ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

DATA SCIENCE AND COMPUTATION

Ciclo 37

Settore Concorsuale: 05/I1 - GENETICA

Settore Scientifico Disciplinare: BIO/18 - GENETICA

MULTI-OMICS AND MACHINE LEARNING APPROACHES FOR COMPLEX DISEASES

Presentata da: Giacomo Cavalca

Coordinatore Dottorato

Daniele Bonacorsi

Supervisore

Paolo Uva

Co-supervisore

Andrea Cavalli

Esame finale anno 2026

Outline

Abstract

Complex diseases arise from an intricate interplay of genetic, environmental, and lifestyle factors, making them difficult to decipher. Integrating multiple molecular layers such as genomics, epigenomics, transcriptomics, and proteomics enables a more comprehensive understanding of disease mechanisms. While this multi-omics perspective offers unprecedented opportunities, it also introduces major challenges in data integration and interpretation. In this context, machine learning (ML) represents a powerful tool to analyze heterogeneous datasets, uncover hidden patterns, and prioritize features most relevant for diagnosis, prognosis, and therapeutic development.

This thesis is structured around three projects aimed at exploring how multi-omics data and machine learning can be combined to advance the study of complex diseases.

In the first project, we investigated stochastic epigenetic mutations (SEMs) as biomarkers of disease reactivation in Juvenile Idiopathic Arthritis (JIA). We demonstrated that SEM burden can both reflect and precede clinical reactivation among patients with otherwise indistinguishable clinical profiles. Furthermore, by adopting a multi-omics approach, we identified candidate epigenetically-driven differentially expressed genes and prioritized potential therapeutic targets.

In the second project, we developed machine learning models to predict cell type-specific regulatory variants contributing to Alzheimer's disease heritability. These models enabled the identification of novel candidate expression Quantitative Trait Loci (eQTLs) and risk genes, offering new perspectives on the molecular mechanisms underlying the disease.

In the third project, we investigated Syndrome of Undifferentiated Recurrent Fever (SURF) through a hierarchical integration of genome-wide analyses and multi-omics QTL approaches. This integrative strategy enabled the identification of candidate molecular mechanisms linking genetic variation to inflammatory pathways and refining disease-specific biomarkers.

1. Background

1.1. Complex diseases

Complex diseases, also known as multifactorial diseases, are pathological conditions that arise from a combination of genetic, environmental, and lifestyle factors [1]. Unlike monogenic diseases, which are caused by a single genetic mutation, these conditions do not follow simple Mendelian inheritance patterns, and inherited risk factors alone are not sufficient to predict disease development [2].

Examples of complex diseases include Alzheimer's disease, diabetes, various forms of cancer, autoimmune and autoinflammatory disorders such as arthritis and certain recurrent fever syndromes, as well as neurodevelopmental conditions such as autism. What characterizes all these conditions is the intricate interplay between inherited genetic predisposition and environmental exposures (e.g., diet, smoking, infections), which may act through epigenetic and regulatory mechanisms.

A major challenge in studying multifactorial diseases is that the effect of any single factor may be obscured or confounded by others [1,2]. This complexity arises from the multifaceted nature of disease etiology, encompassing several phenomena that interact among each other and complicate genetic analysis.

The polygenicity of these diseases implies that multiple genetic variants, each conferring a modest effect, must be present before a pathological condition manifests [3]. At the same time, different genetic loci can converge on common molecular pathways, leading to similar disease phenotypes [4].

Moreover, epistatic interactions can modulate the overall effect of genetic variants, producing context-dependent outcomes that vary across individuals. Environmental vulnerability and gene-environment interactions add another layer of complexity, where genetic predispositions may only manifest in the presence of specific environmental triggers or conditions.

Furthermore, common and rare variants together shape the genetic architecture of complex traits, where numerous small-effect alleles perturb interconnected biological networks rather

than single genes [5]. This network-based view suggests that disease emerges from subtle, system-level perturbations rather than isolated mutations.

Additionally, developmental and age-related gene expression patterns imply that the timing of gene activity during critical life stages can significantly influence disease manifestation, while general aging processes can alter gene function and mutation rates over time. Each of these factors can act independently or in combination, often in a context-dependent manner, making it difficult to isolate the contribution of individual genes [2,3].

To address these limitations, research on complex diseases should move toward integrative approaches to assess the joint contribution of multiple genetic, molecular, and environmental components. Rather than focusing on individual elements, this comprehensive view aims to capture their combined and interacting effects, which ultimately determine disease susceptibility and progression. Such a holistic approach is essential for deciphering the biological complexity that characterizes complex diseases.

1.2. Genetic approaches

1.2.1. Genome-wide association studies

Genome-wide association studies (GWAS) have represented a significant breakthrough in identifying common genetic variants associated with complex traits [6]. GWAS are hypothesis-free methods that aim to detect statistical associations between genetic variants and phenotypic traits by testing whether allele frequencies differ between ancestrally similar but phenotypically distinct individuals [7]. Although GWAS can in principle include different forms of genomic variation, such as copy number variants (CNVs) or sequence variations, they most commonly focus on single-nucleotide polymorphisms (SNPs), which are the most abundant form of variation in the human genome.

Since the first large-scale GWAS by the Wellcome Trust Case Control Consortium [8], more than 5,700 studies conducted across over 3,300 traits have been published [7,9]. These efforts have revealed tens of thousands of robust associations, clarifying the polygenic architecture of complex diseases and providing mechanistic insights into disease biology. Early studies revealed key associations such as *CFH* in macular degeneration [10], *PTPN22* in autoimmune diseases [11,12], *FTO* in obesity [13], and *IL-12/IL-23* signaling in Crohn's disease [14],

exemplifying the clinical impact of GWAS discoveries [15]. Collectively, GWAS findings have been instrumental in highlighting disease-relevant genes and pathways, and supporting drug target discovery and repurposing for complex diseases [16].

However, GWAS have notable limitations. These studies mostly rely on SNP arrays that interrogate only a pre-selected subset of genetic variants, predominantly common variants (typically with minor allele frequency > 5%), making them underpowered for discovering rare variants with potentially larger effect sizes [3]. Moreover, many GWAS hits fall in non-coding regions, complicating functional interpretation [17], and often encompass multiple variants in linkage disequilibrium (LD), which hinders the identification of causal variants [18]. Additionally, GWAS typically explain only a modest proportion of disease heritability, giving rise to the so-called "missing heritability" problem [19]. The proportion of variance explained by identified variants varies substantially across traits and for complex disorders is often less than 10% [20]. This discrepancy suggests that a considerable portion of genetic risk remains unexplained, possibly residing in rare variants, structural variations (such as CNVs, insertions, deletions, or inversions), gene-gene or gene-environment interactions, and other molecular mechanisms not captured by standard GWAS approaches.

