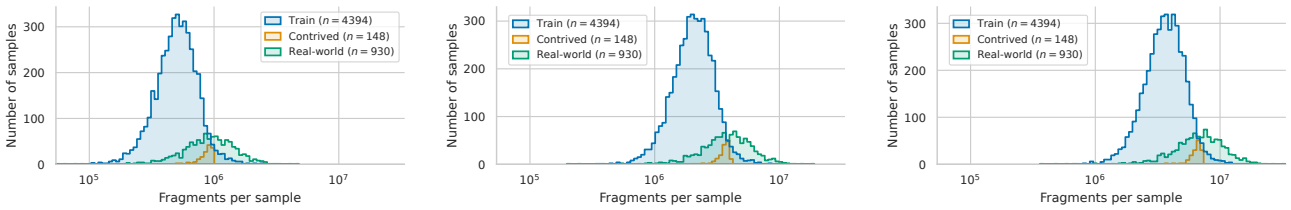
Widrich, M. Long short-term memory and convolutional neural networks for SNV-based phenotype prediction. Master's thesis, JOHANNES KEPLER UNIVERSITY LINZ, 2016.

Widrich, M., Schäfl, B., Pavlović, M., Ramsauer, H., Gruber, L., Holzleitner, M., Brandstetter, J., Sandve, G. K., Greiff, V., Hochreiter, S., and Klambauer, G. Modern Hopfield networks and attention for immune repertoire classification. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18832–18845. Curran Associates, Inc., 2020.

Widrich, M., Hofmarcher, M., Patil, V. P., Bitto-Nemling, A., and Hochreiter, S. Modern hopfield networks for return decomposition for delayed rewards. In *Deep RL Workshop NeurIPS 2021*, 2021.

Yamaguchi, S., Asao, T., Marcet-Palacios, M., Al-Kasspooles, M., Lee, J., Weitz, J., Ambrosini, G., Ju, J., Wiley, E., and Bland, K. High frequency of dap-kinase gene promoter methylation in colorectal cancer specimens and its identification in serum. *Cancer Letters*, 194 (1):99–105, 2003.

A. Fragment count distributions for different input space denoising strategies



(a) Prior Denoised *(b)* CRC Panel *(c)* Expanded Region

Figure A1. Fragment count distributions for the Prior Denoised (PD), CRC Panel (C), and Expanded Region (ER) training sets vs. the challenging contrived and real-world clinical test sets. Input space size increases from (a) to (b) to (c).

Fig. A1 illustrates the distribution of fragment counts per sample, for training and held-out test datasets, across the three input space configurations described in Sec. 3 and Sec. D. As the model's input space expands, more fragents are in scope, and we observe the expected rightward shift in the fragment count distributions. The PD-FLDL configuration, which applies aggressive biological filtering to remove noisy regions, yields the smallest input space and the lowest fragment counts. The C-FLDL configuration uses the CRC panel without biological prior filtering, resulting in an intermediate increase in input space and typical fragment count. Finally, the ER-FLDL configuration, which uses the full capture panel, leads to the largest number of instances, with fragment counts almost always exceeding 10^6 per sample and sometimes reaching 10^7 fragments. Of note, both the real-world clincial and challenging contrived hold-out test sets are generated using an improved wet-lab sample processing pipeline; as a result, they show a marked increase in fragment count per sample compared to the training set. The pipeline improvements are intended to increase accuracy of future blood test versions, but as a side effect, the distribution shift also provides a rigorous testbed for evaluating model generalization in the face of an evolving input distribution.

B. Implementation details

B.1. Details on Fragment Representation and Embedding

Each cfDNA fragment s_i is represented using five input modalities. **Sequence and CpG Methylation:** the nucleotide sequence and per-CpG methylation status of the fragment, represented as a one-hot encoded matrix with channels for sequenced nucleotides, reference nucleotides, and CpG methylation status. **Methylation Statistics:** a vector of per-fragment summarized methylation statistics, including the sequence length and length-normalized counts of methylated, unmethylated, and total CpGs. **Foundation Model Embedding:** a learned representation based on the HyenaDNA foundation model (Nguyen et al., 2023). This embedding is derived from large-scale training on the human reference genome and captures long-range genomic dependencies and high-order sequence motifs. **Genomic Position:** the numerical location of the fragment start position in a concatenated version of the human reference genome. **Strand:** a boolean indicator for the DNA strand. To ensure a balanced contribution from the heterogeneous data modalities, each modality is projected to a latent embedding of uniform dimension m via using a dedicated modality-specific encoder. Sequence features are processed via a 1D Convolutional Neural Network (CNN) followed by max pooling and a Multi-Layer Perceptron (MLP). The foundation model embeddings are projected to m dimensions using an MLP. Scalar features such as methylation statistics and genomic position are encoded using triangular encoding (Widrich, 2016) followed by MLPs, while the boolean strand feature is processed directly by an MLP. Lastly, the individual modality embeddings, each of size m, are concatenated and processed by an MLP to form a unified fragment representation $\boldsymbol{h}_i \in \mathbb{R}^m$. This is illustrated in Fig. A2.

