Multi-View Feature Analysis of Cell-free DNA Data for Cancer Prediction

by

Lucia Navarčíková

In partial fulfilment of the requirements for the degree of
Master of Science
at Delft University of Technology,
to be defended publicly on **29. 6. 2026**

Faculty: **Faculty of Electrical Engineering, Mathematics and Computer Science**
Department: **Intelligent Systems**
Programme: **Data Science and Artificial Intelligence Technology**

Mentors / Supervisors: Marcel Reinders (Thesis Advisor)
Bram Pronk (Daily Co-Supervisor)

Graduation committee: Marcel Reinders (Chair)
Willem-Paul Brinkman (Core Member 2)
Bram Pronk (Advisor)

Master Thesis, 2026, pp. 1–24

doi:

Multi-View Feature Analysis of Cell-Free DNA Data for Cancer Prediction

Lucia Navarčíková,[1] Bram Pronk[1] and Marcel Reinders[1]

[1]Bioinformatics Lab, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE, Delft, The Netherlands

Abstract

Cell-free DNA (cfDNA) fragmentomics using liquid biopsy data has emerged as a promising minimally invasive approach for cancer detection. While multiple fragmentomic features, including the Fragment Short Long Ratio (FSLR), Motif Diversity Score (MDS), and Copy Number Alterations (CNA) have individually demonstrated predictive value, their complementarity and optimal integration remain insufficiently explored. In this study, we investigate the relationships between these fragmentomic features in a genome-wide setting and evaluate their complementarity through multi-view intermediate integration for the binary classification task. Variance decomposition with correlation analysis showed that FSLR, MDS, and CNA capture partially non-redundant aspects of tumour-derived cfDNA signal, with only 7.2% overlap among outliers. This biological complementarity did not translate into drastically improved predictive performance. The strongest downstream model was the PCA-based concatenation baseline using all three views, achieving an AUC of 0.969, with only marginal gains over CNA alone. In contrast, MOFA+ did not improve classification performance, reflecting a mismatch between its variance-maximization objective and cancer–healthy discrimination in cfDNA data. Similarly, Contrastive Multi-View Kernel Learning (CMK) failed to yield separable representations under an unsupervised objective, with meaningful class structure emerging only when supervision was introduced, yet still not surpassing the concatenation baseline. Across all methods, CNA was consistently the most discriminative single feature, while FSLR provided an additional independent signal. MDS performed worst in all settings and contributed little to predictive performance. This limited contribution may reflect the chosen 5 Mb resolution rather than an inherent lack of biological signal, suggesting that feature-specific resolution optimization should be considered prior to integration.

Key words: cell-free DNA, fragmentomics, whole genome sequencing, classification

Introduction

Cancer is one of the leading causes of death in the population, with as many as nearly one in six deaths attributed to cancer according to the World Health Organization [1]. This group of diseases manifests through the uncontrolled growth of malignant cells, forming tumours and later spreading to the whole body. Due to the variation in symptoms across cancer types with many cancers being asymptotic in the early stages [2] or exhibiting symptoms of other common diseases, reliable cancer classification remains an active field of research [3].

Liquid biopsy is a minimally invasive diagnostic approach that analyzes biomarkers obtained from bodily fluids, most commonly blood. A blood sample consists of approximately 55% plasma and 45% red blood cells [4]. The plasma component has been found to contain shorter DNA segments, called circulating cell-free DNA (cfDNA). cfDNA is typically released into circulating blood following cell death [5], either through a variety of programmed cell deaths such as apoptosis or as a result of accidental and non-regulated process such as necrosis. Consequently, any pathology with increased cell death releases higher levels of cfDNA from the affected tissue into circulation [6], with tumour-derived fragments in cancer specifically referred to as circulating tumour DNA (ctDNA).

Unlike tissue biopsies, liquid biopsy does not require tumour localisation and has lower associated costs [7], making it applicable at any stage of treatment. The short half-life of cfDNA in circulation, estimated at minutes to 1–2 hours [8], leads to cfDNA composition reflecting the current balance of tumour and healthy cell-derived fragments, making liquid biopsy well-suited for real-time monitoring of tumour burden and treatment response [9]. Nevertheless, reliably detecting cancer-derived signal remains non-trivial, motivating the development of feature engineering for classification.

Fragmentomics is the study of fragmentation patterns in cfDNA. Using whole genome sequencing (WGS) the liquid biopsy sample is processed into reads, which correspond to cfDNA fragments. The reads can then be analyzed to obtain sample measurements such as fragment length. The length of a fragment can be a discriminative feature for classification [10], as it is typically 166 base pairs (bp) for healthy individuals, with ctDNA fragments exhibiting shorter lengths of approximately 143 bp [6].

Moving beyond fragment length derived features, the findings of Jiang et al. [11] showcase an increase of end motif diversity in ctDNA, where motif represents a short nucleotide sequence. These aberrant end motifs have been observed

across multiple cancer types and can be indicative of tissue of origin [11], making them a suitable cancer biomarker in prior literature [12], [11], [13]. In addition, copy number alterations (CNA), referring to gains or losses of a genomic region such as a chromosome arm, are extremely common in cancer [14], making them a suitable cancer biomarker with particular cancer types such as breast or ovarian cancer being more affected than others [15].

While early studies such as DELFI [16] established cfDNA fragmentation as a valuable source of cancer-related information, subsequent work explored whether combining multiple fragmentomic features improves classification performance over single feature approaches. Chabon et al. [17] combined fragment size, copy number, and mutation-based features for early lung cancer detection, and Hou et al. [18] showed that an ensemble classifier integrating ten fragmentonic patttens improved cancer detection over any single pattern alone. Furthermore, multi-modal frameworks such as SPOT-MAS [19] and MESA [20] demonstrated that integrating multiple fragmentation-derived features across modalities can further achieve higher classification performance.

However, the integration strategies employed in these works are limited to early or late fusion. SPOT-MAS concatenates multimodal features into a single dataframe prior to classification, while MESA trains separate models per modality and learns an optimal combination of their output probabilities. Many commonly used fragmentomic features originate from the same underlying biological processes related to chromatin organization and nucleosome positioning [3], suggesting that they may contain complementary and shared information about tumor-derived cfDNA. Previous works [19, 18, 20] analyzed features independently prior to their integration, consequently making the relationships between fragmentomics features and their methods of integration vastly underexplored. More specifically, no previous approaches have explored their complementarity through *intermediate integration*, in which a shared latent representation is learned jointly across feature views prior to classification. Exploring such approaches could improve classification performance by leveraging dependencies between feature types, while simultaneously providing insight into how different manifestations of cfDNA fragmentation reflect underlying cancer-associated biological processes.

Multiple features extracted from WGS data translate naturally to the multi-view domain, where each view offers differing information on the same sample. In the genome-wide setting, each feature computed across genomic bins constitutes a view, and the views jointly characterize the sample. Further suitability of multi-view learning for cfDNA classification comes from its two principles: *consensus* and *complementarity* [21]. The consensus principle reflects the agreement between views: since features are computed from the same sample cohort, they share discriminative information for cancer/healthy separation and should be integrated in a way that promotes view alignment. The complementarity principle captures the view-specific information, which must be preserved to achieve performance gains over single view approaches.

Consensus principle has been widely implemented in the multi-view domain. Canonical Correlation Analysis (CCA) [22] seeks to maximize the correlation between two projected views, directly encoding view agreement. Kernel CCA extends this to nonlinear settings by first mapping samples into a kernel Hilbert space, and has been used together with SVM in SVM-2K [23] to enforce hyperplane agreement across views. Multi-objective formulations such as MVMED [24] and MED-2C [25] balance both principles simultaneously. MED-2C integrates classification directly into the feature transformation, resulting in an augmented representation of higher dimensionality than the original features, which is undesirable given the small sample sizes typical of this domain.

The multi-view setting naturally gives itself to shared variation as fragmentomics features are computed on the same patient cohort, where shared variation translates to consensus across the views. MOFA+, originally developed for multi-modal single-cell data, untangles the shared and individual variation across modalities and groups through its sparsity inducing priors on both factors and weights [26]. The modelling of individual variation in this case captures the complementary information between the fragmentomics features. Furthermore, the reduced dimensionality of latent factors and the modality specific weights offers interpretability, which is crucial for further understanding of underlying biological changes in cancer. Overall, by integrating fragmentomic features through the lens of shared and individual variation, MOFA+ facilitates the interpretation of their complementarity while producing low-dimensional latent factors that are suitable for downstream classification tasks.

Aside from variance-based approaches such as MOFA+, contrastive learning explicitly encourages agreement between representations of the same sample across views [27], thereby aligning with the consensus principle in multi-view learning. Contrastive Multi-View Kernel Learning by Liu et al. [28] extends this paradigm with a second training stage that incorporates a spectral objective to promote cluster separation among samples. By aligning fragmentomic views in a shared latent space, this framework can capture shared biological signal while utilizing complementary information through intermediate integration. Furthermore, by jointly training with contrastive and spectral objectives, they demonstrated that the resulting sample representation achieves high classification performance [28], making the approach well-suited for intermediate integration method for cancer classification using liquid biopsy data.

Thereby in this work, we aim to address the gap of missing complementarity analysis between the commonly fragmentonics features in cancer classification using WGS liquid biopsy data. First, we asses their complementarity through outlier overlap, trends in variation across cancer/healthy classes and genome-wide correlation between feature pairs. We go on to utilize MOFA+ as an interpretable tool to jointly model shared and individual variation across fragmentonics features. Furthermore, we investigate their complementarity through the lens of classification performance, to assess whether integrating fragmentonics features adds discriminative signal for cancer/healthy separation or whether it primarily introduces noise. Lastly, we apply architecture-wise distinct intermediate integration approaches, MOFA+ and Contrastive Multi-View Kernel Learning, to assess the benefits of intermediate integration compared to a concatenation-based baseline in this domain.

Results

To assess the complementarity of fragmentomics features and intermediate integration approaches, we utilized the patient cohort from Cristiano et. al [16]. The experiments were conducted on 204 healthy and 203 cancer samples with a separate group of 50 healthy samples used as a Panel of Normals (PON).

The fragmentomics features investigated in this study are Fragment Short Long Ratio (FSLR) [16], Motif Diversity Score (MDS) [11] and Copy Number Alterations (CNA) [29]. While FSLR captures the size discrepancy of fragments at a specific region, CNA reflects deviations to normal baseline based on large regions such as chromosome arms deletions or amplifications. With MDS quantifing the diversity of DNA cleavage site motifs within a genomic bin, these fragmentomics offer distinct views on cancer signal in liquid biopsy data. By genome-wide integration of these fragmentomics features, we expect higher cancer/healthy separation compared to utilizing the features individually.

Complementarity

To assess the complementarity of fragmentomics features, we first examined the raw values of FSLR, MDS and CNA across the genomic bins. Liquid biopsy data contains the fragments of healthy cells as well as cancer cells, therefore outliers relative to the healthy class are assumed to have strong cancer signal. Consequently, examining the overlap in identified outliers to the PON group across the fragmentomics views gives insights into the their complementarity.

Table 1 quantifies the overlap in cancer outliers relative to PON group across the fragmentomics features. Relating lower outlier overlap to higher potential complementarity, 7.2% consensus when utilizing all 3 fragmentomics features indicates possible performance gains for downstream cancer classification. Conversely, the highest consensus across cancer types is present for for breast, pancreatic and colorectal cancer, with lung cancer showing high consensus (66.7%) for the MDS CNA pair.

Aside from complementarity inspection through mean fragmentomics value, we extend the feature analysis to a genome-wide setting by investigating correlation to PON group and sample visualization per fragmentomics feature. Cancer samples show consistently elevated FSLR values and greater inter-sample variability (Supplementary Figure 1). Pancreatic cancer samples with genome-wide correlation below 0.6 can be identified as outliers in other views, suggesting limited complementarity gains in for this cancer group in concordance with our previous findings. Bile duct cancer samples are visually indistinguishable from the healthy group in MDS (Supplementary Figure 2), despite one sample deviating clearly in FSLR, highlighting a case where the two features provide non-redundant information. Conversely, the two lowest lung cancer samples in MDS were not flagged in FSLR, while a bile duct sample with 0.540 MDS correlation achieved 0.929 correlation in CNA (Supplementary Figure 3), further illustrating that poor signal in one feature does not preclude strong signal in another.

To investigate the class differences genome-wide, we aggregated the fragmentomics metrics FSLR, MDS and CNA as an average per bin after Z-score normalization with the PON group. CNA and MDS exhibit greater variance for the cancer class compared to FSLR (Supplementary Figure 4), with healthy variance staying consistent throughout fragmentomics features. The greater variance in cancer samples for MDS and CNA drives the class separation of class averages, affirming the usage of these fragmentomics features for binary classification.

To assess the degree of shared information between fragmentomic views, Pearson correlations were computed between all feature pairs (Supplementary Figure 5). Inspecting the correlation between MDS and FSLR, the correlation between these fragmentomics features seem low to moderate (-0.25 to 0.25), but whether this is noise or complementary signal will depend on the downstream classification performance. On the other hand, the increased correlation in FSLR and CNA, especially in the cancer class indicate that these two views share the same information. The amplification or deletion in cancer would affect the mapping of read fragments, thereby also affecting the FSLR as a ratio of short to long fragments. This consequently leads to higher correlation across these features for affected chromosome locations. Upon closer inspection the correlation peaks between FSLR and CNA (chromosomes 7, 12, 13 and 17) are driven by different subsets of cancer types across chromosomes, with no single cancer type dominating across all regions. The last feature pair, MDS and CNA show low whole-genome correlation in both the cancer (-0.4) and healthy (-0.01) groups, with only minor peaks in the cancer class at chromosomes 1, 3 and 7.

The consistently low healthy class correlation across all pairs reflects the biologically distinct processes underlying each feature, while the cancer-specific peaks in FSLR CNA indicate that shared signal between these views arises from tumour-driven genomic alterations. Taken together, the outlier overlap and correlation analyses consistently indicate that FSLR, MDS, and CNA capture partially non-redundant aspects of tumour-derived cfDNA signal, providing a rationale for their integration in downstream classification.

MOFA+

To evaluate the complementarity of the fragmentomic views by decomposing their shared and view-specific sources of variation, we applied MOFA+ to jointly model the FSLR, MDS and CNA data. We analyzed the sample cohort excluding the PON group and examined the proportion of variance explained (R^2) by each latent factor within each view and sample group (Figure 1). Six latent factors were identified.

This analysis revealed two dominant patterns of variation across the fragmentomic feature sets. First, MDS contributed strongly to the latent representation, with Factor 1 capturing a major source of variation present in both healthy and cancer samples. In contrast, multiple factors were driven jointly by the FSLR and CNA views, indicating a substantial amount of shared information between fragment length profiles and copy number alterations. This finding is consistent with the previously observed correlations between these feature sets at specific genomic bins.

