University of Salerno
Department of Medical Sciences
PhD Program in Translational Medicine – XXXVIII Cycle

Multi-omics approaches to breast cancer and development of general-purpose tensor methods in R

Candidate: Alessandro Giordano
Supervisor: Prof. Giovanni Nassa
Co-supervisors: Prof. Giorgio Giurato and Prof. Francesca Rizzo

Academic Year 2025–2026

Abstract

Endocrine therapy resistance in ERα-positive breast cancer is a complex regulatory phenotype that rarely arises from alterations confined to a single molecular layer. Instead, resistance emerges through coordinated changes across transcriptional programs, chromatin state, and epigenetic regulation, motivating integrative multi-omics approaches that constrain interpretation by requiring concordant evidence across complementary assays. This thesis pursues two objectives: (i) to investigate ERα-associated regulatory dependencies in luminal breast cancer using integrative functional genomics, including the development of a pathway-first framework for cross-model methylation–expression concordance and (ii) to contribute general-purpose methodology for supervised analysis and interpretation of longitudinal multi-omics data with an intrinsic multi-way (tensor) structure.

In Part I, BRPF1 is investigated as a chromatin-associated coregulator connected to ERα signalling. By integrating ERα and BRPF1 ChIP-seq, BRPF1 ChIP-seq after ERα knockdown, ATAC-seq under BRPF1 perturbation, and RNA-seq, the analyses support an ERα-dependent recruitment of BRPF1 at a subset of shared regulatory loci and show that BRPF1 inhibition is associated with widespread accessibility losses enriched in ERα-linked regions, accompanied by coherent transcriptional attenuation. Together, these multi-layer readouts motivate BRPF1 as a candidate regulatory dependency within ERα-associated circuitry.

In Part II, endocrine resistance across heterogeneous ERα-positive cell line models is investigated by integrating EPIC DNA methylation profiling with RNA-seq. Because global methylation structure is dominated by cell line identity, the analysis shifts from CpG-level heterogeneity to pathway-level convergence, combining Hallmark GSEA, consensus leading-edge gene cores, and promoter methylation summaries to test whether recurrent functional programs show *enriched* epigenetic support in the expected inverse direction. Direction-preserving filters further assess robustness by enforcing increasing cross-model consistency at both the pathway and gene-core levels, yielding conservative but interpretable sets of recurrent pathway signatures and their associated consensus gene cores.

In Part III, `TensorPLS` is introduced as an `R` package for supervised analy-

sis of three-way longitudinal omics datasets represented as tensors. The workflow combines tensor-native preprocessing and Tucker-3 imputation with PLS-based discrimination, and provides mode-resolved interpretability, including time point contribution analysis and time-resolved feature importance through VIP2D, with permutation-based robustness assessment. Evaluated on longitudinal case–control data from the TEDDY cohort across multiple omics modalities, `TensorPLS` illustrates how preserving tensor structure in preprocessing and interpretation enables structured feature prioritisation and identification of temporal windows associated with class separation.

Taken together, the chapters demonstrate that combining cross-layer evidence with structure-aware analytical workflows improves interpretability and supports robust prioritisation of candidate mechanisms and biomarkers for subsequent validation.

Contents

1

Introduction

Endocrine therapy remains the cornerstone of treatment for ERα-positive breast cancer, yet disease relapse and progression under endocrine pressure remain major clinical challenges. In many cases, resistance does not reflect a single deterministic event at the receptor level, but rather an adaptive process in which tumour cells preserve key proliferative programs by rewiring the regulatory circuitry that controls ERα output. This plasticity motivates a regulatory view of endocrine resistance, where altered transcriptional programs are coupled to changes in chromatin accessibility and epigenetic state [52, 22].

Within this framework, chromatin-associated coregulators and epigenetic factors emerge as plausible candidate vulnerabilities, because they can shape ERα function by modulating enhancer activity, chromatin accessibility, and the transcriptional response to hormone signalling [19, 21]. However, any single assay captures only one facet of this regulatory architecture, and signals that appear compelling in one molecular layer may be ambiguous without supporting evidence from complementary readouts. For this reason, integrative multi-omics analyses are increasingly adopted to strengthen interpretation by aligning findings across layers and prioritising dependencies supported by concordant functional evidence [26, 18].

In parallel to the biological questions, this work places emphasis on analysis frameworks that make multi-omics integration more interpretable and reproducible. A pathway-first integration strategy is used to move from heterogeneous gene-level signals to conserved functional programs supported across omics layers, and a tensor-based supervised framework is developed for longitudinal multi-omics tensors, combining tensor-aware preprocessing and imputation with PLS-based discrimination

and time-resolved interpretability.

1.0.1 Thesis overview

This thesis is organised into three parts.

Part I: BRPF1 as an ERα-associated chromatin coregulator. The first part focuses on integrating multi-omics layers to illustrate how complementary assays reveal different aspects of the same regulatory problem. In particular, we integrate ERα and BRPF1 ChIP-seq to define co-occupied regulatory elements, BRPF1 ChIP-seq after ERα knockdown to assess ERα-dependent recruitment, ATAC-seq to quantify accessibility changes upon BRPF1 inhibition at ERα-linked loci, and RNA-seq to connect accessibility to transcriptional consequences.

Part II: DNA methylation and transcriptomic convergence in endocrine resistance models. The second part focuses on whether heterogeneous endocrine resistance models converge on shared functional programs when analysed at an integrative level. We integrate EPIC DNA methylation profiling with RNA-seq across multiple ERα-positive endocrine resistance models. Because global methylation structure is strongly influenced by cell line identity, we shift from CpG-level heterogeneity to pathway-level convergence by combining Hallmark GSEA signatures with gene-level promoter methylation summaries and testing whether pathway cores are disproportionately supported by promoter methylation changes in the expected inverse direction. We further evaluate robustness by introducing direction-preserving filters that require increasing levels of cross-model consistency.

Part III: TensorPLS. This final chapter is intentionally methodological and independent from the endocrine-resistance case studies presented in Parts I–II. Longitudinal multi-omics datasets often have an intrinsic multi-way structure (e.g., subjects $\times$ features $\times$ time/conditions), yet many widely used supervised omics tools are designed for two-dimensional matrices. To address this gap, we present `TensorPLS`, an R framework emphasizing tensor-aware preprocessing and interpretation for supervised analysis of three-way datasets, including time-resolved feature importance and robustness assessment by permutation testing. The framework is evaluated on a longitudinal cohort as a realistic testbed.

The next chapter introduces the biological and conceptual context required to interpret the Results chapter.

2

Background

2.0.1 Breast cancer: heterogeneity and clinically relevant classification

Breast cancer is a heterogeneous disease comprising tumour entities that differ in cellular origin, regulatory programs, clinical behaviour, and therapeutic vulnerabilities [48, 12]. Most breast cancers arise from epithelial cells of the ducts or lobules, and variation in tissue architecture and lineage context contributes to pathological and molecular diversity [12]. This heterogeneity has direct clinical implications, as tumours with distinct biological drivers show different patterns of progression and response to therapy.

In clinical and research settings, breast cancer is commonly described using complementary classification frameworks. Histological classification captures tumour morphology and tissue architecture, stage-based classification summarises disease extent (typically using the TNM system), and molecular classification groups tumours according to biomarkers and signalling programs that strongly influence prognosis and treatment response [49, 11]. In this thesis, the molecular perspective is central because it directly connects tumour biology to endocrine therapy and to the problem of endocrine therapy resistance.

In practice, molecular subtype assignment is often approximated using immunohistochemistry (IHC) markers—primarily estrogen receptor (ER), progesterone receptor (PR), and HER2—together with proliferation readouts such as Ki-67 [16]. These markers are routinely used to guide treatment decisions and to stratify patients into clinically actionable groups [57].

2.0.2 Molecular subtypes and clinical surrogates

Expression-based "intrinsic" subtypes (e.g., PAM50) capture distinct transcriptional programs and are widely used in research, whereas clinical practice frequently relies on IHC-based surrogate definitions that approximate these categories using ER/PR/HER2 status and Ki-67 [57]. Within this framework, breast cancers are commonly grouped into four major subtype families:

- Luminal A: typically ER-positive (often also PR-positive), with lower proliferation and generally a more favorable prognosis. In IHC-based surrogate definitions, Luminal A–like tumours are usually ER-positive and HER2-negative with relatively low Ki-67 [16, 57].
- Luminal B: also ER-positive but generally characterized by higher proliferation and/or additional aggressive features, with a less favorable prognosis compared with Luminal A. In surrogate definitions, Luminal B–like tumours often show higher Ki-67 and/or lower PR expression and may be HER2-negative or HER2-positive [16, 57].
- HER2-enriched: characterized by activation of the HER2 pathway and often associated with aggressive behaviour, but frequently responsive to HER2-targeted therapies. Intrinsic HER2-enriched status does not always coincide with HER2 positivity by IHC or FISH, although HER2 remains the key clinical biomarker guiding therapy [57].
- Triple-negative (Basal-like): negative for ER, PR, and HER2. Because it lacks these targets, this subtype does not respond to endocrine therapy or HER2-targeted drugs and is primarily treated with chemotherapy and, in selected contexts, other targeted or immunotherapeutic approaches [50].

Luminal (ERα-positive) breast cancers account for the majority of diagnoses and represent the main clinical context for endocrine therapy. Although endocrine treatments are highly effective in many patients, resistance frequently emerges over time, making ERα-positive disease a central setting in which transcriptional and epigenetic regulation is directly relevant to patient outcomes.

2.0.3 Breast cancer in numbers

Breast cancer is the most commonly diagnosed cancer in women and remains a leading cause of cancer-related mortality worldwide [31, 34]. In 2022, an estimated

2,296,840 new cases and 666,103 deaths were reported globally [31]. If current incidence and mortality trends persist, projections suggest that by 2050 annual burden may rise by approximately 38% for incidence and 68% for mortality [34], corresponding to about 3.17 million cases and 1.12 million deaths per year. This would imply an increase of roughly +452,950 deaths/year (about +68%) compared with 2022.

At the population level, breast cancer subtype composition depends on the classification scheme and cohort, but broad estimates consistently indicate that the majority of tumours fall within the hormone receptor-positive spectrum (around 70%) [**?**], whereas HER2-positive disease accounts for about 15–20% of cases and triple-negative breast cancer for about 10–15% [51]. Consequently, a large fraction of diagnoses lie within the luminal ERα-positive context where endocrine therapy is a cornerstone of treatment.

Despite major clinical benefit, relapse under endocrine pressure remains a key limitation. In ERα-positive disease, long-term risk of distant recurrence persists even after completing 5 years of adjuvant endocrine therapy, varying substantially with initial stage and reaching approximately 10% in lower-risk groups up to about 40% in higher-risk node-positive disease over years 5–20 [53]. Moreover, in the advanced/metastatic setting, endocrine resistance and progression under endocrine-based treatments are frequent, with estimates indicating that a substantial fraction of HR-positive metastatic patients develop endocrine resistance over the course of therapy [**?**]. This clinical observation motivates the focus of the next section on ERα biology, endocrine therapy mechanisms, and endocrine resistance as an adaptive regulatory phenotype.

2.0.4 ERα-positive breast cancer and endocrine therapy

Estrogen receptor alpha (ERα) is a nuclear hormone receptor that plays a central role in normal breast development and in the growth of a large fraction of breast tumours. In ERα-positive disease, estrogen signalling provides a strong proliferative and survival advantage by activating transcriptional programs that promote cell-cycle progression and limit apoptosis [52].

At the molecular level, ERα functions as a ligand-activated transcription factor. Upon estrogen binding, ERα undergoes conformational changes, dimerizes, translocates to the nucleus, and binds regulatory DNA elements (including estrogen response elements, EREs). ERα then cooperates with coregulators and chromatin-associated factors to modulate transcriptional output in a context-dependent manner [25, 19]. This implies that ERα activity is not determined solely by receptor

presence, but also by the regulatory environment in which the receptor operates.

Endocrine therapy exploits the dependence of ERα-positive breast cancers on estrogen signalling. Clinically, endocrine treatments act either by reducing estrogen availability or by directly interfering with receptor function. Major classes include aromatase inhibitors (which reduce estrogen synthesis), selective estrogen receptor modulators (SERMs), and selective estrogen receptor degraders (SERDs) [52]. A prototypical SERM is tamoxifen, widely used in ERα-positive breast cancer. Tamoxifen binds the ligand-binding domain of ERα and induces a receptor conformation that alters coregulator recruitment, thereby attenuating estrogen-driven transcription in breast tissue [61, 24].

The clinical impact of endocrine therapy is substantial; however, endocrine treatment is not curative in all cases. A significant fraction of patients show de novo non-response or relapse after a period of clinical benefit, establishing endocrine resistance as a major unresolved challenge in ERα-positive breast cancer management [52, 53].

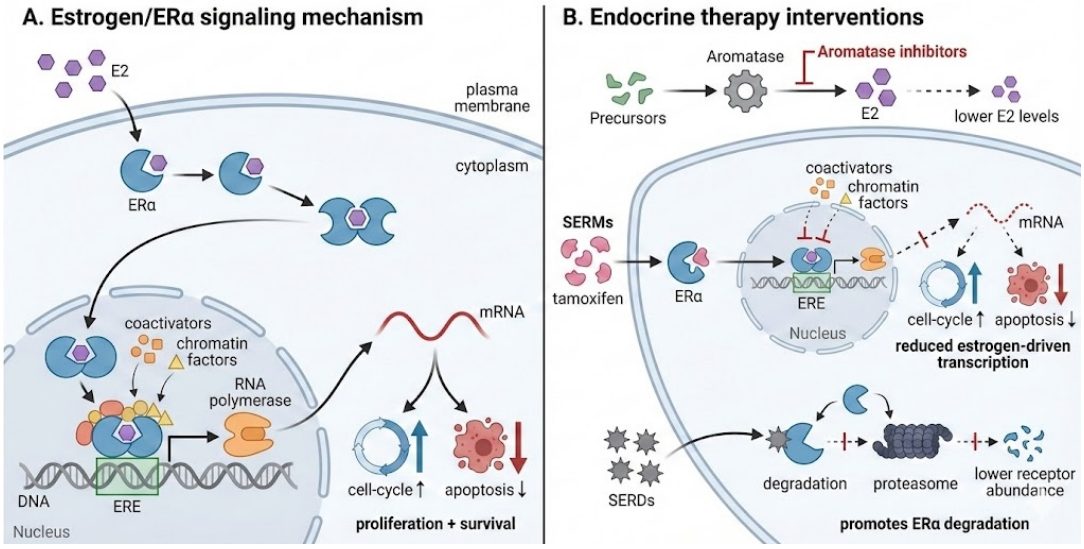


Figure 2.1: **ERα signalling and endocrine therapy in ERα-positive breast cancer. (A)** Estrogen (E2) binding activates ERα, promotes nuclear localization and binding to estrogen response elements (EREs), and recruits coregulators and chromatin-associated factors to drive transcriptional programs supporting proliferation and survival. **(B)** Endocrine therapy interventions include aromatase inhibitors (reduced E2 synthesis), selective estrogen receptor modulators (SERMs; e.g., tamoxifen), which alter coregulator recruitment and reduce estrogen-driven transcription, and selective estrogen receptor degraders (SERDs; e.g., fulvestrant), which promote ERα degradation.

2.0.5 Endocrine therapy resistance as an adaptive regulatory phenotype

Endocrine therapy resistance is commonly classified as *intrinsic* (primary) resistance, when tumours fail to respond from the outset, and *acquired* resistance, when tumours relapse after an initial period of sensitivity. Although these scenarios may involve distinct initiating events, both reflect the ability of cancer cells to maintain proliferative and survival programs under sustained therapeutic pressure.

A key observation is that resistance does not necessarily coincide with loss of ERα expression. Many resistant tumours retain ERα and preserve elements of ER-associated transcriptional activity, suggesting that resistance can emerge through changes in regulatory context rather than through elimination of the receptor itself [55, 52]. In this view, endocrine resistance is best interpreted as an adaptive phenotype shaped by regulatory rewiring: cancer cells reconfigure the conditions under which ERα operates, enabling estrogen-independent or estrogen-like transcriptional outputs even in the presence of endocrine therapy.

Several mechanisms can contribute to this rewiring. These include altered abundance or usage of ERα coregulators, signalling pathway crosstalk that converges on transcriptional control, and changes in chromatin state that modify the accessibility and activity of regulatory elements. Importantly, these processes do not act in isolation: resistance typically reflects coordinated changes across molecular layers, which motivates analytical strategies capable of integrating multiple omics readouts.

This regulatory perspective has direct translational implications. If ERα signalling is sustained through a reconfigured chromatin and coregulator landscape, then chromatin-associated factors and epigenetic enzymes can become functional dependencies and candidate therapeutic vulnerabilities in resistant disease [19, 68]. This rationale underlies the focus of this thesis on ERα-associated chromatin regulation and on multiple approaches to detect convergent functional signals across heterogeneous models.

2.0.6 Chromatin and epigenetic regulation in ERα-positive breast cancer

Transcriptional regulation occurs within chromatin, and ERα-dependent transcription is intrinsically chromatin-dependent. The ability of ERα to engage regulatory DNA, recruit coregulators, and activate or repress target genes depends on the accessibility and regulatory state of cis-regulatory elements, including promoters and enhancers [25, 19]. Changes in chromatin accessibility can therefore reshape which

ERα-bound regions are functional under specific conditions, including therapeutic pressure.

Chromatin regulation is tightly linked to histone modifications and the enzymatic machinery that writes, erases, and interprets these marks. Many chromatin-modifying enzymes act as coregulators rather than sequence-specific DNA-binding factors, and their effects depend on where and how they are recruited. In ERα-positive breast cancer, chromatin-associated coregulators can influence ERα output by stabilizing transcriptional complexes, facilitating enhancer–promoter communication, and modulating local chromatin structure. Under endocrine therapy, shifts in coregulator usage or chromatin state can enable alternative regulatory configurations that preserve proliferation and survival.

Epigenetic regulation adds an additional layer of stability to these adaptive processes. DNA methylation and histone modification patterns can reinforce transcriptional states and contribute to long-term regulatory memory in resistant cells. While promoter DNA methylation is often discussed in relation to inverse transcriptional regulation, its interpretability depends on regulatory context and genomic scale. In this thesis, methylation–expression integration is performed at the gene level by summarising differential methylation within promoter regions, a choice that provides a direct, biologically interpretable link to transcription and enables consistent comparisons across heterogeneous models. At the same time, promoter-centric summaries do not capture distal regulatory mechanisms, including enhancer-driven regulation, chromatin accessibility changes, and transcription factor activity, which can modulate gene expression independently of promoter methylation.

From a therapeutic perspective, this broader regulatory architecture motivates interest in chromatin regulators and epigenetic enzymes as candidate vulnerabilities: regulatory dependencies may be pharmacologically tractable and, in principle, reversible. However, identifying which regulatory nodes are functionally relevant in endocrine resistance typically requires evidence beyond single-layer association. Functional genomics and integrative multi-omics profiling provide complementary angles to connect candidate regulators to chromatin state and transcriptional output, strengthening mechanistic inference and prioritisation.

2.1 Experimental and omics readouts used in this thesis

To study endocrine therapy resistance as a regulatory phenotype, this thesis integrates multiple omics layers that probe complementary aspects of gene regulation.

Each assay provides a partial readout of the same regulatory system; therefore, interpretation is strongest when signals are considered jointly and with awareness of the assumptions and limitations of each modality.

2.1.1 ChIP-seq: chromatin occupancy and regulatory locus definition.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) maps genome-wide enrichment of a protein of interest at specific genomic loci, typically reported as *peaks*. Conceptually, a peak reflects an excess of sequencing reads at a locus compared with an appropriate control, consistent with factor occupancy in a population of cells [54, 20]. In ERα-positive breast cancer, ChIP-seq can be used to define the ERα cistrome and to identify candidate regulatory elements where ERα co-localizes with chromatin-associated coregulators [25, 19]. In this thesis, peak intersections are used to define co-occupied regions as candidate sites of functional cooperation. Importantly, occupancy alone does not imply regulatory activity or causal effects on transcription: a bound site may be context-dependent, redundant, or non-functional, and additional readouts are required to connect binding to chromatin state and transcriptional output [62, 3].

2.1.2 ATAC-seq: accessibility as a proxy for regulatory permissiveness.

ATAC-seq profiles open chromatin by capturing genomic regions that are accessible to transposase integration, providing a quantitative proxy for regulatory permissiveness [9, 10]. Increased accessibility is often associated with active regulatory elements, including promoters and enhancers, whereas decreased accessibility can reflect chromatin compaction and reduced capacity for transcription factor engagement [14]. However, chromatin accessibility does not uniquely identify enhancers, does not establish target genes, and does not guarantee transcriptional activation; rather, it reports a regulatory state that can enable or restrict transcription factor binding in a context-dependent manner [35]. In this thesis, differential accessibility is used to test whether the perturbation of chromatin regulators preferentially reshapes accessibility at ERα-associated loci, supporting integration with factor occupancy and transcriptional readouts.

2.1.3 RNA-seq: transcriptional consequences and pathway-level convergence.

RNA-seq quantifies transcript abundance and provides a direct functional readout of downstream transcriptional output [71, 13]. Because endocrine resistance can involve heterogeneous gene-level responses across models, pathway-level approaches such as Gene Set Enrichment Analysis (GSEA) are useful to summarise coherent biological programs and enable cross-model comparisons even when individual differentially expressed genes are not shared [64]. In this thesis, RNA-seq is used both to quantify transcriptional responses to perturbations and to derive pathway-level signatures for integration with epigenetic layers; in particular, Hallmark gene sets provide a compact representation of recurrent transcriptional programs [41, 42].

2.1.4 DNA methylation profiling with Illumina Infinium arrays (450K, EPIC 850K, EPIC v2).

Illumina Infinium DNA methylation microarrays provide genome-wide, single-CpG measurements from bisulfite-converted DNA, yielding quantitative methylation estimates commonly reported as β values (and corresponding M values). The widely used Infinium HumanMethylation450 BeadChip ("450K") interrogates more than 450,000 CpG sites across diverse genomic contexts [28]. The Infinium MethylationEPIC BeadChip ("EPIC 850K") expanded coverage to over 850,000 CpGs while retaining the majority of 450K content, with additional probes enriching distal regulatory regions, including enhancer-associated CpGs curated from major regulatory resources [56, 46]. More recently, the Infinium MethylationEPIC v2.0 BeadChip updated probe content and targets over 935,000 CpGs (often referred to as ~950K in practice), while maintaining compatibility with established Infinium workflows [29, 30]. Importantly, the DNA methylation analyses in this thesis were generated using the Infinium MethylationEPIC v2.0 array [29].

2.2 Methodological background: tensors, Tucker decomposition, and PLS-based supervised modelling

This section introduces the methodological concepts required to interpret Part III. We briefly motivate why time-course and multi-condition multi-omics data are natu-

rally multi-way, and summarise the tensor representations and PLS-based supervised modelling choices used throughout the framework.

2.2.1 Motivation: longitudinal multi-omics as multi-way data

Longitudinal and multi-condition studies naturally generate structured datasets in which the same subjects are measured repeatedly across time points, conditions, or assay blocks. When the same feature space is observed across these blocks, the resulting data can be represented as a three-way array (tensor) with dimensions *subjects* $\times$ *features* $\times$ *time/blocks* [36].

From a practical standpoint, analysing longitudinal omics data as tensors is also motivated by the current software landscape. Multi-way PLS implementations exist and directly model three-way arrays (e.g., N-PLS and sparse N-PLS) [27], while several R packages provide multi-way component models such as Tucker decompositions [7, 67]. However, these tools often address only parts of the end-to-end workflow, and it is less common to find a single reproducible pipeline that explicitly integrates tensor-aware preprocessing, principled tensor-based imputation, supervised discrimination, and mode-resolved interpretability. This gap motivates the design choices of `TensorPLS`.