B.2. Implementation, training, ensembling

Training on millions of fragments per sample presents significant computational challenges. We implement a two-stage fragment subsampling strategy to mitigate runtime and GPU memory consumption. During training and inference for early stopping, random dropout of fragments is followed by further attention-based subsampling that retains the fragments with the highest current attention scores. For test-time inference, we omit random dropout and only apply the attention-based subsampling. We train the model end-to-end in PyTorch (Paszke et al., 2019) using the AdamW optimizer (Loshchilov & Hutter, 2017). To prevent exploitation of confounding effects such as collection batch or patient age, we employ a dynamic minibatch balancing scheme that pairs samples with opposite class labels but similar confounding characteristics.

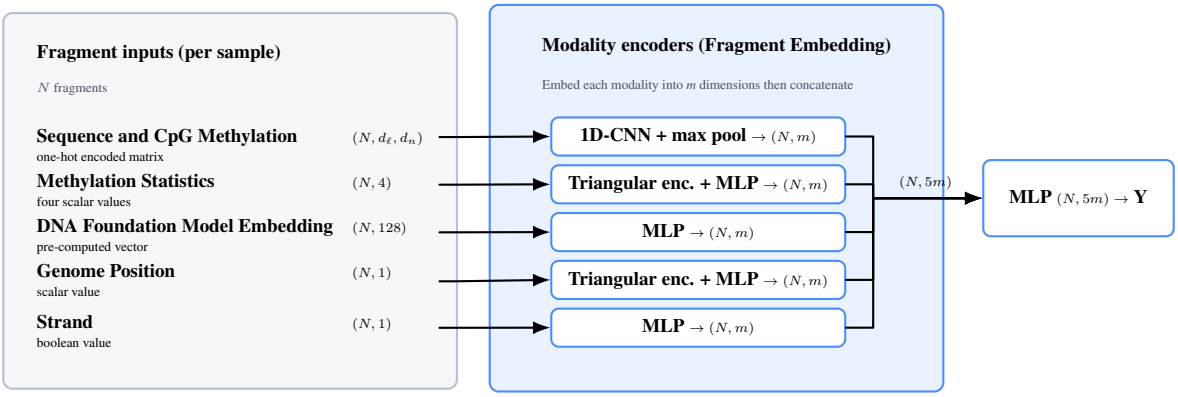


Figure A2. FLDL Fragment Embedding: Each input modality is embedded by a dedicated encoder. After initial embedding, the per-modality embedded vectors are concatenated and processed using an MLP.

To improve generalization, the model is also trained with auxiliary tasks: predicting relevant clinical data and biological characteristics alongside the primary binary disease status.

Due to the relatively low number of training samples and low signal-to-noise ratio (as a result of low witness rate), the representation learned may vary among trained models based on the random weight initialization and the order in which samples are selected for use during training. To address this, we ensemble 25 models obtained from 5 random restarts of 5-fold cross validation (CV). Specifically, we apply 5-fold cross validation with 5 random restarts to obtain 25 members for the final ensemble model (Fig. A3). In each case, 4 different hyperparameter settings are considered, and the optimal hyperparameter choice for that iteration is the one that maximizes the minimum of training and tuning set accuracy. (While uncommon, this maximin metric provides stronger regularization, which is helpful given the small tuning set sample size.) The final model score for a sample is the average of the scores from the 25 models in the ensemble.