In addition, a subset of factors captured variation largely restricted to cancer samples, indicating the expected increased heterogeneity within the cancer group and a disproportionate contribution of cancer-specific variability to the latent space. This aligns with the higher cancer group variance observed in prior exploratory analyses.

Inspection of the factor weights showed broadly distributed contributions across the genome, with no strong localization to specific chromosomal regions. This suggests that the corresponding factors reflect genome-wide signals rather than

Disease	FSLR MDS (%)	FSLR CNA (%)	MDS CNA (%)	FSRL MDS CNA (%)
Bile duct cancer	0.0	7.1	14.3	0.0
Breast cancer	35.0	34.8	33.3	16.7
Lung cancer	0.0	0.0	66.7	0.0
Pancreatic cancer	25.0	16.7	33.3	16.7
Colorectal cancer	20.0	58.3	22.2	16.7
Gastric cancer	0.0	7.1	14.3	0.0
Ovarian cancer	0.0	50.0	0.0	0.0
Average	**11.4**	**24.9**	**26.6**	**7.2**

Table 1. Pairwise and three-way Jaccard index overlap (%) of outlier samples identified by each fragmentomics view (FSLR/MDS/CNA), where outliers are defined as samples exceeding 1.5 IQR from the PON reference distribution averaged across the whole genome, stratified by cancer type.

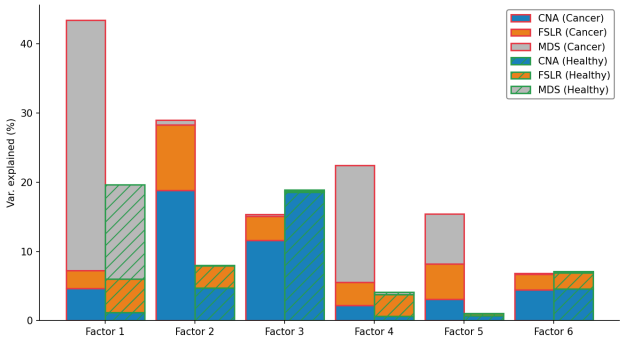


Fig. 1. MOFA+ Latent Factors (1-6) ranked by variance explained per view (FSLR/MDS/CNA) and group (Healthy/Cancer). MOFA+ was applied on all three fragmentonics views at 5Mb resolution allowing for up to 10 factors to be found with Gaussian likelihoods and required explained factor variance (R2) at 0.001.

region-specific effects, supporting the genome-wide approaches in this field.

Inspecting the factor values per sample, Figure 2 showcases the spread across Factor 1. The distributions of healthy and cancer samples show substantial overlap, indicating that this latent factor alone does not provide sufficient discriminative power to reliably distinguish healthy individuals from cancer patients. The remaining factors in Supplementary Figure 6 follow this trend, with healthy samples clustering more tightly around zero, while cancer samples display greater dispersion across the latent dimensions.

To assess class separability in the MOFA+ latent space for the downstream classification task, we computed the silhouette scores using the binary cancer/healthy label. The overall score was low (0.068), indicating limited cluster separation. The healthy samples formed a relatively coherent cluster (silhouette score = 0.370), while the cancer samples were dispersed on the periphery, as visualized using UMAP in Supplementary Figure 7. Computing silhouette scores across varying cancer types confirmed that cancer samples do not cluster based on cancer subtype. Even though the cancer samples do not form a coherent cluster or small clusters based on cancer subtype, their location on the periphery suggest that a downstream classifier can separate the classes based on outlier detection. These findings further support the use of a downstream supervised classifier rather than relying on unsupervised clustering-based approaches for classification.

Intermediate Integration for Classification

While variance-based complementarity analysis gives insight into shared information between the fragmentomics features, it does not directly establish whether combining these views is beneficial for the cancer classification task. Classification performance provides a direct empirical test of this complementarity: if the views capture distinct, *informative* signal, their integration is expected to improve discrimination between cancer and healthy samples relative to any single view used in isolation. We therefore evaluate integration approaches for classification, first establishing baselines (Table 1) as the reference point against which the intermediate integration methods, MOFA+ and Contrastive Multi-View Kernel Learning are compared.

Baseline

For the single-view baseline, each view is reduced independently via PCA and classified using Support Vector Machine and Logistic Regression classifier under 5-fold nested cross-validation (Table 2). For the multi-view baseline, the per-view PCA representations are concatenated prior to classification, representing a naive form of integration with no explicit modeling of cross-view structure.

In the single-view setting, CNA achieves the best overall performance (AUC 0.962, Accuracy 0.881), with FSLR close behind and MDS notably weaker across all metrics. MDS also shows the highest variance in accuracy and the lowest mean sensitivity at 95% specificity (0.444), despite the class separation previously indicated for this view (Supplementary Figure 4). This suggests that while MDS carries some separation signal, the high variance of this fragmentonic feature could be hindering its translation into classification performance gains.

In the pairwise setting, combinations involving CNA perform best overall. Improvements over the single-view baselines are generally small across most metrics, but M.S. at 95% specificity shows the most notable gains, particularly for FSLR CNA pair. Adding MDS to FSLR, by contrast, decreases accuracy and AUC relative to FSLR alone, with only a marginal precision gain. To further understand this, we compared the LR coefficients for FSLR and FSLR MDS (Figure 3): the shared FSLR dimensions show high correlation between the two models (Pearson r = 0.921) and carry the largest magnitude weights, indicating that FSLR drives the FSLR MDS model and MDS contributes little additional discriminative information.

Combining all three fragmentomic features (FSLR MDS CNA, LR) achieves the highest accuracy of the baseline experiments, though the improvement over the best pairwise

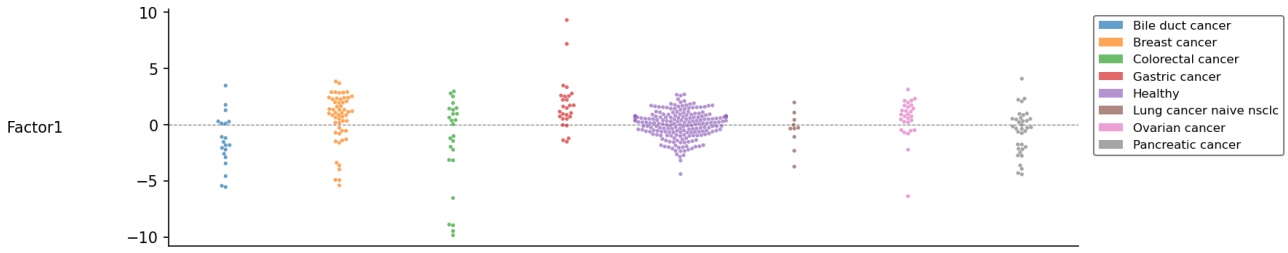


Fig. 2. MOFA+ Factor 1 sample visualization across healthy and cancer subtypes. Factor 1 was extracted using all three fragmentonics views (FSLR, MDS, CNA) at 5 Mb resolution with sample normalization and Z-scoring with PON group.

Table 2. Baseline Classification performance (mean ± std) across 5-fold nested cross-validation. PCA was applied separately per view with concatenation in multi-view settings. Each combination was evaluated using both Logistic Regression (LR) and Support Vector Machine (SVM) classifiers, specified in the Clf column. M.S. 95% stands for mean sensitivity at 95% specificity.

Model	Clf	Accuracy	AUC	Precision	Recall	M.S. 95%
FSLR	SVM	0.840 ± (0.025)	0.912 ± (0.024)	0.877 ± (0.035)	0.793 ± (0.066)	0.680
MDS	SVM	0.695 ± (0.039)	0.768 ± (0.031)	0.794 ± (0.126)	0.568 ± (0.158)	0.444
CNA	SVM	0.871 ± (0.023)	0.959 ± (0.016)	**0.911** ± (0.020)	0.819 ± (0.055)	**0.824**
FSLR	LR	0.848 ± (0.014)	0.919 ± (0.015)	0.887 ± (0.022)	0.798 ± (0.051)	0.734
MDS	LR	0.725 ± (0.043)	0.798 ± (0.031)	0.800 ± (0.020)	0.602 ± (0.088)	0.448
CNA	LR	**0.881** ± (0.029)	**0.962** ± (0.017)	0.893 ± (0.036)	**0.864** ± (0.057)	0.793
FSLR MDS	SVM	0.833 ± (0.027)	0.907 ± (0.021)	0.884 ± (0.068)	0.774 ± (0.066)	0.729
FSLR CNA	SVM	0.881 ± (0.031)	0.967 ± (0.016)	0.919 ± (0.039)	0.834 ± (0.066)	0.839
MDS CNA	SVM	0.881 ± (0.033)	0.960 ± (0.018)	**0.939** ± (0.038)	0.814 ± (0.079)	0.834
FSLR MDS	LR	0.833 ± (0.017)	0.916 ± (0.010)	0.867 ± (0.029)	0.789 ± (0.067)	0.715
FSLR CNA	LR	**0.901** ± (0.041)	**0.972** ± (0.015)	**0.939** ± (0.045)	**0.854** ± (0.049)	**0.844**
MDS CNA	LR	0.881 ± (0.029)	0.962 ± (0.012)	0.927 ± (0.025)	0.824 ± (0.054)	0.824
FSLR MDS CNA	SVM	0.896 ± (0.030)	0.965 ± (0.010)	0.948 ± (0.014)	0.834 ± (0.059)	0.869
FSLR MDS CNA	LR	**0.903** ± (0.030)	**0.969** ± (0.007)	**0.949** ± (0.013)	**0.849** ± (0.062)	**0.889**

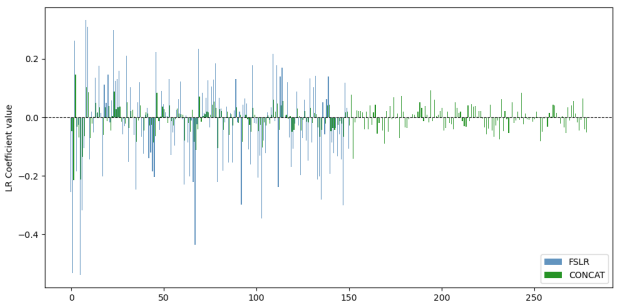


Fig. 3. Comparison of Coefficients of Logistic Regression Classifier for FSLR and PCA concatenation of FSLR and MDS pair (CONCAT).

model is marginal. The clearer gain is again in M.S. at 95% specificity, which reaches its highest value across all baseline configurations (0.889).

Overall, the single-view baselines already achieve strong performance, and multi-view integration yields only incremental improvements in accuracy and AUC. CNA is consistently the strongest individual view, while MDS is consistently the weakest. The most consistent benefit of combining views is seen in M.S. at 95% specificity, which improves progressively from single views to pairs to the full three-view combination.

MOFA+

Having established that simple concatenation yields only incremental gains over single view baselines, we next assess whether MOFA+'s joint latent representation offers a meaningful improvement. The classification performance using MOFA+ sample embeddings summarized in Table 3 demonstrates that the classes are not linearly separable based on Logistic Regression classifier performance.

In the single-view setting, CNA peforms the best. Notably the best accuracy for the single-view MOFA+ setting is comparable to the worst accuracy in single-view baseline (MDS). Variance across folds also increases relative to the baseline, likely due to the test embedding approximation step.

Consistent with the baseline results, the best pairing of fragmentomics features is the FSLR CNA pair. In MOFA+ setting, MDS does not hurt the performance of the other fragmentomics feature, but the performance benefits are too small for a meaningful contribution. Using MOFA+ does not bring performance benefits, on the contrary, the accuracy of FSLR and CNA pair is decreased by 0.137. This performance decrease is likely caused by the reduced sample representation, as MOFA+ drastically prunes the number of factors, leaving 4 factors as opposed to 317 dimensions in PCA-concatenated baseline.

Combining all three views yields the best MOFA+ results, as can be seen in Figure 4, with accuracy 0.819 and AUC 0.900, but this still falls short of the FSLR and CNA single-view

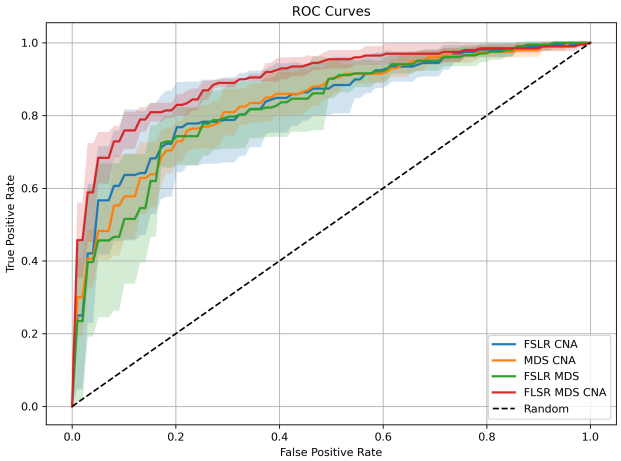


Fig. 4. ROC curves for MOFA+ models using SVM classifiers across all fragmentomic pairwise combinations and the three-view integration, averaged over the five outer folds of nested cross-validation.

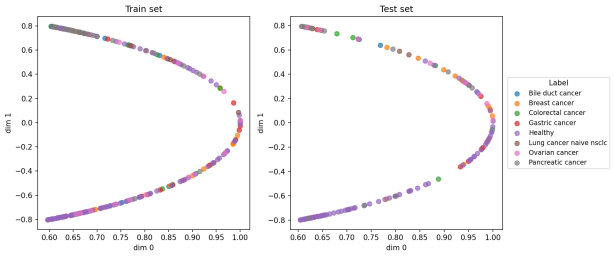


Fig. 5. Two-dimensional unsupervised CMK sample embeddings for outer fold 3, obtained by projecting the unified kernel space onto its two leading principal components (PC1: 41.98% variance, PC2: 18.38% variance), with training and test samples shown separately using FSLR and MDS views.

baselines, and the increased variance suggests instability across folds as the learned factors vary with training composition.

Mean Sensitivity at 95% specificity follows the same pattern but with much larger relative jumps: from 0.311 (best view) to 0.567 (best pair) to 0.684 (all three views, SVM), proportionally far larger increases than seen in the baseline. However, even this MOFA+ configuration falls short of the FSLR baseline with (M.S. 95% = 0.734), let alone the full baseline integration (0.889).

These results suggest that while MOFA+ factors are optimized to capture overall variance across views and groups, this does not necessarily align with the cancer/healthy separation. This is likely caused by the fact that the group-aware structure does not directly enforce separation, and the small number of retained factors may further limit the capacity to represent class-relevant signal. This performance contradicts the previous class separability claims when using MOFA+ for variance interpretation with silhouette scores, suggesting that a variance-driven decomposition might be not well suited as a pre-processing step for this classification task in small dataset settings.