1.2.2. Beyond GWAS: next-generation sequencing and rare variant analysis

To address the limitations of GWAS, next-generation sequencing (NGS) technologies [21] have revolutionized the landscape of genetic research by enabling comprehensive, unbiased assessment of genomic variation. Unlike array-based methods that interrogate only pre-selected variants, NGS technologies directly read DNA sequences, allowing the detection of both common and rare variants, as well as structural variations, and novel mutations not represented on genotyping arrays.

Within the NGS framework, whole-exome sequencing (WES) selectively targets and sequences the protein-coding regions of the genome (exome), comprising approximately 1-2% of the total genome. By focusing on exons, WES provides a cost-effective strategy to capture variants most likely to directly affect protein structure and function, making it particularly powerful for identifying rare mutations in complex diseases [22]. In contrast, whole-genome sequencing (WGS) captures both coding and non-coding regions across the entire genome,

enabling a comprehensive assessment of genomic variation [23]. Compared with SNP array-based approaches, these sequencing technologies allow the detection of both common and rare variants without prior selection.

A wide variety of statistical methods have been specifically developed for rare-variant association analysis, differing in their underlying assumptions and in the type of data they can incorporate. Burden tests [24–26] aggregate the information from multiple variants within a gene or functional region into a single summary score, assuming that all variants influence the phenotype in the same direction. Weighted burden tests [27] additionally assign weights to variants based on their allele frequency or predicted functional impact, while adaptive burden tests [28–30] aim to accommodate bidirectional effects by selecting optimal weighting schemes. Variance-component (kernel) tests, such as C-alpha [31] and SKAT [32], also allow for bidirectional effects, but may be underpowered when most variants have effects in the same direction or when many variants are causal within a locus. To balance these trade-offs, unified approaches such as SKAT-O [33] combine burden and variance-component tests to achieve robust performance across a wide range of genetic architectures.

To further refine association signals and address the challenge of pinpointing causal variants within associated loci, fine-mapping approaches have been developed. These methods aim to disentangle the effects of variants in LD by estimating the posterior probability that each variant is truly causal, given the observed association statistics and the local LD structure [18,34]. This process helps distinguish causal variants from non-functional tag SNPs, enabling more precise biological interpretation and functional validation.

Despite these advancements, a substantial fraction of heritability remains unexplained, highlighting the need to incorporate regulatory, epigenetic, and environmental dimensions of genomic variation into disease models [35,36].

1.2.3. Quantitative Trait Locus analysis

Insights from GWAS revealed that many disease-associated variants map to noncoding regions, often overlapping with putative regulatory elements such as enhancers or promoters [17]. This observation highlighted the need for approaches that could decipher the functional effects of genetic variants beyond traditional protein-coding changes.

Quantitative trait locus (QTL) analysis provides a framework to investigate how genetic variants influence phenotypic variability [37]. A QTL is defined as a genomic locus that contributes to variation in a quantitative trait. When the trait being measured is molecular in nature, we refer to it as a molecular QTL (molQTL). Several types of molQTL have been characterized, each capturing different layers of biological regulation: RNA expression (eQTL), protein expression (pQTL), splicing (sQTL), methylation level (meQTL), chromatin accessibility (caQTL), histone modification (hQTL), and many others [38].

Molecular QTL analyses have proven crucial for linking genetic variation to intermediate molecular phenotypes, thereby offering mechanistic insights into how genetic variants exert their effects on complex traits and diseases. However, given that multiple traits may be associated with the same genomic region, a major challenge lies in disentangling whether a molecular QTL and a disease association signal share the same causal variant, or instead reflect distinct variants in LD [39].

To address this challenge, statistical colocalization methods have been developed [40–42]. These approaches aim to test whether the same genetic variant is likely responsible for two distinct associations, typically one with a molecular trait (e.g. DNA methylation, chromatin accessibility, gene expression) and the other with a complex phenotype (e.g., disease status from GWAS). Colocalization helps prioritize candidate regulatory elements and effector genes that mediate the genetic association, thereby providing a biologically interpretable mechanism.

In parallel, statistical approaches such as Mendelian randomization (MR) have been developed to assess whether molecular traits exert a causal effect on disease [43–46]. MR leverages genetic variants associated with a molecular trait as instrumental variables to infer the direction and magnitude of causal effects on disease risk. When supported by colocalization evidence indicating that the same variant underlies both associations, MR provides stronger evidence for causal mediation rather than mere correlation due to LD, enhancing the biological interpretability of molecular QTL-disease relationships.

Large consortia like GTEx [47] and eQTLGen [48] have enabled the systematic identification of molecular QTLs across diverse tissues, developmental stages, and cell types. These large-scale efforts have revealed that regulatory effects are highly context-specific. For instance, a genetic variant may act as an eQTL, meQTL, or caQTL only in certain immune cell types, developmental windows, or environmental conditions. This context-dependency has major implications for interpreting GWAS findings, especially in complex diseases with strong tissue

or cell-type specificity, such as autoimmune, neuropsychiatric, or metabolic disorders [47,49,50].

Building upon these tissue-level insights, recent developments in single-cell sequencing technologies have further extended QTL mapping to cellular resolution. Single-cell QTL studies integrating genotype with RNA-sequencing (scRNA-seq), ATAC-seq (scATAC-seq), or other multi-omic assays, have uncovered cell type-specific and context-dependent regulatory variants, revealing fine-grained mechanisms that are often masked in bulk analyses [51,52].

To translate these molecular insights into a functional interpretation of GWAS findings, several integrative frameworks have been developed. FUMA [53] maps GWAS signals to genes by integrating functional annotations, eQTL, and chromatin interaction data. S-PrediXcan [54] and SMR [55] leverage eQTL information to link predicted gene expression levels to complex traits, providing evidence of regulatory mediation. Similarly, eMAGMA [56] extends the MAGMA framework [57] by incorporating tissue-specific eQTL data to prioritize genes whose expression is likely influenced by trait-associated variants. Collectively, these post-GWAS approaches bridge statistical associations and molecular mechanisms, facilitating the biological interpretation of genetic signals in a tissue- and cell-type–specific context.

1.3. The “multi-omics” approach

The recognition that genetic variation alone cannot fully explain complex disease susceptibility has led to the emergence of multi-omics approaches, which integrate diverse molecular layers to obtain a more comprehensive view of disease biology [58]. The suffix “-omics” denotes the large-scale characterization of a set of molecules, thus “multi-omics” refers to the simultaneous analysis of multiple high-throughput data types, including genomics, transcriptomics, epigenomics, proteomics, metabolomics, and microbiomics, among many other emerging disciplines.