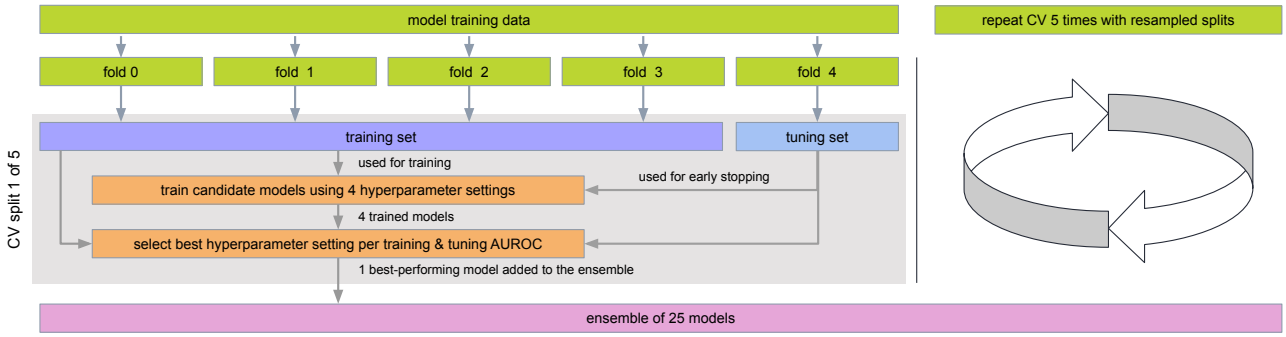


Figure A3. Model training and ensemble construction methodology using 5-fold cross validation with 5 random restarts.

C. Dataset details

Our **training set** consists of 4,394 samples, of which 925 are positive cases, comprising 377 CRCs and 577 of the more difficult APLs, and 3,469 are negative controls. To assess classification accuracy on data that are representative of future use cases, we utilize two complementary independent hold-out test sets: a) a **real-world clinical sample set** with 930 samples, containing 331 negative and 599 positive samples (388 APLs, 211 CRCs) collected independently of the training dataset; and b) a **challenging contrived set**, to probe the detection of samples with low witness rates in a controlled setting and with much higher replication. The contrived set consists of 148 replicates created by mixing material from a single advanced CRC donor into plasma from a healthy donor pool to yield a ctDNA level just above the state-of-the-art machine learning model's detection limit. To inspect the learned representations of the FLDL model's sample-level embeddings, we leverage two additional hold-out test sets: c) a **high-signal contrived set** consisting of 36 replicates of a blend similar to the challenging contrived set but with ctDNA level well above the machine learning model's detection limit, and d) an **artificial no-methylation set** consisting of 92 synthetically unmethylated samples, where any residual methylation signal can be

attributed solely to technical noise.

D. Compared Methods

FLDL variants. The wet-lab sample processing pipeline's full capture panel targets >500 kilobases (kb) of differentially methylated genomic regions, identified through an iterative process leveraging public and internal DNA methylation data. While the full panel captures signals relevant to multiple cancer types, a subset is tailored to maximize CRC signal detection and is referred to as the CRC panel. In order to detect sparse signals necessary for early cancer detection, both the full and CRC panels retain certain loci that can exhibit sporadic background methylation in healthy individuals, which requires models to denoise, i.e., ignore such loci so that non-ctDNA fragments from noisy regions do not lead to false positives. While the machine learning baseline incorporates an explicit denoising step during training (described below), the FLDL architecture does not. FLDL can rely on attention for implicit denoising, and biologically informed pre-filtering can assist in denoising when training data are limited.

To investigate the effect of input space denoising on model performance at the current training dataset size, we evaluate three distinct FLDL configurations:

Prior Denoised (PD-FLDL): A variant incorporating a pre-filter step outside of model training to reduce noise in input data. We first profile methylation signal in the CRC panel using a hold-out cohort of 825 healthy individuals. Regions exhibiting elevated background hypermethylation in these controls are aggressively filtered, yielding a reduced, higher-signal-to-noise input space. We refer to this data-driven pre-filtration of the feature space as denoising with *explicit biological priors*.

CRC Panel (C-FLDL): The FLDL model operating on the CRC panel without prior filtering. This configuration tests the architecture's capacity to implicitly denoise the input space and identify cancer-specific signals within the CRC panel, which is smaller than the full panel but still includes noisy regions.

Expanded Region (ER-FLDL): A scalability benchmark extending the input space to the full capture panel. This configuration further challenges the model to isolate CRC signals within a substantially larger search space that includes a higher fraction of noisy features.

Note that PD-FLDL regions are a subset of C-FLDL regions, which are a subset of ER-FLDL regions. Using a larger panel with more regions results in more fragments per sample (Appendix Fig. A1).