2.2.2 From matrices to tensors: notation and unfolding

A matrix $X \in \mathbb{R}^{n \times p}$ represents two-way data (e.g., n subjects and p features). A three-way tensor extends this structure to $X \in \mathbb{R}^{n \times p \times k}$, where k denotes time points or experimental blocks [36]. The slice $X_{::t}$ corresponds to the matrix of measurements at time point (or block) t.

A common strategy to apply matrix-based methods to tensor data is *unfolding* (matricization), which reshapes the tensor into a two-dimensional matrix. Unfolding along the subject mode yields

$$X_{(1)} \in \mathbb{R}^{n \times (p \cdot k)},$$

where each subject is represented by the concatenation of its features across all time points/blocks. This representation enables the use of established matrix-based supervised methods while retaining the ability to map results back to the original multi-way structure for interpretation.

2.2.3 PLS-based supervised modelling

Partial Least Squares (PLS) is a supervised latent-variable method designed to model relationships between a predictor matrix $X \in \mathbb{R}^{n \times p}$ and a response matrix $Y \in \mathbb{R}^{n \times q}$ by extracting components that capture directions in X that are maximally associated with Y [72]. In practice, PLS constructs score vectors (latent components) that summarise the predictors while emphasizing covariance with the response, which is particularly useful in omics where predictors are high-dimensional and strongly collinear.

In classification settings, PLS can be used in a discriminant-analysis framework (PLS-DA) by encoding class membership in Y (e.g., dummy variables) and estimating components that emphasize separation between groups in the latent space [4, 60]. Many omics workflows further rely on sparse variants of PLS that incorporate variable selection through penalization, improving interpretability by focusing on a subset of discriminant features; such approaches are implemented in widely used software, including the `mixOmics` ecosystem [38, 59].

For context, PLS(-DA) is sometimes compared with Principal Component Analysis (PCA). PCA is an unsupervised dimension-reduction method that identifies directions in X that maximize explained variance without using outcome information [33]. By contrast, PLS(-DA) is outcome-informed: its components are estimated to capture variation in X that is most relevant to Y. As a consequence, PCA components may not align with group structure, whereas PLS-DA components are explicitly oriented toward discrimination [60].

2.2.4 Multi-way PLS and Tucker-3 decomposition

Multi-way extensions of PLS exist, including N-way PLS (N-PLS), which generalizes PLS to multi-way arrays and aims to preserve multi-way structure in the modelling step [7]. A widely used practical alternative is to reshape a three-way tensor into a two-dimensional matrix via unfolding and then apply a standard PLS model on the unfolded representation, while retaining the original tensor structure for downstream organization and interpretation. In this thesis, supervised modelling follows this unfolding-based strategy, whereas tensor structure is preserved where it is most beneficial for data handling.

In longitudinal multi-omics, several preprocessing tasks also benefit from operating directly on the tensor structure, particularly missing-data handling. Tucker-3 decomposition (three-mode factor analysis) provides a principled low-rank representation of a three-way array, $X \approx G \times_1 A \times_2 B \times_3 C$, where A, B, and C are

mode-specific factor matrices and G is a core tensor capturing interactions among components [69, 36]. Tucker models are commonly estimated using alternating least squares (ALS) [36], and they can be leveraged in iterative imputation schemes in which missing entries are updated from Tucker reconstructions while observed values are kept fixed, yielding imputations that remain coherent across subjects, features, and time/blocks. The following Methods chapter details the specific workflows, software, and parameters used throughout the thesis.

3

Methods

This chapter describes the computational and statistical methods used in this thesis to generate the results reported in the subsequent chapters. For each analysis, methods are presented in a reproducibility-oriented manner by specifying the input data, the processing and quality-control steps, the software tools and their versions, and the key parameters and decision criteria adopted to obtain the final outputs.

The Results section is organised into three complementary parts: BRPF1 (integrative functional genomics), DNA methylation–RNA-seq integration in endocrine resistance models (pathway-first framework), and TensorPLS (methodological development for tensor-structured longitudinal data). Accordingly, this Methods chapter is structured into corresponding sections that detail the workflows used to derive the results for each part, including any shared preprocessing steps and downstream analyses specific to each study component.

3.1 BRPF1 multi-omics analyses

This section describes the computational workflow used to analyse BRPF1 and ERα ChIP-seq binding profiles and ATAC-seq chromatin accessibility data generated under BRPF1 perturbation. The aim is to (i) define the BRPF1 cistrome genome-wide, (ii) quantify overlap with ERα regulatory elements, and (iii) test whether BRPF1 inhibition induces accessibility changes concentrated at ERα-associated loci. All analyses were performed on the hg38 genome build.

3.1.1 ChIP-seq analyses (BRPF1 and ERα)

Data sources. ERα ChIP-seq data were obtained from Nassa *et al.* [47]. BRPF1 ChIP-seq analyses were performed starting from raw FASTQ files.

Quality control, alignment, and duplicate removal. Read-level quality control was performed using FastQC [2]. Reads were aligned to the hg38 reference genome using Bowtie2 [37], allowing up to one mismatch and retaining uniquely mapping reads. PCR duplicates were removed using Picard tools v2.9.0 (`MarkDuplicates`) [8].

Peak calling and replicate integration. For each biological replicate, peak calling was performed using MACS2 [75] using the corresponding control sample and a p-value threshold of 0.05. Replicate peak sets were combined using MuSERA [45] (replicate type: *biological*), retaining peaks supported by at least two biological replicates (C:2). When multiple regions from one sample intersected regions from another sample, the lowest p-value was retained.

Peak overlap and co-occupancy definition. Overlap between ERα and BRPF1 peak sets was computed using `bedtools intersect` [58]. Peaks were considered co-occupied if their genomic intervals overlapped by at least 1 bp. The resulting intersected regions defined the ERα–BRPF1 co-bound subset used for downstream analyses.

Annotation, quantification, and motif enrichment. Peak annotation was performed using HOMER (`annotatePeaks.pl`) [23]. Comparison, integration, and quantification of BRPF1 and ERα binding signals across peak sets were performed using seqMINER [73] with K-means clustering ($k = 3$, selected by the elbow method). Overrepresented sequence motifs were determined using PScan-ChIP [74] with motif descriptors from the Transfac database; only motifs with p-value ≤ 0.05 were considered.

3.1.2 ATAC-seq analyses (BRPF1 inhibition)

Quality control, trimming, and alignment. Quality control of ATAC-seq reads was performed using FastQC, and adapter sequences were removed using Trimmomatic v0.38 [6]. Reads were aligned to hg38 using Bowtie2 [37] with parameters `-very-sensitive -no-discordant -X 2000`. Duplicate reads were removed using Picard `MarkDuplicates`, and mitochondrial reads were removed using Samtools [39].

Peak calling (MACS2) and filtering. Open chromatin regions were identified using MACS2 [75] with parameters `-q 0.05 -nomodel -shift 75 -extsize 150 -call-summits -keep-dup all`. Peaks overlapping ENCODE blacklist regions were excluded [1].

Differential accessibility. Differential accessibility analysis was performed using DiffBind [63]. Only regions with FDR ≤ 0.05 were retained for downstream analyses.

3.1.3 RNA-seq analyses (BRPF1 perturbation)

Reference genome and gene model consistency. RNA-seq reads were aligned to the same reference genome assembly used for ATAC-seq and ChIP-seq analyses (hg38). Gene models were defined using GENCODE Release 40, ensuring consistency of gene identifiers for integration with peak-to-gene assignments.

Read processing, quantification, and differential expression. Raw FASTQ files underwent quality control using FastQC, and adapter sequences were removed using Trimmomatic v0.38. Filtered reads were aligned to hg38 using STAR v2.7.10b [17] with standard parameters considering genes present in GENCODE Release 40. Gene-level quantification was performed using featureCounts [40]. Differential expression was computed using DESeq2 [43]. A transcript was considered expressed if supported by at least 10 raw reads. Differential expression was defined as $|\mathrm{FC}| \geq 1.5$ with Benjamini–Hochberg adjusted p-value ≤ 0.05. RNA-seq results were integrated with ATAC-seq and ChIP-seq by matching gene identifiers derived from GENCODE Release 40.

3.2 DNA methylation and gene expression in endocrine resistance models

This section integrates genome-wide DNA methylation (EPIC v2) and transcriptomic (RNA-seq) profiles from estrogen receptor–positive breast cancer cell line models encompassing both tamoxifen-derived resistance models and fulvestrant/ICI-associated conditions. Global methylation analyses (including the directionality of events and their genomic localization) were performed across both endocrine therapy contexts to capture shared and context-specific features of resistance. By contrast, pathway-level analyses based on RNA-seq (GSEA and leading-edge–based integration with promoter methylation) were restricted to the tamoxifen-derived resistant

models, which serve here as a focused case study to demonstrate the pathway-first cross-omics framework. Importantly, the resulting workflow is modular and generalizable, and can be directly extended to other endocrine resistance settings (including fulvestrant/ICI) when appropriately matched transcriptomic datasets are available.

DNA methylation profiling was generated across two independent EPIC array runs. Run membership and model nomenclature (standardized as `CellLine/Condition`) are summarised in Table 3.1. All methylation preprocessing, normalization, and downstream analyses were performed within-run using an identical workflow to minimize run-specific technical variability. RNA-seq data were available for all models considered here except `MCF7L/TAMR`. RNA-seq profiles originated from two independent runs: the contrast between tamoxifen-resistant MCF7/TAMR and parental MCF7/WT under tamoxifen treatment was produced in Run 1, whereas all remaining RNA-seq comparisons were produced in Run 2. Because no cell line/condition was profiled with biological replicates across both RNA-seq runs (i.e., no bridging samples), run effects are not identifiable separately from biological differences; therefore, differential expression and downstream analyses were performed within-run to avoid confounding-driven batch correction.

Table 3.1: Cell line models included in this chapter, grouped by EPIC array run. Nomenclature is standardized as `CellLine/Condition`.

Model (standard)	**Therapy context**	**RNA-seq**
EPIC Run 1		
BT474/TAM1	Tamoxifen	Yes
ZR-75-1/TAM2	Tamoxifen	Yes
T47D/TAM2	Tamoxifen	Yes
MCF7/TAMR	Tamoxifen	Yes
MCF7L/TAMR	Tamoxifen	No
MCF7/NA-ICIR	Fulvestrant/ICI	Yes
T47D/NA-ICIR	Fulvestrant/ICI	Yes
EPIC Run 2		
T47D/TR1	Tamoxifen	Yes
MCF7/TAMR4	Tamoxifen	Yes
MCF7/164R4	Fulvestrant/ICI	Yes
MCF7/182R-1	Fulvestrant/ICI	Yes
MCF7/182R-6	Fulvestrant/ICI	Yes

3.2.1 DNA methylation data preprocessing (Illumina Infinium MethylationEPIC v2.0 (EPIC v2) arrays)

Genome-wide DNA methylation profiling was performed using Illumina Infinium MethylationEPIC arrays (EPICv2; ∼950K probes), and analyses were carried out on the **hg38** genome build. For each run, raw `IDAT` files were imported and processed in `R` using the `ChAMP` pipeline. Import was performed starting from raw red and green channel intensities, generating methylated and unmethylated signal matrices and deriving both β values and M values. In this study, probes mapping to autosomes and sex chromosomes were retained (i.e., probes on chromosomes `X` and `Y` were included), allowing genome-wide summaries to reflect the full EPIC probe content.

Quality filtering was applied to each run separately using `ChAMP`'s `champ.filter()` after import. Probes were removed if they failed the detection P-value criterion (`detP` > 0.01), if they had low beadcount (`beadcount` < 3 in at least 5% of samples), if they were non-CpG probes (`filterNoCG`), or if they overlapped common SNPs according to the general mask options provided by `ChAMP`'s EPICv2 annotation (`filterSNPs`). As part of the filtering step, β values were additionally constrained to the open interval $(0, 1)$ by replacing values ≤ 0 with the smallest positive value and values ≥ 1 with the largest value below 1, preventing downstream numerical instabilities. Probe filtering retained 890,311 probes in Run 1 and 892,520 probes in Run 2, defining the run-specific methylome coverage used in all downstream analyses.

Normalization was performed on β values using BMIQ (Beta MIxture Quantile normalization) [65], as implemented in ChAMP, to reduce bias between type I and type II probe designs on EPIC arrays. After normalization, potential batch effects were evaluated separately within each run using ChAMP's SVD-based diagnostics (`champ.SVD`). For Run 1, `champ.SVD` indicated a run-internal technical effect associated with the Sentrix Slide variable; therefore, ComBat [32] was applied to the BMIQ-normalized methylation values using Sentrix Slide as the batch covariate. For Run 2, ComBat adjustment was not applied because it empirically worsened the inferred data structure and reduced within-model coherence in `champ.SVD`-guided assessments; therefore, Run 2 analyses were conducted on BMIQ-normalized values without ComBat correction. Importantly, the two runs were processed using the same import, filtering, and normalization workflow, and all downstream analyses were performed within-run on the corresponding filtered and BMIQ-normalized probe sets.

3.2.2 Unsupervised analyses and global methylation summaries

To characterize global relationships among samples and assess replicate coherence, unsupervised hierarchical clustering was performed within each EPIC run using BMIQ-normalized β values.

For computational efficiency and to focus on the most informative signal, clustering was computed on the top N most variable CpGs, with $N = \min(20{,}000,\ P)$ where P is the number of probes available after preprocessing within each run. CpG-level variance was computed across samples, and the N CpGs with the highest variance were retained. Missing β values within the selected CpGs were imputed using the CpG-wise mean across samples. Hierarchical clustering was then performed on the sample-by-CpG matrix using correlation distance (defined as $1 - \rho$, where ρ is the Pearson correlation between sample methylation profiles) computed via `pdist(metric="correlation")`, and agglomeration was carried out using *complete* linkage (`linkage(method="complete")`) [70].

In addition to clustering, global summaries were computed to describe the directional bias of large methylation changes across comparisons. For each comparison, we first restricted to statistically significant events (q-value < 0.05) and then defined as *extreme* only those with large effect size, i.e., $\Delta\beta > 0.20$ (hypermethylated) or $\Delta\beta < -0.20$ (hypomethylated). Significant events not exceeding these $\Delta\beta$ thresholds were not considered *extreme* and were therefore excluded from the proportion calculation.

For each comparison, the total number of extreme events was computed as $N_{\text{extreme}} = N_{\text{hyper}} + N_{\text{hypo}}$, where N_{hyper} is the number of extreme hypermethylated events ($\Delta\beta > 0.20$) and N_{hypo} is the number of extreme hypomethylated events ($\Delta\beta < -0.20$). The proportions reported in the corresponding figures represent the relative contribution of each class among extreme events: $\%\text{Hyper} = (N_{\text{hyper}}/N_{\text{extreme}}) \times 100$ and $\%\text{Hypo} = (N_{\text{hypo}}/N_{\text{extreme}}) \times 100$.

3.2.3 Genomic annotation of differential methylation

To characterize the genomic context of differential methylation events, CpG probes (and, where applicable, aggregated regions) were annotated using **HOMER** (`annotatePeaks.pl`) with the **hg38** reference genome. Annotation was performed in two complementary steps: (i) assignment to the nearest transcription start site (TSS), and (ii) classification of the genomic feature overlapping the *center* of each locus. For the TSS-based assignment, HOMER computes the distance to the closest TSS (negative values upstream and positive values downstream) and links each locus to the correspond-

ing gene identifier(s). Genomic feature labels were obtained through HOMER's basic genome annotation, yielding categories such as promoter/TSS, TTS, exonic, intronic, and intergenic.

In this work, promoter regions were defined according to HOMER's standard TSS window, i.e., from -1 **kb to** $+100$ **bp** relative to the annotated TSS. Importantly, rather than relying solely on HOMER's default RefSeq TSS definitions, we provided a custom transcript annotation using the GTF option. Specifically, we used `gencode.v47.basic.annotation.gtf` as the reference gene model (`-gtf gencode.v47.basic.annotation.gtf`), so that TSS positions and gene-structure features (TSS/TTS/exons/introns) used in the basic annotation were derived from GENCODE v47. This ensured uniform gene mapping across methylation- and transcriptomics-based analyses.

3.2.4 RNA-seq read processing, quantification, and differential expression

RNA-seq analyses were performed starting from raw `FASTQ` files. Adapter trimming was carried out using `cutadapt` [44], and reads shorter than 20 nucleotides after trimming were discarded to reduce potential alignment artifacts. Trimmed paired-end reads were aligned to the **hg38** reference genome using `STAR`. Gene models were provided via the same transcript annotation adopted for methylation analyses, `gencode.v47.basic.annotation.gtf`, ensuring a consistent reference gene definition across omics layers. Unless otherwise stated, STAR was run with standard settings and produced unsorted BAM alignments (`-outSAMtype BAM Unsorted`).

Aligned reads were summarised at the gene level using `featureCounts`, using the same GTF file (`gencode.v47.basic.annotation.gtf`) to assign reads to annotated genes. This procedure produced a gene-by-sample raw count matrix for each RNA-seq batch/run. Given the run-specific structure of the RNA-seq dataset and the absence of bridging biological replicates spanning runs (i.e., no sample type was represented with replicates in both runs), differential expression analyses were performed within-run.

Differential gene expression was computed using the standard `DESeq2` workflow. Biological replicates were modeled as belonging to the same experimental condition, and contrasts were defined according to the resistant versus matched control comparisons described in the Dataset Overview. `DESeq2` size-factor normalization and dispersion estimation were applied, and differential expression was tested using the Wald test. For each comparison, `DESeq2` outputs included $\log_2$ fold changes, Wald test statistics, and Benjamini–Hochberg adjusted q-values. Ranked gene lists for

preranked pathway analyses were derived from the `stat` column (Wald statistic) of the corresponding `DESeq2` result tables, as detailed in the following subsection.

3.2.5 Pathway analysis and construction of consensus leading-edge gene sets in tamoxifen-resistant models

To identify biological processes recurrently associated with tamoxifen resistance despite inter–cell line heterogeneity at the single-gene level, we leveraged a fully automated analysis pipeline developed in-house to support the entire RNA-seq–to–methylation integration workflow described in this chapter. Within this pipeline, preranked GSEA is used as the entry point to define pathway-level signatures that are subsequently connected to promoter DNA methylation and tested for cross-omics concordance. Pathway-level analyses were intentionally restricted to the tamoxifen-derived resistant models, consistent with the main biological focus of this thesis, while the overall workflow is general and can be extended to additional endocrine settings when matched transcriptomic data are available.

For each tamoxifen-resistant comparison, a preranked gene list (`.rnk`) was generated from DESeq2 results using the Wald test statistic (`stat`) as the ranking metric. The pipeline then executed `GSEAPreranked` in batch mode across all ranking files using the MSigDB Hallmark collection (GMT file `h.all.v2025.1.Hs.symbols.gmt`), producing standardized enrichment reports for each cell line comparison. GSEA results were filtered at FDR < 0.25 [64] and stratified by direction of enrichment into positively enriched pathways (Pos; NES > 0) and negatively enriched pathways (Neg; NES < 0). To prioritize signals conserved across models, we defined as *conserved/recurrent pathways* those that were significant (FDR < 0.25) and showed the same enrichment direction (Pos or Neg) in at least three tamoxifen-resistant cell lines.

To build pathway-specific gene cores for downstream integration, the same pipeline extracted leading-edge genes for each pathway and each cell line by parsing the corresponding GSEA output tables and selecting genes annotated as `CORE ENRICHMENT = Yes`. For each conserved pathway, gene-level frequencies (i.e., the number of cell lines in which a gene appears in the leading edge) were computed across all tamoxifen-resistant models. Finally, *consensus leading-edge gene sets* were defined by retaining only genes appearing as leading-edge in at least two cell lines for that pathway (MIN_SUPPORT ≥ 2). These consensus sets represent robust pathway cores and constitute the gene-level input used in subsequent steps of the pipeline to integrate promoter methylation and gene expression and to test whether cross-omics concordant genes are enriched within conserved pathways.

Collapsing promoter DMPs to gene-level promoter summaries. To enable cross-omics integration at the gene level, probe-level differential methylation results were collapsed into a *gene-promoter* representation for each methylation comparison. First, significant CpG-level events (FDR-controlled) were restricted to probes annotated to promoter/TSS regions according to the HOMER definition used in this thesis (TSS window: -1 kb to $+100$ bp; hg38; GENCODE v47 GTF). For each gene and comparison, all promoter-associated DMPs were then aggregated to produce a compact promoter summary including: (i) the number of significant promoter DMPs linked to the gene (`n_promoter_DMP`); (ii) the strongest promoter-level statistical evidence, defined as the minimum adjusted p-value/FDR among the gene's promoter DMPs (`min_FDR_promoter`); and (iii) the maximum absolute effect size observed in the promoter (`max_abs_deltaBeta_promoter`).

Representative promoter effect size and direction. Because multiple CpGs can map to the same promoter, a single *representative* promoter methylation change was defined to capture directionality for integration. Specifically, for each gene and comparison, we selected the promoter DMP showing the largest absolute methylation difference ($|\Delta\beta|$) and used its signed $\Delta\beta$ as the representative promoter change (`deltaBeta_rep`); the corresponding adjusted significance value was stored as `adjP_rep`. This choice yields a directionally interpretable summary that reflects the strongest promoter-associated methylation shift for the gene while retaining the associated statistical support. Based on the sign (and, where applicable, a minimal magnitude criterion), each gene was classified into a promoter direction category (`promoter_status`): `hyper` (positive `deltaBeta_rep`), `hypo` (negative `deltaBeta_rep`), or `small` (remaining cases with limited promoter effect), and downstream integration focused primarily on the `hyper` and `hypo` classes.

For reproducible joins with RNA-seq and pathway gene sets, gene identifiers were standardized by trimming whitespace and converting to upper case (`gene_norm`), and all gene-level promoter summaries across comparisons were stored in an aggregated table (`ALL_COMPARISONS__gene_promoter.tsv`) used as input for the integration workflow.

3.2.6 Pathway-level enrichment of concordant genes

The integration analysis described above identifies genes showing concordant promoter methylation–expression changes (e.g., `hyper+down` or `hypo+up`) within each comparison [15, 5]. Here, we address a pathway-driven question: are concordant genes overrepresented within the conserved pathway cores identified by GSEA (con-

sensus leading-edge gene sets), compared with what would be expected by chance given the set of genes measurable in both omics layers? This enrichment framework allows us to prioritize pathways whose leading-edge (i.e., the transcriptional core driving the GSEA signal) is disproportionately supported by promoter methylation changes in the expected direction.

Background universe. To avoid bias from genes not measurable in one of the two platforms, the statistical background universe (U) was defined as the intersection between (i) all genes contained in the RNA preranked lists used for GSEA (union of genes across the `.rnk` files) and (ii) all genes present in the gene-level promoter methylation summary table. All subsequent sets (leading-edge cores and concordant genes) were restricted to U.

Pathway gene sets (consensus leading-edge). For each conserved pathway, the pathway core A was defined as the set of *consensus leading-edge* genes, constructed by retaining only genes that appeared as leading-edge (`CORE ENRICHMENT = Yes`) in at least two tamoxifen-resistant cell lines. The resulting consensus sets were then intersected with U prior to testing.

Definition of concordant gene sets. Concordant genes (C) were obtained from the gene-level RNA-seq/methylation integration under explicit RNA-seq filtering criteria: genes were considered differentially expressed if `baseMean` > 1, `padj` < 0.05, and $|\log_2 \mathrm{FC}| \geq 0.5$ (upregulated if $\log_2 \mathrm{FC} \geq 0.5$; downregulated if $\log_2 \mathrm{FC} \leq -0.5$). These DE calls were combined with promoter methylation direction to define concordance classes (`hypo+up` and `hyper+down`). Pathway enrichment was tested using a direction-specific concordant set depending on the analysis branch; for example, for negatively enriched pathways (Leading Neg) we tested enrichment of `hyper+down` genes. Concordant genes were finally restricted to U.