Zitnik et al. [59] distinguished between two types of omics integration: horizontal integration, which analyzes the same omics data across different sample groups, and vertical integration, which examines multiple omics variables within the same samples.

Multi-omics studies often overlook critical issues related to horizontal integration, despite it being one of the first aspects that should be addressed, particularly in multi-center studies or

when data are generated at different time points or with different technologies. These challenges include batch effects between sample groups, technical variability across datasets, and difficulties in ensuring data comparability when identical omics measurements are performed under varying experimental conditions or across different laboratories.

Vertical integration, on the other hand, seeks to leverage the complementary information provided by different omics layers to build a more complete picture of biological processes. This integrative framework acknowledges that disease phenotypes result from dynamic interactions across multiple biological layers, spanning from DNA sequence variation and epigenetic modifications to gene expression, protein abundance, and metabolite levels. By capturing these interconnected molecular systems, multi-omics provides a holistic perspective of disease mechanisms and represents a significant advancement in the study of complex diseases [58]. Compared with single-omics analyses, it enables a systems-level understanding of pathological processes and opens new opportunities for biomarker discovery, patient stratification, and precision medicine.

1.3.1. Benefits of multi-omics integration

The integration of multiple omics layers offers substantial benefits for biomedical research by enabling a more comprehensive exploration of disease biology. Unlike single-omics studies, which are limited to one molecular dimension, multi-omics approaches allow researchers to trace the flow of information from the initial causes of disease, whether genetic, environmental, or developmental, to their functional consequences. This perspective provides a deeper understanding of how different biological processes interact and converge in disease pathogenesis [58].

By combining different molecular data types, it becomes possible to identify potential causal changes that contribute to disease onset and progression. Such integration is essential for distinguishing between processes that are directly involved in pathogenesis and those that merely represent downstream or reactive effects [60]. Furthermore, multi-omics facilitates the identification of reliable biomarkers for diagnosis and prognosis, as well as the identification of driver genes that serve as critical points for therapeutic intervention. These discoveries

enable more refined patient stratification, allowing the delineation of molecular subtypes that inform personalized therapeutic strategies and drug repurposing efforts [61].

Another major advantage lies in the discovery of novel drug targets. Integrative approaches that merge GWAS findings with QTL maps, for example, can highlight high-confidence effector genes that might otherwise remain undetected in single-layer analyses [60]. Additionally, multi-omics has proven particularly powerful in addressing disease heterogeneity. By uncovering molecular subgroups or refining existing classifications, it provides a framework for understanding and treating highly heterogeneous conditions, where variability across patients poses a central challenge to effective management.

Recent analyses emphasize how rapidly the multi-omics field is expanding [62], with the number of publications on multi-omics integration more than doubling between 2022 and 2023 compared to its first mention in 2002. This surge reflects not only the promise of multi-omics for deciphering disease mechanisms, but also its growing role in precision medicine, from early disease detection to the design of individualized therapeutic strategies.

1.3.2. Challenges and Future Directions

While multi-omics integration has opened new avenues for understanding complex diseases, it continues to face several challenges that require careful consideration and methodological innovation. A central issue is the heterogeneity and complexity of the data itself. Since omics layers are generated using diverse technologies and platforms, the resulting datasets often differ in structure, scale, and format. Rigorous preprocessing steps, including filtering, normalization, and correction for batch effects, are indispensable but can also introduce additional layers of complexity [61]. Another methodological limitation is the high dimensionality of multi-omics datasets, where the number of features is much higher than the number of samples, creating what is known as the "curse of dimensionality" [63]. This issue is further exacerbated in studies of rare complex diseases, where limited sample sizes amplify the imbalance between the number of features and the number of individuals. Because omics analyses involve testing vast numbers of variables simultaneously, statistical power is strongly dependent on both the number of individuals included and the magnitude of observed effects. Human studies are further complicated by lifestyle- and environment-related confounders, which make large and well-characterized cohorts essential for obtaining reliable results and minimizing false

discoveries [58]. Beyond statistical considerations, the challenge of distinguishing causal alterations from reactive or secondary changes persists. While genomic variants can often be assumed to precede disease onset, establishing causality in transcriptomic, epigenomic, or metabolomic alterations remains far more elusive, particularly in cross-sectional studies.

The rapid growth of the field has also highlighted the need for robust analytical methodologies. Despite the availability of general computational frameworks, the absence of universally accepted gold-standard pipelines means that analyses often require customized strategies. These strategies must be tailored not only to the dataset but also to the specific research questions, and should be carefully designed before data generation begins. Moreover, biological considerations can limit the feasibility of multi-omics studies. For many complex diseases, such as osteoarthritis, Alzheimer's disease, or type 2 diabetes, access to disease-relevant primary tissues is often highly restricted, relying on post-mortem material or invasive surgical procedures [60].

Furthermore, issues of representation and diversity remain pressing. To date, most genetic and molecular studies have disproportionately focused on cohorts of European ancestry, which have led to biases in the interpretation and translation of findings. Addressing this imbalance will be critical for future research, not only to improve the global applicability of scientific discoveries but also to ensure equitable access to the benefits of precision medicine across populations [60]. Expanding the inclusion of under-represented groups will therefore be essential for generating a more complete and accurate landscape of human molecular variation.

Emerging perspectives highlight further aspects that will shape the future of multi-omics. Mohr et al. [62] emphasized the need for integrative infrastructures capable of handling longitudinal data, and the development of AI-based tools for extracting clinically actionable insights. At the same time, ethical considerations, data privacy, and security remain central issues, with novel technologies offering potential solutions for safeguarding sensitive information. Meeting these challenges will require coordinated efforts across disciplines, from computational biology to clinical sciences, to translate multi-omics discoveries into routine healthcare applications.

1.4. Machine learning for omics data

The integration of multi-omics data presents both unprecedented opportunities and significant analytical challenges. The heterogeneity, complexity, and scale of these datasets often exceed the capacity of traditional statistical approaches, making machine learning an essential tool in this field.

Machine learning (ML) is a branch of artificial intelligence that enables algorithms to learn patterns and make predictions from data without being explicitly programmed. Within ML, deep learning (DL) refers to a family of methods based on artificial neural networks with multiple layers, capable of modeling highly complex and non-linear relationships. There are several paradigms of learning approaches, but the main subdivision is between supervised and unsupervised. The key distinction is that supervised learning uses labeled data to train a model for making predictions by learning from known inputs and outputs. In contrast, unsupervised learning uses unlabeled data to discover hidden patterns, structures, or anomalies within the data, without any explicit guidance.