Deep learning via max pooling. To isolate the contribution of the MHN attention mechanism, we evaluate an alternative deep learning model that retains the FLDL fragment feature extractor but replaces Hopfield Pooling with a fragment aggregation strategy based on maximum embedded feature activations. Starting with $\boldsymbol{Y} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N]$ as defined in Section 2, the MaxPool approach constructs a sample embedding $\boldsymbol{z} \in \mathbb{R}^{m^2}$ by identifying m representative fragments, one for each fragment embedding dimension. Specifically, for each dimension j, we determine the fragment $\boldsymbol{h}_{i^*(j)}$ that maximizes the activation of that dimension: $i^*(j) = \arg\max_i h_i^j$. We then define $\boldsymbol{r}_j = \boldsymbol{h}_{i^*(j)} \in \mathbb{R}^m$. The final sample embedding $\boldsymbol{z}$ is obtained by concatenating $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_m$ to yield $\boldsymbol{z} \in \mathbb{R}^{m^2}$. This pooling strategy is motivated by two key considerations: a) it acts as a discrete analogue to attention-based pooling, where the maximum embedded feature activation serves as a proxy for importance; and b) by preserving the intact embedded fragment $\boldsymbol{h}_{i^*(j)}$ rather than creating a pseudo-fragment from dimension-wise scalar maxima, the model maintains the co-occurrence of features within the representative fragments. This in turn ensures that the sample representation is also constructed from intact embedded fragments. We evaluate this architecture on the Prior Denoised (**PD-MaxPool**) and CRC (**C-MaxPool**) input spaces, omitting the Expanded Region configuration due to poor performance in preliminary studies.

State-of-the-art machine learning model (ML Baseline). This model operates by identifying hypermethylated fragments (HMFs) in the CRC panel, where HMFs are defined as cfDNA fragments showing significantly more CpG methylation than typically seen in similar fragments obtained from individuals without disease (Shaukat et al., 2025). In practice, the model first subdivides the CRC panel into small, similarly sized bins. For each bin b, fragments derived from healthy samples are compared to fragments derived from cases, and a per-bin methylated CpG threshold t_b is learned. Fragments intersecting bin b and with a methylated CpG count at or above t_b are deemed to be hypermethylated. The model aggregates HMF counts observed throughout the regions of interest in order to compute an overall score for the sample. This approach includes an explicit denoising step in the training process: only fragments with methylated CpG counts exceeding that seen in training set control samples are allowed to contribute to the classification score. As for FLDL, an ensemble of 25 such models, each trained to different subsets of the training data, produces the final binary classification.

E. Additional Results

E.1. Performance on challenging contrived test set

The challenging contrived test set provides control over the true ctDNA level, and the blending process provides enough plasma volume for substantially more replication than is possible with a conventional clinical sample. Using the same model-specific 90% specificity classification thresholds as in Table 1 we report each model's proportion of positive predictions among the 148 replicates.

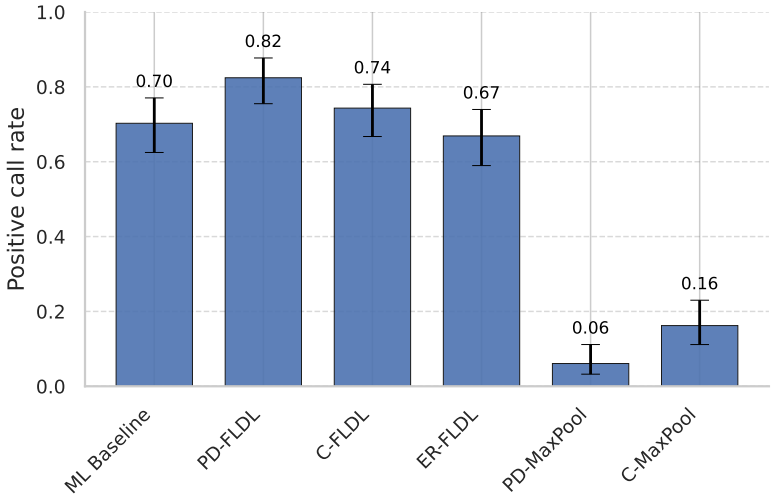


Figure A4. Positive call rates among 148 replicates of the challenging contrived test set, at 90% specificity. Whiskers indicate Wilson's method 95% CIs.The attention-based PD-FLDL outperforms the ML Baseline and implicitly denoised variants, while max pooling models fail to capture the sparse tumor signal in this low signal-to-noise regime.

E.2. Scaling behavior with increasing number of training samples

To understand data requirements for the CRC detection task and assess potential performance plateaus we train the C-FLDL model on nested subsets of our full training dataset and report the resulting models' positive call rate on the challenging contrived test samples. To preserve the ratio of positives to negatives, we subsample the positive cases (CRCs and APLs) and negative controls separately, to fractions ranging from 20% to 100% in 20% increments. To reduce computational effort during training, we only ensemble 5 models from a single round of 5-fold cross validation.