Contrastive Multi-View Kernel Learning

To move beyond variance-based integration, we apply Contrastive Multi-View Kernel Learning (CMK) [28], which learns a unified embedding space across views by encouraging view-specific representations of the same sample to align (contrastive loss), while jointly optimizing for cluster structure (spectral loss) relevant to downstream classification.

In the original work by Liu et al. [28], per-view kernel matrices are computed from the view-specific sample embeddings and averaged into a single combined similarity kernel. The sample representation used for clustering evaluation is then obtained via PCA on this combined kernel, with dimensionality n set to the number of classes, therefore implicitly assuming that the largest sources of variance in the combined kernel correspond to class separation. For binary classification, this results in a 2-dimensional embedding.

Figure 5 shows this 2-dimensional embedding obtained by projecting samples onto the two principal components with the highest eigenvalues of the averaged kernel for FSLR and MDS. The first dimension, which captures the most variance (41.98%), is uninformative for class separation, while the second principal component (18.38%) achieves the best separation. While this figure validates the chosen projection method for the test samples, it ultimately yields only one informative dimension for downstream classification.

Consequently, we chose to utilize the view-specific latent embeddings for classification by concatenating them, resulting in dimensionality of 128. Figure 6 shows 2 distinct clusters in both FSLR and MDS embeddings, as a result of training phase 2 with the spectral loss. While each of the clusters clearly has a dominant class, nevertheless due to the class overlap in each cluster, this representation is unlikely to result in acceptable classification performance. These results lead us to conclude that in the unsupervised setting CMK achieves distinct clusters with a degree of class separation, but is unlikely to improve over the baselines, which is confirmed by Table 4.

Furthermore, given that the dataset comprises of seven distinct cancer types alongside healthy samples, we extended the binary setting by choosing $n = 8$. Restricting the embedding to two dimensions risks discarding cancer-type-specific signal, while increasing n may yield a richer sample representation for downstream classification. However, as shown in Supplementary Figure 8, the additional dimensions do not improve cancer/healthy separation in the unsupervised setting. The spectral loss now optimizes for eight clusters rather than binary separation of the cancer healthy classes. Nevertheless, the 8-class unsupervised setting achieves the best unsupervised CMK performance using FSLR and MDS (M.S. 95% = 0.684, Table 4), performing on par with MOFA+ across all three features. This suggests that the richer embedding captures additional discriminative signal despite the misaligned clustering objective. These results suggest that unsupervised CMK, regardless of the number of classes, is not directly optimized for the binary classification objective. We therefore introduce a supervised variant by augmenting the contrastive loss following Khosla et al. [30], explicitly aligning the learned embedding with the cancer/healthy classification task.

The supervised loss formulation extends the set of positive pairs to include any point from the same class, regardless of the view. By encouraging higher similarity across points from the same class and enforcing 2 clusters, we hypothesize a better class separation in the latent embedding space leading to better downstream classification performance.

This is confirmed by classification performance in Table 4 and Figure 7. While the dimension 0 remains uninformative, there is greater class separation across the dim 1 for both

Table 3. MOFA+ Classification performance (mean ± std) across 5-fold nested cross-validation. MOFA+ was applied to individual views, all pairwise view combinations, and the combination of all three views. Each configuration was evaluated using Logistic Regression (LR) and Support Vector Machine (SVM) classifiers, specified in the Clf column. The Factors column reports the number of latent factors (mode) extracted by MOFA+. M.S. 95% denotes the mean sensitivity at 95% specificity.

Model	Clf	Factors	Accuracy	AUC	Precision	Recall	M.S. 95%
FSLR	SVM	4	0.695 ± (0.032)	0.716 ± (0.091)	0.696 ± (0.041)	**0.694** ± (0.032)	0.251
MDS	SVM	2	0.585 ± (0.048)	0.576 ± (0.056)	0.582 ± (0.047)	0.606 ± (0.061)	0.113
CNA	SVM	5	**0.710** ± (0.047)	**0.780** ± (0.054)	**0.719** ± (0.048)	0.678 ± (0.088)	**0.311**
FSLR	LR	4	0.496 ± (0.009)	0.455 ± (0.068)	0.086 ± (0.192)	0.030 ± (0.067)	0.051
MDS	LR	2	0.501 ± (0.005)	0.470 ± (0.067)	0.000 ± (0.000)	0.000 ± (0.000)	0.048
CNA	LR	5	0.506 ± (0.004)	0.446 ± (0.080)	0.000 ± (0.000)	0.000 ± (0.000)	0.128
FSLR MDS	SVM	4	0.754 ± (0.063)	0.824 ± (0.069)	0.754 ± (0.075)	**0.758** ± (0.058)	0.456
CNA FSLR	SVM	6	**0.764** ± (0.076)	**0.839** ± (0.073)	**0.765** ± (0.082)	**0.758** ± (0.086)	**0.567**
CNA MDS	SVM	5	0.762 ± (0.023)	0.834 ±(0.024)	0.763 ± (0.023)	0.754 ± (0.073)	0.482
FSLR MDS	LR	4	0.519 ± (0.031)	0.473 ± (0.091)	0.223 ± (0.305)	0.165 ± (0.246)	0.121
CNA FSLR	LR	6	0.506 ± (0.004)	0.537 ± (0.079)	0.000 ± (0.000)	0.000 ± (0.000)	0.157
CNA MDS	LR	5	0.506 ± (0.004)	0.524 ± (0.059)	0.000 ± (0.000)	0.000 ± (0.000)	0.167
CNA FSLR MDS	SVM	7	**0.819** ± (0.028)	**0.900** ±(0.022)	**0.827** ± (0.066)	**0.809** ± (0.048)	**0.684**
CNA FSLR MDS	LR	7	0.506 ± (0.004)	0.512 ± (0.056)	0.000 ± (0.000)	0.000 ± (0.000)	0.181

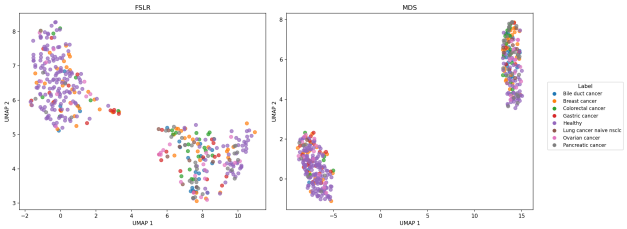


Fig. 6. UMAP visualization of latent embeddings obtained from unsupervised CMK applied to FSLR and MDS views, colored by disease label.

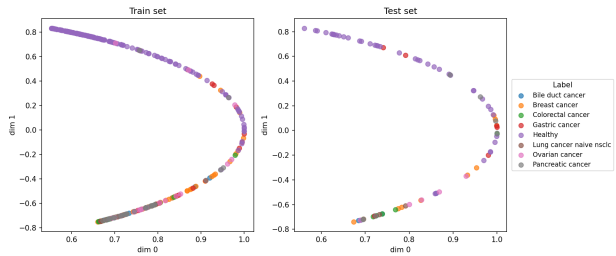


Fig. 7. Two-dimensional supervised CMK sample embeddings for outer fold 3, obtained by projecting the unified kernel space onto its two leading principal components,with training and test samples shown separately using FSLR and MDS views.

training and test set. Notably while in the unsupervised setting increasing the number of classes resulted in performance benefits, this does not hold in the supervised setting. This is likely due to the mismatch of 8 class separation while testing on binary classification.

Having determined the most suitable CMK setting for this task, we now evaluate binary supervised CMK across all fragmentomics feature combinations (Table 5) to assess their complementarity. FSLR CNA remains the best performative pair of fragmentomics features, consistent with MOFA+ and baselines, however in this setting utilizing all three fragmentomics features leads to performance degradation. It is unlikely that this downgrade is caused by the kernel aggregation, as in the baseline all three fragmentomics features are represented equally or that it comes from drastic dimensionality reduction, as the classification was evaluated on the latent 128 dimensional embeddings.

Ultimately, supervised CMK with FSLR and CNA approaches, but does not surpass the concatenation baseline (AUC 0.966 vs 0.972, M.S. 95% 0.845 vs 0.844). For FSLR and MDS, where inter-view correlation is low and potential complementarity is higher, supervised CMK similarly fails to translate this into classification gains, suggesting that the 5 Mb resolution limits the discriminative content of MDS regardless of the integration strategy. The failure of the three-view combination to improve over FSLR and CNA alone is further consistent with MDS contributing noise rather than complementary signal at this resolution.

Taken together, the results reveal a consistent picture across all integration strategies. CNA is the dominant single view, FSLR contributes independent discriminative signal, and MDS consistently underperforms. However, neither MOFA+ nor CMK improve upon simple concatenation (Figure 8), despite being specifically designed to exploit cross-view structure. This suggest that the discriminative signal available in these views at 5 Mb resolution is already sufficient captured by PCA-reduced concatenation, and the additional cross-view alignment objectives do not unlock complementary information beyond what naive combination achieves.

Discussion

This paper sets out to analyze the complementarity of commonly used fragmentomics features - FSLR, MDS and CNA - alongside the intermediate integration with MOFA+ and Contrastive Multi-view Kernel learning, both unsupervised and supervised, for improving cancer classification using cfDNA data obtained through liquid biopsy. While the results confirm that the three features capture distinct biological signals, the intermediate integration methods fail to outperform simple concatenation baseline, raising questions about the suitability of (un)supervised multi-view frameworks

Table 4. Contrastive Multi-View Kernel Learning classification performance (mean ± std) using FSLR and MDS across 5-fold nested cross-validation. Results are reported for both supervised and unsupervised settings, as well as for binary and 8-class classification tasks (PC = 8 denotes the 8-class setting using PCA representations). The Latent setting refers to models trained on concatenated view-specific embeddings.

Representation	Clf	Accuracy	AUC	Precision	Recall	M.S. at 95%
Unsupervised						
PC = 2	LR	0.708 ± (0.042)	0.783 ± (0.035)	0.723 ± (0.067)	0.686 ± (0.081)	0.310
PC = 2	SVM	0.690 ± (0.035)	0.770 ± (0.053)	0.718 ± (0.070)	0.652 ± (0.133)	0.316
Latent	LR	0.774 ± (0.070)	0.845 ± (0.070)	0.787 ± (0.082)	0.758 ± (0.089)	0.556
Latent	SVM	0.754 ± (0.097)	0.839 ± (0.053)	0.755 ± (0.106)	0.778 ± (0.088)	0.502
PC = 8	LR	0.789 ± (0.072)	0.867 ± (0.034)	0.809 ± (0.088)	0.764 ± (0.117)	0.610
PC = 8	SVM	0.786 ± (0.060)	0.856 ± (0.034)	0.795 ± (0.078)	0.784 ± (0.099)	0.605
Latent	LR	0.791 ± (0.064)	**0.887** ± (0.047)	0.815 ± (0.089)	0.764 ± (0.109)	0.660
Latent	SVM	**0.821** ± 0.0640	0.885 ± (0.031)	**0.837** ± (0.076)	**0.799** ± (0.086)	**0.684**
Supervised						
PC = 2	LR	0.833 ± (0.032)	**0.919** ± (0.020)	0.827 ± (0.064)	0.853 ± (0.071)	**0.704**
PC = 2	SVM	0.816 ± (0.046)	0.879 ± (0.042)	0.805 ± (0.060)	0.842 ± (0.096)	0.526
Latent	LR	0.828 ± (0.032)	0.917 ± (0.022)	0.819 ± (0.047)	0.848 ± (0.078)	0.695
Latent	SVM	**0.843** ± (0.040)	**0.919** ± (0.027)	**0.841** ± (0.072)	**0.857** ± (0.068)	0.685
PC = 8	LR	0.754 ± (0.051)	0.851 ± (0.043)	0.760 ± (0.039)	0.750 ± (0.147)	0.582
PC = 8	SVM	0.740 ± (0.054)	0.823 ± (0.042)	0.754 ± (0.052)	0.726 ± (0.158)	0.503
Latent	LR	0.764 ± (0.057)	0.872 ± (0.031)	0.781 ± (0.064)	0.745 ± (0.126)	0.592
Latent	SVM	0.754 ± (0.034)	0.860 ± (0.031)	0.773 ± (0.046)	0.730 ± (0.121)	0.562

Table 5. Supervised Contrastive Multi-View Kernel Learning classification performance (mean ± std) for the two-class setting across all fragmentomic feature pairs and the three-view integration.

Representation	Clf	Accuracy	AUC	Precision	Recall	M.S. at 95%
FSLR CNA						
PC = 2	LR	**0.906** ± (0.040)	0.962 ± (0.023)	**0.922** ± (0.039)	0.884 ± (0.061)	0.799
PC = 2	SVM	**0.906** ± (0.037)	0.960 ± (0.024)	0.909 ± (0.042)	0.900 ± (0.047)	0.799
Latent	LR	0.896 ± (0.024)	0.963 ± (0.026)	0.883 ± (0.007)	**0.910** ± (0.058)	0.804
Latent	SVM	0.896 ± (0.034)	**0.966** ± (0.025)	0.904 ± (0.039)	0.884 ± (0.068)	**0.845**
MDS CNA						
PC = 2	LR	0.848 ± (0.066)	0.912 ± (0.055)	0.836 ± (0.067)	0.864 ± (0.072)	0.628
PC = 2	SVM	0.848 ± (0.060)	0.873 ± (0.088)	0.832 ± (0.071)	0.875 ± (0.070)	0.470
Latent	LR	**0.866** ± (0.054)	0.931 ± (0.031)	0.843 ± (0.065)	**0.900** ± (0.056)	0.663
Latent	SVM	0.861 ± (0.046)	**0.930** ± (0.025)	**0.849** ± (0.042)	0.874 ± (0.069)	**0.698**
FSLR MDS CNA						
Latent	LR	0.888 ± (0.047)	**0.960** ± (0.024)	0.895 ± (0.061)	**0.880** ± (0.069)	**0.774**
Latent	SVM	**0.901** ± (0.054)	0.955 ± (0.032)	**0.918** ±(0.066)	**0.880** ± (0.091)	0.764

for supervised classification tasks using WGS liquid biopsy data.

The cross-view outlier analysis provided strong evidence that FSLR, MDS, and CNA capture partially non-overlapping aspects of tumour-derived cfDNA signal. Detecting cancer outliers using the PON statistics for each fragmentomic feature and comparing their overlap provides a quantifiable measure of complementarity. However, liquid biopsy samples contain a substantial proportion of fragments originating from healthy cells, which may prevent cancer samples with early-stage tumours from being identified as outliers. Consequently, the number of detected outliers can be small, potentially inflating overlap-based measures such as the Jaccard index.

Importantly, the observed biological complementarity does not necessarily imply improved prediction performance. While the outlier and correlation analyses demonstrated that FSLR, MDS and CNA capture distinct aspects of tumour-derived cfDNA biology, not all unique variation contributed to cancer/healthy discrimination. Consequently, integration methods may successfully preserve complementary information while failing to improve classification performance. This distinction is particularly relevant in liquid biopsy settings, where a substantial proportion of the measured signal originates from healthy cells and inter-individual variation. As a result, biologically meaningful sources of variation may dominate the learned representation despite being weakly related to the downstream classification objective.