Fisher exact test. For each pathway, a 2×2 contingency table was constructed using A (consensus leading-edge genes in U) and C (concordant genes in U). Let $a = |A \cap C|$ denote the number of concordant genes within the pathway core, $b = |A \setminus C|$ the number of pathway-core genes that are not concordant, $c = |C \setminus A|$ the number of concordant genes outside the pathway core, and $d = |U \setminus (A \cup C)|$ the remaining genes in the universe. Enrichment of concordant genes within the pathway core was assessed using a one-sided Fisher exact test (*alternative* = "greater"), i.e., testing whether a is larger than expected given the margins. Effect size was summarised

as an odds ratio; to ensure numerical stability when any cell count was zero, odds ratios were computed with a small pseudocount (0.5) added to all cells.

Multiple testing correction and reporting. p-values were adjusted across pathways using the Benjamini–Hochberg procedure, yielding BH-FDR values. For each pathway, we reported the size of the consensus leading-edge set, the number of concordant genes observed within the pathway, the size of the concordant gene set, the universe size, and the list of concordant genes contributing to the enrichment signal.

3.2.7 Cross-line directionality branches

The pathway-level enrichment framework above treats concordance as an event that can be observed within at least one model. However, gene-level promoter methylation–expression relationships may vary across cell lines even when pathways are recurrent. Here, we address the robustness question: *do the pathway enrichments persist when we require that gene-level concordance is directionally consistent across multiple cell line comparisons?* To answer this, we introduced branch-based gene filters that impose increasingly stringent cross-line directionality constraints before repeating the same Fisher enrichment analysis.

Input and unit of cross-line support. Branching was applied to the gene-level concordance table produced by the integration pipeline, which reports for each gene the concordance class (`hyper+down` or `hypo+up`) and the corresponding model-specific comparison identifier. Cross-line support was evaluated using the comparison identifier (i.e., each `methylation_comparison` was treated as one cell line–specific unit). For each gene, we counted in how many distinct comparisons it was observed as `hyper+down` ($n_{\text{hyperdown}}$) and in how many distinct comparisons it was observed as `hypo+up` (n_{hypoup}), and defined $n_{\text{total}} = n_{\text{hyperdown}} + n_{\text{hypoup}}$.

Branch definitions. Three directionality branches were defined:

- **Strict.** A gene was retained if it showed concordance in at least two comparisons ($n_{\text{total}} \geq 2$) and remained non-divergent (no evidence of the opposite concordant direction in any other comparison). This defines a high-confidence set with directionally stable concordance across models.
- **Majority.** A gene was retained if it showed concordance in at least two comparisons ($n_{\text{total}} \geq 2$) and a single concordant direction was dominant across

comparisons. Dominance was defined by requiring that the most frequent direction (`hyper+down` or `hypo+up`) accounts for at least a fraction $\geq 2/3$ of the gene's concordant occurrences ($\max(n_{\text{hyperdown}}, n_{\text{hypoup}})/n_{\text{total}} \geq 2/3$), excluding ties.

For each branch, we derived gene sets in three variants: a combined set including both directions, and two direction-specific sets (`hyper+down` only and `hypo+up` only), enabling branch-aware enrichment tests aligned to the pathway direction (e.g., `hyper+down` for Leading Neg pathways and `hypo+up` for Leading Pos pathways).

Branch-aware enrichment testing. For each branch-derived gene set, we repeated the pathway-level enrichment analysis using the same framework as in Section 3.2.6. The background universe U was kept identical (genes present in the RNA preranked lists used for GSEA intersected with genes present in the promoter methylation gene-level table), and pathway cores A were defined as consensus leading-edge gene sets (obtained by retaining only those leading-edge genes supported by at least two comparisons) restricted to U. For each pathway, a one-sided Fisher exact test (*greater*) was computed on the corresponding 2×2 contingency table, odds ratios were calculated with a 0.5 pseudocount for numerical stability, and p-values were corrected across pathways using the Benjamini–Hochberg procedure. This branching strategy therefore acts as a robustness analysis: pathways that remain significant under stricter branches identify a conservative set of signals, and the corresponding branch-retained genes define a high-confidence *core* of directionally stable concordant genes. Together, these retained pathways and their conserved gene cores represent the most robust cross-omics evidence, supported consistently across multiple cell line comparisons.

3.2.8 Software, parameters, and reproducibility

Computational environment and software versions. All analyses were performed using R and Python, combining established bioinformatics tools for methylation arrays, RNA-seq processing, and pathway analysis. Key external software and versions were: `ChAMP` v2.29.1 [66] (EPIC v2 processing and filtering), `HOMER` v5.1 (`annotatePeaks.pl` for genomic annotation on hg38 with `gencode.v47.basic.gtf`), `cutadapt` v5.1 (adapter trimming), `STAR` v2.7.11b (RNA-seq alignment to hg38 using `gencode.v47.basic.annotation.gtf`), and `featureCounts` v2.0.3 (gene-level read counting with the same GTF). Preranked pathway enrichment was performed using the GSEA command-line v4.3.2 (`GSEAPreranked`) with the MSigDB Hallmark gene sets (`h.all.v2025.1.Hs.symbols.gmt`).

Key parameters. Unless otherwise stated, the main thresholds and settings used throughout this chapter were: detection P-value filtering at 0.01; beadcount filtering < 3 in at least 5% of samples; BMIQ normalization; inclusion of sex-chromosome probes; extreme methylation defined as q-value < 0.05 and $|\Delta\beta| > 0.20$; unsupervised clustering on the top 20,000 most variable CpGs using correlation distance $(1 - \rho)$ and complete linkage with CpG-wise mean imputation for missing values; RNA-seq adapter trimming with minimum read length 20, alignment with STAR to hg38, and gene-level counting with featureCounts. For gene-level cross-omics integration, RNA-seq genes were filtered as `baseMean`> 1, `padj`< 0.05, and $|\log_2 \mathrm{FC}| \geq 0.5$. Conserved pathways were defined as significant (FDR < 0.25) with the same enrichment direction in at least three cell line models, and consensus leading-edge genes were defined as present in at least two models (MIN_SUPPORT ≥ 2). Enrichment tests were performed using one-sided Fisher exact tests (*greater*) with Benjamini–Hochberg correction, and odds ratios were computed with a 0.5 pseudocount for numerical stability.

Code availability and workflow traceability. All scripts used to generate the integration results—including construction of preranked inputs, automated GSEA execution, extraction and consensus construction of leading-edge gene sets, promoter-level methylation collapsing to gene-level summaries, RNA-seq/methylation concordance classification, Fisher enrichment analyses, and the branch-based robustness framework—are provided in a dedicated Git repository (`https://github.com/alejanner/MultiEnrich`). This repository contains the complete codebase required to reproduce each figure and table reported in the integration part of this thesis, together with the corresponding configuration files and standardized intermediate outputs (e.g., `.tsv` and `.rnk` tables) enabling end-to-end traceability.

3.3 TensorPLS: development and evaluation

This section describes the methodological development and evaluation of `TensorPLS`, an R package developed in this thesis to support general-purpose PLS-based discriminant analysis on longitudinal and multi-omics data naturally represented as three-way tensors (e.g., *subjects* $\times$ *features* $\times$ *time*). The package is designed to address five practical questions in supervised multi-way settings: (i) whether groups can be discriminated, (ii) which variables drive discrimination, (iii) how cross-validated model behaviour (Q^2) changes under resampling for tuning and comparison, (iv) how much variation is captured (R^2), and (v) which time points or blocks contribute most

to the discriminant components.

3.3.1 Software availability, dependencies, and installation

`TensorPLS` is distributed as an R package and can be installed from GitHub using standard development tools. The package depends on several R packages that are installed automatically during installation; however, `mixOmics` is distributed via Bioconductor and must be installed prior to installing `TensorPLS`. The complete source code, documentation, and example datasets are available in the project repository `https://github.com/alejanner/TensorPLS/`.

3.3.2 Data structures and tensor construction

`TensorPLS` operates natively on 3D arrays (tensors) encoding three modes, with the default convention *subjects* $\times$ *features* $\times$ *time/blocks*. Input omics tables are converted into tensors using a dedicated preprocessing function (`prepare_omics`), which standardizes identifiers, aligns time points, and (optionally) filters samples based on an external cohort metadata table. The function supports common input formats (e.g., CSV) and includes utilities to handle wide versus long layouts (including forced transposition when needed) and robust numeric coercion. The output of this step is a three-way array X with dimensions matching the expected multi-way structure.

In supervised settings, the outcome Y is encoded as a 3D array aligned to X, with dimensions *subjects* $\times$ 1 $\times$ *time*. In the binary case, Y represents class membership (e.g., 0/1) and preserves the temporal axis to enable consistent multi-way handling across the workflow.

3.3.3 Datasets and preprocessing for evaluation

`TensorPLS` was evaluated on four longitudinal datasets derived from the TEDDY study, each organised as a three-way tensor (*subjects* $\times$ *variables* $\times$ *time*). Across modalities, measurements were available at five aligned time points (0, -3, -6, -9, -12 months); time 0 was defined as the date of IA confirmation for each case, and negative values indicate months preceding this event. To limit missingness-driven uncertainty, only individuals with observations in at least 3 of the 5 time points were retained. The analysed modalities were: (i) whole-blood gene expression, (ii) plasma metabolomics (GCTOFX), (iii) positive dietary biomarkers, and (iv) negative dietary biomarkers.

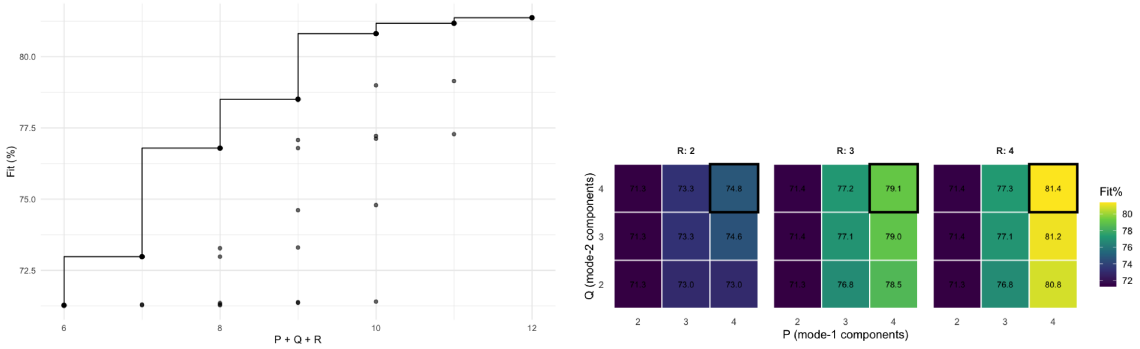
Missing entries were imputed on the three-way tensors using Tucker-3 imputation (Section 3.3.4).

3.3.4 Missing data handling and Tucker-based imputation

Longitudinal multi-omics tensors often contain missing entries due to incomplete follow-up or assay-specific dropouts. To address this, `TensorPLS` implements an imputation step based on an *iterative Tucker-3 decomposition* fitted via ALS (Alternating Least Squares). Importantly, imputation is performed directly on the 3D structure, thereby preserving the multi-way organization when estimating missing values.

Given an input tensor X (*subjects* $\times$ *variables* $\times$ *time*), missing values are first initialized by sampling from a normal distribution whose mean and standard deviation are estimated from the observed entries (with safeguards for degenerate variance). The algorithm then iteratively (i) fits a Tucker-3 model with a user-specified number of components for each mode (`fac` $= (P, Q, R)$), (ii) reconstructs the full tensor $\hat{X}$, and (iii) updates only the missing positions of X with the corresponding entries of $\hat{X}$. Iterations continue until the relative change in the Tucker model fit falls below a convergence threshold (`conver`, default 10^{-7}) or a maximum number of iterations is reached (`max.iter`, default 1000). For reproducibility, an optional random seed can be provided (`seed`). The number of components is automatically clipped to the tensor dimensions to prevent rank specifications exceeding the size of any mode.

The selection of Tucker mode ranks represents a key modeling choice. We used two complementary diagnostics. First, a Pareto (elbow) plot summarises the trade-off between overall model complexity and reconstruction quality by evaluating candidate models across increasing values of the total number of components ($P+Q+R$); the elbow region indicates where additional complexity yields diminishing returns (Fig. 3.1). Second, a heatmap-based grid search was used to select *which* ranks to allocate to each mode by comparing candidate triplets (P, Q, R) and choosing a balanced configuration that achieves high fit without unnecessary complexity (Fig. 3.1).



(a) Pareto (elbow) plot guiding the choice of the overall number of components ($P + Q + R$).

(b) Heatmap grid search guiding the choice of the specific Tucker ranks (P, Q, R).

Figure 3.1: Diagnostics used to select Tucker-3 imputation ranks. The Pareto plot supports the choice of *how many* total components to use, whereas the heatmap supports the choice of *which* ranks to allocate to each mode.

3.3.5 N-PLS-DA performance metrics and VIP representations

To quantify discriminant performance and obtain interpretable feature importance measures on tensor-structured data, `TensorPLS` provides the function `nplsda_vips()`, which implements a unified N-PLS-DA workflow and returns (i) cross-validated performance summaries (Q^2), (ii) explained variance (R^2), and (iii) multiple VIP (Variable Importance in Projection) representations. This enables answering four core questions in a single run: *(1) can the model discriminate groups? (2) how the model behaves under cross-validation (Q^2)? (3) how much variation is captured (R^2)?* and *(4) which features contribute most to discrimination (VIPs)?*

Input structure and model fitting. The predictors X are provided as a 3D array of size $n \times p \times k$ (*subjects* $\times$ *features* $\times$ *time/blocks*). The response Y can be supplied either as a matrix ($n \times q$) or as a 3D array; the function supports both the classic N-PLS-DA setting (single outcome) and an extended "regression class 2" setting (multi-response Y), with additional validation and error handling. Prior to model fitting, `nplsda_vips()` validates that X and Y share the same number of subjects, contain no missing values, and that the number of latent components satisfies $2 \leq$ `ncomp` ≤ 10 and does not exceed the statistical limit $\min(n-1, p)$. For model fitting, the tensor X is unfolded into a two-dimensional matrix ($n \times (p \cdot k)$), and a PLS model is fitted to the corresponding unfolded representation of Y using an internal PLS core (`pls_reg()`), returning fitted scores/loadings, cross-validated

predictions, and performance statistics.

For clarity, this is an *N-PLS-DA-style (tensor-input) workflow* in which the discriminant model is fitted on the unfolded representation using a standard PLS-DA core, and tensor-aware outputs are obtained by re-indexing results back to the feature–time structure.

Cross-validated performance (Q^2). Model performance is summarised using Q^2 estimated by cross-validation. In this thesis, Q^2 is used primarily as a resampling-based criterion for component tuning and for comparing modelling configurations. The function returns both a global unfolded Q^2 and a time-resolved summary computed slice-by-slice; slice-specific Q^2 curves are aggregated across time points to obtain mean 3D summaries (`NPLSDAQ2mean3D` and `NPLSDAQ2cummean3D`).

Explained variance (R^2). Explained variance is reported for both X and Y. In addition to global unfolded R^2 values (incremental and cumulative; `explvar`), the function computes a time-resolved explained-variance summary by fitting PLS models independently on each time slice and calculating incremental R^2_X and R^2_Y per component; these are then averaged across valid slices to produce a 3D mean summary (`Explvar3D`). Together, these outputs describe how much of the predictor structure (X) and class structure (Y) is captured by the latent components, both globally and across time.

VIP representations and interpretability. Feature importance is quantified using VIP scores in three complementary representations, designed to address distinct interpretability questions. In the PLS core used by `TensorPLS`, VIP values are computed from the PLS weight matrix $W = [w_{k,a}]$ and a component-wise measure of explained Y-variation, $R2y_a = \mathrm{mean}_q(\mathrm{cor}(Y_{\cdot q}, t_a)^2)$, where t_a denotes the score vector of component a. The VIP for feature k at component depth h is computed cumulatively as:

$$VIP_{k,h} = \sqrt{p \cdot \frac{\sum_{a=1}^{h} w_{k,a}^2 \, R2y_a}{\sum_{a=1}^{h} R2y_a}},$$

where p is the number of predictors. This formulation assigns higher VIP scores to features that have large weights on components that contribute strongly to explaining the response.

- **VIP2D** (`VIP2D`): a matrix indexed by *feature–time* pairs and latent components. Each row corresponds to a specific feature f at a specific time point t (encoded as `feature_time`), and each column corresponds to a component

h. VIP2D therefore answers: *on component h, how important is feature f at time t?*

- **VIP3D Model 1** (`VIP3Dmodel1`): a feature-by-component matrix obtained by aggregating VIP2D across time (by default averaging across a component-aligned subset of time points). VIP3D Model 1 answers: *on component h, how important is feature f on average across time?* This representation is useful to summarise component-specific drivers while reducing time granularity.

- **VIP3D Model 2** (`VIP3Dmodel2`): a feature-by-time matrix obtained by aggregating VIP2D across components (by default summing VIPs across components). VIP3D Model 2 answers: *at time t, how important is feature f overall across the discriminant components?* This representation is useful to highlight time windows in which specific features contribute strongly, irrespective of component identity.

An optional `slice_vip` mode is available to compute VIP summaries from separate PLS fits performed independently at each time slice, which can be used to emphasize time-specific importance patterns. In all cases, the returned VIP objects are aligned to the original tensor dimension names to support downstream feature selection, visualization, and robustness analyses.

3.3.6 Component tuning and model evaluation (R^2 and Q^2)

While the Tucker-3 imputation ranks `fac` $= (P, Q, R)$ were selected as discussed in Section 3.3.4, here we tune the number of latent components `ncomp` used in the discriminant PLS(-DA) step. The number of latent components (`ncomp`) is tuned to balance discriminative structure, model parsimony, and resampling-based predictive behaviour. `TensorPLS` provides functions that estimate Q^2 trajectories as a function of the number of components; the selected `ncomp` corresponds to the elbow region where additional components yield diminishing gains in Q^2.

Model quality is summarised using explained variance measures (R^2) and Q^2 computed under cross-validation.

To visualize class separation, `TensorPLS` computes N-PLS variates (scores) and produces score plots for selected component pairs. These plots represent each subject in the latent space and enable qualitative assessment of discrimination, complementing the quantitative R^2/Q^2 metrics.

3.3.7 VIP-based feature selection

To improve interpretability and, in high-dimensional settings, potentially enhance discriminant performance, `TensorPLS` implements a VIP-driven feature selection procedure through the `feature_selection()` function. The method is designed to be agnostic to the VIP representation: it can be applied to `VIP2D`, `VIP3Dmodel1`, `VIP3Dmodel2`, or to any compatible VIP matrix (including VIP outputs from `mixOmics`). The procedure is recommended especially for large omics datasets, where many variables are weakly informative and may dilute the discriminant signal.

Given a VIP matrix with variables in rows and one or more columns (e.g., components or time points), feature selection is performed by thresholding VIP scores at a chosen percentile thr $\in [0, 100]$ (default: 95th percentile). Optionally, selection can be restricted to a subset of columns (parameter `cols`), for example to focus on a specific component in `VIP2D`. For each selected column, the thr-percentile VIP value is computed and each variable is marked as "passing" if its VIP score is greater than or equal to that threshold. When multiple columns are considered, variable retention is determined by an aggregation rule (`aggregator`): (i) `any` retains a variable if it passes the threshold in at least one column (logical OR), (ii) `all` retains a variable only if it passes in all selected columns (logical AND), and (iii) `mean` computes the row-wise mean VIP across the selected columns and retains variables whose mean VIP lies above the thr-percentile of the mean distribution. This design enables both permissive selection (capturing variables that are strongly important in at least one component/time) and stringent selection (prioritising variables consistently important across multiple columns).

Because time-resolved VIP representations may encode variables with a time suffix (e.g., `gene|t3` or similar conventions), `feature_selection()` supports optional harmonization of feature identifiers by removing a user-defined suffix via a regular expression (`strip_time = TRUE`). By default, a pattern matching the terminal time tag is removed, and the resulting feature list is deduplicated. This step facilitates comparison and combination of selected sets across VIP views, such as computing intersections or unions between VIP2D-derived and VIP3D-derived selections.

In the evaluation workflow used in this thesis, VIP-based selections were used to define reduced feature sets that were then used to refit N-PLS-DA models and to compare model behaviour (R^2 and Q^2) across alternative selections (e.g., intersection/union between VIP-derived sets). This approach supports a transparent trade-off between model compactness and interpretability while retaining time-aware interpretability when VIP2D is used.

3.3.8 Time contribution analysis

In longitudinal tensor data, discrimination between groups may be driven more strongly by specific time points than by others. To quantify when the discriminant signal is most prominent, `TensorPLS` evaluates the contribution of the temporal mode (Mode 3) through the time loadings estimated by the N-PLS-DA model. Conceptually, the model learns latent components jointly across subjects and variables, while assigning each time point a coordinate (loading) on each component, thereby capturing how the temporal slices contribute to the discriminant structure.

Time contributions were assessed using the Mode 3 loadings (time weights), denoted W_{Mode3}, which provide one loading per time point and component. To facilitate interpretation, time points were projected onto the space of two selected components (by default components 1 and 2) and visualized as labelled points (`plot_nplsda_blockX_mode3`). In this representation, time points farther from the origin have a larger influence in shaping the discriminant components and therefore contribute more strongly to group separation, whereas time points near the origin contribute minimally. The relative position (direction and sign) of time point coordinates indicates whether time points contribute similarly or oppositely to the latent components, enabling identification of temporal windows that most strongly drive discrimination in each omics modality.

3.3.9 VIP2D robustness assessment by permutation testing

Time-resolved VIP2D scores provide a ranked list of features for each time point and latent component, but high VIP values can occasionally arise from sampling noise or model-specific fluctuations. To quantify the stability of time-specific VIP2D signals, `TensorPLS` implements a permutation-based robustness assessment integrated in the visualization workflow (`plot_vip2d_with_groups_nogaps`).

For a chosen component, VIP2D values are first stratified by time point and the Top–N features with the highest VIP2D scores are selected within each time point (*with ties removed*). Robustness is then assessed by repeatedly permuting the sample labels in Y (i.e., shuffling class membership across subjects while keeping X unchanged), refitting the N-PLS-DA model, and recomputing VIP2D. At each permutation, Top–N features are re-extracted *within each time point* using the same rule as in the original fit. To reduce memory usage and focus the test on the reported results, the procedure tracks only whether the originally selected feature–time pairs reappear among the permuted Top–N lists.

For each feature–time pair in the observed Top–N, a Monte Carlo permutation

p-value is computed as

$$p_{\text{perm}} = \frac{c+1}{R+1},$$

where R is the number of permutations and c is the number of permutations in which the same feature–time pair appears again among the Top–N for that time point. Feature–time pairs with $p_{\text{perm}} < \alpha_{\text{perm}}$ (default $\alpha_{\text{perm}} = 0.05$) are highlighted as robust, indicating that their time-specific importance is unlikely to be recovered under random label assignments.

In addition to VIP ranking, the plotting function provides a compact directionality summary by comparing the mean feature value between the two groups (case vs. control) at each time point. This complementary panel is used only to indicate the direction of group differences underlying the VIP2D ranking and does not alter the VIP computation. Overall, this permutation framework enables transparent reporting of time-specific biomarkers by separating features that are highly ranked in a single fit from those that remain consistently recovered under label randomization.

4

Results

This chapter is organised into three sections. First, we present a multi-omics case study centred on BRPF1 to illustrate how integrating chromatin occupancy, accessibility, and transcription can reveal mechanistic features of ERα-linked regulation that are not evident from a single omics layer.

Second, we report a cross-omics framework applied to endocrine-resistance models that follows a *pathway-first* strategy: rather than prioritising individual genes in each model, we identify recurrently enriched pathways across models, define *consensus leading-edge* gene cores, and then test whether these pathway cores show disproportionate support from concordant promoter methylation–expression changes under explicit robustness constraints. This design aims to extract signals that are comparable across heterogeneous systems while reducing gene-level noise and model-specific variability.