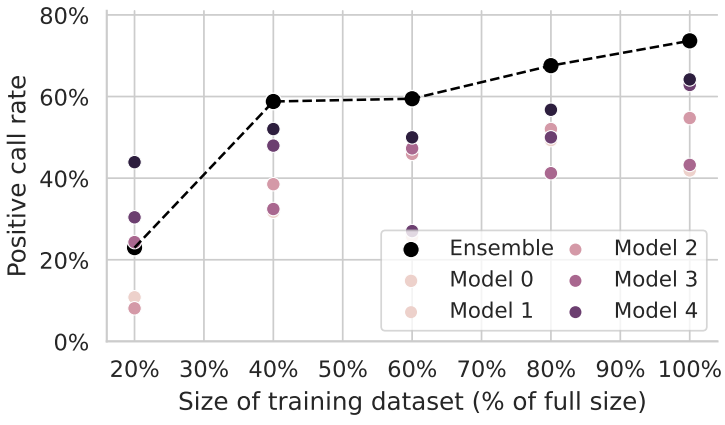


Figure A5. Positive call rates among 148 replicates of the challenging contrived test set, at 90% specificity, for C-FLDL models trained on subsets (20% to 100%) of the full training dataset.

Results in Appendix Figure A5 shown for the ensemble of 5 models as well as for individual ensemble members (Model 0 to Model 4). The ensemble (black) largely outperforms individual folds (colored) and shows consistent performance gains without a plateau suggesting benefits from additional data.

E.3. Attention value analysis for fragment interpretability

For FLDL interpretability at the fragment level, we analyze the model's attention values. To identify the genomic loci driving the C-FLDL model's predictions, we analyze the distribution of high-attention fragments within the CRC panel. The starting points for this analysis are the fragments retained after the attention-based subsampling of the Hopfield Pooling layer. To focus on those fragments with the most meaningful contributions to the prediction, we apply a sample-specific filtering step: Given the attention matrix $\boldsymbol{A} \in \mathbb{R}^{K \times N}$, for each of the K state patterns we retain only those fragments with an attention value above threshold $t_i = \tau \cdot \max_j(A_i^j)$ where $i \in \{1, \ldots, K\}$, $j \in \{1, \ldots, N\}$, and $\tau = 0.1$. Next, we partition the genomic regions interrogated by the CRC panel into b non-overlapping bins, maintaining consistency with the bins used by the ML Baseline model. The filtered fragments are aligned to these bins, and the relevance count for a bin is incremented for each fragment that at least partially overlaps the bin. This aggregation identifies the genomic regions prioritized by the C-FLDL model for its classification decisions. We compare this to the ML Baseline model's per-bin HMF counts.

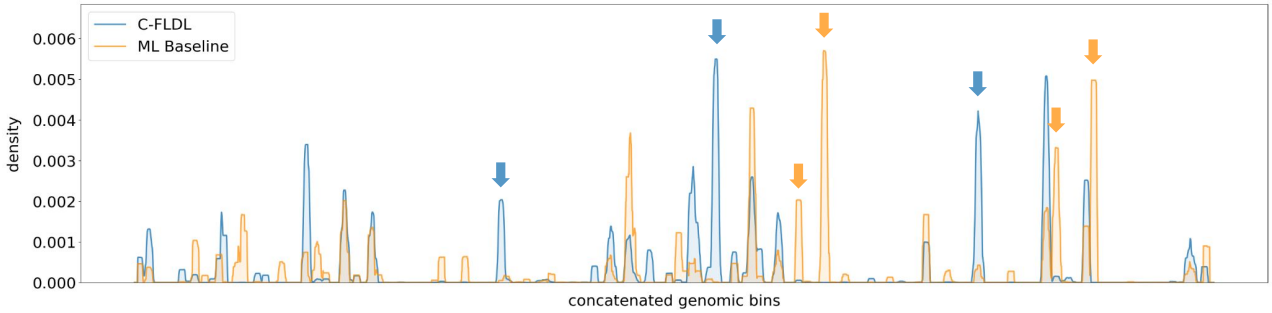


Figure A6. Classification relevance for genomic bins comprising the CRC panel, in 13 correctly predicted late-stage CRC samples from the real-world clinical test set. Blue curve: density of high-attention fragments from C-FLDL. Orange curve: density of HMFs identified by ML Baseline. Arrows indicate regions prioritized by one model but not the other. Both curves are smoothed with a moving average.

Fig. A6 illustrates this comparison using data from 13 correctly classified late-stage CRC samples from the real-world clinical test set. While the models exhibit high-frequency variation at the individual bin level, they demonstrate substantial agreement after smoothing with a moving average, as seen by the overlaps between the orange and the blue curves. Note that a subset of each model's prioritized genomic loci is ignored by the other model (arrows). This complementarity may suggest that neither model has fully saturated the available signal, or that the models have made different but largely equivalent sparsification choices when presented with sets of loci showing correlated signal.