Having presented numerous performance metrics for evaluating downstream classification performance, it is crucial to establish that mean sensitivity at 95% specificity (M.S. 95%) is the most clinically relevant for cancer classification in this

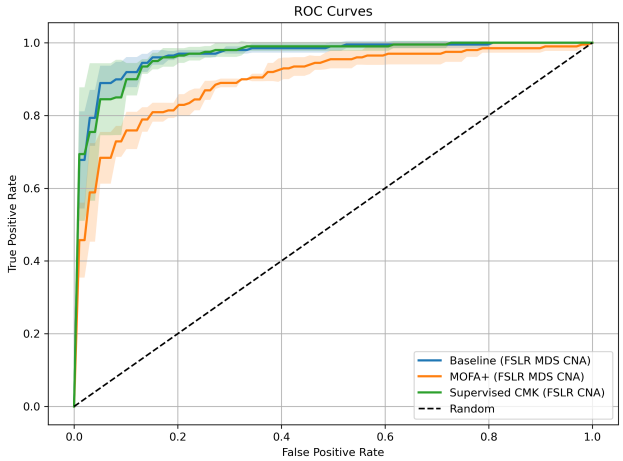


Fig. 8. ROC curves for the best-performing classifier per integration strategy, averaged across five outer folds of nested cross-validation (shaded regions indicate ±1 standard deviation). The baseline uses PCA-reduced concatenation of all three views with LR; MOFA+ uses SVM on seven-factor embeddings of all three views; supervised CMK uses SVM on latent embeddings of FSLR and CNA.

context. This follows from the intended application of cfDNA-based classifiers: population-level cancer screening. Because cancer prevalence in a general screening population is low, even a modest reduction in specificity translates into a large number of false positives [31]. For the patient, a false positive triggers further diagnostic workup, often including invasive procedures, and carries psychological burden disproportionate to the eventual negative outcome. For the healthcare system, the resulting work is costly and adds burden to the diagnostic capacity, particularly as multi-cancer early detection tests move toward broader clinical adoption [32]. A classifier with a high false positive rate is therefore unusable in screening practice. Evaluating classifiers at a fixed high-specificity operating point is consistent with how this trade-off is handled in the cfDNA literature [16, 19, 32], and we focus on M.S. 95% accordingly when interpreting the suitability of the intermediate integration methods.

CNA was the most discriminative single feature (M.S. 95% = 0.824), consistent with its established role as one of the most robust genome-wide cancer biomarkers detectable in cfDNA [14]. Unlike fragmentation-based features, CNA directly reflects large-scale genomic instability, a hallmark of many cancers, making it less susceptible to small-scale local variation in chromatin structure or sequence composition. FSLR performed competitively (M.S. 95% = 0.734), suggesting that fragment length distributions contain substantial cancer-associated information independent of copy number state. The improved performance of the FSLR CNA combination (M.S. 95% = 0.844) indicates that these features capture partially distinct aspects of tumour biology. While CNA reflects genomic alterations, FSLR is thought to capture changes in nucleosome positioning, providing complementary biological information.

In contrast, MDS consistently underperformed compared to both CNA and FSLR, and contributed little additional predictive value when combined with other features, contradicting the results of Hou et al. [18]. We hypothesize that the 5 Mb resolution used in this study is too coarse to fully exploit sequence motif diversity patterns. Averaging motif frequencies over large genomic regions likely compresses local variability and weakens cancer-specific signal. Consequently, the limited contribution of MDS observed across all integration strategies may reflect the chosen resolution granularity rather than an inherent lack of biological relevance.

MOFA+ consistently failed to improve upon single feature baselines with M.S. 95% at 0.684, which is drastically lower than previous works for cancer detection in clinical applications [33]. We hypothesize that the decrease in performance comes from no clear group separation objective in the method as well as much lower dimensionality. By integrating multiple fragmentomics features, the number of dimensions increases, adding more signal for downstream classification. Based on MOFA+ factor vizualizations, multiple of these dimensions were not discriminative at large for cancer/healthy separation, but only model the shared variance across both groups. Furthermore, the test embedding approximation via Linear Regression introduced cross-fold instability during classifier training, as the assumed linear mapping from raw features to MOFA+ latent space varies with training set composition, contributing to the elevated variance observed across folds. Because cfDNA in plasma is dominated by DNA originating from healthy tissues, especially in low tumour fraction samples, the dominant sources of variance may reflect individual variability rather than disease status. Therefore poor classification performance should not necessarily be interpreted as a failure of the method itself, but rather as evidence that variance-explaining factors are not equivalent to discriminative features in cfDNA cancer detection. These findings are also consistent with the original design objectives of MOFA+, which was developed primarily as an integrative analysis framework for inferring a low-dimensional representation through latent factors that capture variance across multiple data modalities rather than maximizing predictive performance [26]. Therefore while MOFA+ remains a valuable framework for variance interpretation to improve the understanding of relationships between fragmentomics features and their complementarity, its use for downstream classification is not recommended based on provided results.

Unsupervised CMK using FSLR and MDS views yielded only one informative projection dimension following the Liu et al. [28] experiment setting. The primary variance axis was uninformative for class separation, reflecting the healthy cell signal dominating over the healthy-cancer boundary in liquid biopsy data. Supervised contrastive loss substantially improved the performance in the setting where CMK assumption of class separation by variance in aggregated similarity matrix does not hold. By extending positive pairs to include all samples sharing the same class label, supervised CMK encouraged stronger class clustering in the shared latent space, directly addressing the weak cancer signal challenge at low tumour fractions. This made a notable improvement to M.S. 95% 0.704, which ultimately failed to improve over the standalone FSLR baseline. When CNA was incorporated, supervised CMK with FSLR and CNA achieved M.S. 95% of 0.845, approaching the best concatenation baseline of the pair (0.889). However, differing from the baseline and MOFA+ results, the three-view combination did not improve further. With the low MDS contributions in previous experiments, we hypothesize that the equal weighing of kernels during aggregation leads to lower performance in CMK when MDS is added.

More broadly, the limited benefit of both MOFA+ and CMK may stem from the nature of the views themselves. In many successful multi-omics applications [19, 20], the integrated modalities originate from fundamentally different

biological measurement technologies, such as gene expression or methylation. In contrast, FSLR, MDS and CNA are all derived from the same underlying whole-genome sequencing experiment. Consequently, the opportunity for sophisticated integration methods to uncover additional predictive signal may be inherently smaller than in traditional multi-omics settings.

One of key limitations of this work was the Panel of Normals sample composition. While all the samples were healthy, the characteristics such as age, gender, genetic predisposition or lifestyle affect their cfDNA profiles. Therefore, any statistics calculated for outlier detection or normalization are prone to change depending on the PON group composition, affecting the generalizability of results in downstream classification. Another key limitation was the cancer group composition. While the Cristiano et al. cohort is balanced for cancer/healthy split, the cancer types are not represented equally. This skewed cancer sample distribution directly affects variance-based analysis such as MOFA+ factors or class aggregates per genomic position, with more represented cancer type dominating the analysis, limiting the conclusions about specific cancer types.

A methodological limitation of this study is that only a single representative framework was evaluated for each intermediate integration paradigm. While MOFA+ and CMK are conceptually distinct approaches, the lack of performance improvements should not be interpreted as evidence that all intermediate integration methods are unsuitable for fragmentomics data. Alternative approaches, particularly methods specifically designed for supervised multi-view classification such as MVMED[24], may yield different results and warrant future investigation.

Consistent with other works in this field, the small dataset size remains a bottleneck for generalizable approaches for reliable cancer classification using liquid biopsy data. There has been an increased attention to predicting the missing modalities, such as FinaleME [34] for methylation, in an effort to increase informative signal at a lower cost. Certain methods such as bisulfite sequencing destroy the collected sample, making it infeasible to obtain a multi-modal view. However, the diversity of sequencing technologies during sample collection as well as pre-processing protocols make a large scale dataset integration challenging and remain one of key directions to be explored in future work.

Future research should explore how genomic resolution influences classification performance when combining complementary fragmentomic features through intermediate integration approaches. As previously discussed, the limited performance of MDS may be partly attributed to the 5 Mb resolution, which could obscure discriminative patterns captured by motif diversity. Furthermore, the optimal resolution may differ across fragmentomic features, reflecting the distinct biological signals they encode. Exploring feature-specific resolutions within an integration framework could therefore provide valuable insight into how these signals can be most effectively combined for cancer classification.

Overall, the results present a consistent picture across all evaluated integration strategies. FSLR, MDS and CNA capture complementary aspects of cfDNA biology, yet this biological complementarity does not translate directly into improved classification performance through intermediate integration. The strongest performance was achieved by simple PCA-based concatenation of FSLR, MDS and CNA, while MOFA+ underperformed and CMK only approached this performance after introducing supervised objectives. These findings suggest that the primary challenge in fragmentomics integration may not be the absence of complementary information, but rather the difficulty of transforming that information into clinically relevant discriminative signal. Future work should therefore focus not only on increasingly sophisticated integration architectures, but also on developing feature representations that better preserve cancer-specific information while suppressing the substantial healthy background signal present in liquid biopsy data.

Methods

Fragmentomics Features

The bam files of sample cohort from study by Cristiano et al. [16] were processed using the code available at comp-cfdna[1]. The cohort consisted of 204 healthy samples, 203 cancer samples and 50 healthy samples serving as a Panel of Normals (PON). The cancer group spanned seven cancer types, including breast (n = 54), pancreatic (n = 35), ovarian (n = 28), gastric (n = 27), colorectal (n = 25), bile duct (n = 24), and lung cancer (n = 10). All cancer samples were preoperative treatment naive at the time of blood sample collection. The features were calculated by aligning the reads from samples with bwa-mem library to hg38 reference at 5Mb bin resolution with GC-correction. Parts of the human genome were filtered out for their low quality and field standards such as the blacklisted Duke regions [35]. The bed file containing the included bases is based on a previous work by Mathios et al. [36]. Furthermore, reads were excluded if they failed to meet a mapping quality threshold of 30 or fell outside the acceptable fragment size range of 100–220 bp.

Fragment Short Long Ratio

Fragment Short Long Ratio (FSLR) at a bin position is calculated as a log2 transformed ratio of short to long reads. Namely, given a bin i, the FSLR value is calculated as:

$$\mathrm{FSLR}_i = \log_2\left(\frac{S_i}{L_i}\right)$$

where S_i stands for the number of short reads (100 to 150 base pairs) and L_i stands for the number of long reads (151 to 220 base pair).

Motif Diversity Score

Another fragmentomics feature utilized in this paper is the Motif Diversity Score (MDS). For a specific bin i, the MDS is calculated as:

$$\mathrm{MDS_i} = \frac{\sum_k -P_k \log(P_k)}{\log(N)}$$

where P_k is the frequency of motif k at bin i, and N is the total number of possible motifs, in this case 64 as the considered motifs are 3-mers . The motifs used for the MDS calculation are the first 3 bases on a forward strand for a read aligned to bin i.

Copy Number Alteration

Copy number alterations were estimated from cfDNA sequencing data using ichorCNA library[2], a tool designed for ultra-low-pass whole-genome sequencing (ULP-WGS). ichorCNA models the genome as a mixture of tumour and normal cells, jointly estimating tumour fraction and copy number state per genomic bin via a hidden Markov model (HMM).

For each sample, read depth was computed in fixed-size genomic bins at 1Mb and corrected for GC content of the sample. The corrected per-bin $\log_2$ copy ratio was extracted from ichorCNA's `log2_TNratio_corrected` output field, defined as:

$$\log R_b = \log_2\left(\frac{d_b^{\mathrm{tumor}}}{d_b^{\mathrm{ref}}}\right) \quad (1)$$

where b indexes each genomic bin, d_b^{tumor} is the GC-corrected read depth in the tumour sample, and d_b^{ref} is the expected depth under a diploid normal reference. The 1 Mb bin values were subsequently aggregated by averaging to match the 5 Mb resolution used for FSLR and MDS. Bins with low mappability or falling within known artefact-prone regions [35] were excluded prior to downstream analysis, similarly to FSLR and MDS. The resulting genome-wide $\log R$ profile served as the CNA feature vector for each sample.

Data Transformations

To improve comparability across samples and enhance separation between the cancer and healthy classes, the following pre-processing transformations were applied.

Sample Normalization

For each sample, bin values were normalized to a standard normal distribution across the genome, following Cristiano et al. [16]. Each fragmentomics feature was standardized bin-wise using the sample mean and variance computed across all genomic bins.

Z-scoring with Panel of Normals

To further differentiate between the healthy and cancer samples, we separated 50 healthy samples to serve as a panel of normals (PON). These samples were used only for the Z-scoring and did not take part in any model training or evaluation. As previously described, we first calculated fragmentomics feature mean along with variance and applied sample normalization. We then computed the mean μ and variance σ in PON group per bin to obtain the PON statistics. These per bin statistics (μ and σ), calculated separately for FSLR, MDS and CNA view, were then used to transform the normalized bin values for all samples in order to quantify their deviation from PON samples.

Baseline

For each view, Principal Component Analysis (PCA) [37] was applied retaining components that cumulatively explained 90% of the variance, yielding a lower-dimensional sample representation. In the multi-view setting, PCA was applied independently to each view and the resulting representations were concatenated prior to classification. In all cases, PCA was fitted exclusively on the outer training fold and applied to the test fold to prevent data leakage.

MOFA+

Multi-Omics Factor Analysis v2 (MOFA+) is a statistical framework originally designed by Argelaguet et al.[26] in order to integrate single cell multi-modal data. Similarly to principal components in PCA [37], it recovers lower dimensional representation of the observed data through extracting latent *factors*, which depict the sources of variation in the data. Specifically, given a series of observation matrices Y_{gm}, where m determines the modality (or view) and g determines the group, MOFA+ decomposes the matrices as:

$$Y_{gm} = Z_g W_m^T + \epsilon_g m \quad (2)$$

[1] https://github.com/lnavarcikova/comp-cfdna

[2] https://github.com/broadinstitute/ichorCNA/tree/master

where W_m is the weight matrix of modality m and Z_g contains the lower dimensional representations of group g expressed through recovered factors [26].

Although the sample cohort contains 7 distinct cancer types, in this paper we focus on binary classification between healthy and cancer, leading to application of MOFA+ on the combination of fragmentomics views (FSLR, MDS and CNA) with 2 groups (healthy and cancer). MOFA+ was applied in fast converge mode allowing for up to 10 factors to be found with Gaussian likelihoods with required explained factor variance (R2) at 0.001. The downstream analysis of MOFA+ results was done using the mofax package [3]. The MOFA+ implementation used in this paper can be found on GitHub[4].