Third, we introduce and evaluate `TensorPLS`, a general-purpose methodological contribution for PLS-based discriminant analysis on longitudinal multi-omics data represented as tensors, with an emphasis on time-aware interpretability and feature prioritisation.

4.1 BRPF1 as a multi-omics case study of ERα chromatin coregulation

BRPF1 was prioritized as an ERα-associated essential factor through a previously established pipeline integrating luminal breast cancer dependency profiles, ERα-centric regulatory information, and tumour expression patterns. In this thesis, BRPF1 is treated as a pre-selected candidate and is used as a case study to illustrate how integrating orthogonal omics layers can reveal regulatory mechanisms that are not detectable from a single dataset.

The specific question addressed here is whether BRPF1 supports ERα signalling through a chromatin-mediated mechanism, and whether perturbing BRPF1 leads to a coordinated collapse of ERα-linked regulatory activity at the levels of chromatin occupancy, accessibility, and transcription. To answer this question, we integrated: (i) BRPF1 and ERα ChIP-seq to map co-occupied regulatory elements, (ii) BRPF1 ChIP-seq after ERα knockdown to test recruitment dependency, (iii) ATAC-seq to quantify changes in chromatin accessibility upon BRPF1 inhibition, and (iv) RNA-seq to measure transcriptional consequences and link them to accessibility changes.

4.1.1 BRPF1 co-occupies ERα regulatory elements genome-wide

ChIP-seq profiling of BRPF1 identified 20,049 genomic binding regions, among which 6,332 were co-occupied by ERα, indicating widespread overlap across the genome. Motif enrichment analysis showed that the subset of ERα–BRPF1 co-occupied regions was significantly enriched for canonical estrogen response elements (EREs) and ERE-like motifs, consistent with recruitment at ERα-bound regulatory loci.

As shown in Figure 4.1 genomic distribution analysis further showed that ERα–BRPF1 co-bound regions preferentially localize to promoters and enhancers, i.e., regulatory elements central to estrogen-responsive transcription. To assess whether ERα–BRPF1 co-occupied regions occur in specific genomic contexts rather than being randomly distributed, we performed a genomic partition enrichment analysis. Peaks were assigned to genomic partitions (e.g., promoter core, promoter proximal, exonic, intronic, UTRs, intergenic) and enrichment was quantified as $\log_{10}$(Obs/Exp), where *Obs* denotes the observed fraction of regions falling in each partition and *Exp* the expected fraction given the genome-wide partition distribution. This analysis showed that the ERα–BRPF1 co-occupied set is preferentially

enriched in promoter-related regulatory partitions, supporting the view that BRPF1 concentrates at regulatory elements central to ERα-dependent transcription. Together, these results are consistent with BRPF1 acting as a chromatin-associated coregulator within the ERα transcriptional network.

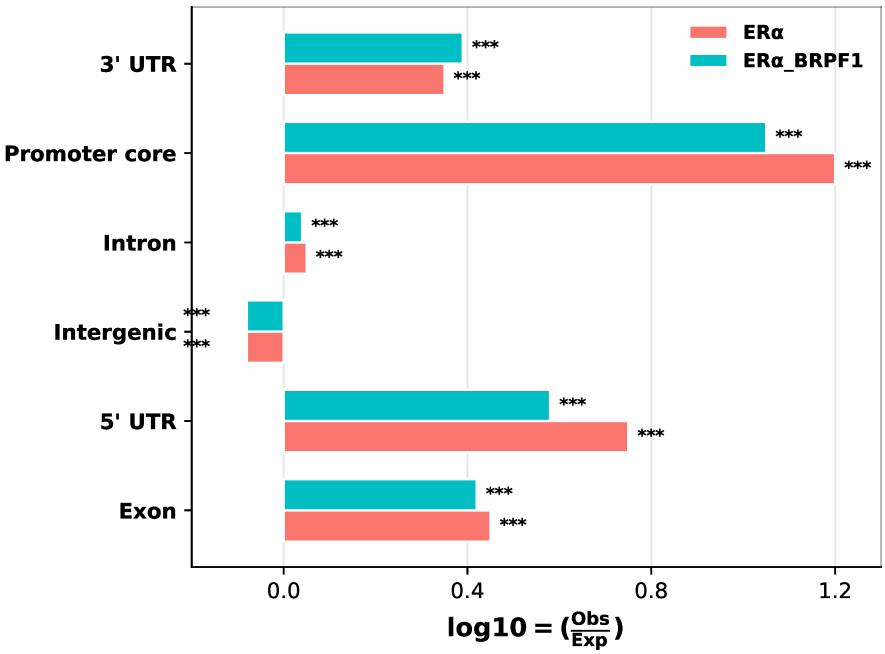


Figure 4.1: Genomic partition enrichment of ERα peaks and ERα–BRPF1 co-occupied regions. Bars report $\log_{10}$(Obs/Exp), where *Obs* is the observed fraction of peaks in each genomic partition and *Exp* is the expected fraction given the genome-wide partition distribution (hg38). Positive values indicate enrichment and negative values indicate depletion. Asterisks denote partition-wise significance from χ^2 tests of independence (*** $p < 0.001$, ** $0.001 \leq p < 0.01$, * $0.01 \leq p < 0.05$, ns $p \geq 0.05$).

4.1.2 Is BRPF1 recruitment to chromatin ERα-dependent?

Because BRPF1 does not bind DNA directly, a key mechanistic question is whether its association with chromatin reflects recruitment by DNA-bound factors such as ERα. To test ERα dependency, we profiled BRPF1 binding by ChIP-seq after ERα silencing (siERα) and compared it with control conditions (siCTRL).

To distinguish ERα-dependent from ERα-independent BRPF1 binding events, BRPF1 peaks were partitioned into: (i) *shared* sites, defined as BRPF1 peaks overlapping ERα peaks (genomic interval intersection), and (ii) *ERα-independent* sites, defined as BRPF1 peaks with no overlap with ERα peaks. We then compared BRPF1 ChIP-seq signal between siCTRL and siERα conditions.

Following ERα knockdown, BRPF1 showed a marked loss of occupancy at the

subset of chromatin regions shared with ERα, consistent with ERα-dependent recruitment at co-occupied regulatory loci (Fig. 4.2). In contrast, BRPF1 binding was largely preserved at ERα-independent sites, indicating that a distinct component of BRPF1 chromatin association persists even in the absence of the receptor. Together, these results support a model in which ERα specifically targets BRPF1 to a subset of chromatin locations, consistent with functional cooperation between BRPF1 and ERα on breast cancer cell chromatin.

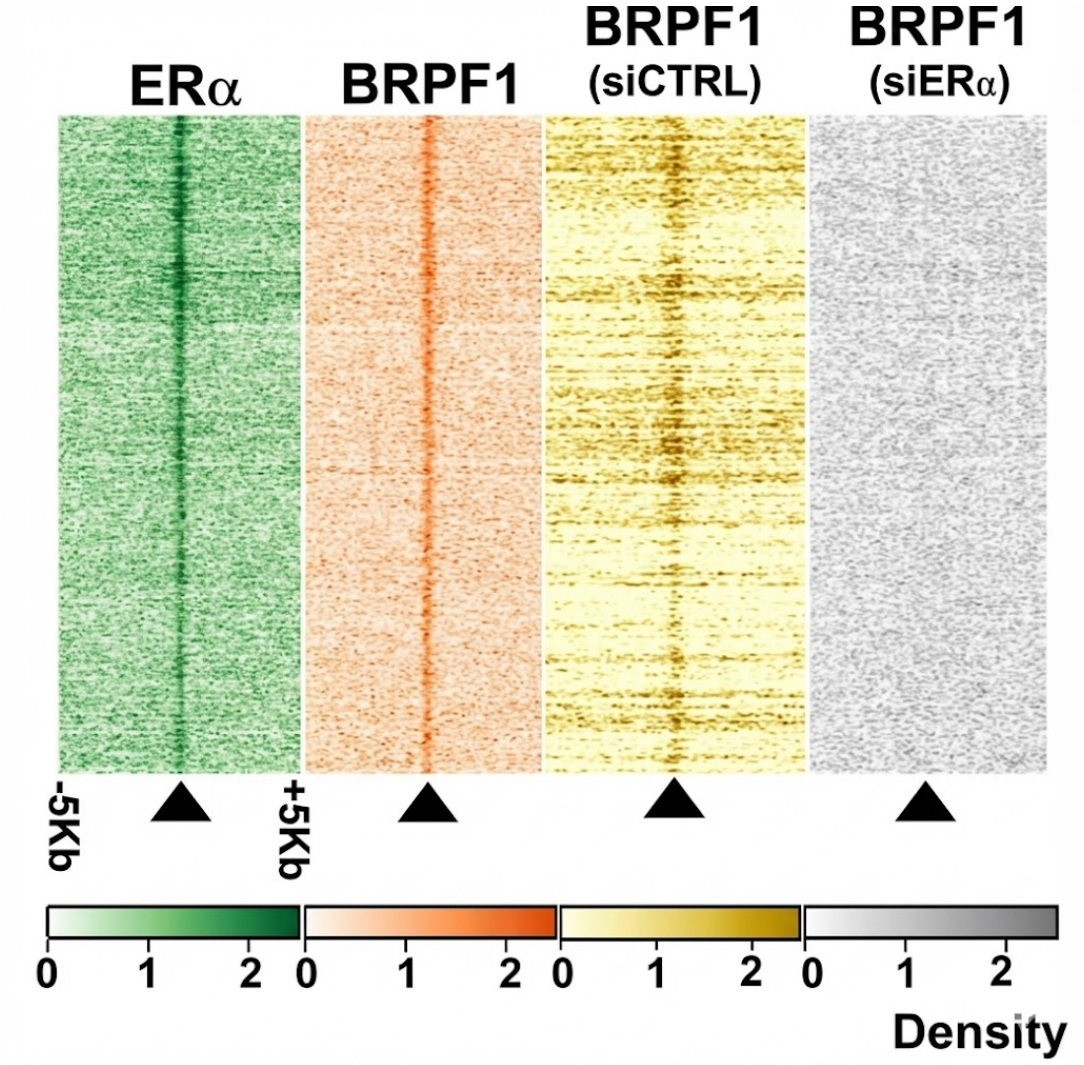


Figure 4.2: ERα knockdown reduces BRPF1 occupancy at ERα-associated regulatory sites.

These findings link BRPF1 chromatin occupancy to ERα availability and motivate the subsequent functional question: whether BRPF1 activity is required to maintain an accessible chromatin state at ERα regulatory loci.

4.1.3 Does BRPF1 inhibition preferentially reduce chromatin accessibility at ERα regulatory loci?

Having established that BRPF1 recruitment to co-occupied chromatin regions is largely ERα-dependent, we next asked whether BRPF1 activity is functionally required to maintain an open chromatin configuration at estrogen receptor regulatory loci. To address this, we analysed ATAC-seq profiles generated under pharmacological BRPF1 inhibition (active compound) and an inactive analogue (control), and quantified differential chromatin accessibility genome-wide.

Differential accessibility analysis identified 615 regions with significant changes upon BRPF1 inhibition (FDR ≤ 0.05). Notably, the response was strongly asymmetric: approximately 80% of the differentially accessible regions showed reduced accessibility, indicating a global trend toward chromatin compaction following BRPF1 inhibition.

To determine whether accessibility losses were preferentially localized at ERα-related regulatory regions rather than distributed as expected by chance, we intersected ATAC-seq loss regions with ERα ChIP-seq peaks and with the subset of ERα–BRPF1 co-occupied sites. Enrichment was assessed using a 2×2 contingency table against the background of all ATAC-seq peaks retained for differential testing. Regions losing accessibility were significantly enriched at ERα-associated loci (chi-square test, $p = 9.9 \times 10^{-5}$), whereas regions gaining accessibility showed no comparable enrichment.

Together, these results indicate that BRPF1 inhibition does not induce a uniform collapse of chromatin accessibility, but preferentially affects regulatory regions embedded in the ERα transcriptional network. This is consistent with a model in which BRPF1 contributes to the maintenance of an accessible chromatin state at ERα regulatory elements, thereby supporting estrogen-responsive transcription. Differentially accessible regions were assigned to the nearest transcription start site (TSS), and gene-level expression changes were extracted from RNA-seq differential expression results generated on the same hg38 reference build. When comparing accessibility changes to transcript-level changes, we observed a positive coupling between the two layers: loci that lost accessibility tended to be associated with genes showing reduced expression, whereas loci gaining accessibility tended to map to genes with increased expression (Fig. 4.3). This relationship was modest but statistically significant (Pearson $r = 0.13$, $p = 0.042$), consistent with a coordinated response in which BRPF1 inhibition weakens transcription primarily by reducing chromatin openness at regulatory loci.

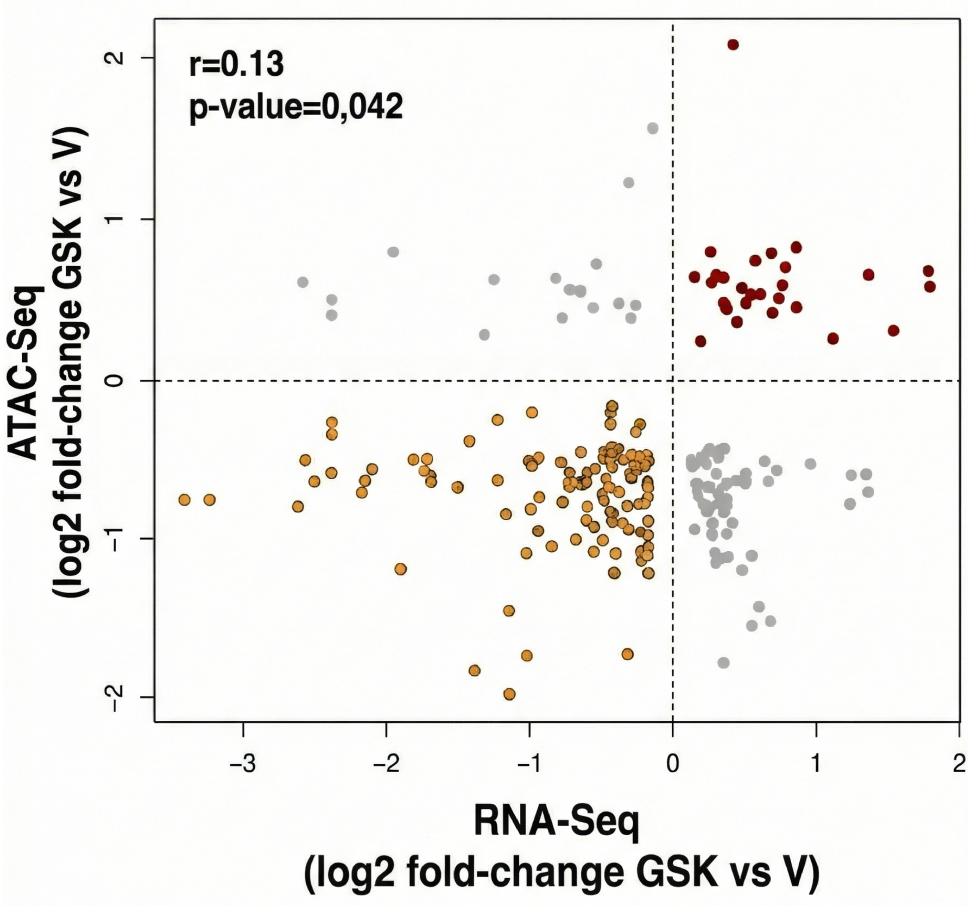


Figure 4.3: Integration of chromatin accessibility and transcriptional output after BRPF1 inhibition. Each point represents a locus linked to a nearby gene, with the x-axis reporting RNA-seq differential expression ($\log_2$ fold-change, GSK vs vehicle) and the y-axis reporting ATAC-seq accessibility change ($\log_2$ fold-change, GSK vs vehicle). The dashed lines indicate no change in either layer. Overall, accessibility and expression changes show a modest but significant positive association ($r = 0.13$, $p = 0.042$), consistent with coordinated downregulation at loci losing accessibility.

Moreover, several genes associated with regions showing altered accessibility were functionally connected to key survival and stress-response programs, including TP53 transcriptional control, DNA damage repair, and PI3K/AKT signalling. In multiple cases, these genes also harbored BRPF1 and/or ERα–BRPF1 binding events in their regulatory regions, supporting the interpretation that BRPF1-dependent chromatin remodeling impacts transcriptional programs relevant to cell survival and therapy response.

4.1.4 Summary

Overall, the results support a model in which BRPF1 operates as an epigenetic coregulator within the ERα transcriptional network. BRPF1 is recruited to ERα-

associated regulatory loci in an ERα-dependent manner, and BRPF1 inhibition preferentially disrupts chromatin accessibility at ERα regulatory regions with a concordant impact on transcriptional output. Functionally, the affected genes converge on survival and stress-response programs, consistent with attenuation of ERα-driven signalling. Importantly, these effects extend across luminal breast cancer contexts and are supported in patient-derived organoid models, highlighting BRPF1 as a therapeutically actionable node to weaken ERα signalling in endocrine therapy settings, including resistant disease that remains estrogen responsive.

4.1.5 Limitations and future directions.

Peak-to-gene assignment in the ATAC–RNA integration relied on proximity-based linking (nearest-gene annotation), which is practical but may miss distal regulatory relationships (e.g., enhancer–promoter pairs acting over tens to hundreds of kb). A natural bioinformatic extension would therefore be to refine regulatory attribution using curated enhancer–gene maps (including ERα-relevant enhancer catalogues) and to re-run the ATAC–RNA coupling analyses under enhancer-aware gene linking.

In addition, integrating three-dimensional genome information would strengthen causal interpretation of distal regulatory events. Public Hi-C/HiChIP/PLAC-seq resources in MCF-7 (or closely related ERα-positive contexts) could be used to connect BRPF1/ERα-associated accessibility changes to promoter contacts, enabling contact-informed assignment of differential ATAC regions to target genes and prioritisation of regulatory loops most perturbed by BRPF1 inhibition. Finally, these enhancer- and contact-aware annotations could be combined with motif and cistromic information to identify which ERα-linked regulatory circuits are most affected at the chromatin level.

4.2 A Cross-omics Framework applied to Endocrine Resistance Models

A key question motivating this section is whether endocrine resistance is a dominant source of variation in genome-wide DNA methylation across models, or whether methylation profiles are primarily driven by baseline cell line identity. This assessment is useful because it clarifies what can and cannot be expected from unsupervised methylome structure in multi-model settings. In parallel, the aim of this section is explicitly methodological: to present a reproducible cross-omics framework for extracting robust, comparable signals across heterogeneous resistance models. The framework begins at the functional level by using pathway-level transcriptomic signals as the entry point, defining robust pathway cores (consensus leading-edge genes), and then testing whether these cores are supported by concordant promoter methylation–expression changes under explicit robustness constraints. Individual pathways and genes are therefore reported as framework outputs that provide a prioritized shortlist for downstream follow-up, rather than as the final biological endpoint of this analysis.

For clarity, fulvestrant/ICI-associated models are used here only for descriptive methylome summaries (directionality and genomic localization of DMPs), whereas the full pathway-first cross-omics framework is demonstrated on tamoxifen-derived models, for which matched RNA-seq comparisons are available.

4.2.1 Dataset overview and global sample structure

Genome-wide DNA methylation profiling was generated using EPIC v2 arrays across a panel of endocrine therapy models profiled in biological replicates and processed in two independent array runs. Preprocessing and normalization were performed separately for each run (Methods); accordingly, all unsupervised analyses and differential methylation tests in this section are performed within-run. After filtering and BMIQ normalization, 890,311 probes were retained in Run 1 and 892,520 probes in Run 2.

To assess replicate coherence and global sample relationships, we performed unsupervised hierarchical clustering within each run using BMIQ-normalized β values on the top 20,000 most variable probes. As shown in Fig. 4.4 and Fig. 4.5, biological replicates cluster closely within each model, supporting technical reproducibility. However, samples segregate primarily by cell line/model identity rather than by a universal resistant-versus-parental separation across models. This indicates that

baseline methylation differences between cell lines represent the dominant axis of genome-wide variation, whereas resistance-associated methylation changes are superimposed on model-specific methylation landscapes.

These observations motivated a pathway-first strategy to recover convergent signals across heterogeneous models. We therefore analyse differential methylation within matched comparisons and focus on functional summaries downstream. Global methylation summaries (directionality and genomic localization of DMPs) are reported across both tamoxifen-derived and fulvestrant/ICI-associated models to provide an overview of methylation changes across endocrine therapy contexts. By contrast, the pathway-driven, cross-omics component of the framework is centred on tamoxifen resistance, reflecting the primary biological focus of this chapter: we use pathway-level transcriptomic signals to define consensus pathway cores (leading-edge genes) and then test whether these cores show disproportionate support from concordant promoter methylation–expression changes, with robustness assessed via direction-preserving branches. In the next section, we first summarise the directionality and genomic context of DMPs across comparisons.

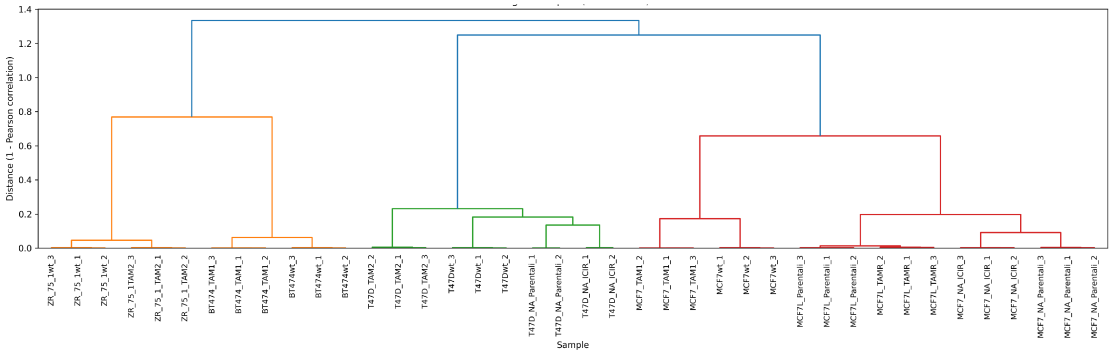


Figure 4.4: Hierarchical clustering of Run 1 samples based on BMIQ-normalized β values from EPIC v2 arrays. Distances were computed as $1-\rho$ (Pearson correlation) across the top 20,000 most variable probes, and clustering was performed using complete linkage (Methods).

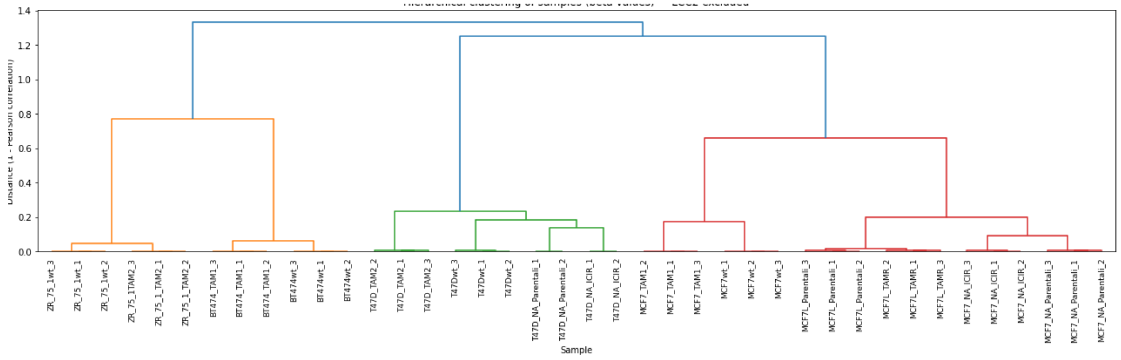


Figure 4.5: Hierarchical clustering of Run 2 samples based on BMIQ-normalized β values from EPIC v2 arrays. Distances were computed as $1-\rho$ (Pearson correlation) across the top 20,000 most variable probes, and clustering was performed using complete linkage (Methods).