While ML is indispensable for analyzing individual omics layers, it is particularly powerful for integrating heterogeneous molecular profiles and capturing the intricate cross-talk between biological systems [64,65].

1.4.1. Dimensionality reduction methods

One first way to tackle the high dimensionality issue of multi-omics datasets is the dimensionality reduction, an optional step that can be used prior to the integration, aimed at reducing the number of variables, thereby decreasing dataset dimensionality and noise. There are two types of dimensionality reduction: feature selection and feature extraction.

Feature selection directly removes noisy and redundant variables, retaining a smaller set of features that are presumed to hold most of the relevant information. It enhances computing efficiency, reduces complexity, minimizes overfitting, and improves model interpretability. Feature selection methods are categorized into filter-based (e.g. RCA [66], ReliefF [67]), which use statistical measures independent of ML models; wrapper-based (e.g., Recursive Feature Elimination [68]), which repeatedly apply a predictive ML model on different feature subsets and keep the ones showing the best performance; and embedded methods (e.g., tree-based

feature importance [69], regularization methods like LASSO, Ridge and Elastic Net [70]), which incorporate feature selection directly into the algorithm.

On the other hand, feature extraction methods transform the original input features into a new, smaller set of variables, which are linear or non-linear combinations of the initial features. The goal is to retain relevant information while reducing noise and redundancy, thereby decreasing complexity and improving computational efficiency. A drawback is that the new features may be less interpretable as they are no longer direct biological measurements. The most popular feature extraction method is the Principal Component Analysis (PCA) [71], which transforms the original features into a new set of uncorrelated variables called principal components, that explain the most variation in the dataset. Extensions like Kernel PCA [72], Bayesian PCA [73], Correspondence Analysis (CA) [74], and Independent Component Analysis (ICA) [75] address PCA's limitations (e.g., sensitivity to outliers, inability to handle non-linear trends). Most feature extraction methods include sparsity constraints, using regularization methods in order to remove useless or redundant features (e.g., Sparse PCA [76] and Sparse Canonical Correlation Analysis [77]).

1.4.2. Multi-omics integration methods

Several classifications of multi-omics integration strategies have been proposed in the literature. A widely adopted framework was introduced by Ritchie et al. [78], who distinguished three main categories: concatenation-based integration, where datasets are merged prior to analysis, transformation-based integration, in which each dataset is mapped or transformed into a new representation before analysis, and model-based integration, where each omics type is analyzed separately and the results are subsequently combined. Building on this foundation, Picard et al. [65] refined the terminology and extended the framework to five distinct integration strategies, renaming them as early (concatenation-based), mixed (transformation-based), and late (model-based), and introducing two additional classes: intermediate and hierarchical.

- Early integration is based on the concatenation of all datasets into a single large matrix. This process drastically increases the number of variables while keeping the number of observations constant. This exacerbates challenges related to complexity, noisiness, and high dimensionality of the data, making learning difficult. Differences in size between

omics datasets can also create a learning imbalance, with algorithms spending more time on omics with more variables. Furthermore, this strategy ignores the specific data distributions of each omics, potentially leading ML models to find irrelevant patterns. Despite these drawbacks, early integration remains common due to its simplicity, ease of implementation, and the ability of ML models to directly uncover interactions between different biological layers. It often requires prior dimensionality reduction or feature selection. DL models have been frequently used in this context due to their flexibility and power in detecting relevant patterns. Chaudhary et al. [79] proposed an autoencoder framework that reduces multi-omics data dimensionality to extract compact DL-derived features predictive of liver cancer survival. Similarly, Xie et al. [80] integrated multi-omics and clinical data into a deep neural network architecture connected to a Cox proportional hazard model, incorporating group lasso regularization to enhance survival prediction accuracy in cancer patients. However, a major challenge with DL models is their "black box" nature, which makes interpretation difficult, a crucial aspect in biomedical studies.

- The mixed integration strategy aims to overcome the limitations of early integration by independently transforming each omics dataset into a simpler, less dimensional, and less noisy representation. This new representation reduces heterogeneities between datasets, such as differences in type or size, and allows their combined analysis through classical ML models. Several transformation approaches have been proposed. Kernel-based methods exploit implicit mappings to high-dimensional feature spaces, enabling linear models to capture inherently non-linear relationships. For instance, Multiple Kernel learning (MKL) [81] combines kernels calculated for each omics dataset to produce a global similarity matrix. Graph-based approaches, on the other hand, represent each omics dataset as a network. Integration can be achieved by fusing these networks into a homogeneous structure, as in Similarity Network Fusion (SNF), where patients are represented as nodes connected by weighted similarity edges [82]. Alternatively, omics can be modeled as multi-layer (or multiplex) networks, where each layer corresponds to a molecular level and inter-layer connections are either inferred or obtained from external knowledge, enabling the exploration of network topology to identify relevant molecules and perturbed pathways [83]. Graph representations can also be used to derive low-dimensional embeddings from each network, which are then combined and analyzed through classical ML models, while more advanced methods

such as Graph Neural Networks (GNN) or Graph Convolutional Neural Networks (GCN) are specifically designed to process graph-structured data [84]. Finally, artificial neural networks (ANNs) provide another strategy to learn meaningful latent representations from omics datasets. By processing each dataset in separate layers, ANNs can extract deep features, new variables learned by the hidden layers of the model. Unsupervised neural architectures such as autoencoders compress the data into bottleneck layers, generating compact representations that can then be concatenated or integrated through shared layers in more complex neural networks for downstream analysis [85].

- Intermediate strategies jointly integrate multi-omics datasets without requiring either independent preliminary transformation or simple concatenation. They generally output new constructed representations, one common to all omics and some specific to each omics, upon which further analysis can be performed. This step reduces dimensionality and complexity, but these methods are often used after feature selection and robust pre-processing due to dataset heterogeneity. They are often formulated with the assumption that different datasets share a common latent space that can reveal underlying biological mechanisms. Examples include extensions of Non-negative Matrix Factorization (NMF) [86], such as joint and integrative NMF [87], which infer a common matrix describing latent relationships between each omics dataset. Other methods include Canonical Correlation Analysis (CCA) [88] and Co-Inertia Analysis (CIA) [89], which infer omics-specific factors while maximizing a joint measure like correlation. Among the most widely adopted approaches are latent factor models, including Multi-Omics Factor Analysis (MOFA) [90] and its single-cell extension MOFA+ [91], which learn hidden factors that capture both shared and modality-specific sources of variation, thereby disentangling biological from technical heterogeneity. These factors provide versatile low-dimensional representations that can be used for downstream tasks such as subgroup identification, outlier detection, data imputation, and, in the case of MOFA+, scalable integration of large single-cell multi-omics data with flexible sparsity constraints. A related approach is DIABLO (Data Integration Analysis for Biomarker discovery using Latent cOmponents) [92], which identifies correlated latent components across multiple omics while simultaneously selecting molecular features that discriminate between phenotypic groups, thereby enabling predictive modeling in new samples. Similarly, iCluster [93] employs a joint latent variable model to capture

associations across data types and variance-covariance structures within them, facilitating the discovery of clinically relevant subtypes (e.g., tumor subgroups) directly from integrated profiles. Overall, the main strength of intermediate methods lies in their ability to uncover shared inter-omics structures while retaining complementary information unique to each modality.