Contrastive Multi-View Kernel Learning

Inspired by the work of Liu et al. [28], we apply a two-phase training paradigm: phase one enforces the consensus principle by aligning the views in a shared latent space, and phase two enforces sample separation through spectral clustering. In phase one, the fragmentomics views are each mapped into the shared latent space through a separately learned linear projection, after which a Gaussian kernel is applied per view to define the contrastive loss quantifying inter-view agreement. The following equations directly reproduce the formulation of Liu et al. [28] and are included for clarity.

The contrastive loss of a ith data sample in view v for the multi-view setting becomes:

$$\ell_{i,v} = \frac{1}{V-1} \sum_{\substack{v'=1 \\ v' \neq v}}^{V} -\log \frac{\exp\left(k_z\left(z_i^v,\, z_i^{v'}\right)\right)}{\sum_{(j,v'')\in\mathcal{A}_{i,v}} \exp\left(k_z\left(z_i^v,\, z_j^{v''}\right)\right)} \tag{3}$$

where k_z is the Gaussian kernel function and z_i^v is the L2 normalized embedding of data point x_i after the linear projection in the shared view space. The similarity of the positive pair $(k_z(z_i^v, z_i^{v'}))$ is normalized by the positive and negative pair similarities, whose combinations are denoted by $A_{i,v}$ such that

$$A_{i,v} = \{1, 2, ..., N\}\{1, 2, ..., V\} \setminus (i, v) \tag{4}$$

with N being the number of samples in the dataset and V denoting the number of views. The overall contrastive loss l_c is then determined by the summation of all possible $l_{i,v}$ divided by NV as shown:

$$l_c = \frac{1}{NV} \sum_{i,v=1}^{N,V} l_{i,v} \tag{5}$$

The only trainable parameters in this approach are the weights of the linear projection, optimized through stochastic gradient descent (SDG) based on the contrastive loss l_c.

Deviating from Liu et al. [28], stopping criterion is defined by number of epochs, namely 20% of the training set is held out as a validation set for early stopping to prevent overfitting. Training terminates when the validation loss fails to decrease by at least 1% over 20 consecutive epochs, at which point the model weights from 20 epochs prior are restored before proceeding to the next phase with same termination conditions.

Upon completion of phase 1, phase 2 introduces the spectral loss, which quantifies the separability of the clusters. Based on equation (36) from Liu et al. [28], the spectral loss is defined as the Multiple Kernel k-means objective over the averaged combined kernel:

$$\ell_s = -\frac{1}{N} \sum_{v=1}^{V} \beta_v \operatorname{Tr}\left[\mathbf{K}_v^c \left(\mathbf{I}_N - \mathbf{F}\mathbf{F}^\top\right)\right] \tag{6}$$

where $\mathbf{K}_v^c$ is the kernel matrix for view v computed from the projected embeddings, $\mathbf{F} \in \mathbb{R}^{N\times k}$ is the soft cluster assignment matrix with k equal to the number of classes, $\mathbf{I}_N$ is the identity matrix, and $\beta_v = 1/V$ weights each view equally. The soft label $\mathbf{F}$ is obtained as the horizontal concatenation of the k eigenvectors corresponding to the k largest eigenvalues of $\sum_{v=1}^{V} \mathbf{K}_v^c$. The overall training objective in phase two combines the contrastive and spectral losses with equal weighting:

$$\ell = \ell_c + \lambda \ell_s, \quad \lambda = 1 \tag{7}$$

Since the kernel PCA decomposition is computed on the training set only, test samples must be projected into the learned eigenspace for downstream classification. Unlike Liu et al. [28], who do not perform the train/test split when evaluating performance, the classification setting requires an explicit out-of-sample embedding.

To project test samples into the eigenspace learned on the training set, we follow the Nyström out-of-sample extension proposed by Bengio et al. [38]. Given the eigenvectors $\mathbf{U}_k$ and eigenvalues $\boldsymbol{\lambda}_k$ of the averaged training kernel matrix, the test embedding is obtained as:

$$\mathbf{Z}_{\text{test}} = \frac{\mathbf{K}_{\text{cross}}\mathbf{U}_k}{\boldsymbol{\lambda}_k} \tag{8}$$

where $\mathbf{K}_{\text{cross}} \in \mathbb{R}^{N_{\text{test}}\times N_{\text{train}}}$ is the averaged Gaussian kernel matrix computed between test and training samples.

Supervised Contrastive Multi-View Kernel Learning

To further leverage the available class label information, we extend the contrastive loss with the supervised contrastive objective proposed by Khosla et al.[30]. Rather than defining positive pairs solely based on view correspondence (i.e., FSLR and MDS from the same patient), we additionally treat all samples sharing the same class label as positives. The supervised contrastive loss, in the formulation of Khosle et al. [30], for anchor i in view v then becomes:

$$\mathcal{L}_{i,v}^{\text{sup}} = \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp\, k_z(z_i^v,\, z_p^{v'})}{\sum_{(j,\, v'')\in A_{i,v}} \exp\, k_z(z_i^v,\, z_j^{v''})} \tag{9}$$

where $P(i)$ is the set of all samples in the batch sharing the same label as anchor i, and $A_{i,v}$ retains the same definition as in Eq. 4. The overall supervised contrastive loss is then computed analogously to Eq. 5. All other components of the two-phase training pipeline, including the linear projection, Gaussian kernel, and spectral loss in Phase 2 remain unchanged.

[3] https://github.com/bioFAM/mofax

[4] https://github.com/bioFAM/mofapy2

Downstream Classification

In order to evaluate generalization performance as well as determine the best classifier hyperparameters on a small-scale dataset, we employ nested cross validation. In the following section all the performed folds are stratified based on healthy and a specific cancer-type labels for a fair split - therefore even though we treat the classification task as binary, we still keep the cancer-subtypes balanced across the splits. The randomness seed is fixed, ensuring identical dataset splits across the whole evaluation.

We perform 5 outer folds and inside each outer fold we run the hyperparameter grid. Per hyperparameter combination, we perform 5 inner splits to reliably gauge the effect of classifier hyperparameters on the performance, quantified in accuracy and area under the ROC curve (AUC). The best configuration is determined by mean accuracy on the inner fold test set and is then used for the evaluation of the outer fold performance.

The dataset X was first split into X_{train} and X_{test} per each outer fold. X_{train} is then passed to the inner fold, splitting into X_{train_inner} and X_{test_inner}. In the case of CMK, 20% of the X_{train_inner} was used as validation set for early stopping. We evaluate classifier hyperparameter combination performance on X_{test_inner} and return the best configuration on hyper parameters per outer fold.

For the SVM classifier, the hyperparameter grid consisted of the `rbf`, `polynomial` and `linear` kernel with the regularization parameter C in the list of [0.001, 0.01, 0.1, 1, 10,100]. For the rbf and polynomial kernels, the SVM classifier *gamma* parameter was one of ['scale', 'auto', 0.01, 0.1, 1].

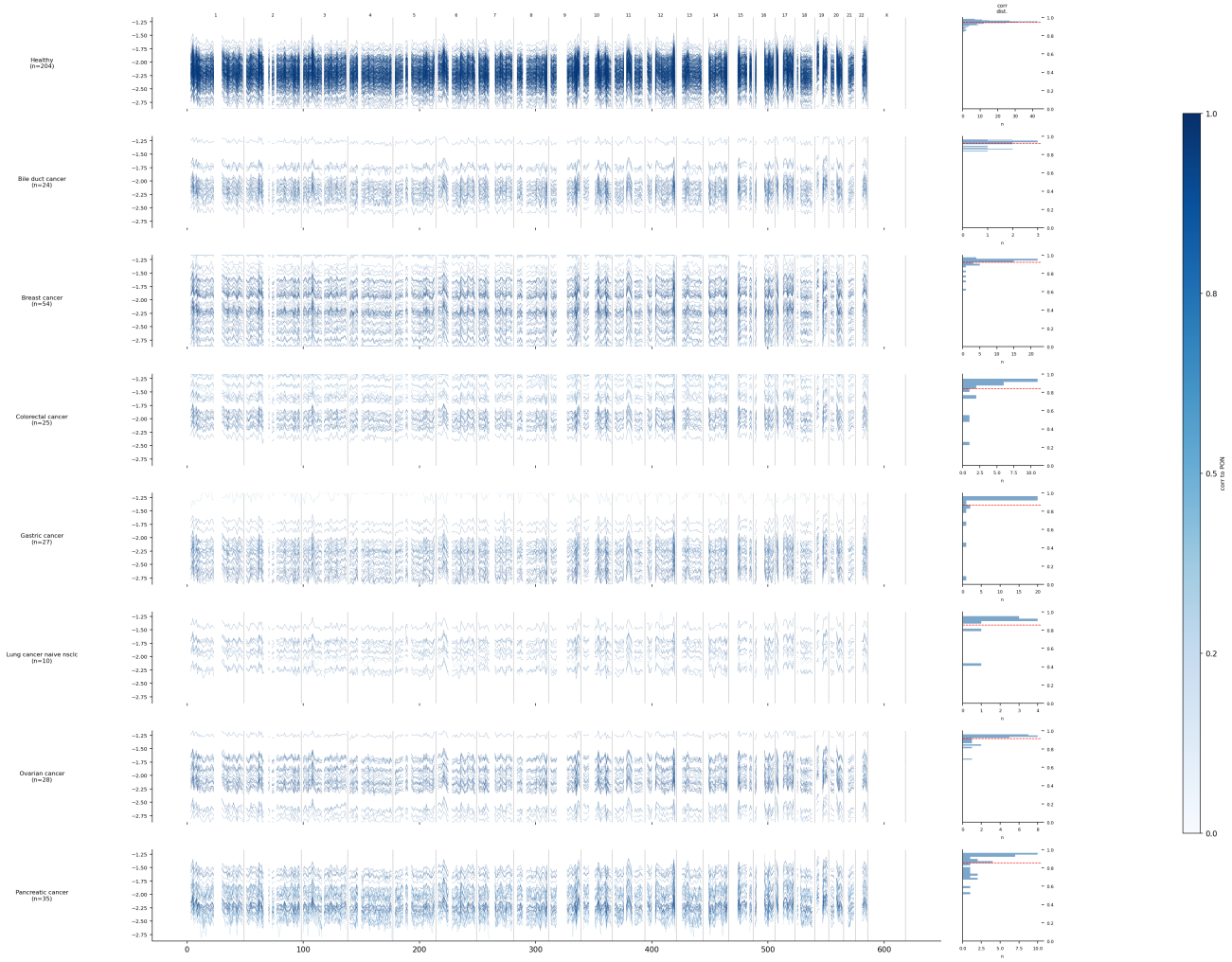
For the LogisticRegression classifier, the regularization parameter C was in the same range as for the SVM, with a maximum number of iterations set to 10000. The hyperparameter grid also included the different solvers, namely 'lbfgs','liblinear' and 'saga' for both L1 and L2 regularization.

Due to high dimensionality of fragmentomics views (477) compared to the relatively small dataset (407 samples), we applied PCA with variance explained threshold at 90% for the baseline, consistent with field standards as seen in [36], [39]. When using multiple fragmentomics features, PCA was first applied independently and the resulting sample representations were concatenated. By applying PCA independently per view, we avoid variance from one fragmentomics feature dominating the sample embedding.

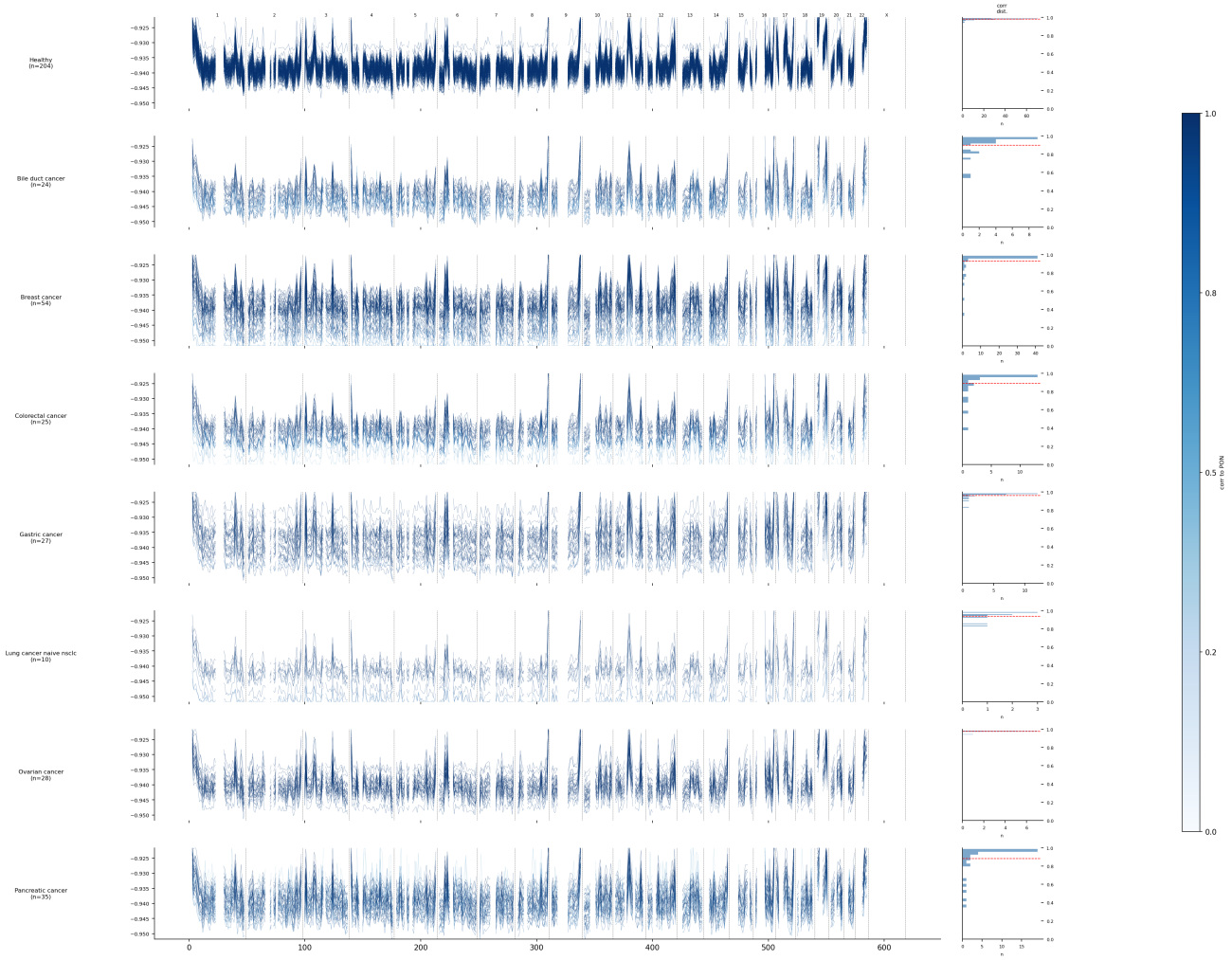
For MOFA+, we normalize the MOFA+ embeddings with StandardScaler in order to mitigate range differences between factors. In order to avoid data leakage, MOFA+ is trained independently on each of 5 folds in nested cross validation. As MOFA+ optimizes sample representations Z during training, test samples are not directly provided by the method. Consistent with MOFA+ evaluation by Makrodimitris et al.[40], a Linear Regression model was fitted to map concatenated feature vectors,formed by horizontally concatenating FSLR, MDS, and CNA measurements, to the learned factor matrix Z, and subsequently applied to project test samples into the same latent space.