4.2.2 Descriptive methylome overview across endocrine contexts

After preprocessing steps, we quantified the global directionality of methylation changes across endocrine-resistant breast cancer cell line comparisons. To focus on higher-amplitude methylation shifts, we restricted the analysis to differentially methylated positions (DMPs), defined as CpG sites with $\Delta\beta > 0.20$ (hypermethylation) or $\Delta\beta < -0.20$ (hypomethylation). For each comparison, the proportion of hyper- and hypomethylated DMPs was calculated relative to the total number of DMPs (hyper + hypo), thereby capturing the directional bias among the strongest methylation changes.

Figure 4.6 summarises the balance between hyper and hypomethylated DMPs across the 12 comparisons. Notably, 11 out of 12 models showed a higher fraction of hypermethylated than hypomethylated DMPs, indicating that methylation changes are predominantly hypermethylation-driven in the vast majority of endocrine-resistant settings considered.

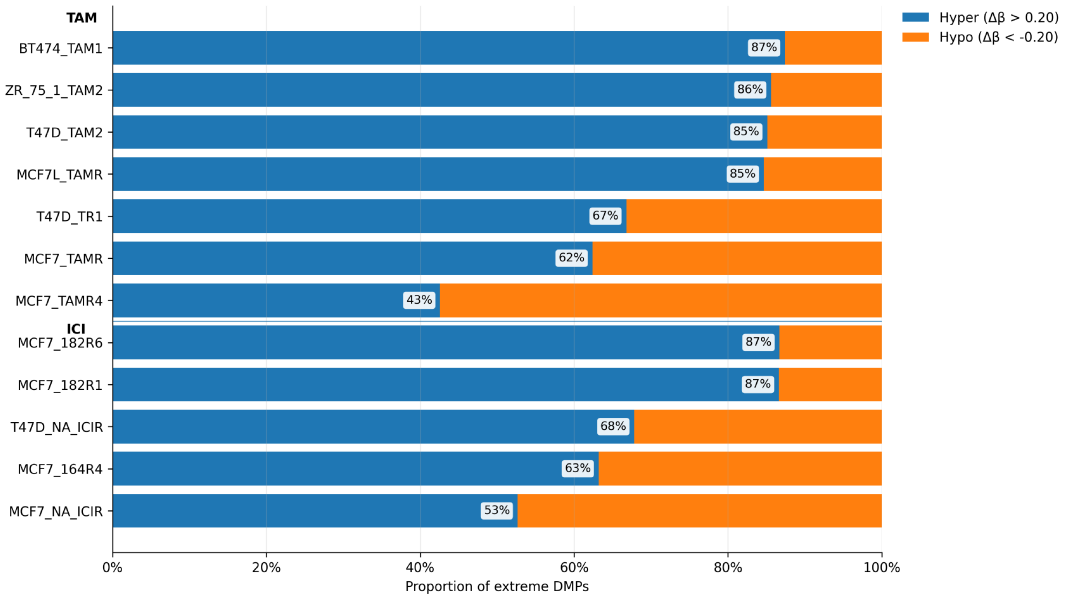


Figure 4.6: Directional balance of differentially methylated positions (DMPs) across endocrine-resistant breast cancer cell line comparisons measured on EPIC v2 arrays. Only CpG sites with $\Delta\beta > 0.20$ (hypermethylation); $\Delta\beta < -0.20$ (hypomethylation) were considered. Bars show the proportion of hyper- and hypomethylated DMPs within each comparison. Tamoxifen-derived comparisons are shown first, followed by fulvestrant/ICI-associated models. Percent labels indicate the fraction of hypermethylated DMPs.

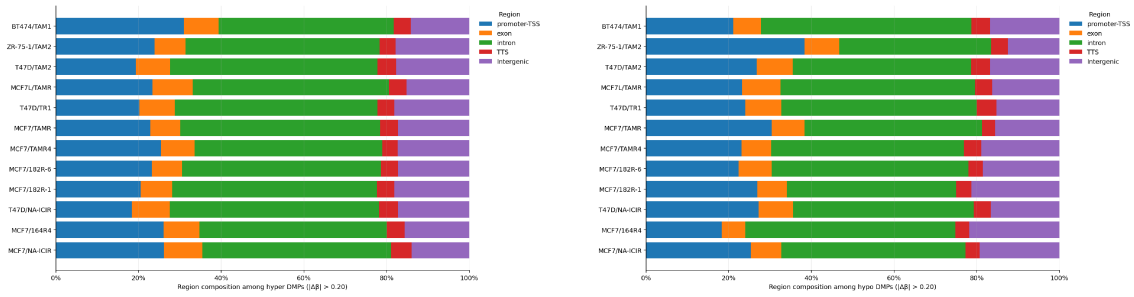
Comparisons are grouped by experimental condition (tamoxifen-derived models first, followed by fulvestrant/ICI-associated models). Within the tamoxifen-derived group, the `BT474/TAM1`, `ZR-75-1/TAM2`, `T47D/TAM2`, and `MCF7L/TAMR` cell lines showed the highest hypermethylation fractions (approximately 85–87%), indicating that the largest methylation shifts in these models are predominantly driven by gains in methylation. In contrast, other tamoxifen-related comparisons were more balanced (e.g., `T47D/TR1` ∼67% hyper; `MCF7/TAMR` ∼62% hyper), and one model (`MCF7/TAMR4`) exhibited a relative predominance of hypomethylation among these DMPs (only ∼43% hyper).

In the fulvestrant/ICI-associated group, the directionality was similarly heterogeneous. Two comparisons involving the `MCF7/182R-1` and `MCF7/182R-6` cell lines showed a hypermethylation fraction comparable to the most hypermethylated tamoxifen-derived models (both ∼87% hyper), whereas the remaining comparisons displayed more moderate hypermethylation bias (`T47D/NA-ICIR` ∼68%; `MCF7/164R4` ∼63%; `MCF7/NA-ICIR` ∼53%). Together, these results indicate that while hypermethylation dominates the methylation landscape in most models, the magnitude of this bias varies substantially across comparisons.

4.2.3 Genomic localization of DMPs

To determine whether methylation changes preferentially occur in specific genomic contexts, we annotated CpG probes to genomic features using HOMER (`annotatePeaks.pl`) with the hg38 reference as described in Methods (Section 3.2.3).

Figure 4.7 shows, for each comparison, the genomic region composition of DMPs stratified by direction, considering hypermethylated ($\Delta\beta > 0.20$) and hypomethylated ($\Delta\beta < -0.20$) CpGs. Across models, DMPs mapped predominantly to intronic regions. The remaining events were distributed across promoter–TSS and exonic regions (with smaller fractions at *TTS* and *intergenic* space), and this overall pattern was broadly consistent across tamoxifen-derived and fulvestrant/ICI-associated comparisons.



(a) Hypermethylated DMPs genomic locations.

(b) Hypomethylated DMPs genomic locations

Figure 4.7: Genomic region composition of differentially methylated positions across endocrine-resistant breast cancer cell line comparisons. (A) Hypermethylated events and (B) hypomethylated events are shown separately. Bars represent, for each comparison, the fraction of events mapping to promoter–TSS, exon, intron, TTS, or intergenic regions, based on HOMER annotation (hg38).

4.2.4 Pathway-level convergence across tamoxifen-resistant models

To investigate whether tamoxifen resistance converges on shared functional programs despite the strong cell line-specific structure observed in unsupervised methylation analyses, we performed a preranked Gene Set Enrichment Analysis (GSEA) independently for each tamoxifen-resistant model using DESeq2-derived gene rankings (Section 3.2.5).

Hallmark gene sets were classified as positively enriched (Pos; NES > 0) or negatively enriched (Neg; NES < 0). Pathways were considered significant at FDR < 0.25 and defined as recurrent when significant in at least three out of six tamoxifen-resistant cell lines.

Using this criterion, we identified 29 recurrent Pos Hallmark pathways and 15 recurrent Neg Hallmark pathways. To summarise cross-model convergence while retaining information on recurrence, effect size, and significance, recurrent pathways are visualized as lollipop/bubble plots (Figs. 4.8–4.9). Pathways are ordered by decreasing recurrence (number of supporting cell lines) and then by median |NES|. The x-axis reports the number of supporting models; dot size encodes median |NES|, and dot color encodes $-\log_{10}$(median FDR) (lower FDR corresponds to higher $-\log_{10}$ values).

Together, these results indicate that, although resistance-associated changes are largely model-specific at the genome-wide methylation level, tamoxifen-resistant models share a conserved set of transcriptional programs at the pathway level.

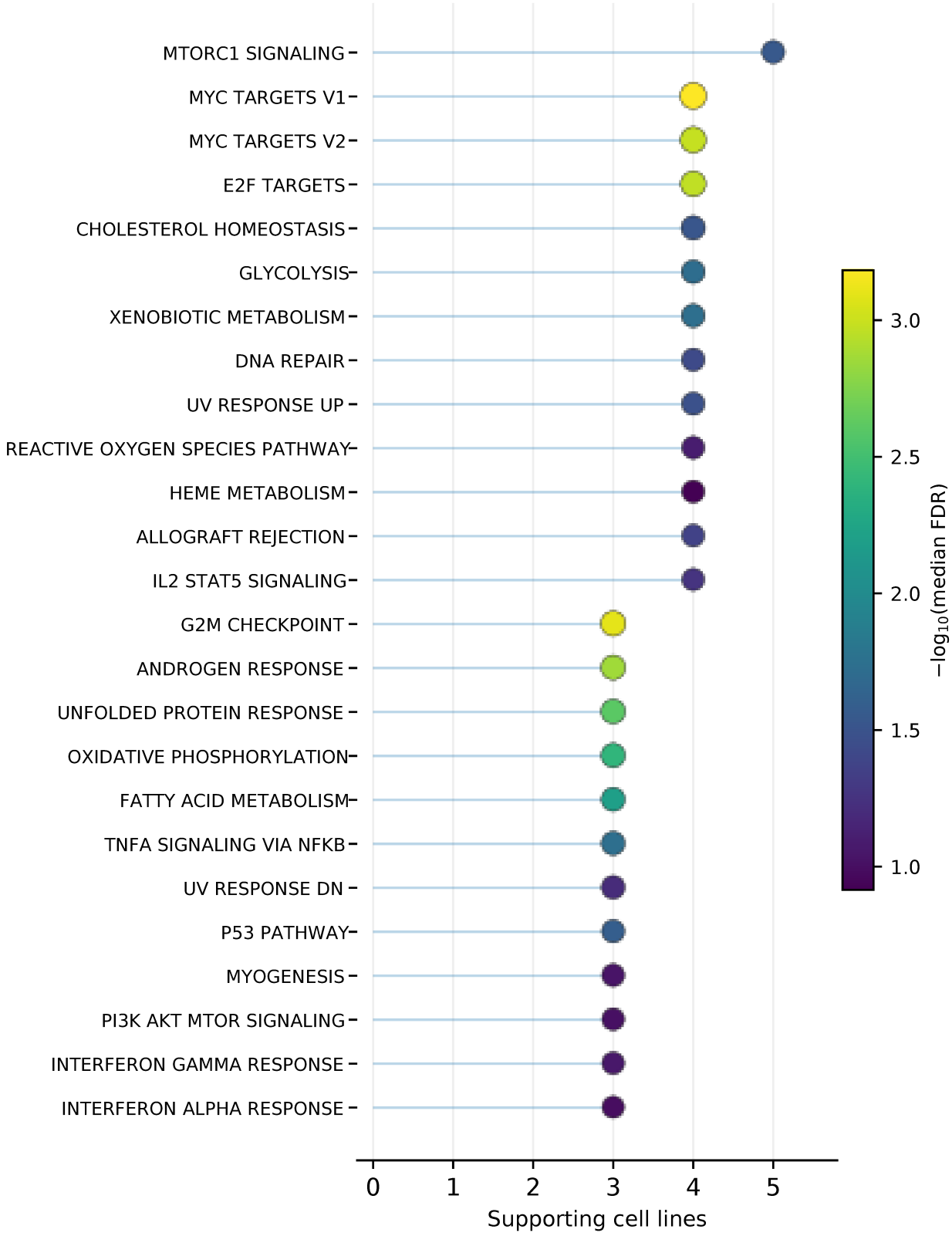


Figure 4.8: Recurrently positively enriched Hallmark pathways (NES > 0) across tamoxifen-resistant models. The x-axis indicates the number of supporting cell lines (FDR < 0.25). Dot size represents the median |NES|, and dot color represents $-\log_{10}$(median FDR). Pathways are ordered by decreasing recurrence and then by median |NES|.

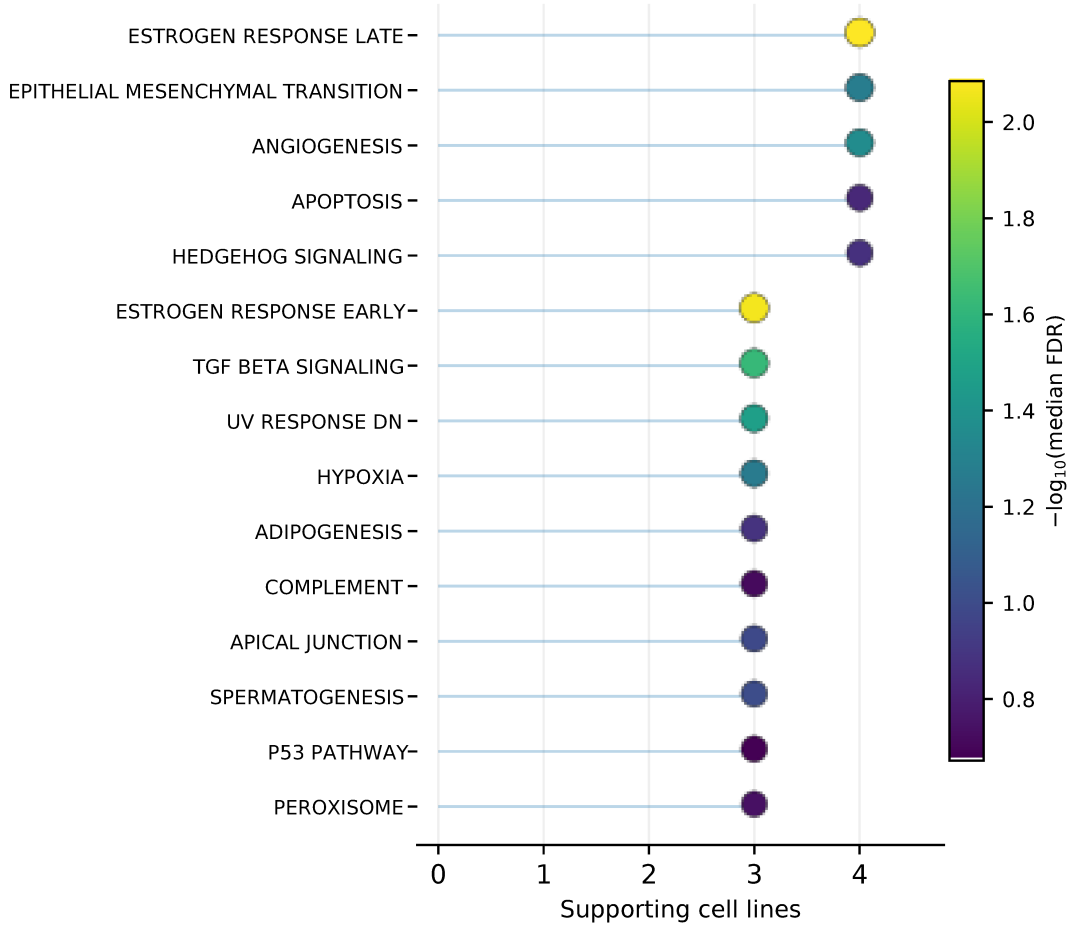


Figure 4.9: Recurrently negatively enriched Hallmark pathways (NES < 0) across tamoxifen-resistant models. The x-axis indicates the number of supporting cell lines (FDR < 0.25). Dot size represents the median |NES|, and dot color represents $-\log_{10}$(median FDR). Pathways are ordered by decreasing recurrence and then by median |NES|.

4.2.5 Consensus leading-edge genes in recurrent Hallmark pathways

After identifying recurrent Hallmark pathways across TAMR models (*conserved pathways* defined as significant with FDR < 0.25 in at least 3 cell lines), we investigated which genes robustly drive pathway enrichment across different models.

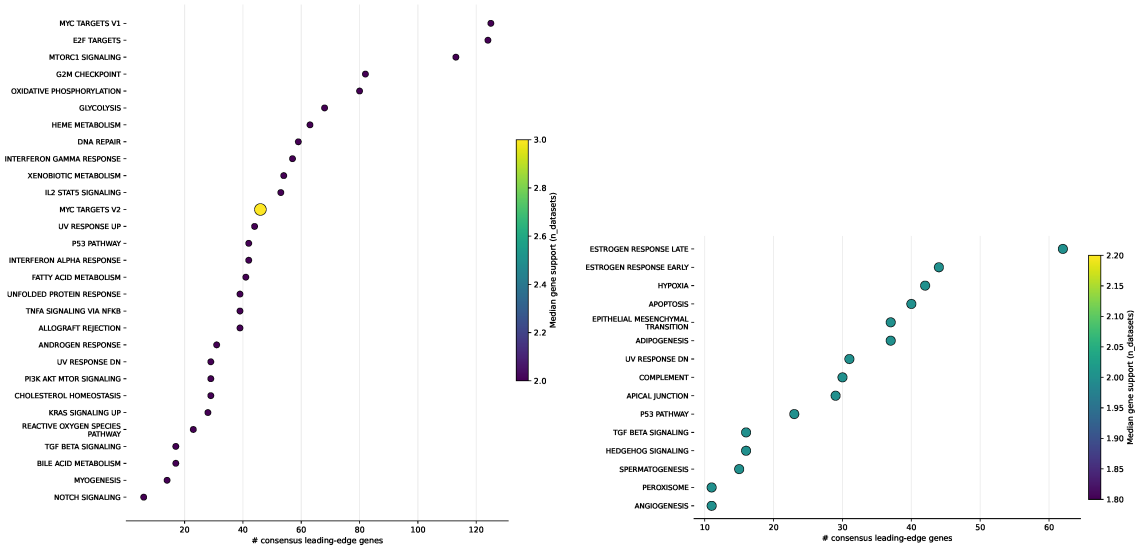
For each pathway, *leading-edge genes* were extracted from the GSEA results for each cell line. Subsequently, a *consensus* set was defined by retaining only genes that appear in the leading-edge in at least two cell lines, in order to reduce inter-model noise and prioritize reproducible signals.

Figure 4.10 summarises, separately for positively (POS; NES > 0) and negatively (NEG; NES < 0) enriched pathways, the size of the *consensus leading-edge* gene set

associated with each recurrent pathway. In each panel:

- the x-axis represents the number of *consensus leading-edge* genes identified for that pathway;
- the y-axis reports the names of the Hallmark pathways;
- the color of the point represents the *median support* of genes within the pathway, i.e., the median number of cell lines in which each gene of the consensus set appears as a leading-edge gene.

Accordingly, pathways characterized by a large number of *consensus* genes and higher median support indicate not only recurrence at the GSEA level, but also stronger convergence at the gene-specific level across TAMR models.



(a) POS pathways (NES > 0): number of *consensus leading-edge* genes per recurrent pathway.

(b) NEG pathways (NES < 0): number of *consensus leading-edge* genes per recurrent pathway.

Figure 4.10: Summary of *consensus leading-edge genes* in recurrent Hallmark pathways across TAMR cell lines. Each point corresponds to a conserved Hallmark pathway (significant with FDR < 0.25 in at least 3 cell lines). The x-axis indicates the number of *consensus leading-edge* genes associated with the pathway (minimum support ≥ 2 cell lines), while color represents the median gene support (median number of cell lines in which genes appear in the leading-edge).

4.2.6 Global and leading-edge integration of promoter methylation and gene expression

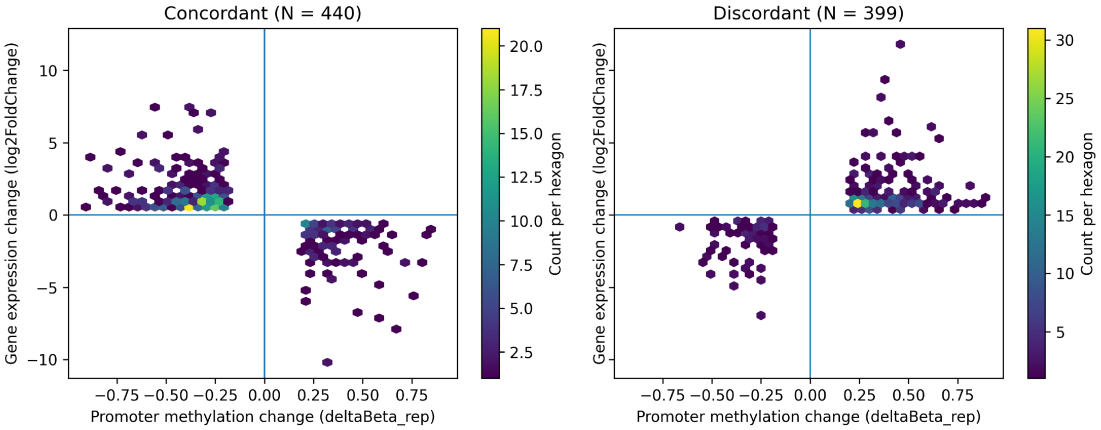
To investigate whether transcriptional changes are supported by epigenetic alterations, we integrated RNA-seq differential expression with promoter DNA methy-

lation at the gene level. For each gene shared between the two omics layers, we considered the promoter methylation change ($\Delta\beta$, reported as `deltaBeta_rep` on the x-axis) together with the gene expression change (DESeq2 $\log_2$ fold change on the y-axis). We then classified genes according to the expected inverse relationship between promoter methylation and transcription: (i) *concordant* genes, where promoter hypermethylation is associated with downregulation or promoter hypomethylation is associated with upregulation (top-left and bottom-right quadrants), and (ii) *discordant* genes, where methylation and expression change in the same direction (top-right and bottom-left quadrants). In the density (hexbin) plots, the color intensity represents the number of genes per hexagon, allowing a compact visualization of the global trend while preserving directionality through the four quadrants.

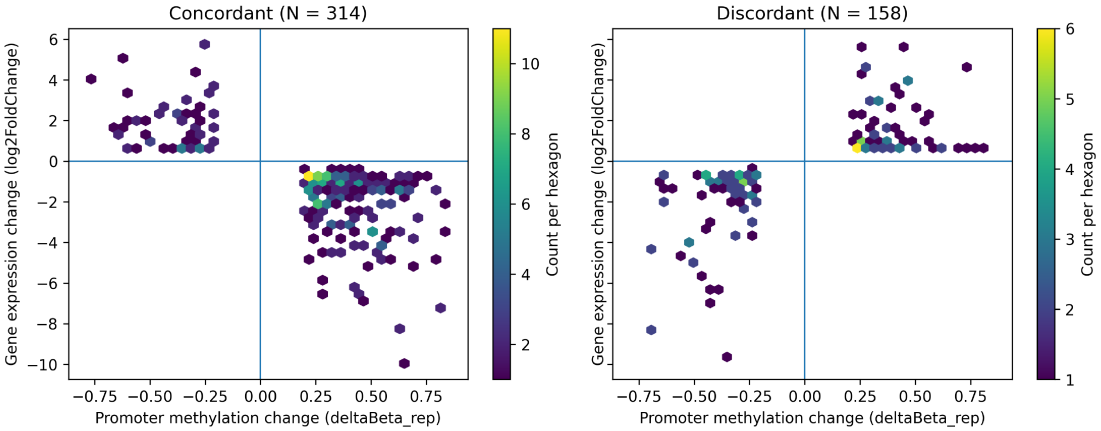
This integration was performed at three levels: (1) on consensus leading-edge genes from positively enriched pathways (Leading Pos), (2) on consensus leading-edge genes from negatively enriched pathways (Leading Neg), and (3) on the full background set of genes shared between methylation and expression. Across all three analyses, concordant genes outnumbered discordant genes (Table 4.1), indicating an overall tendency for promoter methylation changes to be coupled to gene expression in the expected inverse direction.