- Late integration is the most straightforward strategy for handling multi-omics datasets, where ML models are trained independently on each dataset and their predictions are subsequently combined. The main advantage of this approach is its simplicity: it leverages existing tools optimized for individual omics types and avoids the difficulties associated with directly merging heterogeneous data. For example, Sun et al. [94] trained neural networks separately on gene expression, copy number variation, and clinical information, and then aggregated their outputs linearly to generate a single prediction for cancer prognosis. A more sophisticated example is MOGONET [95], which employs GCNs to learn from both omics features and sample similarity networks for omics-specific classification, then integrates these predictions through a cross-omics discovery tensor for final classification, explicitly modeling correlations between omics-specific prediction spaces. However, late integration cannot capture inter-omics interactions, since each model is trained in isolation and no information is shared across omics.

- Hierarchical strategies integrate multi-omics data by explicitly incorporating prior knowledge of regulatory relationships between molecular layers, reflecting the sequential nature of biological processes. For example, genetic variations at the nucleotide level can alter gene expression, which may lead to functional changes in proteins and ultimately manifest as phenotypic differences. By embedding this kind of prior biological knowledge, hierarchical methods aim to better capture the causal flow of information across omics. This approach often relies on external information from interaction databases and the literature, allowing the challenges of integration to be addressed separately for each dataset. Methods in this category include iBAG [96], a framework developed to investigate the associations between methylation, gene expression and patient survival. Robust Network [97] focuses on modeling the regulatory effects of copy number variations on gene expression, distinguishing between dominant cis-acting and weaker trans-acting influences. Hierarchical integration can also be used to infer gene regulatory networks, as shown by Zarayeneh

et al. [98], who combined gene expression, CNVs, and DNA methylation to improve regulatory interaction prediction. Finally, Fortelny and Bock [99] proposed a deep neural network where nodes represent biological entities such as genes or proteins and edges represent known interactions, with layers reflecting the flow of information in the cell. A key advantage of this framework is its interpretability, as activated nodes directly highlight the molecular drivers underlying the prediction.

No single strategy can be considered universally optimal, as the suitability of each approach strongly depends on the characteristics of the datasets and the specific research objective. This highlights the importance of carefully balancing methodological complexity, interpretability, and predictive performance when designing multi-omics studies.

2. From theory to practice: ML approaches applied to multi-omics datasets in complex diseases

2.1. Project 1. Stochastic epigenetic mutation profiles as biomarkers of clinical activity in juvenile idiopathic arthritis: a multi-omic machine learning approach for gene prioritization.

This study was conceived within the Clinical Bioinformatics Unit at Giannina Gaslini Institute, in close collaboration with Dr. Giovanni Fiorito, with whom I jointly developed the initial idea and analytical framework. Under his scientific supervision, I designed and conducted all computational analyses, spanning from preprocessing of DNA methylation and transcriptomic data to the implementation of multi-omics feature selection pipelines and external validation.

The work was carried out in close collaboration with clinical experts from the Giannina Gaslini Institute, who contributed providing domain expertise and clinical insights, as well as with external collaborators who shared datasets for validation purposes.

In this study, we investigated juvenile idiopathic arthritis (JIA), a complex autoimmune disease, by combining multi-omics data and machine learning methods to uncover molecular mechanisms linked to disease reactivation and to identify potential therapeutic targets.

2.1.1. Introduction

Juvenile idiopathic arthritis (JIA) is a broad term that describes a group of arthritis conditions of unknown origin, characterized by onset before the age of 16 and lasting for more than six weeks [100,101]. Its pathogenesis is thought to result from a complex interplay between genetic and environmental factors [102]. Epigenetic mechanisms, particularly DNA methylation (DNAm), have been described as mediators in gene-environment interactions, potentially contributing to autoimmune disease onset and progression [103]. Emerging evidence suggests

that specific DNAm profiles could serve as biomarkers for predicting therapeutic responses in autoimmune diseases, including JIA [104].

Despite advances in treatment, a substantial proportion of JIA patients experience relapse after anti-tumor necrosis factor (anti-TNF) therapy withdrawal [105]. Since patients who maintain inactive disease (ID) are often clinically indistinguishable from those who relapse, identifying molecular signatures that anticipate disease reactivation remains a crucial unmet need in clinical management.

Traditional approaches in epigenome-wide association studies (EWAS) commonly rely on linear models to detect differentially methylated regions between groups. However, these approaches may overlook critical biomolecular mechanisms driven by rare epigenetic events. To address this limitation, Gentilini et al. [106] proposed a methodological approach to identify rare stochastic epigenetic mutations (SEM) not shared among subjects. The total number of SEMs per subject, termed epigenetic mutation load (EML) [107], has been proposed as a biomarker of cumulative DNA damage resulting from endogenous and exogenous exposures throughout life [108]. SEMs and EML have been shown to increase exponentially with age and to associate with immune dysregulation and higher risk of complex diseases in prospective studies [109].

Additionally, other DNAm-based biomarkers of biological aging such as epigenetic clocks [110,111], have recently gained popularity.

These models estimate an individual's "epigenetic age" from a subset of CpG sites, and the deviation from chronological age provides a proxy of biological aging acceleration or deceleration [112]. Since the development of the original epigenetic clock by Horvath [110], several refined versions have been proposed. Current literature distinguishes between first-generation clocks, aimed to predict chronological age, and next-generation clocks, which are trained to predict healthspan- or mortality-related processes [113]. In adult rheumatoid arthritis (RA), accelerated epigenetic aging has been observed compared to healthy controls [114]. However, evidence in pediatric autoimmune diseases remains scarce, as most clocks were trained on adult cohorts [114]. A few studies have suggested a link between accelerated epigenetic aging and neurodevelopmental disorders [115,116], but their relevance to pediatric autoimmune disease remains unexplored.