CMK was likewise trained independently on each outer training fold to prevent data leakage. Two output representations were evaluated for downstream classification: the concatenated per-view latent embeddings produced by the linear projection of the views, and the PCA-reduced representation of the summed kernel matrix, consistent with the original formulation of Liu et al. [28]. CMK was trained using SGD with momentum 0.9 and an initial learning rate of 1.0 with cosine annealing decay. A Gaussian kernel with bandwidth $t = 1.0$ was used throughout, and the projection heads mapped each view to a latent dimension of 64 with ℓ_2 normalization applied. Training followed a two-phase schedule: the first phase trained on contrastive loss alone, after which the spectral clustering objective was introduced with a trade-off weight of $\lambda = 1.0$. Early stopping was applied independently in each phase, triggering when validation loss failed to improve by at least 1% over 20 consecutive epochs. An internal validation split of 20% of the outer training fold was used for early stopping.

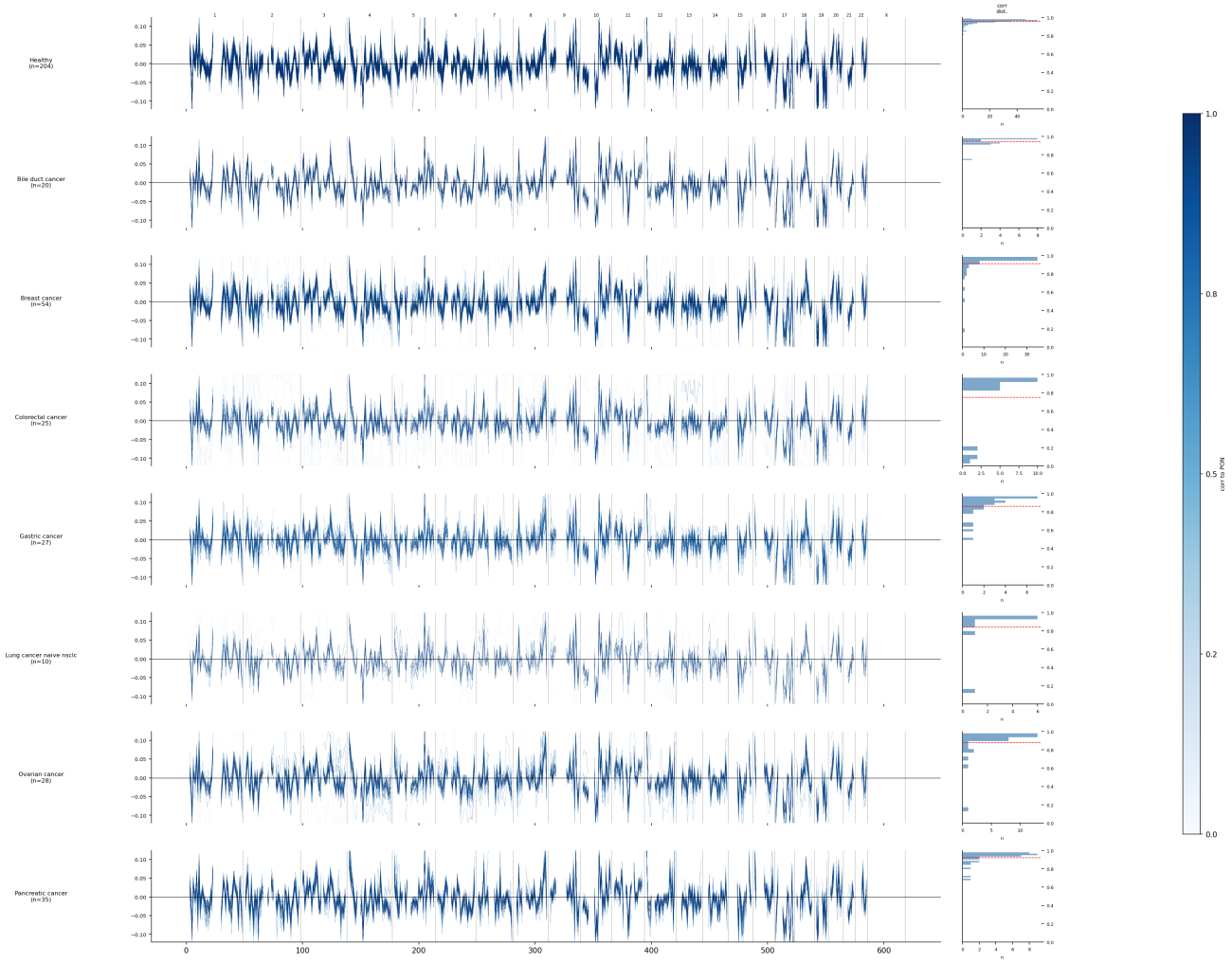
Supplementary Information



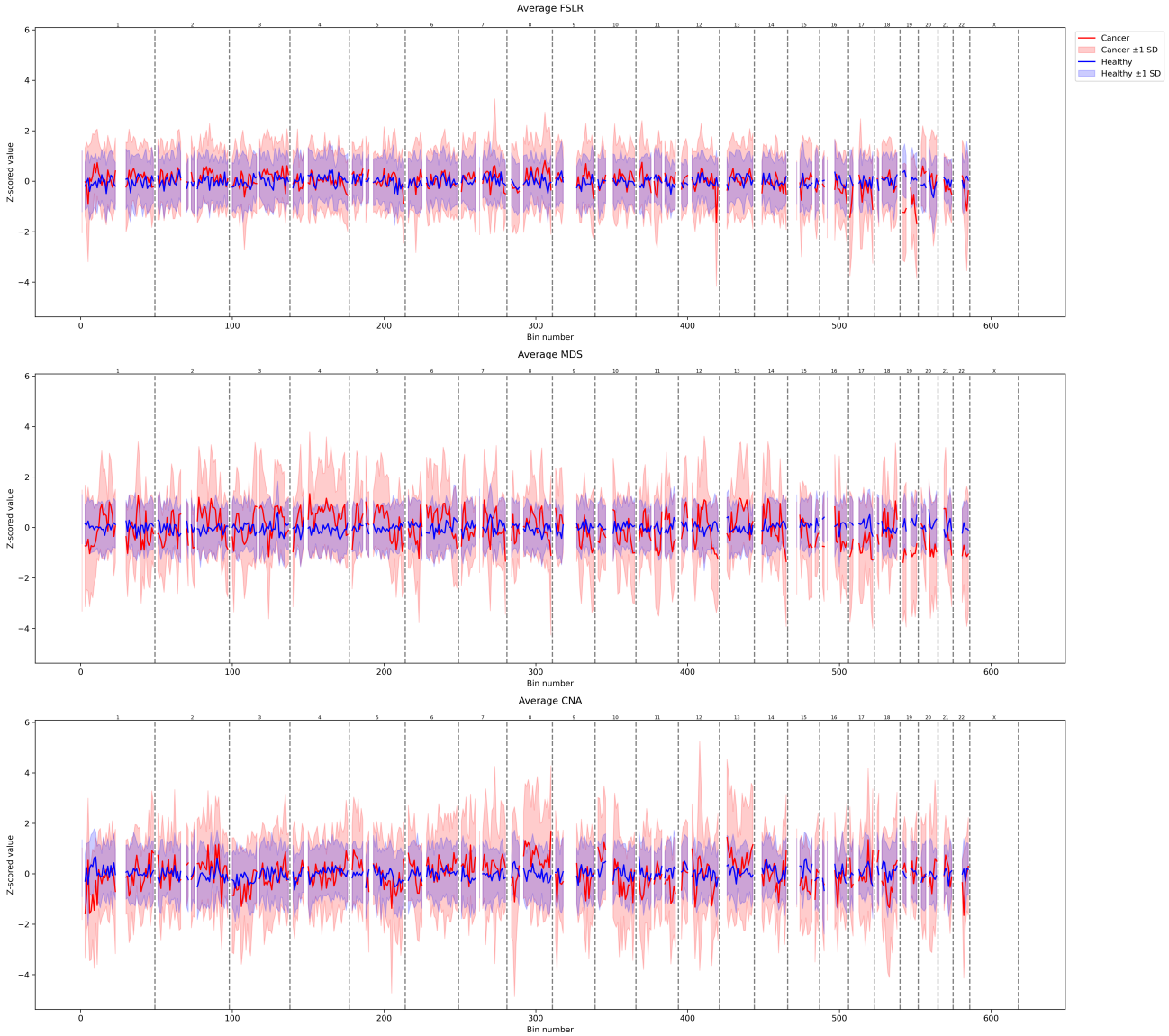
Supplementary Figure 1. Sample FSLR colored by genome-wide correlation to the mean PON FSLR value. The coloring is based on absolute correlation value as indicated by the scale between 0 and 1. Histograms on the right side show distribution of correlation values with red dashed line indicating the mean correlation per group. The raw FSLR values are clipped to 99th percentile across all samples.



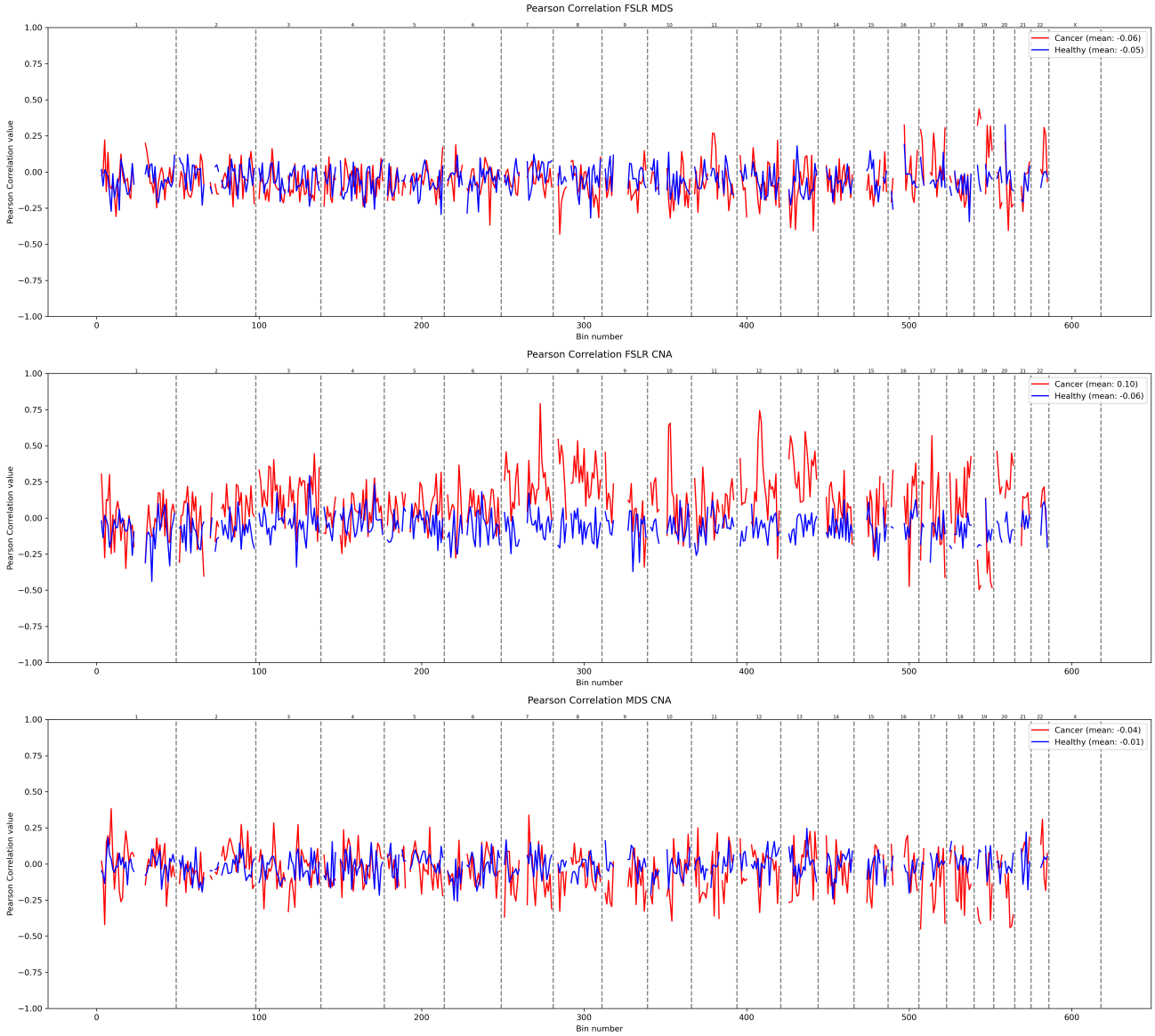
Supplementary Figure 2. Sample MDS colored by genome-wide correlation to the mean PON MDS value. The coloring is based on absolute correlation value as indicated by the scale between 0 and 1. Histograms on the right side show distribution of correlation values with red dashed line indicating the mean correlation per group. The raw FSLR values are clipped to 99th percentile across all samples.