Table 4.1: Concordant vs discordant genes across analyses.

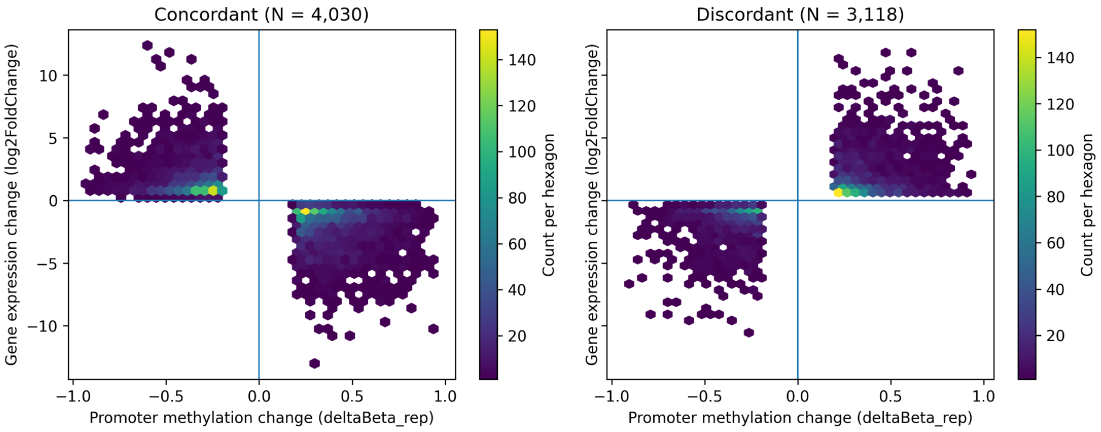
Analysis	Concordant (N)	Discordant (N)
Leading Pos	440	399
Leading Neg	314	158
All shared genes	4,030	3,118



(a) Integration restricted to consensus Leading Pos genes.



(b) Integration restricted to consensus Leading Neg genes.



(c) Integration across all genes shared between methylation and expression (background).

Figure 4.11: Gene-level integration of promoter DNA methylation and gene expression

4.2.7 Pathway-level enrichment of concordant methylation–expression support among leading-edge genes

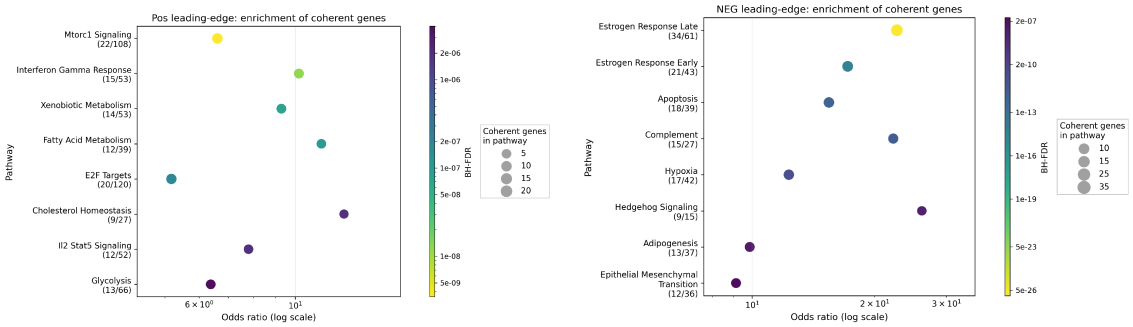
The integration analysis in Fig. 4.11 suggests that concordant promoter methylation–expression changes are frequent overall. Here, we ask a more specific, pathway-driven question: *for a given conserved pathway, are concordant genes enriched among its consensus leading-edge genes compared with the cross-omics background?*

Formally, for each pathway we treat the set of consensus leading-edge genes as the pathway-specific target set and test whether it contains an excess of concordant genes relative to the universe of genes measurable in both omics layers. This enrichment framework enables us to prioritize pathways whose transcriptional core is disproportionately supported by promoter methylation changes in the expected direction, beyond the global baseline concordance observed across all shared genes.

To make this question concrete, we applied the enrichment test separately to (1) consensus leading-edge genes from positively enriched pathways (Leading Pos) and (2) consensus leading-edge genes from negatively enriched pathways (Leading Neg). Because promoter DNA methylation is expected to regulate transcription inversely, we defined *concordant* gene-level events within each cell line comparison as: `hypo+up` (promoter hypomethylation together with gene upregulation) or `hyper+down` (promoter hypermethylation together with gene downregulation). In this first-level analysis, concordance is evaluated *within each individual cell line comparison* (i.e., a gene is counted as concordant whenever it shows a concordant pattern in at least one comparison), without yet enforcing that the same direction is maintained across all lines in which the gene appears. Therefore, this initial test answers whether consensus leading-edge genes are enriched for *any* concordant promoter methylation–expression evidence, rather than whether the direction is conserved across models.

The resulting pathway-level enrichments are summarised in Fig. 4.12. For Leading Pos pathways (Fig. 4.12a), enrichment was evaluated using the concordant `hypo+up` signal, consistent with pathways driven by increased expression. Conversely, for Leading Neg pathways (Fig. 4.12b), enrichment was evaluated using the concordant `hyper+down` signal, consistent with pathways driven by decreased expression. In each dot plot, the x-axis reports the odds ratio (log scale) from the one-sided Fisher test, and bubble size reflects the number of concordant genes observed within the pathway's consensus leading-edge. Color encodes the BH-adjusted FDR, highlighting the pathways whose leading-edge is most disproportionately supported by promoter methylation changes in the expected direction.

Importantly, because this first-level definition does not penalize cross-line directional disagreement, a gene can contribute to pathway enrichment even if it shows a concordant pattern in one line but an opposite pattern in another. For this reason, we next introduce stricter direction-preserving branches (non-divergent, conserved-majority, and conserved-strict) to assess whether the pathway-level enrichments remain when requiring increasing levels of directional consistency across cell line models.



(a) Leading Pos: Fisher enrichment of concordant `hypo+up` genes among consensus leading-edge genes.

(b) Leading Neg: Fisher enrichment of concordant `hyper+down` genes among consensus leading-edge genes.

Figure 4.12: Pathway-level enrichment of concordant promoter methylation–expression changes within consensus leading-edge genes. Odds ratios are from one-sided Fisher exact tests (greater), with BH-FDR correction across pathways. Bubble size reflects the number of concordant genes observed within each pathway's consensus leading-edge.

4.2.8 Cross-line directionality branches: assessing the robustness of concordant support

The first-level enrichment analysis in Fig. 4.12 treats concordance as an event that can be observed within *any* single cell line comparison. As a consequence, a gene can contribute to pathway enrichment even if it exhibits a concordant pattern in one model but an opposite (or different) pattern in another. To evaluate whether pathway-level enrichments are robust to such cross-line directional variability, we introduced stricter, direction-preserving *branches* that operate at the gene level across all comparisons in which the gene is observed.

In this section, we focus on two increasingly stringent branches. The **conserved-majority** branch retains a gene if it shows concordance in at least two cell line comparisons and if a *single* concordant direction is dominant across those occurrences (i.e., the gene displays a clear majority pattern, such as predominantly

`hypo+up` or predominantly `hyper+down`). This rule allows occasional discordant or non-informative observations, but requires an overall consistent direction. The **conserved-strict** branch further strengthens this requirement by retaining only genes that are concordant in at least two comparisons *and never* display the opposite concordant direction in any other comparison. Thus, conserved-strict isolates a higher-confidence subset of genes whose promoter methylation–expression relationship is directionally stable across the analysed models.

For each branch, we repeated the pathway-level Fisher enrichment test using the same background universe and the same pathway-specific consensus leading-edge target sets as in the baseline analysis. We then asked a retention question: *which pathways remain significantly enriched (BH-FDR* < 0.05*) under the branch definition, and therefore remain supported even after enforcing cross-line directional consistency?* Retained pathways are visualized by comparing baseline odds ratios (open circles) with branch odds ratios (filled circles), where filled-circle color encodes the branch BH-FDR and point size reflects the number of branch-retained concordant genes within the pathway leading-edge.

Applying the **conserved-majority** branch, we retained 14 positively enriched pathways and 9 negatively enriched pathways (BH-FDR < 0.05). Under the more stringent **conserved-strict** branch, the retained set contracted to 9 pathways for Leading Pos and 5 pathways for Leading Neg (BH-FDR < 0.05). This progressive reduction is expected: enforcing cross-line agreement filters out genes whose concordant evidence is driven by a single comparison or is directionally unstable across cell lines, thereby prioritising pathways supported by a more reproducible cross-omics signal.

These branches also highlight an important statistical trade-off. Compared with the baseline definition, majority/strict rules yield a smaller but more "clean" set of concordant genes. In Fisher-based enrichment, this often leads to (i) a relative increase in estimated effect sizes (odds ratios), because retained genes tend to be more concentrated within the pathway leading-edge, but (ii) potentially less extreme p-values/FDR in some cases, because the reduction in the total number of retained concordant genes decreases the effective sample size and therefore statistical power. Consequently, the branch framework should be interpreted as a robustness analysis: pathways that remain significant under stricter rules represent a conservative, higher-confidence core, whereas pathways retained under majority but not strict may reflect signals that are real but more context-dependent across models.

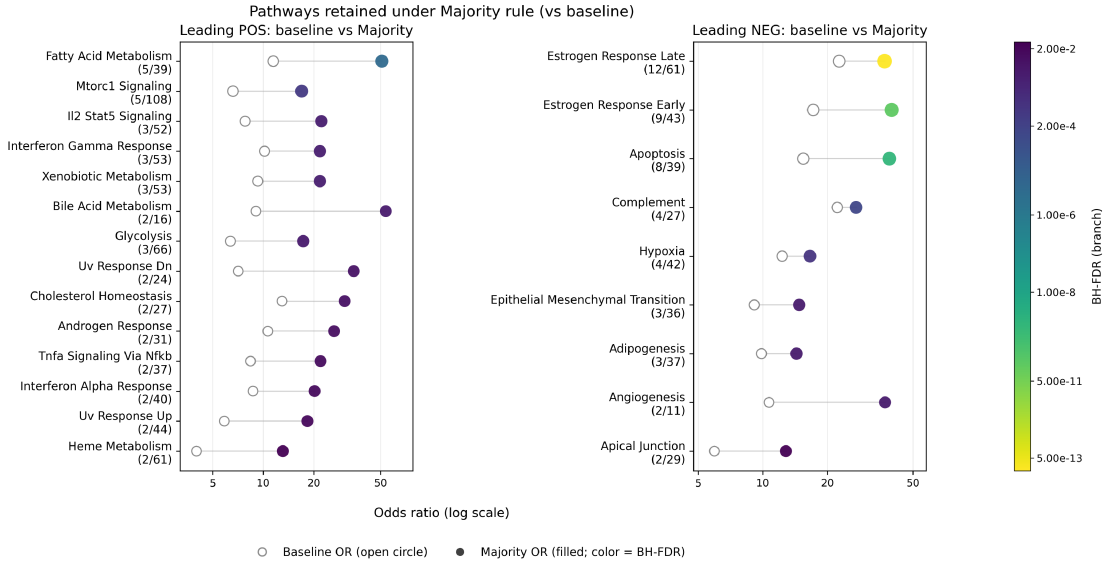


Figure 4.13: Pathways retained under the conserved-majority branch (BH-FDR < 0.05), shown as a baseline-vs-branch comparison. The left panel reports Leading Pos pathways and the right panel reports Leading Neg pathways. For each retained pathway, open circles denote baseline odds ratios (first-level definition; concordance within any single comparison), while filled circles denote majority-branch odds ratios (cross-line majority-consistent concordance). Filled-circle color encodes the branch BH-FDR.

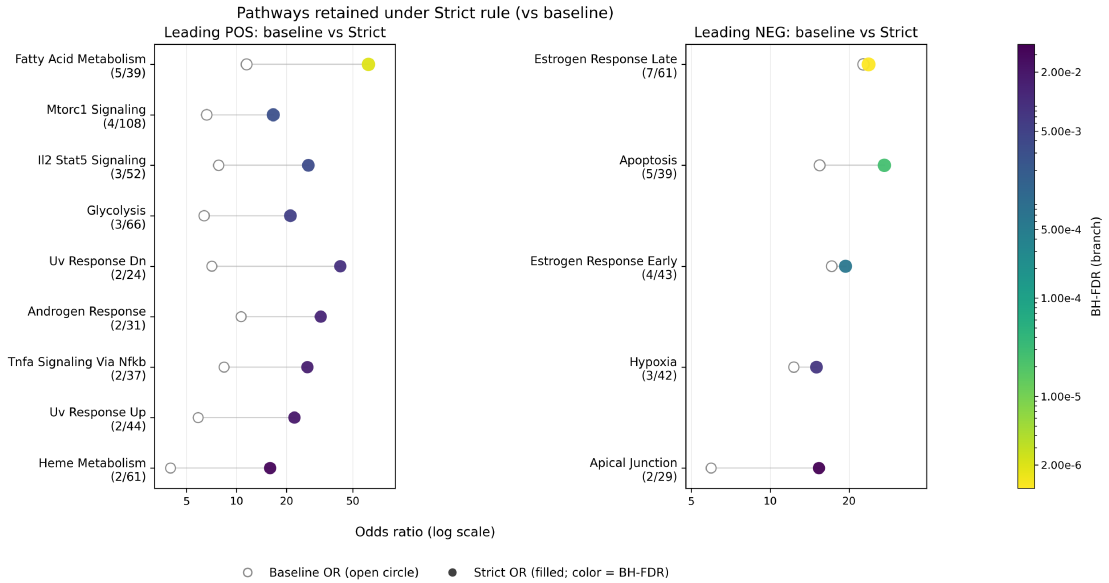


Figure 4.14: Pathways retained under the conserved-strict branch (BH-FDR < 0.05), shown as a baseline-vs-branch comparison. The left panel reports Leading Pos pathways and the right panel reports Leading Neg pathways. Open circles denote baseline odds ratios, while filled circles denote strict-branch odds ratios under the requirement that concordant direction never flips across cell line comparisons. Filled-circle color encodes the branch BH-FDR.

4.2.9 Summary, limitations, and future directions

Summary. In this section, we combined genome-wide promoter DNA methylation profiles and RNA-seq to assess whether recurrent transcriptional programs associated with endocrine resistance are supported by epigenetic changes in the expected inverse direction. Although global methylation variability was dominated by cell line identity, pathway-level analyses revealed cross-model convergence in transcriptional programs. We therefore moved from CpG-level heterogeneity to functional convergence by focusing on recurrent Hallmark pathways and their *consensus leading-edge* gene sets.

By integrating promoter methylation changes ($\Delta\beta$) with gene expression changes (DESeq2 $\log_2$FC), we observed a global predominance of concordant promoter methylation–expression patterns, both within consensus leading-edge genes (Leading Pos and Leading Neg) and across the full cross-omics background. We then formalized this observation with a pathway-level enrichment framework that tests whether concordant genes are overrepresented among pathway-specific consensus leading-edge genes relative to a background universe defined by genes measurable in both omics layers. This step refines the qualitative global trend into a pathway-resolved prioritisation of transcriptional cores that are disproportionately supported by promoter methylation.

Finally, we evaluated robustness to cross-model heterogeneity by introducing cross-line directionality branches, which require increasing levels of directional consistency across cell line comparisons. Under the *conserved-majority* branch, 14 positively enriched and 9 negatively enriched pathways remained significant (BH-FDR < 0.05), while the stricter *conserved-strict* branch retained 9 (Pos) and 5 (Neg) pathways. Beyond pathway prioritisation, the branch framework also provides gene-level prioritisation: genes retained under majority and especially strict rules define a progressively higher-confidence subset of *cross-line conserved* concordant genes within the leading-edge, representing candidate pathway cores with directionally stable cross-omics evidence.

Limitations. A few practical considerations apply when interpreting these results. First, promoter methylation was summarised at the gene level, which improves interpretability and comparability across models but may mask locus-specific regulatory complexity. Second, the concordant/discordant classification is based on an expected inverse promoter methylation–expression relationship; while broadly reasonable for promoter regulation, it does not capture all mechanisms shaping transcription (e.g., distal enhancers, transcription factor activity, chromatin architec-

ture, or post-transcriptional regulation). Third, the Fisher enrichment test quantifies over-representation of concordant genes in leading-edge sets, but does not model within-pathway gene dependence and does not establish causal directionality. These points mainly affect how the results should be used: as a robust and interpretable prioritisation layer rather than as a locus-resolved mechanistic explanation.

Future directions. This framework provides a general and portable strategy to prioritize both (i) pathways whose transcriptional cores are disproportionately supported by promoter methylation changes and (ii) a ranked subset of pathway core genes showing cross-line conserved concordant evidence. Several extensions follow naturally. (i) The gene-level integration can be refined to locus-level resolution by mapping which promoter CpGs or promoter sub-regions drive each gene call and by integrating additional annotations (e.g., regulatory features, transcription factor binding, or chromatin marks). (ii) The same enrichment and branch logic can be applied to other regulatory layers (e.g., enhancer methylation, chromatin accessibility, or histone modifications) to distinguish promoter-driven and enhancer-driven transcriptional programs. (iii) The pathways and genes retained under the *conserved-majority* and especially *conserved-strict* rules define high-confidence candidates for downstream focused analyses, such as prioritising pathway core genes for targeted validation, constructing compact cross-omics signatures, or testing reproducibility in additional models and independent cohorts.

4.3 Development of a General-Purpose Tensor-Based Method in Omics analyses

The first two sections of the Results chapter focused on biological questions related to endocrine therapy response in breast cancer, addressed through integrative analysis of multiple omics layers. These studies highlighted how complex phenotypes, such as the acquisition of drug resistance, emerge from the coordinated action of transcriptional, epigenetic, and chromatin-level regulatory mechanisms. They also underscored a broader methodological theme: multi-omics investigations increasingly require analytical frameworks capable of capturing structure, interactions, and variation across several data modalities simultaneously.

Motivated by this need, the final part of this thesis shifts from a biological case study to a methodological contribution. Here, I present the development of `TensorPLS`, an R package designed to extend Partial Least Squares (PLS) methodology to data naturally represented as tensors. Unlike classical PLS implementations—which operate on two-dimensional matrices—`TensorPLS` provides a workflow for three-way longitudinal datasets, combining tensor-native preprocessing/imputation with PLS-based discrimination and mode-resolved interpretability outputs. This capability is particularly valuable in modern biomedical research, where longitudinal, multi-omic, or multi-condition datasets often exhibit three-dimensional organization.

Although the method was evaluated on data from the TEDDY cohort—a large-scale longitudinal study of type 1 diabetes—its formulation is not specific to any tissue, disease, or experimental context. The approach is general-purpose: any setting involving samples measured across multiple variables and time points (or other modes) can be analysed within this framework. In a translational perspective, this includes applications to breast cancer, where multi-omic datasets increasingly integrate gene expression, methylation, chromatin accessibility and proteomic layers, often across different conditions or temporal windows.

The goal of `TensorPLS` is not only dimensionality reduction but also interpretability. By exploiting tensor structure, the method identifies patterns of variables and time-specific effects that discriminate between biological states—for example between sensitive and resistant cells, between treated and untreated conditions, or among patient subgroups.

In the following sections, I describe the formulation of the method, its implementation in R, validation on real-world datasets, and potential applications in translational research, including contexts such as breast cancer where multi-omics

tensors are increasingly becoming available.

4.3.1 Dataset Overview and Analysis Objectives

We applied `TensorPLS` to four TEDDY-derived longitudinal case–control datasets organised as three-way tensors (*subjects* × *variables* × *time*), as described in Section 3.3.3. Briefly, measurements were aligned to five time points (0, −3, −6, −9, −12 months) relative to IA confirmation (time 0), and missing values were handled via Tucker-3 imputation.

These data structures enable three evaluation objectives: (i) assess whether temporal omics profiles discriminate cases from controls, (ii) identify discriminant features (including time-resolved importance via VIP2D), and (iii) quantify which time points contribute most to the discriminant components.

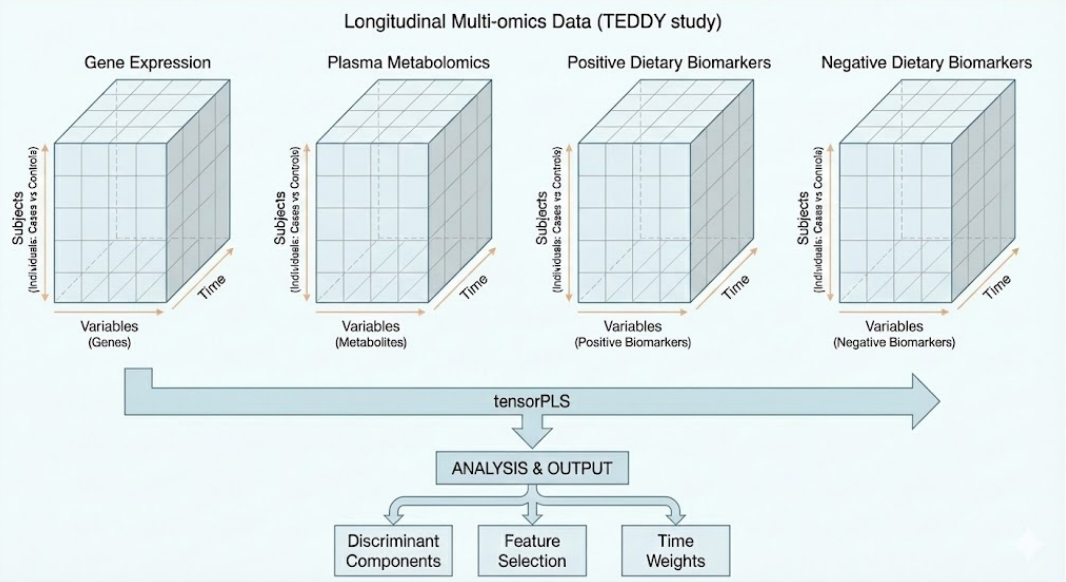
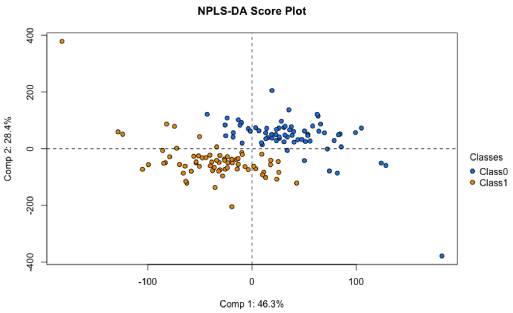


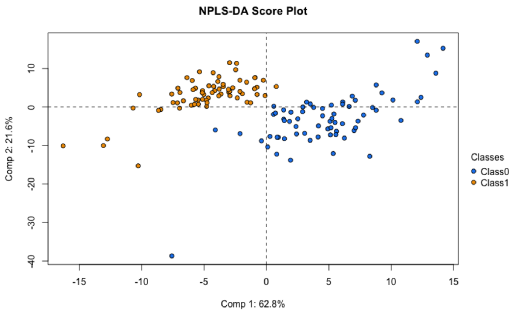
Figure 4.15: Schematic of the TEDDY longitudinal multi-omics datasets analysed with `TensorPLS`. Each modality is represented as a three-way tensor (*subjects* × *variables* × *time*).