Building on these observations, we investigated differences in EML and several epigenetic clocks between JIA patients who maintained inactive disease (ID) status eight months after anti-TNF therapy withdrawal and those who did not (NO ID). Our aim was to identify epigenetic biomarkers that are predictive of clinical prognosis, as well as those that change as a consequence of disease reactivation.

Furthermore, through a multi-omics approach, we leveraged ML methods to prioritize genes with a higher number of SEMs in NO ID compared to ID JIA patients, with the ultimate goal of uncovering novel molecular mechanisms underlying disease reactivation and proposing potential therapeutic candidates.

2.1.2. Methods

Dataset composition and preprocessing

We reanalyzed a publicly available dataset of 68 JIA patients (56 polyarticular and 12 extended oligoarticular) from the GEO repository (GSE89253) [117]. Whole-genome DNAm and gene expression profiles were obtained from $CD4^+$ T cells using the Illumina HumanMethylation450 BeadChip and the Illumina HumanHT-12 v4.0 expression BeadChip arrays, respectively. Samples were collected at two time points: at anti-TNF therapy withdrawal (T_0), when all

patients had inactive disease (ID), and eight months after the baseline (T_{end}), when 14 out of 68 patients did not maintain inactive disease (NO ID) status (**Table 1.1**).

	ID	**NO ID**
# of JIA patients	30 (68%)	14 (32%)
# of Females (%)	21 (70%)	10 (71%)
Age at T0: mean (std. dev.)	9.91 (3.55)	13.42 (3.69)
Disease form:		
# of oligoarticular extended (%)	5 (17%)	1 (7%)
# of polyarticular (%)	25 (83%)	13 (93%)

Table 1.1 - Characteristics of the study population stratified by clinical activity at T_{end}. No significant differences in the distribution of sex by clinical activity (chi-squared test p = 0.99). Statistically significant differences in the distribution of age by clinical activity (T test p = 0.04). No significant differences in the distribution of disease form by clinical activity (chi-squared test p = 0.69).

Twenty-four out of 68 JIA patients experienced acute disease flares at T_{end}. Based on the findings of the original study [117], which reported strong epigenetic similarities between ID and flare patients, these individuals were excluded from our analyses. Instead, we focused our comparison on ID versus NO ID patients to increase the likelihood of identifying biomarkers of disease reactivation at an early stage. All wet-lab procedures are detailed in the original publication [117].

As a sensitivity analysis, we repeated the main analyses after aggregating flare patients with the ID group (ID + Flare vs. NO ID), following Spreafico et al. [117].

DNAm data were loaded as β-values, defined as the ratio between the fluorescence intensity of methylated probes and the total probe intensity. CpG sites with detection p-values greater than 0.01 were considered unreliable and set as missing. Probes and samples with a call rate below 95% were excluded from further analyses, and the remaining missing values were imputed using the median across all samples. After quality control and filtering, the dataset comprised 338,688 CpGs across 88 samples, with no samples discarded during preprocessing.

Transcriptomic data underwent an analogous preprocessing workflow. Transcripts and samples exhibiting more than 5% missing values were removed, while residual missing ones were imputed using the median across patients. Following these steps, 10,243 transcripts were retained for downstream analyses, and no samples were excluded.

To exclude the presence of potential patient clusters with aberrant epigenetic or transcriptomic profiles, we performed preliminary exploratory analyses using Uniform Manifold Approximation and Projection (UMAP) [118] on both DNAm and transcriptomic data (see **Figure S1** and **Figure S2** in the original publication [119]).

Study design

The overall workflow is summarized in **Figure 1.1**. After preprocessing, we identified SEMs and computed EML (natural logarithm of the total number of SEMs), along with 19 epigenetic clocks for each patient at the two time points.

Linear regression models were then applied to assess associations between epigenetic biomarkers (EML and clocks) and disease activity (ID vs. NO ID), adjusting for age and sex (**Figure 1.1a and 1.1b**). EML emerged as the most robust and statistically significant epigenetic biomarker distinguishing ID from NO ID patients.

We annotated epimutated CpG sites on transcription factor binding sites (TFBSs) [120] and KEGG pathways [121] (**Figure 1.1c**) and we performed supervised feature selection using (i) SKAT-O [33], (ii) Elastic Net Regularized Regression [122] and (iii) eXtreme Gradient Boosting (XGBoost) [123] (**Figure 1.1d**).

Given the modest sample size, a ML-based strategy was preferred over classical regression to minimize type II errors and enhance biological interpretability. Features consistently identified across all three algorithms were retained as robust and biologically meaningful candidates.

Genes mapping to multiple selected TFBSs or KEGG pathways were defined as hub genes and further filtered by differential expression analysis between ID and NO ID patients (**Figure 1.1e**). This process yielded a final list of genes where the higher number of epimutations in NO ID patients corresponded to differential gene expression (named epigenetically driven up- and down-regulated genes in NO ID patients).

Subsequently, the list of candidate up- and down-regulated genes was used as the input for the Connectivity Map (CMAP) tool [124] (**Figure 1.1f**) to identify compounds capable of reversing the disease-associated transcriptional profile. CMAP compares user-defined gene signatures with expression profiles induced by chemical perturbations in multiple cell lines, highlighting candidate drugs with opposite regulation patterns.

Finally, results were validated (**Figure 1.1g**) using three independent datasets:

1) Genome-wide DNAm profiles from $CD4^+$ T cells of 56 JIA patients and 57 age- and sex-matched healthy controls (Illumina 450K BeadChip);

2) Gene expression data from peripheral blood mononuclear cells (PBMCs) of 17 children with rheumatoid factor–negative (RF^-) polyarticular JIA, profiled using Illumina HumanWG-6 v3.0 expression BeadChip at two time points: baseline (all patients with inactive disease) and follow-up (8 with active disease, 9 with continued inactive disease);

3) RNA-seq data from PBMCs of 85 JIA patients spanning three clinical categories: inactive treated, active treated, and active untreated.