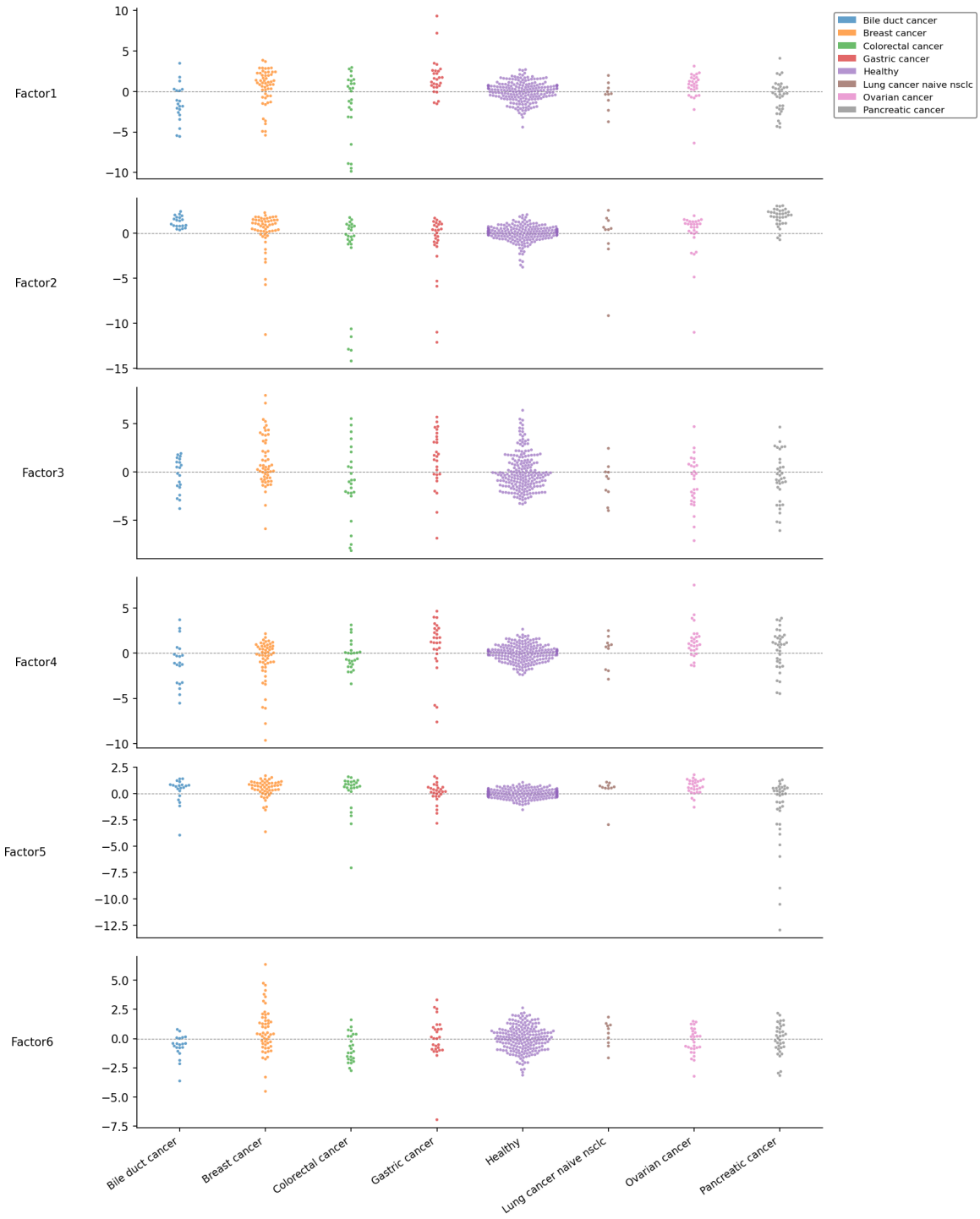
Supplementary Figure 3. Sample CNA from ichorCNA colored by genome-wide correlation to the mean PON CNA value. The coloring is based on absolute correlation value as indicated by the scale between 0 and 1. Histograms on the right side show distribution of correlation values with red dashed line indicating the mean correlation per group. The raw CNA values are clipped to 99th percentile across all samples



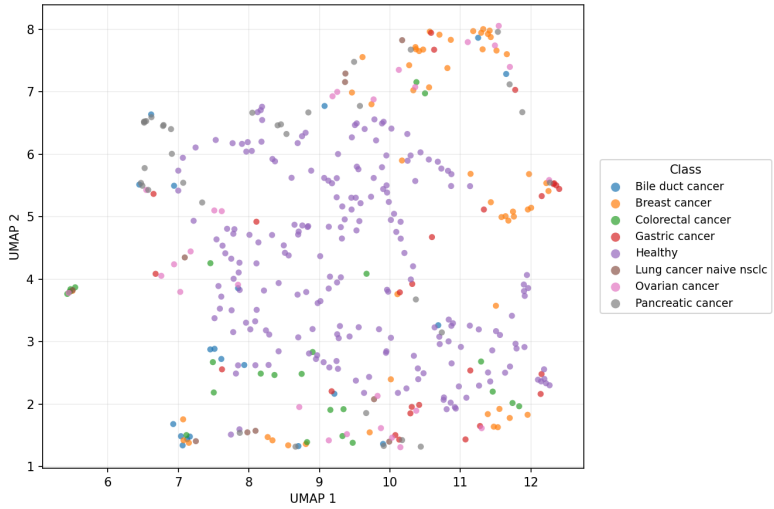
Supplementary Figure 4. Class-aggregated genomic profiles of FSLR, MDS, and CNA across 5 Mb bins, showing the mean signal per bin for healthy controls and each cancer type, with shaded regions indicating one standard deviation. The vertical dashed lines indicate chromozome boundaries, with the chromozome number depicted at the top of the graph. The gaps correspond to filtered bin regions.



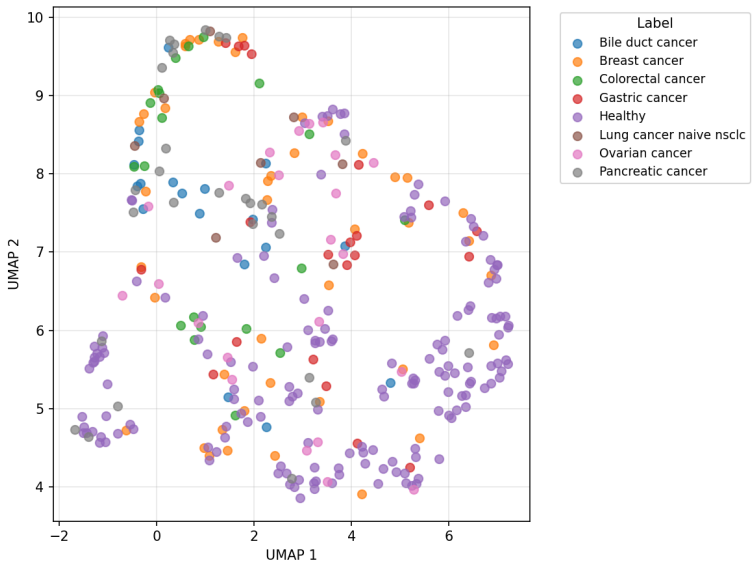
Supplementary Figure 5. Per-bin Pearson correlation between view pairs (FSLR MDS, FSLR CNA, MDS CNA) computed across all samples, shown separately for cancer (red) and healthy (blue) classes at 5 Mb bin resolution with Z-scoring against the PON. Dashed vertical lines indicate chromosome boundaries.



Supplementary Figure 6. MOFA+ sample representations across all extracted factors for healthy and cancer subtypes, extracted using FSLR, MDS, and CNA fragmentomics views at 5 Mb bin resolution with Z-scoring against the PON. Each panel corresponds to one factor, with samples coloured by disease label.



Supplementary Figure 7. UMAP visualization of MOFA+ latent sample representations derived from FSLR, MDS, and CNA fragmentomics views at 5 Mb bin resolution with Z-scoring against the PON. Samples are coloured by disease label.



Supplementary Figure 8. UMAP visualization of 8-dimensional latent embeddings obtained from unsupervised CMK using FSLR and MDS views, colored by disease label for the 8-class setting.

Generative AI Usage

The usage of Generative AI technology in this section was as follows. During the Literature review stages of this project, ChatGPT was used to clarify condense scientific writing in the bioinformatics domain or breaking down equations, while remaining critical of generated outputs with verifications when needed. In terms of coding support, GitHub Copilot was used for small coding tasks, such as plotting function generation or line completion. No generative AI was used for experiment design setup with the exception of initial hyperparameter grid generation, which was subsequently augmented. In terms of the writing assistance, Claude (Sonnet 4.6) was used for quality improvements, in terms of style, flow and grammar. Furthermore, it was used for feedback in terms of text structuring and paraphrasing. Overall, Generative AI tools were used only as a supporting tool, where all outputs were critically evaluated and in many cases rejected.

Acknowledgments

I'd like to thank my fellow authors Prof.dr.ir M.J.T Reinders and Bram Pronk for their support and extensive feedback during our collaboration on this project. I'd also like to thank the Delft Bioinformatics Lab for their friendly welcome and enthusiasm during lab initiatives such as BioTalk or poster sessions. Furthermore, I'd like to express my heartfelt gratitude for the support of my family and friends.

References

1. World Health Organization. Cancer, April 2026. URL: https://www.who.int/news-room/fact-sheets/detail/cancer.
2. Ahtisham Abbasi, Muhammad Sajjad, Muhammad Asim, Sebastian Vollmer, and Andreas Dengel. *From Cancer Molecular Subtype to AI Hype: Benchmarking AI in Cancer Molecular Subtyping*. March 2025. doi:10.1101/2025.03.10.642355.
3. W.H. Adrian Tsui, Peiyong Jiang, and Y.M. Dennis Lo. Cell-free DNA fragmentomics in cancer. *Cancer Cell*, 43(10):1792–1814, October 2025. URL: https://linkinghub.elsevier.com/retrieve/pii/S1535610825003988, doi:10.1016/j.ccell.2025.09.006.
4. Joscilin Mathew, Parvathy Sankar, and Matthew A. Varacallo. Physiology, Blood Plasma. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2026. URL: http://www.ncbi.nlm.nih.gov/books/NBK531504/.
5. Teppei Hashimoto, Kohsuke Yoshida, Akira Hashiramoto, and Kiyoshi Matsui. Cell-Free DNA in Rheumatoid Arthritis. *International Journal of Molecular Sciences*, 22(16):8941, August 2021. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8396202/, doi:10.3390/ijms22168941.
6. Y. M. Dennis Lo, Diana S. C. Han, Peiyong Jiang, and Rossa W. K. Chiu. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science*, 372(6538):eaaw3616, April 2021. URL: https://www.science.org/doi/10.1126/science.aaw3616, doi:10.1126/science.aaw3616.
7. David O'Reilly, Carolyn Moloney, Anthony O'Grady, David Synnott, Daniel Ryan, Michael Emmet O'Brien, Ross E Morgan, Ian Counihan, Sinead Cuffe, Grzegorz Korpanty, Lisa Mary Prior, Stephen Finn, Robert Cummins, Sinead Toomey, Bryan T Hennessy, Jan Sorensen, Kathleen Bennett, Parthiban Nadarajan, Brendan Doyle, and Jarushka Naidoo. Cost and resource utilisation for liquid biopsy vs tissue biopsy genotyping in advanced NSCLC: A micro-costing model.
8. Pavel Stejskal, Hani Goodarzi, Josef Srovnal, Marián Hajdúch, Laura J. van 't Veer, and Mark Jesus M. Magbanua. Circulating tumor nucleic acids: biology, release mechanisms, and clinical relevance. *Molecular Cancer*, 22(1):15, January 2023. doi:10.1186/s12943-022-01710-w.
9. Hao Wang, Yi Zhang, Hao Zhang, Hui Cao, Jinning Mao, Xinxin Chen, Liangchi Wang, Nan Zhang, Peng Luo, Ji Xue, Xiaoya Qi, Xiancheng Dong, Guodong Liu, and Quan Cheng. Liquid biopsy for human cancer: cancer screening, monitoring, and treatment. *MedComm*, 5(6):e564, June 2024. URL: https://onlinelibrary.wiley.com/doi/10.1002/mco2.564, doi:10.1002/mco2.564.
10. Hunter R. Underhill, Jacob O. Kitzman, Sabine Hellwig, Noah C. Welker, Riza Daza, Daniel N. Baker, Keith M. Gligorich, Robert C. Rostomily, Mary P. Bronner, and Jay Shendure. Fragment Length of Circulating Tumor DNA. *PLoS Genetics*, 12(7):e1006162, July 2016. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4948782/, doi:10.1371/journal.pgen.1006162.
11. Peiyong Jiang, Kun Sun, Wenlei Peng, Suk Hang Cheng, Meng Ni, Philip C. Yeung, Macy M.S. Heung, Tingting Xie, Huimin Shang, Ze Zhou, Rebecca W.Y. Chan, John Wong, Vincent W.S. Wong, Liona C. Poon, Tak Yeung Leung, W.K. Jacky Lam, Jason Y.K. Chan, Henry L.Y. Chan, K.C. Allen Chan, Rossa W.K. Chiu, and Y.M. Dennis Lo. Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. *Cancer Discovery*, 10(5):664–673, May 2020. doi:10.1158/2159-8290.CD-19-0622.
12. Ravi Bandaru, Hailu Fu, Haizi Zheng, Jocelyn Liang, Li Wang, Shuchi Gulati, Benjamin H Hinrichs, Mingxiang Teng, Bin Zhang, Marsha Kocherginsky, Dechen Lin, David A. Hildeman, Francis P Worden, Matthew O Old, Neal E Dunlap, John M Kaczmar, Maura Gillison, Dalia El-Gamal, Trisha Wise-Draper, and Yaping Liu. Genome-Wide Variations of End Motif in Cell-Free DNA Fragments Distinguish Immunotherapy Responders from Non-Responders in Head and Neck Cancer: A Multi-Institute Prospective Study. *medRxiv*, page 2026.03.24.26348354, March 2026. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC13060422/, doi:10.64898/2026.03.24.26348354.
13. Juqing Xu, Haiming Chen, Weifei Fan, Mantang Qiu, and Jifeng Feng. Plasma cell-free DNA as a sensitive biomarker for multi-cancer detection and immunotherapy outcomes prediction. *Journal of Cancer Research and Clinical Oncology*, 150(1):7, January 2024. doi:10.1007/s00432-023-05521-4.
14. Rameen Beroukhim, Craig H. Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S. Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin T. Mc Henry, Reid M. Pinchback, Azra H. Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael S. Lawrence, Barbara A. Weir, Kumiko E. Tanaka, Derek Y. Chiang, Adam J. Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J. Kaye, Hidefumi Sasaki, Joel E. Tepper, Jonathan A. Fletcher, Josep Tabernero, Jose Baselga, Ming-Sound Tsao, Francesca DeMichelis, Mark A. Rubin,

Pasi A. Janne, Mark J. Daly, Carmelo Nucera, Ross L. Levine, Benjamin L. Ebert, Stacey Gabriel, Anil K. Rustgi, Cristina R. Antonescu, Marc Ladanyi, Anthony Letai, Levi A. Garraway, Massimo Loda, David G. Beer, Lawrence D. True, Aikou Okamoto, Scott L. Pomeroy, Samuel Singer, Todd R. Golub, Eric S. Lander, Gad Getz, William R. Sellers, and Matthew Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, February 2010. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2826709/, doi:10.1038/nature08822.