4.3.2 Baseline PLS-DA Model Without Feature Selection

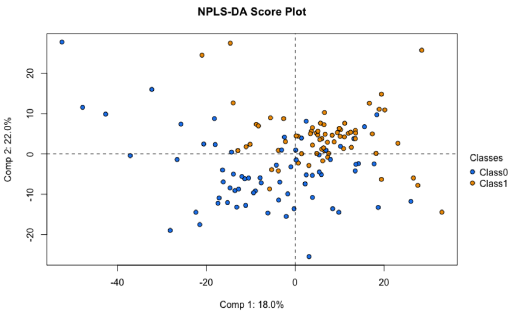
To assess whether the longitudinal multi-omics data contained intrinsic case–control structure prior to any feature selection, we computed baseline PLS-DA models using the full feature set for each dataset: gene expression, metabolomics, positive dietary biomarkers, and negative dietary biomarkers. The corresponding score plots for the first two components are shown in Figures 4.16a – 4.16b – 4.16c – 4.16d.



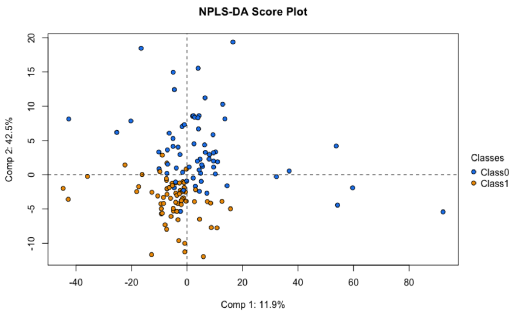
(a) Gene expression.



(b) Metabolomics.



(c) Positive dietary biomarkers.



(d) Negative dietary biomarkers.

Figure 4.16: Baseline PLS-DA score plots obtained using the full feature set prior to feature selection. The first two latent components are shown for (a) gene expression, (b) metabolomics, (c) positive dietary biomarkers, and (d) negative dietary biomarkers. These analyses were performed to evaluate the presence of an intrinsic case–control structure before applying dimensionality reduction or feature selection procedures.

Across the four datasets, the baseline PLS-DA models revealed markedly different degrees of case–control separability.

In the gene expression dataset (Fig. 4.16a), samples showed a clear, although not complete, separation along the first two latent components. This visual pattern was consistent with strong internal performance indicators ($R^2 = 0.91$, $Q^2 = 0.77$), suggesting a pronounced discriminant structure under the applied cross-validation scheme.

The plasma metabolomics dataset (GCTOFX; Fig. 4.16b) displayed a clearer case–control structure than the dietary biomarker panels. The corresponding metrics ($R^2 = 0.84$, $Q^2 = 0.50$) indicate comparatively higher explained variance and substantially stronger cross-validated discrimination among the metabolite-based layers.

By contrast, the negative dietary biomarkers (Fig. 4.16d) showed a weaker discriminant structure, with moderate explained variance but low cross-validated per-

formance ($R^2 = 0.54$, $Q^2 = 0.07$), indicating limited internal predictive stability.

Similarly, the positive dietary biomarkers (Fig. 4.16c) exhibited only minimal separation in score space and very low cross-validated performance ($R^2 = 0.40$, $Q^2 = 0.02$), suggesting a weak and unstable discriminant signal under the current modelling setup.

Overall, these baseline analyses indicate that (i) discriminant structure varies substantially across omics layers; (ii) gene expression shows the strongest signal; (iii) among metabolite-based datasets, GCTOFX provides the clearest discrimination whereas the dietary biomarker panels show limited predictive stability; and (iv) the generally modest Q^2 values in two of the four datasets motivate the use of VIP-based feature selection to improve interpretability and concentrate the analysis on the most informative variables.

In the next section, we apply the VIP-based feature selection procedure implemented in the `TensorPLS` workflow and assess how restricting the analysis to high-VIP variables affects score-space separation and feature prioritisation.

4.3.3 PLS-DA Results After Feature Selection

To enhance interpretability and prioritise a compact set of variables most associated with case–control discrimination, we applied VIP-based feature selection using scores derived from the `TensorPLS` workflow. Variables above the chosen percentile threshold were retained, resulting in a substantial reduction in dimensionality across datasets. In particular, gene expression decreased from $p = 21{,}285$ to $p_{\text{sel}} = 816$ variables, while the metabolite-based datasets decreased from $p = 364$ to $p_{\text{sel}} = 82$ (GCTOFX metabolomics), from $p = 514$ to $p_{\text{sel}} = 64$ (positive dietary biomarkers), and from $p = 414$ to $p_{\text{sel}} = 152$ (negative dietary biomarkers), respectively.

Models were then refitted using the same cross-validation scheme and the same number of components selected in the baseline analysis. The resulting score plots are shown in Figure 4.17. Compared with the baseline models, the reduced-feature models display a clearer visual separation between cases and controls, consistent with a more concentrated representation of the discriminant structure in score space.

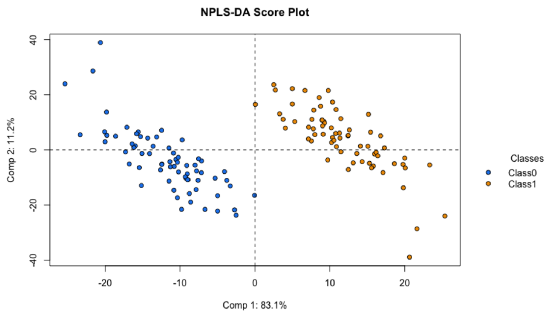
Since feature selection is label-informed, post-selection performance estimates may be optimistic. Accordingly, we report the following *apparent* internal R^2 and Q^2 values after feature selection (Table 4.2).

Overall, post-selection performance indicators were highest for gene expression ($R^2 = 0.98$, $Q^2 = 0.93$) and increased for all metabolite-based datasets (GCTOFX: $Q^2 = 0.63$; positive: $Q^2 = 0.34$; negative: $Q^2 = 0.23$). These reduced-feature models are used primarily to support score-space visualization and feature prioritisation

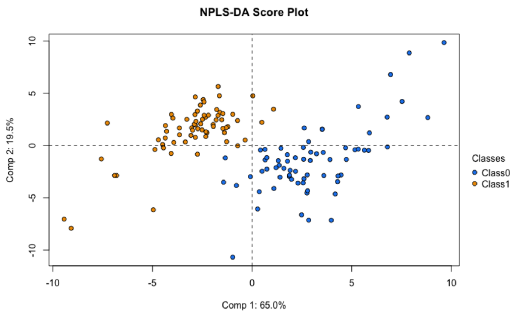
Table 4.2: Dimensionality reduction and apparent PLS-DA performance after VIP-based feature selection. Post-selection metrics are reported as *apparent* internal indicators because feature selection is label-informed.

Dataset	p **(before)**	p_{sel} **(after)**	R^2 **(after)**	Q^2 **(after)**
Gene expression	21,285	816	0.98	0.93
GCTOFX metabolomics	364	82	0.84	0.63
Positive dietary biomarkers	514	64	0.58	0.34
Negative dietary biomarkers	414	152	0.56	0.23

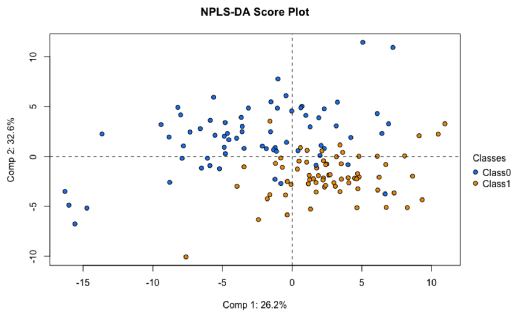
for downstream interpretation rather than as unbiased estimates of out-of-sample predictive performance.



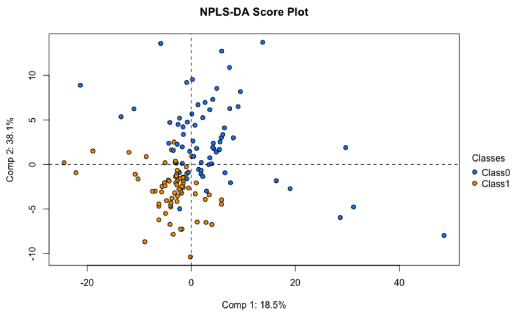
(a) Gene expression after feature selection.



(b) Metabolomics (GCTOFX).



(c) Positive dietary biomarkers.



(d) Negative dietary biomarkers.

Figure 4.17: PLS-DA score plots after VIP-based feature selection for: (a) gene expression, (b) GCTOFX metabolomics, (c) positive dietary biomarkers, and (d) negative dietary biomarkers. Feature selection yields a reduced set of variables that enhances the visualization of case–control separation.

4.3.4 Time Contribution Analysis

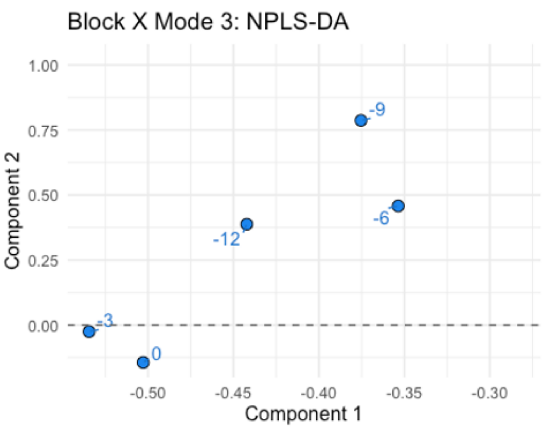
Up to this point, we have shown that feature selection improves class discrimination and enhances the predictive performance of PLS-DA models. However, in longitudinal multi-omics data, it is not only *which* variables matter, but also *when* they

matter. In other words, beyond identifying discriminative molecular features, we also aim to determine *which time points contribute most strongly to the case–control separation.*

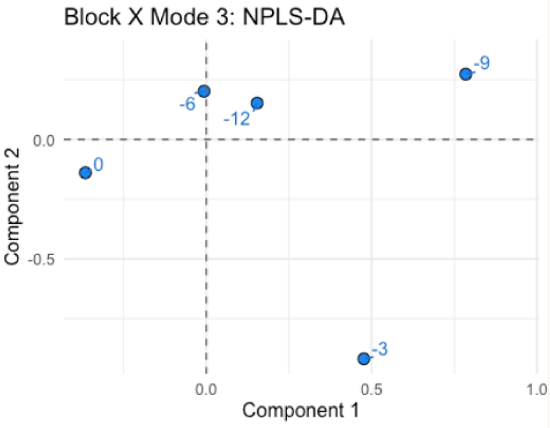
To address this question, we examined the Mode 3 scores returned by the TensorPLS framework. These scores represent the projection of each time point onto the latent components used in the PLS-DA model. The interpretation is straightforward:

- Time points located far from the origin $(0, 0)$ have a strong influence on shaping the discriminant components.
- Time points near the origin contribute minimally to the separation between cases and controls.
- The distance from the origin is proportional to importance.
- The direction indicates which component the time point drives the most.
- Opposite signs reflect opposite temporal phases.
- Clusters of time points indicate similar temporal roles.
- Outliers reveal key time points strongly influencing the model.

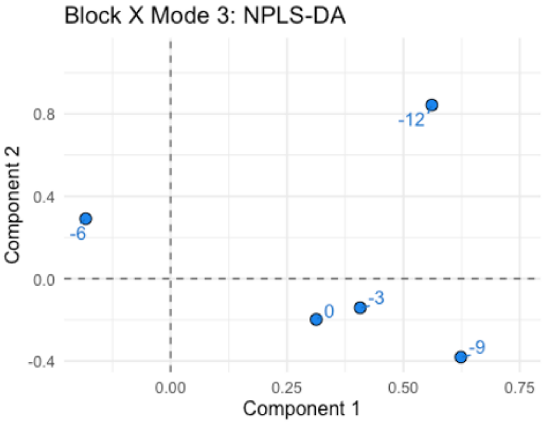
The Mode 3 time point projections for all four datasets are shown in Figure 4.18. These plots allow us to quantify the temporal structure of each omics layer and identify the developmental windows most relevant for class discrimination.



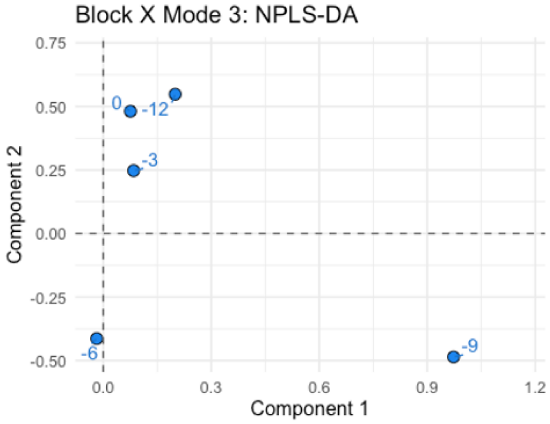
(a) Gene expression.



(b) Metabolomics (GCTOFX).



(c) Positive dietary biomarkers.



(d) Negative dietary biomarkers.

Figure 4.18: Mode–3 time point projections from NPLS-DA across the four omics datasets. Time points located farther from the origin exert a stronger influence on the discriminant components, revealing the temporal windows that most strongly drive case–control separation.

Across datasets, clear temporal signatures emerged:

Gene expression (Fig. 4.18a). The −6 and −9 month time points showed the strongest influence, indicating that transcriptional divergence between cases and controls emerges well before seroconversion.

Metabolomics (GCTOFX) (Fig. 4.18b). Two time points dominated the discriminant structure: −9 and −3 months, suggesting metabolic alterations occurring both early and closer to disease onset.

Negative dietary biomarkers (Fig. 4.18d). The −12 and −9 month time points contributed the most, while other time points clustered near the origin, indicating relatively stable patterns closer to seroconversion.

Positive dietary biomarkers (Fig. 4.18c). Here, the contribution was spread

across −12, −9, and −3 months, revealing both long-term and intermediate dietary signatures associated with disease risk.

In summary, the time contribution analysis highlights that different omics layers encode discriminative information at distinct temporal windows. These findings underline the importance of preserving the temporal mode within tensor-based approaches and demonstrate how TensorPLS enables the identification of biologically meaningful time points driving case–control separation.

4.3.5 Identification of key discriminant features across time

Having established that feature selection enhances class separation and predictive performance, and having identified the time points that contribute most strongly to case–control discrimination, we next examined how discriminant features are distributed across time.

TensorPLS explicitly addresses this question by providing time-resolved measures of feature importance through the VIP2D representation.

VIP2D extends the standard Variable Importance in Projection (VIP) concept by assigning an importance score to each feature at each time point for a given latent component. This representation allows the identification of time-specific discriminant features, avoiding the loss of temporal structure that occurs when features are summarised across the entire time course.

To visualize and interpret these results, we used the `plot_vip2d_with_groups_nogaps()` function. This function extracts the Top–N features with the highest VIP2D scores within each time point for the selected component and displays them in a faceted layout. This time-aware ranking strategy enables a focused inspection of the most informative features at each temporal slice.

A permutation-based robustness assessment was included to evaluate the stability of the selected Top–N features. In this procedure, class labels are randomly permuted, the discriminant model is refitted, and VIP2D scores are recomputed repeatedly. For each feature, a Monte Carlo permutation p-value is estimated based on how often the feature reappears among the Top–N under label permutation. Features with $p_{\text{perm}} < \alpha_{\text{perm}}$ are highlighted as robust, as their importance is unlikely to arise by chance.

Each figure consists of two aligned panels. The left panel displays the VIP2D scores as lollipop plots, where features are ranked by their importance within each time point. The Case/Control panel is included as a descriptive aid to support interpretation of the VIP2D ranking. For each selected feature–time pair, it reports the direction of the mean difference between groups (i.e., whether the feature tends

to be higher in cases or in controls at that specific time point). This information helps translate an importance score into an interpretable pattern, allowing the user to quickly understand the sign of the contrast associated with high-VIP features and to assess whether the direction is consistent across time points. Importantly, the Case/Control summary is computed post hoc from the observed data, does not enter model fitting, and does not affect VIP estimation; it should therefore be interpreted as a visualization-oriented summary of group differences rather than as evidence of a mechanistic or causal effect.

The results for the gene expression dataset are shown in Figure 4.19. In this modality, some Top–N features per time point are identified as robust according to permutation testing, indicating stable, time-specific contributions to class separation.

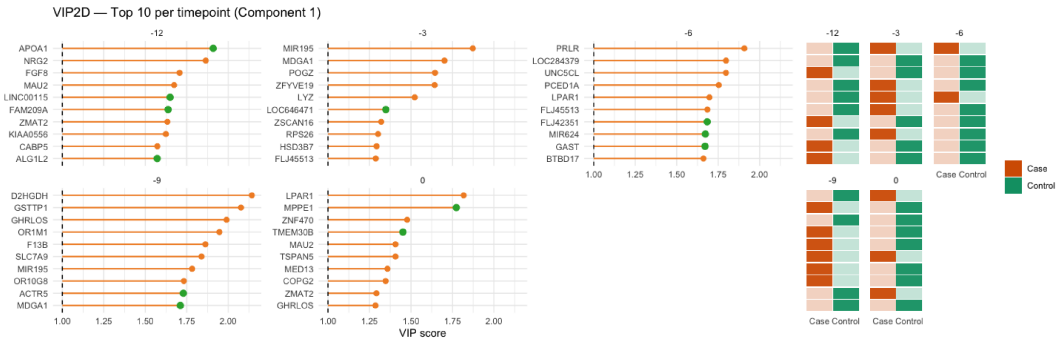


Figure 4.19: VIP2D-based feature selection across time for the gene expression dataset. Top–N features are shown for each time point. Green markers indicate features that remain significant under permutation testing ($p_{\text{perm}} < \alpha_{\text{perm}}$).

The corresponding analysis for the GCTOFX metabolomics dataset is reported in Figure 4.20. In this case, although Top–N features can be ranked by VIP2D score, only one passes the permutation significance threshold, suggesting weaker or less stable discriminant signals.

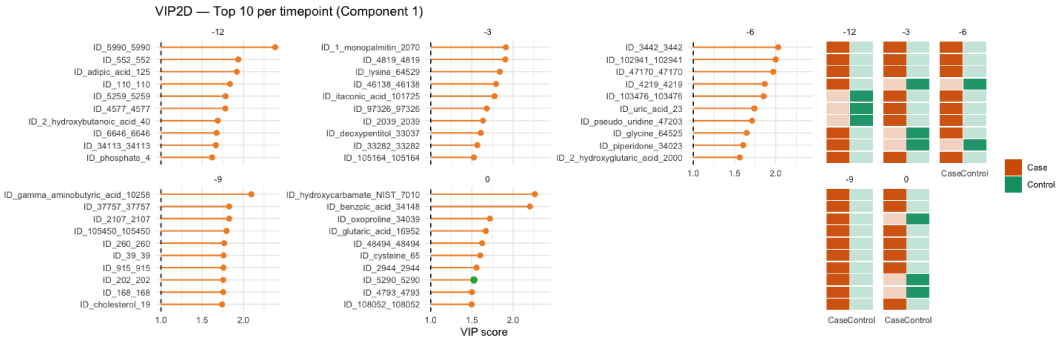


Figure 4.20: VIP2D-based feature selection across time for the GCTOFX metabolomics dataset. Only one reaches permutation significance, indicating limited robustness of the discriminant signal.

Results for the negative dietary biomarker dataset are shown in Figure 4.21. In this modality, some features identified as permutation-significant across multiple time points, indicating a comparatively more stable discriminant structure.

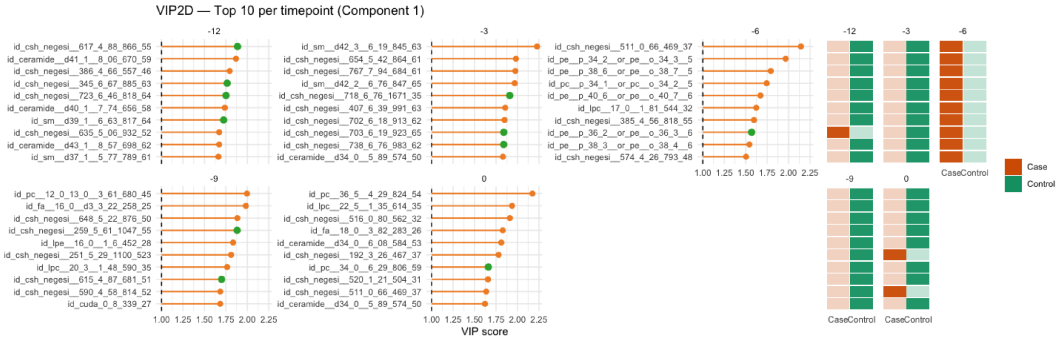


Figure 4.21: VIP2D-based feature selection across time for the negative lipidomics dataset. Several Top–N features show robustness under permutation testing.

Finally, the positive dietary biomarker dataset is reported in Figure 4.22. Similar to the GCTOFX dataset, no one feature is identified as robust, despite clear VIP-based ranking within individual time points.

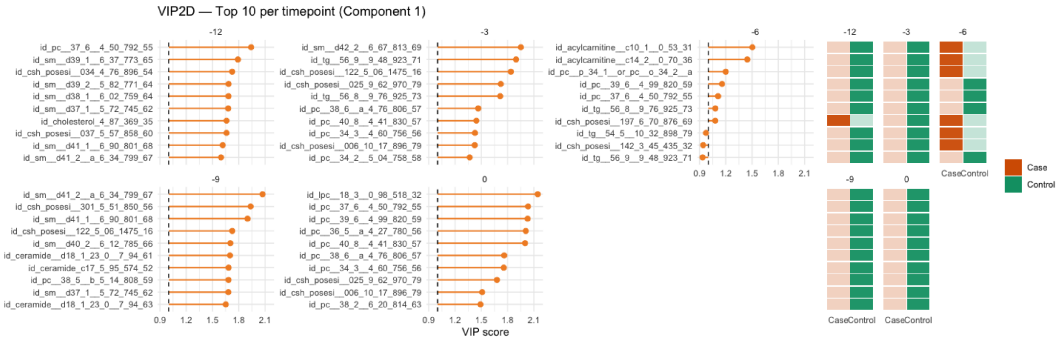


Figure 4.22: VIP2D-based feature selection across time for the positive dietary biomarker dataset.

In summary, this function enables time-specific feature prioritisation by identifying, for each time point, the variables with the highest VIP2D contributions to case–control separation. The aligned Case/Control panel then indicates the direction of the difference (higher in cases vs. controls) for each selected feature at that time point, providing an interpretable description of how separation is achieved. Permutation highlighting serves as a robustness check by flagging features that are less likely to be recovered under random label assignments.

4.3.6 Limitations and Future Directions

Limitations Despite its flexibility and interpretability, `TensorPLS` has different limitations.

First, `TensorPLS` operates on three-way tensors and exploits the native 3D structure in key preprocessing and interpretability steps. In particular, missing data are handled via an iterative Tucker-3 imputation performed directly on the tensor, preserving the multi-way organization when estimating missing entries. By contrast, the discriminant model is currently fitted on an unfolded matrix representation of X using a standard PLS-DA core. Multi-way outputs (e.g., VIP2D and Mode–3 time contributions) are then obtained by re-indexing the fitted weights and scores back to the feature–time layout. Therefore, the workflow is tensor-aware in its input handling and mode-specific interpretation, but it does not implement a fully multi-way discriminant optimization that explicitly models temporal dependence as a structured interaction during model fitting.

Second, as in most omics applications, the combination of high dimensionality ($p \gg n$) and potentially weak signals can lead to overfitting. In this work, Q^2 is computed under cross-validation and is used primarily as a resampling-based indicator for component tuning and for comparing modeling configurations under a

fixed validation scheme. These values should be interpreted as internal performance indicators rather than as definitive evidence of out-of-sample generalization.

Third, when VIP-based feature selection is performed prior to cross-validation, the same data inform both the choice of features and the subsequent evaluation, which can inflate post-selection metrics (including Q^2). A stricter assessment would embed feature selection within each training fold (nested cross-validation) or rely on an external validation set. In this thesis, post-selection models are mainly used to enhance score-space visualization and to prioritize candidate features for downstream interpretation.