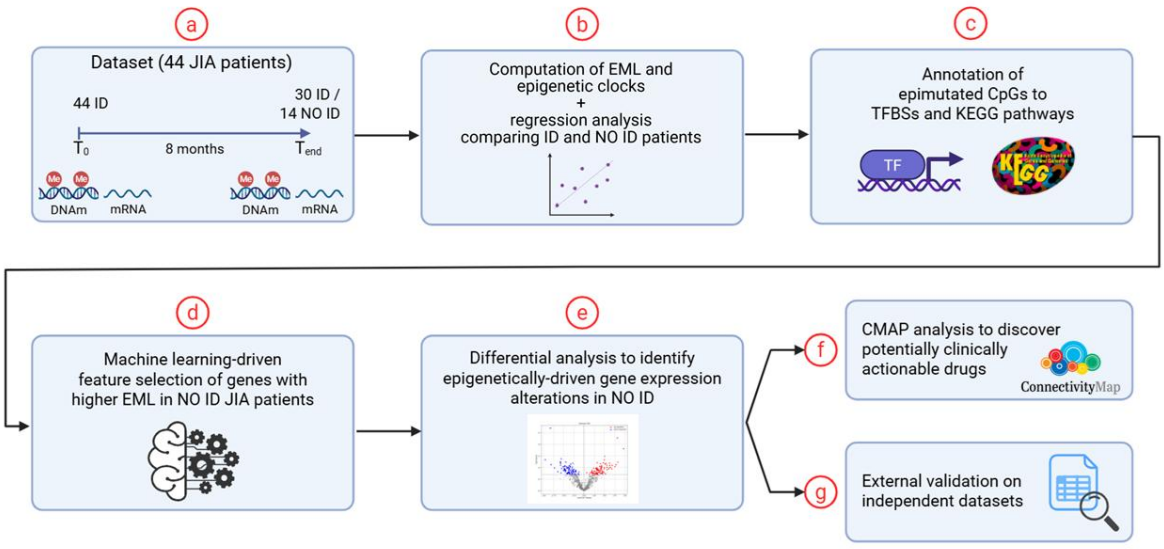

Figure 1.1 – Schematic representation of the workflow: (a) Description of the dataset including

DNAm and mRNA whole-genome data from 68 JIA patients at two time points; (b) Identification of SEMs, computation of EML, and 19 epigenetic clocks for each patient. Regression models to investigate differences between ID and NO ID patients; (c) Annotation of epimutated CpGs to transcription factor binding sites and KEGG pathways; (d) Supervised feature selection using SKAT-O, XGBoost, ElasticNet machine learning to identify a list of candidate epigenetically-driven differential expressed genes in NO ID JIA patients; (e) Differential expression analysis on candidate genes selected by machine learning. (f) Interrogating Connectivity MAP (CMAP) tool for identification of potential therapeutic targets. (g) Signature validation on independent datasets.

Figure created with biorender.com.

Detection of SEMs and computation of EML

Considering DNAm profiles of ID JIA patients at T_0 as the reference, we computed quartiles and interquartile range (IQR) of methylation β-values for each CpG probe. SEMs were then identified as extreme methylation values exceeding three times the IQR: β-values lower than Q1-(3 x IQR) or higher than Q3+(3 x IQR) for all patients at both T_0 and T_{end} time points. The EML was defined as the natural logarithm of the total number of SEMs per individual as described in previous studies [107].

Computation of epigenetic clocks, SEMs and EML

Nineteen epigenetic clocks, including first- and next-generation models, were computed for all JIA patients to investigate their association with clinical activity.

Each epigenetic clock was computed as a linear combination of a specific set of CpG sites with varying number of CpGs and weight based on the specific epigenetic clock. We considered the following epigenetic clocks: (i) principal components adjusted version of Horvath original pan tissue clock (Horvath v1), Horvath skin and blood clock (Horvath v2), Hannum, Levine's DNAmPhenoAge, DNAmGrimAge and DNAmTL as described by Higgins-Chen et al. [125], (ii) Bernabeu and Zhang epigenetic clocks as implemented in Hillary et al. [126], (iii) DNAmAdaptAge, DNAmDamAge and DNAmCausAge epigenetic clocks by Ying et al. [127], (iv) DNAmFitAge epigenetic clock by Jokai et al. [128], the mitotic age tracking of the number

of stem-cell divisions proposed by Teschendorff et al. [129], and (vi) epigenetic clocks developed using non-linear analytical approaches like the Bayesian Neural Network (BNN), the pediatric buccal cells (PedBE) clock, Wu's clock design to predict biological age in children, the Best Linear Unbiased Prediction (BLUP) clock, the Zhang Elastic Net (EN) clock, the predictor of gestational age DNAmGA, as implemented in the *methylclock* R package by Pelegí-Sisó et al. [130]. For each clock, we computed the epigenetic age acceleration (AA) as the residuals of the linear regression of the epigenetic age on chronological age. White blood cell (WBC) proportions were estimated according to the procedure implemented in the *methylclock* R package. The estimated WBC proportions, shown in **Figure 1.2**, confirm that the DNAm data were derived from $CD4^+$ T cells. As a result, we did not adjust for white blood cell (WBC) proportions in our analyses, although it is a common practice in epigenetic studies.