15. Ziyu Tao, Shixiang Wang, Chenxu Wu, Tao Wu, Xiangyu Zhao, Wei Ning, Guangshuai Wang, Jinyu Wang, Jing Chen, Kaixuan Diao, Fuxiang Chen, and Xue-Song Liu. The repertoire of copy number alteration signatures in human cancer. *Briefings in Bioinformatics*, 24(2):bbad053, February 2023. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10025440/, doi:10.1093/bib/bbad053.
16. Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C. Bruhm, Sarah Østrup Jensen, Jamie E. Medina, Carolyn Hruban, James R. White, Doreen N. Palsgrove, Noushin Niknafs, Valsamo Anagnostou, Patrick Forde, Jarushka Naidoo, Kristen Marrone, Julie Brahmer, Brian D. Woodward, Hatim Husain, Karlijn L. Van Rooijen, Mai-Britt Worm Ørntoft, Anders Husted Madsen, Cornelis J. H. Van De Velde, Marcel Verheij, Annemieke Cats, Cornelis J. A. Punt, Geraldine R. Vink, Nicole C. T. Van Grieken, Miriam Koopman, Remond J. A. Fijneman, Julia S. Johansen, Hans Jørgen Nielsen, Gerrit A. Meijer, Claus Lindbjerg Andersen, Robert B. Scharpf, and Victor E. Velculescu. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, 570(7761):385–389, June 2019. URL: https://www.nature.com/articles/s41586-019-1272-6, doi:10.1038/s41586-019-1272-6.
17. Jacob J. Chabon, Emily G. Hamilton, David M. Kurtz, Mohammad S. Esfahani, Everett J. Moding, Henning Stehr, Joseph Schroers-Martin, Barzin Y. Nabet, Binbin Chen, Aadel A. Chaudhuri, Chih Long Liu, Angela B. Hui, Michael C. Jin, Tej D. Azad, Diego Almanza, Young-Jun Jeon, Monica C. Nesselbush, Lyron Co Ting Keh, Rene F. Bonilla, Christopher H. Yoo, Ryan B. Ko, Emily L. Chen, David J. Merriott, Pierre P. Massion, Aaron S. Mansfield, Jin Jen, Hong Z. Ren, Steven H. Lin, Christina L. Costantino, Risa Burr, Robert Tibshirani, Sanjiv S. Gambhir, Gerald J. Berry, Kristin C. Jensen, Robert B. West, Joel W. Neal, Heather A. Wakelee, Billy W. Loo, Christian A. Kunder, Ann N. Leung, Natalie S. Lui, Mark F. Berry, Joseph B. Shrager, Viswam S. Nair, Daniel A. Haber, Lecia V. Sequist, Ash A. Alizadeh, and Maximilian Diehn. Integrating genomic features for non-invasive early lung cancer detection. *Nature*, 580(7802):245–251, April 2020. URL: https://www.nature.com/articles/s41586-020-2140-0, doi:10.1038/s41586-020-2140-0.
18. Yuying Hou, Xiang-Yu Meng, and Xionghui Zhou. Systematically Evaluating Cell-Free DNA Fragmentation Patterns for Cancer Diagnosis and Enhanced Cancer Detection via Integrating Multiple Fragmentation Patterns. *Advanced Science*, 11(30):2308243, June 2024. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11321639/, doi:10.1002/advs.202308243.
19. Van Thien Chi Nguyen, Trong Hieu Nguyen, Nhu Nhat Tan Doan, Thi Mong Quynh Pham, Giang Thi Huong Nguyen, Thanh Dat Nguyen, Thuy Thi Thu Tran, Duy Long Vo, Thanh Hai Phan, Thanh Xuan Jasmine, Van Chu Nguyen, Huu Thinh Nguyen, Trieu Vu Nguyen, Thi Hue Hanh Nguyen, Le Anh Khoa Huynh, Trung Hieu Tran, Quang Thong Dang, Thuy Nguyen Doan, Anh Minh Tran, Viet Hai Nguyen, Vu Tuan Anh Nguyen, Le Minh Quoc Ho, Quang Dat Tran, Thi Thu Thuy Pham, Tan Dat Ho, Bao Toan Nguyen, Thanh Nhan Vo Nguyen, Thanh Dang Nguyen, Dung Thai Bieu Phu, Boi Hoan Huu Phan, Thi Loan Vo, Thi Huong Thoang Nai, Thuy Trang Tran, My Hoang Truong, Ngan Chau Tran, Trung Kien Le, Thanh Huong Thi Tran, Minh Long Duong, Hoai Phuong Thi Bach, Van Vu Kim, The Anh Pham, Duc Huy Tran, Trinh Ngoc An Le, Truong Vinh Ngoc Pham, Minh Triet Le, Dac Ho Vo, Thi Minh Thu Tran, Minh Nguyen Nguyen, Thi Tuong Vi Van, Anh Nhu Nguyen, Thi Trang Tran, Vu Uyen Tran, Minh Phong Le, Thi Thanh Do, Thi Van Phan, Hong-Dang Luu Nguyen, Duy Sinh Nguyen, Van Thinh Cao, Thanh-Thuy Thi Do, Dinh Kiet Truong, Hung Sang Tang, Hoa Giang, Hoai-Nghia Nguyen, Minh-Duy Phan, and Le Son Tran. Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization. *eLife*, 12:RP89083, October 2023. URL: https://elifesciences.org/articles/89083, doi:10.7554/eLife.89083.
20. Yumei Li, Jianfeng Xu, Chaorong Chen, Zhenhai Lu, Desen Wan, Diange Li, Jason S. Li, Allison J. Sorg, Curt C. Roberts, Shivani Mahajan, Maxime A. Gallant, Itai Pinkoviezky, Ya Cui, David J. Taggart, and Wei Li. Multimodal epigenetic sequencing analysis (MESA) of cell-free DNA for non-invasive colorectal cancer detection. *Genome Medicine*, 16(1):9, January 2024. URL: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-023-01280-6, doi:10.1186/s13073-023-01280-6.
21. Zhiwen Yu, Ziyang Dong, Chenchen Yu, Kaixiang Yang, Ziwei Fan, and C. L. Philip Chen. A review on multi-view learning. *Frontiers of Computer Science*, 19(7):197334, July 2025. URL: https://link.springer.com/10.1007/s11704-024-40004-w, doi:10.1007/s11704-024-40004-w.
22. H. Holleling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377, 1936.
23. Jason D.R. Farquhar, David R. Hardoon, Hongying Meng, John Shawe-Taylor, and Sandor Szedmak. Two view learning: SVM-2K, theory and practice. In *Adv. neural inf. proces. syst.*, pages 355–362, 2005. Journal Abbreviation: Adv. neural inf. proces. syst.
24. Shiliang Sun and Guoqing Chao. *Multi-view maximum entropy discrimination.* August 2013. Journal Abbreviation: IJCAI International Joint Conference on Artificial Intelligence Pages: 1712 Publication Title: IJCAI International Joint Conference on Artificial Intelligence.
25. Guoqing Chao and Shiliang Sun. Consensus and complementarity based maximum entropy discrimination for multi-view classification. *Information Sciences*, 367-368:296–310, November 2016. URL: https://linkinghub.elsevier.com/retrieve/pii/S002002551630411X, doi:10.1016/j.ins.2016.06.004.
26. Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni, and Oliver Stegle. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data.

Genome Biology, 21(1):111, December 2020. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02015-1, doi:10.1186/s13059-020-02015-1.

27. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, July 2020. arXiv:2002.05709 [cs]. URL: http://arxiv.org/abs/2002.05709, doi:10.48550/arXiv.2002.05709.
28. Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive Multi-View Kernel Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9552–9566, August 2023. URL: https://ieeexplore.ieee.org/document/10061269/, doi:10.1109/TPAMI.2023.3253211.
29. Viktor A. Adalsteinsson, Gavin Ha, Samuel S. Freeman, Atish D. Choudhury, Daniel G. Stover, Heather A. Parsons, Gregory Gydush, Sarah C. Reed, Denisse Rotem, Justin Rhoades, Denis Loginov, Dimitri Livitz, Daniel Rosebrock, Ignaty Leshchiner, Jaegil Kim, Chip Stewart, Mara Rosenberg, Joshua M. Francis, Cheng-Zhong Zhang, Ofir Cohen, Coyin Oh, Huiming Ding, Paz Polak, Max Lloyd, Sairah Mahmud, Karla Helvie, Margaret S. Merrill, Rebecca A. Santiago, Edward P. O'Connor, Seong H. Jeong, Rachel Leeson, Rachel M. Barry, Joseph F. Kramkowski, Zhenwei Zhang, Laura Polacek, Jens G. Lohr, Molly Schleicher, Emily Lipscomb, Andrea Saltzman, Nelly M. Oliver, Lori Marini, Adrienne G. Waks, Lauren C. Harshman, Sara M. Tolaney, Eliezer M. Van Allen, Eric P. Winer, Nancy U. Lin, Mari Nakabayashi, Mary-Ellen Taplin, Cory M. Johannessen, Levi A. Garraway, Todd R. Golub, Jesse S. Boehm, Nikhil Wagle, Gad Getz, J. Christopher Love, and Matthew Meyerson. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications*, 8(1):1324, November 2017. URL: https://www.nature.com/articles/s41467-017-00965-y, doi:10.1038/s41467-017-00965-y.
30. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning, March 2021. arXiv:2004.11362 [cs.LG]. URL: http://arxiv.org/abs/2004.11362, doi:10.48550/arXiv.2004.11362.
31. Alexander M. Aravanis, Mark Lee, and Richard D. Klausner. Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. *Cell*, 168(4):571–574, February 2017. URL: https://www.cell.com/cell/abstract/S0092-8674(17)30115-0, doi:10.1016/j.cell.2017.01.030.
32. Arash Jamshidi, Minetta C. Liu, Eric A. Klein, Oliver Venn, Earl Hubbell, John F. Beausang, Samuel Gross, Collin Melton, Alexander P. Fields, Qinwen Liu, Nan Zhang, Eric T. Fung, Kathryn N. Kurtzman, Hamed Amini, Craig Betts, Daniel Civello, Peter Freese, Robert Calef, Konstantin Davydov, Saniya Fayzullina, Chenlu Hou, Roger Jiang, Byoungsok Jung, Susan Tang, Vasiliki Demas, Joshua Newman, Onur Sakarya, Eric Scott, Archana Shenoy, Seyedmehdi Shojaee, Kristan K. Steffen, Virgil Nicula, Tom C. Chien, Siddhartha Bagaria, Nathan Hunkapiller, Mohini Desai, Zhao Dong, Donald A. Richards, Timothy J. Yeatman, Allen L. Cohn, David D. Thiel, Donald A. Berry, Mohan K. Tummala, Kristi McIntyre, Mikkael A. Sekeres, Alan Bryce, Alexander M. Aravanis, Michael V. Seiden, and Charles Swanton. Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell*, 40(12):1537–1549.e12, December 2022. URL: https://www.cell.com/cancer-cell/abstract/S1535-6108(22)00513-X, doi:10.1016/j.ccell.2022.10.022.
33. Xingdong Chen, Jeffrey Gole, Athurva Gore, Qiye He, Ming Lu, Jun Min, Ziyu Yuan, Xiaorong Yang, Yanfeng Jiang, Tiejun Zhang, Chen Suo, Xiaojie Li, Lei Cheng, Zhenhua Zhang, Hongyu Niu, Zhe Li, Zhen Xie, Han Shi, Xiang Zhang, Min Fan, Xiaofeng Wang, Yajun Yang, Justin Dang, Catie McConnell, Juan Zhang, Jiucun Wang, Shunzhang Yu, Weimin Ye, Yuan Gao, Kun Zhang, Rui Liu, and Li Jin. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nature Communications*, 11(1):3475, July 2020. URL: https://www.nature.com/articles/s41467-020-17316-z, doi:10.1038/s41467-020-17316-z.
34. Yaping Liu, Sarah C. Reed, Christopher Lo, Atish D. Choudhury, Heather A. Parsons, Daniel G. Stover, Gavin Ha, Gregory Gydush, Justin Rhoades, Denisse Rotem, Samuel Freeman, David W. Katz, Ravi Bandaru, Haizi Zheng, Hailu Fu, Viktor A. Adalsteinsson, and Manolis Kellis. FinaleMe: Predicting DNA methylation by the fragmentation patterns of plasma cell-free DNA. *Nature Communications*, 15(1):2790, March 2024. URL: https://www.nature.com/articles/s41467-024-47196-6, doi:10.1038/s41467-024-47196-6.
35. Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, 9(1):9354, June 2019. URL: https://www.nature.com/articles/s41598-019-45839-z, doi:10.1038/s41598-019-45839-z.
36. Dimitrios Mathios, Jakob Sidenius Johansen, Stephen Cristiano, Jamie E. Medina, Jillian Phallen, Klaus R. Larsen, Daniel C. Bruhm, Noushin Niknafs, Leonardo Ferreira, Vilmos Adleff, Jia Yuee Chiao, Alessandro Leal, Michael Noe, James R. White, Adith S. Arun, Carolyn Hruban, Akshaya V. Annapragada, Sarah Østrup Jensen, Mai-Britt Worm Ørntoft, Anders Husted Madsen, Beatriz Carvalho, Meike de Wit, Jacob Carey, Nicholas C. Dracopoli, Tara Maddala, Kenneth C. Fang, Anne-Renee Hartman, Patrick M. Forde, Valsamo Anagnostou, Julie R. Brahmer, Remond J. A. Fijneman, Hans Jørgen Nielsen, Gerrit A. Meijer, Claus Lindbjerg Andersen, Anders Mellemgaard, Stig E. Bojesen, Robert B. Scharpf, and Victor E. Velculescu. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nature Communications*, 12(1):5060, August 2021. URL: https://www.nature.com/articles/s41467-021-24994-w, doi:10.1038/s41467-021-24994-w.
37. Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (PCA). *Computers & Geosciences*, 19(3):303–342, March 1993. URL: https://linkinghub.elsevier.com/retrieve/pii/009830049390090R, doi:10.1016/0098-3004(93)90090-R.
38. Yoshua Bengio, Jean-françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas L Roux, and Marie Ouimet. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering.
39. R. Taylor Sundby, Jeffrey J. Szymanski, Alexander C. Pan, Paul A. Jones, Sana Z. Mahmood, Olivia H. Reid, Divya Srihari, Amy E. Armstrong, Stacey Chamberlain, Sanita Burgic, Kara Weekley, Béga Murray, Sneh Patel, Faridi Qaium, Andrea N. Lucas, Margaret Fagan, Anne Dufek, Christian F. Meyer, Natalie B. Collins,

Christine A. Pratilas, Eva Dombi, Andrea M. Gross, AeRang Kim, John S.A. Chrisinger, Carina A. Dehner, Brigitte C. Widemann, Angela C. Hirbe, Aadel A. Chaudhuri, and Jack F. Shern. Early Detection of Malignant and Premalignant Peripheral Nerve Tumors Using Cell-Free DNA Fragmentomics. *Clinical Cancer Research*, 30(19):4363–4376, October 2024. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11443212/, doi:10.1158/1078-0432.CCR-24-0797.

40. Stavros Makrodimitris, Bram Pronk, Tamim Abdelaal, and Marcel Reinders. An in-depth comparison of linear and non-linear joint embedding methods for bulk and single-cell multi-omics. *Briefings in Bioinformatics*, 25(1):bbad416, November 2023. URL: https://academic.oup.com/bib/article/doi/10.1093/bib/bbad416/7450271, doi:10.1093/bib/bbad416.