Future Directions Two practical extensions could further strengthen the `TensorPLS` framework. A first extension is to provide leakage-safe performance estimation when VIP-based feature selection is applied. This can be achieved by embedding feature selection (and any label-informed step) within the training folds of the resampling scheme (nested cross-validation). A second extension concerns downstream biological interpretation. Beyond discrimination, the availability of multiple VIP representations (global VIPs, component-specific VIPs, and time-resolved VIP2D) supports structured interpretation of the selected features. For example, features can be mapped to pathway-level enrichment analyses to summarise signals at the level of biological processes rather than individual variables, and robust time-specific signatures can be used to highlight temporal windows linked to specific functional programs.

4.3.7 Conclusions

In this chapter, we introduced and evaluated `TensorPLS` as a general-purpose framework for analysing longitudinal multi-omics data represented as three-way tensors. Across multiple datasets, the workflow shows how PLS-based discriminant analysis can be combined with tensor-aware outputs to improve interpretability beyond standard two-dimensional analyses.

A key strength of the framework is its multi-resolution interpretability. By leveraging complementary VIP representations, `TensorPLS` supports the identification of (i) globally discriminant variables, (ii) component-specific drivers of separation, and (iii) time-resolved feature contributions (VIP2D), together with a direct assessment of which time points contribute most strongly to the discriminant components (Mode–3). Importantly, this time-aware view does not only indicate *which* variables are discriminant, but also highlights *when* the separation emerges, helping to prioritise temporal windows that most strongly drive case–control discrimination.

Feature selection guided by VIP scores further enables a compact representation of the signal driving case–control separation. In practice, reduced-feature models yielded clearer separation in score space and improved internal cross-validated indicators (R^2 and Q^2) in the analysed datasets. As discussed in the Limitations section, when feature selection is performed prior to cross-validation, post-selection performance estimates may be optimistic; nevertheless, the reduced models remain valuable for visualisation and for prioritising candidate features for downstream interpretation.

Overall, `TensorPLS` provides a flexible and reproducible workflow for exploratory analysis and feature prioritisation in longitudinal case–control multi-omics studies, with outputs that retain a time-aware interpretation. This makes the framework a useful starting point for downstream biological interpretation layers (e.g., pathway enrichment, network analysis, and validation in independent cohorts) in translational and systems-level applications.

5

Discussion

The results presented in this thesis support the view that endocrine therapy resistance in ERα-positive breast cancer is best interpreted as a regulatory phenotype emerging from coordinated changes across multiple molecular layers, rather than as a purely receptor-centric event. Within this framework, the BRPF1 case study provides a mechanistically oriented example of how integrative multi-omics evidence can strengthen inference beyond any single dataset. BRPF1 was treated here as a pre-selected ERα-associated candidate, and the central question was whether it contributes to ERα signalling through chromatin-mediated mechanisms. The ChIP-seq analyses establish that BRPF1 occupies thousands of genomic regions and shows substantial overlap with ERα binding, with co-occupied sites enriched for canonical ERE/ERE-like motifs and preferentially localized in promoter- and enhancer-associated partitions. This genome-wide co-occupancy pattern is consistent with BRPF1 acting within an ERα-connected regulatory network, but co-binding alone does not establish dependency or functional relevance. A key mechanistic step is therefore provided by the ERα knockdown experiment: the marked loss of BRPF1 occupancy specifically at ERα-shared loci, coupled with preservation of BRPF1 binding at ERα-independent sites, supports a model in which ERα is required to recruit BRPF1 to a subset of regulatory elements. This observation aligns with the fact that BRPF1 is not a sequence-specific DNA-binding factor, and it narrows interpretation from generic co-localization to a more specific recruitment relationship at ERα regulatory loci. Importantly, the functional relevance of this recruitment is supported by the accessibility and transcriptional readouts. Pharmacological inhibition of BRPF1 induces a strongly asymmetric accessibility response, with the

majority of differentially accessible regions losing accessibility, indicating a global tendency toward chromatin compaction under BRPF1 perturbation. The enrichment of accessibility losses at ERα-associated loci further argues against a uniform chromatin effect and instead supports preferential disruption within the ERα regulatory landscape. Consistent with this, accessibility changes show a modest but significant positive coupling with RNA-seq responses, suggesting that reduced chromatin openness contributes to transcriptional attenuation in BRPF1-inhibited conditions. Although the correlation is small, its directionality is coherent with the mechanistic hypothesis and is not expected to be large under proximity-based peak-to-gene linking, regulatory redundancy, and the many-to-one relationships between distal elements and target genes. Together, the convergence of three independent layers—(i) ERα-dependent recruitment of BRPF1 at shared loci, (ii) preferential accessibility collapse at ERα-associated regulatory regions upon BRPF1 inhibition, and (iii) concordant transcriptional attenuation—supports the interpretation that BRPF1 contributes to maintaining a permissive chromatin state within ERα-linked transcriptional circuitry. From a translational perspective, these observations motivate BRPF1 as a candidate regulatory dependency whose perturbation can weaken ERα-associated programs even in contexts where ERα expression is retained, which is a defining feature of many endocrine-resistant tumours. At the same time, the results highlight important limitations and opportunities for refinement. Peak-to-gene attribution in the ATAC–RNA integration relied on nearest-gene annotation, a practical and reproducible strategy that can misassign distal enhancer effects and dilute apparent coupling between regulatory and transcriptional layers. A natural extension is therefore to incorporate enhancer–gene maps and, where possible, 3D genome resources (Hi-C/HiChIP/PLAC-seq) from ERα-positive contexts such as MCF-7 to enable contact-informed regulatory assignment and more direct prioritisation of the enhancer-promoter loops most perturbed by BRPF1 inhibition. Such enhancer- and contact-aware linking would strengthen causal interpretation of distal regulatory events and allow more precise mapping from co-occupied loci to downstream transcriptional consequences. More broadly, the BRPF1 chapter illustrates the central thesis logic: when independent regulatory readouts converge on the same direction—occupancy, accessibility, and transcription—interpretation becomes more constrained and mechanistically meaningful than any single-layer association, providing a principled basis for prioritising candidate vulnerabilities within ERα-positive endocrine resistance settings.

Building on the locus- and chromatin-centred BRPF1 case study, the DNA methylation and transcriptomic analyses provide a complementary perspective on

endocrine resistance across heterogeneous ERα-positive cell line models, addressing a key practical question in translational epigenomics: whether resistance emerges as a dominant, unsupervised axis of methylome variation, or instead manifests as subtler changes constrained by strong baseline differences between models. The run-stratified EPIC analyses show that genome-wide methylation structure is dominated by cell line identity rather than by a universal "resistant vs sensitive" separation, despite good within-model replicate coherence. This result is important because it sets realistic expectations for multi-model resistance studies: if the baseline methylome of each cell line is the main source of variance, then resistance-associated changes are unlikely to appear as a single global clustering pattern, and cross-model interpretation must move from CpG-level overlap toward more abstract, functionally organised representations. Within this context-aware framework, DMPs show a consistent bias toward hypermethylation across most comparisons and are predominantly located in intronic regions, reinforcing the view that resistance-associated methylation changes extend beyond promoters and are distributed across broader regulatory landscapes. Because CpG-level changes are heterogeneous across models, the analysis shifts to pathway-level transcriptomic convergence using preranked Hallmark GSEA, and then to gene-core recurrence through consensus leading-edge sets, which isolate pathway cores that repeatedly drive enrichment signals across models. The promoter methylation–expression integration connects epigenetic variation to transcriptional output in an interpretable manner and formalizes the question of whether concordant promoter methylation–expression support is concentrated within pathway cores beyond the global background. Finally, the branch-based robustness analysis progressively enforces cross-line directionality constraints, yielding smaller but more defensible sets of retained pathway cores under majority and strict definitions, and making explicit the expected trade-off between robustness and statistical power.

Crucially, the main value of this contribution is methodological rather than the nomination of any single "best" pathway or gene. The central objective is to introduce and validate an analysis framework that can extract robust, interpretable cross-model signals from heterogeneous endocrine resistance datasets by moving from CpG-level variability to functionally structured summaries and by enforcing explicit reproducibility constraints. In this perspective, recurrent pathways and branch-retained gene cores should be viewed primarily as outputs of a general-purpose prioritisation pipeline: they provide a ranked and testable set of candidates that can guide follow-up biological investigation, rather than constituting the final biological claim of the chapter. The core result is therefore the framework itself—run-aware prepro-

cessing, pathway-level convergence via GSEA, consensus leading-edge construction to define stable transcriptional cores, gene-level promoter methylation summarization to enable cross-omics joins, and a pathway-resolved enrichment strategy that quantifies whether concordant promoter methylation–expression support is concentrated within pathway cores beyond the global background. The additional branch logic (baseline versus majority versus strict) is an integral part of this methodological contribution, because it makes cross-line directionality a tunable, transparent requirement and exposes the expected trade-off between robustness and statistical power. Under this design, the specific pathways retained at each stringency level serve as concrete demonstrations that the framework can (i) detect convergence where unsupervised methylome structure is dominated by cell line identity and (ii) isolate a conservative subset of signals that remain supported when directional consistency is required across multiple models.

The final Results chapter is intentionally different in scope from the two breast cancer chapters, yet it addresses a methodological need that emerges directly from them: modern translational studies increasingly generate longitudinal and multi-way omics datasets for which standard matrix-based tools are not designed. The `TensorPLS` contribution responds to this need by providing a practical, reproducible workflow for supervised analysis of three-way data, with particular emphasis on time-resolved interpretability.

A key design choice is explicitly pragmatic. Although the inputs are naturally three-way, the discriminant model is currently fitted on an unfolded two-dimensional representation of X using a standard PLS-DA core. Tensor structure is preserved where it provides the most immediate benefits: (i) tensor-native preprocessing and Tucker-3 imputation for missing data, and (ii) mode-resolved interpretability outputs, including time contributions (Mode–3) and time-resolved feature importance (VIP2D). As a consequence, the workflow is best described as *tensor-aware* rather than a fully multi-way discriminant optimisation: temporal dependence is interpreted through mode-specific outputs, but is not modelled as an explicit higher-order interaction during the optimisation step.

The evaluation on TEDDY-derived longitudinal case–control datasets provides a realistic stress test across omics layers with markedly different signal-to-noise characteristics. Baseline analyses highlight a central point relevant to many $p \gg n$ applications: discriminant structure is not uniform across modalities, and visual separation can coexist with limited cross-validated predictive stability. In `TensorPLS`, this is addressed by reporting both explained variance (R^2) and cross-validated performance indicators (Q^2), which in this thesis are used primarily for component

tuning and for comparing modelling configurations under a fixed validation scheme rather than as definitive evidence of out-of-sample generalization.

VIP-based feature selection was introduced to improve interpretability and concentrate the analysis on a reduced set of high-priority variables. However, because VIP selection is label-informed, when performed prior to cross-validation it can inflate post-selection performance estimates. For this reason, post-selection R^2/Q^2 values are interpreted conservatively as *apparent* internal indicators, and reduced-feature models are used mainly to support clearer score-space visualisation and structured feature prioritisation. Importantly, the framework extends standard PLS-DA interpretability by providing time-resolved importance through VIP2D, enabling the question "which variables matter?" to be decomposed into "which variables matter at which time points?" This is complemented by the Mode–3 time contribution analysis and by a permutation-based robustness procedure that distinguishes features highly ranked in a single fit from those consistently recovered under label randomisation.

These considerations naturally motivate two practical future directions. First, leakage-safe performance estimation can be provided by embedding feature selection (and any label-informed step) within the training folds of a nested cross-validation scheme, or by using external validation sets when available. Second, the structured VIP outputs (global, component-specific, and time-resolved) provide a natural bridge to downstream biological interpretation layers–such as pathway enrichment or network-based analyses—which are beyond the scope of this thesis but are directly enabled by the prioritized feature sets and time windows returned by `TensorPLS`.

Taken together, the three Results chapters share a common methodological lesson: robust biological insight often emerges when analysis choices are adapted to the structure of the data rather than forced into a single resolution. The BRPF1 case study illustrates the value of integrating multiple regulatory layers—binding, accessibility, and transcription—to constrain interpretation beyond what any single readout can support. The methylation/transcriptomic chapter highlights a complementary point: starting from highly granular units (single CpGs) is informative but can be limiting when signals are heterogeneous across models, motivating a shift toward more structured and interpretable representations, such as pathway-level summaries and consensus cores. This change of perspective is not merely a repackaging of results; it can enable alternative analytical routes and, as shown here, lead to the development of new, reusable frameworks.

In this context, the `TensorPLS` chapter should be read as a methodological complement to the biological analyses. It generalises the same need for structure-aware analysis to longitudinal, multi-way omics data, providing a practical workflow that

preserves tensor organisation where it most directly improves data handling and interpretability. Overall, the unifying contribution of this thesis is a consistent analytical philosophy: complex phenotypes require evidence across layers and levels of organisation, and they benefit from frameworks that respect data structure while producing transparent, interpretable summaries that can be assessed for stability and followed up in independent settings.

Bibliography

[1] Heather M. Amemiya, Anshul Kundaje, and Alan P. Boyle. The encode blacklist: Identification of problematic regions of the genome. *Scientific Reports*, 9:9354, 2019.

[2] Simon Andrews. Fastqc: A quality control tool for high throughput sequence data. Babraham Bioinformatics, 2010. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[3] Christopher J. Banks, Anagha Joshi, and Tom Michoel. Functional transcription factor target discovery via compendia of binding and expression profiles. *Scientific Reports*, 6:20649, 2016.

[4] Matt Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003.

[5] Stephen B. Baylin and Peter A. Jones. A decade of exploring the cancer epigenome—biological and translational implications. *Nature Reviews Cancer*, 11:726–734, 2011.

[6] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[7] Rasmus Bro. Multiway calibration. multilinear pls. *Journal of Chemometrics*, 10(1):47–61, 1996.

[8] Broad Institute. Picard toolkit. GitHub Pages documentation, 2019. `https://broadinstitute.github.io/picard/`.

[9] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin. *Nature Methods*, 10(12):1213–1218, 2013.

[10] Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, and William J. Greenleaf. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109:21.29.1–21.29.9, 2015.

[11] Cancer Research UK. Breast cancer stages. Official website. `https://www.cancerresearchuk.org/about-cancer/breast-cancer/stages-grades`.

[12] Centers for Disease Control and Prevention. Breast cancer. Official website. `https://www.cdc.gov/breast-cancer/index.html`.

[13] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal W. Szczesniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17:13, 2016.

[14] M. Ryan Corces, Jeffrey M. Granja, S. Soheil Shams, Brendan H. Louie, Javier A. Seoane, Wanding Zhou, Timothy C. Silva, Coen Groeneveld, Christine K. Wong, Sungkyun W. Cho, Ansuman T. Satpathy, Maxwell R. Mumbach, Katherine A. Hoadley, A. Gordon Robertson, Nicole C. Sheffield, Inanc Felau, M. Alex Castro, Benjamin P. Berman, Louis M. Staudt, Jean Claude Zenklusen, Peter W. Laird, Christina Curtis, William J. Greenleaf, and Howard Y. Chang. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413):eaav1898, 2018.

[15] Aisling M. Deaton and Adrian Bird. Cpg islands and the regulation of transcription. *Genes & Development*, 25(10):1010–1022, 2011.

[16] Paolo Dell'Orto, Giuseppe Viale, Adrian E. Hanlon Newell, Emma Walker, Irene Bai, Geoffrey Harlow, Luca Russo, and Patrick Maisonneuve. Assessing the prognostic and predictive value of ki-67 in breast cancer: A review of the evidence and practical considerations. *Breast Cancer Research*, 16:R65, 2014.

[17] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[18] Mengmeng Dong, Liping Wang, Ning Hu, Yueli Rao, Zhen Wang, Yu Zhang, et al. Integration of multi-omics approaches in exploring intra-tumoral heterogeneity. *Cancer Cell International*, 25:317, 2025.

[19] Anca M. Farcas, Sankari Nagarajan, Sabina Cosulich, and Jason S. Carroll. Genome-wide estrogen receptor activity in breast cancer. *Endocrinology*, 162(2):bqaa224, 2021.

[20] Terrence S. Furey. Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012.

[21] Liliana Garcia-Martinez, Yusheng Zhang, Yuichiro Nakata, Ho Lam Chan, and Lluis Morey. Epigenetic mechanisms in breast cancer therapy and resistance. *Nature Communications*, 12:1786, 2021.

[22] Ariella B. Hanker, Dhivya R. Sudhan, and Carlos L. Arteaga. Overcoming endocrine resistance in breast cancer. *Cancer Cell*, 37(4):496–513, 2020.

[23] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4):576–589, 2010.

[24] Niklas Heldring, Mikael Nilsson, Brian Buehrer, Erik Treuter, Jan-Åke Gustafsson, and Anders C. Wikström. Identification of tamoxifen-induced coregulator interaction surfaces within the ligand-binding domain of estrogen receptors. *Molecular Endocrinology*, 18(5):1088–1100, 2004.

[25] Niklas Heldring, Adrian Pike, Sofia Andersson, James Matthews, Guihua Cheng, Johan Hartman, Manuel Tujague, Anna Ström, Erik Treuter, Margaret Warner, and Jan-Åke Gustafsson. Estrogen receptors: how do they signal and what are their targets. *Physiological Reviews*, 87(3):905–931, 2007.

[26] Enrique Hernández-Lemus and Soledad Ochoa. Methods for multi-omic data integration in cancer research. *Frontiers in Genetics*, 15:1425456, 2024.

[27] D. Hervás et al. Sparse n-way partial least squares with r package snpls. *Computational Statistics & Data Analysis*, 2018.

[28] Illumina. Infinium humanmethylation450 beadchip: Data sheet. Illumina product datasheet (PDF), 2016. `https://www.illumina.com/documents/products/datasheets/datasheet_humanmethylation450.pdf`.

[29] Illumina. Infinium methylationepic v2.0 beadchip: Data sheet. Illumina product datasheet (PDF), 2023. https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/infinium-methylation-epic-data-sheet-m-gl-01156/infinium-methylation-epic-data-sheet-m-gl-01156.pdf.

[30] Illumina. Infinium methylationepic v2.0 kit: Product page. Illumina product page, 2026. https://emea.illumina.com/products/by-type/microarray-kits/infinium-methylation-epic.html.

[31] International Agency for Research on Cancer. Breast cancer fact sheet (GLOBOCAN 2022). Global Cancer Observatory factsheet, 2024. https://gco.iarc.who.int/media/globocan/factsheets/cancers/20-breast-fact-sheet.pdf.

[32] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[33] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002.

[34] Joanne Kim, Andrew Harper, Valerie McCormack, Hyuna Sung, Nehmat Houssami, Eileen Morgan, Miriam Mutebi, Gail Garvey, Isabelle Soerjomataram, and Miranda M. Fidler-Benaoudia. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nature Medicine*, 2025.

[35] Sarah L. Klemm, Zohar Shipony, and William J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Molecular Cell Biology*, 20(4):207–220, 2019.

[36] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[37] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[38] Kim-Anh Lê Cao, Simon Boitard, and Philippe Besse. Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12:253, 2011.

[39] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[40] Yang Liao, Gordon K. Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[41] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database (msigdb) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015.

[42] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[43] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[44] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

[45] Matteo Matteucci, Marco J. Morelli, and Marco Masseroli. Musera: Multiple sample enriched region assessment. *Briefings in Bioinformatics*, 18(3):367–381, 2017.

[46] Sebastian Moran, Clara Arribas, and Manel Esteller. Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389–399, 2016.

[47] Giovanni Nassa, Annamaria Salvati, Roberta Tarallo, Valerio Gigantino, Elena Alexandrova, Domenico Memoli, Assunta Sellitto, Francesca Rizzo, Donatella Malanga, Teresa Mirante, Eugenio Morelli, Matthias Nees, Malin Åkerfelt, Sara Kangaspeska, Tuula A. Nyman, Luciano Milanesi, Giorgio Giurato, and Alessandro Weisz. Inhibition of histone methyltransferase dot1l silences erα gene and blocks proliferation of antiestrogen-resistant breast cancer cells. *Science Advances*, 5(2):eaav5590, 2019.

[48] National Cancer Institute. Breast cancer. Official website. `https://www.cancer.gov/types/breast`.

[49] National Cancer Institute. Breast cancer stages. Official website. `https://www.cancer.gov/types/breast/stages`.

[50] National Cancer Institute. Definition of triple-negative breast cancer. NCI Dictionary of Cancer Terms. `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/triple-negative-breast-cancer`.

[51] Erika Orrantia-Borunda, Pedro Anchondo-Nuñez, Liliana E. Acuña-Aguilar, Fernando O. Gómez-Valles, and Carlos A. Ramírez-Valdespino. Subtypes of breast cancer. In *StatPearls*. StatPearls Publishing, 2022.

[52] C. Kent Osborne and Rachel Schiff. Mechanisms of endocrine resistance in breast cancer. *Annual Review of Medicine*, 62:233–247, 2011.

[53] Hongchao Pan, Richard Gray, Jeremy Braybrooke, Christina Davies, Carolyn Taylor, Paul McGale, Richard Peto, Kathleen I. Pritchard, Jonas Bergh, Mitch Dowsett, and Daniel F. Hayes. 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *New England Journal of Medicine*, 377(19):1836–1846, 2017.

[54] Peter J. Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 2009.

[55] Rima Patel, Paula Klein, Amy Tiersten, and Joseph A. Sparano. An emerging generation of endocrine therapies in breast cancer: a clinical perspective. *npj Breast Cancer*, 9(1):20, 2023.

[56] Ruth Pidsley, Elena Zotenko, Timothy J. Peters, Mitchell G. Lawrence, Gail P. Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J. Clark. Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome Biology*, 17(1):208, 2016.

[57] Aleix Prat, Estela Pineda, Barbara Adamo, Patricia Galván, Aránzazu Fernández, Lidia Gaba, Marc Diez, Margarita Viladot, Ana Arance, and Montserrat Muñoz. Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast*, 24(Suppl 2):S26–S35, 2015.

[58] Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[59] Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11):e1005752, 2017.

[60] Daniel Ruiz-Perez, Haibin Guan, Purnima Madhivanan, Kalai Mathee, and Giri Narasimhan. So you think you can pls-da? *BMC Bioinformatics*, 21(Suppl 1):2, 2020.

[61] Andrew K. Shiau, David Barstad, Paul M. Loria, Lin Cheng, Pamela J. Kushner, David A. Agard, and Geoffrey L. Greene. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 95(7):927–937, 1998.

[62] Mikhail Spivakov. Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays*, 36(8):798–806, 2014.

[63] Rory Stark and Gordon Brown. Diffbind: Differential binding analysis of chip-seq peak data. Bioconductor package, 2011. `https://bioconductor.org/packages/release/bioc/html/DiffBind.html`.

[64] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[65] Andrew E. Teschendorff, Francesco Marabita, Martin Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450k dna methylation data. *Bioinformatics*, 29(2):189–196, 2013.

[66] Y. Tian, T. J. Morris, A. P. Webster, Z. Yang, S. Beck, A. Feber, and A. E. Teschendorff. Champ: updated methylation analysis pipeline for illumina beadchips. *Bioinformatics*, 33(24):3982–3984, 2017.

[67] Valentin Todorov, Valeria Simonacci, Maria A. Di Palma, and Michele Gallo. Robust tools for three-way component analysis of compositional data: The r package rrcov3way. *Behaviormetrika*, 2025.

[68] Eleni Toska et al. Epigenetic mechanisms of cancer progression and therapy resistance in breast cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 2024.

[69] Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[70] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020.

[71] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[72] Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[73] Tao Ye, Andreas R. Krebs, Marie-Ange Choukrallah, Céline Keime, Frédéric Plewniak, Irwin Davidson, and László Tora. seqminer: an integrated chip-seq data interpretation platform. *Nucleic Acids Research*, 39(6):e35, 2011.

[74] Federico Zambelli, Graziano Pesole, and Giulio Pavesi. Pscanchip: finding over-represented transcription factor binding site motifs in sequences from chip-seq experiments. *Nucleic Acids Research*, 41(W1):W535–W543, 2013.

[75] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9):R137, 2008.