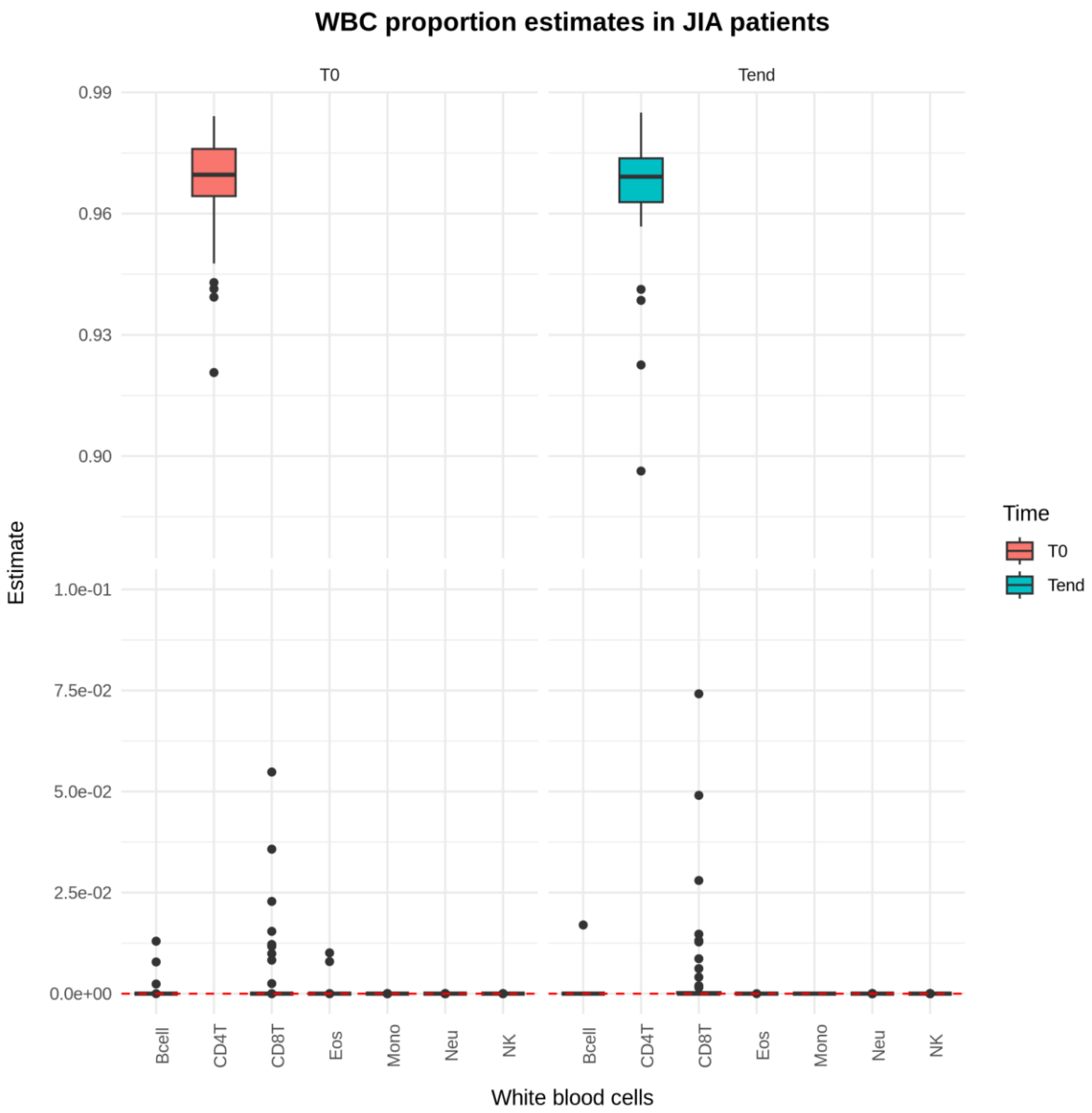

Figure 1.2 – Boxplot of the white blood cell (WBC) proportion estimates according with the Houseman et al. approach (methylclock R package) at T0 and Tend.

Association of clinical activity with epigenetic biomarkers

The association between JIA clinical activity and epigenetic biomarkers (clocks and EML) was assessed using linear regression models adjusted for age and sex. Associations were evaluated separately at T_0 and T_{end}, with each epigenetic biomarker treated as the dependent variable and clinical activity as the predictor. Effect sizes were expressed as standardized differences across

all epigenetic biomarkers, allowing direct comparison among measures originally expressed in different units.

ML-based feature selection

The first step in the feature selection process involved annotating CpG sites on TFBSs using ChIP-seq data generated from $CD4^+$ T cells within the ENCODE project, via the *MeinteR* R package [131], and associating them with KEGG pathways using the *KEGGREST* R package. This resulted in aggregated data, where for each sample, we obtained the number of epimutations associated with 159 TFBS and 358 KEGG pathways.

We then applied three different machine learning approaches (SKAT-O, XGBoost, and Elastic Net Regression) to identify the most relevant TFBSs and KEGG pathways that exhibited a higher number of SEMs in NO ID patients at T_{end}. To verify the expression of selected TFs in $CD4^+$ T cells, we queried the GSE107011 dataset, which contains 114 RNA-seq samples of sorted immune cell populations, using the R package *celldex* [132].

The SKAT-O [33] procedure is a statistical method originally developed to test associations between rare genetic variants and phenotypes. In this study, we adapted it for rare epigenetic mutations, assuming similar underlying biological mechanisms, where the accumulation of epimutations leads to altered gene regulation and, potentially, disease progression.

Specifically, the algorithm computes an empirical p-value by evaluating multiple values of ρ, representing the weight of a given feature. We used a grid search including 15 possible values of ρ equally distributed within the [0,1] interval. Empirical p-values were computed using 'boostrap' option in the SKAT function from the *SKAT* R package [33] with n.Resampling = 10^6.

For Elastic Net Regression and eXtreme Gradient Boosting (XGBoost), we used the *LogisticRegression* class from the *linear_model* Python module of scikit-learn (*sklearn*) [133] and the *XGBClassifier* from the Python *xgboost* library [123], respectively.

Before model training, we normalized features to have average 0 and standard deviation 1, using *StandardScaler* from *sklearn.preprocessing*). Both models were trained using a stratified 5-fold cross-validation (*StratifiedKFold* from Python *sklearn.model_selection*), ensuring class balance within each fold. *GridSearchCV* was used for hyperparameter tuning, with the

“average_precision” metric to select the best-performing model based on predefined parameter grids.

For ElasticNet, the LogisticRegression hyperparameters were tuned as follows: *penalty* was set to "elasticnet" and *l1_ratio* to 0.5, ensuring a balanced combination of L1 and L2 regularizations. The *solver* hyperparameter was specified as "saga" [134], being the only solver compatible with the elasticnet *penalty*. The regularization strength (*C*) was tested at values of 1 (default), 5, and 10. Stopping tolerance (*tol*) was evaluated at 10^{-4} and 10^{-3}, and the maximum number of iterations (*max_iter*) was tested at 5,000 and 10,000 to allow the solver to converge. Finally, *class_weight* was set to "balanced" to adjust the weights to balance the different sample sizes in the two output classes.

Feature importance for ElasticNet was derived from the absolute values of the model's coefficients, with non-zero coefficients indicating significant features.

For XGBoost model, the XGBClassifier hyperparameters were tuned as follows: *n_estimators* was tested with 200 and 500 boosting rounds. The learning rate (*eta*) was tested at 0.3 (default) and 0.1. The L2 regularization term (*lambda*) was validated for 2, 1 and 0 to test regularization at different strengths. The *colsample_bytree* hyperparameter representing the subsample ratio of columns considered by each tree in the model was tested at 0.3, 0.5, and 0.7 rates. Finally, *scale_pos_weight* was computed using the *compute_class_weight* function from *sklearn.utils.class_weight* to handle class imbalance.

Feature importance in XGBoost was determined using the 'weight' metric, reflecting how frequently a feature was used to split a node across trees.

All the statistical analyses have been performed with R v4.3.3 and Python v3.12.4.

External validation of epigenetic signature

To validate the epigenetic component of our findings, we analyzed an independent DNAm dataset from a previously published study by Chávez-Valencia et al. [135]. This dataset comprises genome-wide DNAm profiles from $CD4^+$ T cells of 56 JIA patients and 57 age- and sex-matched healthy controls, generated using the Illumina HumanMethylation450 BeadChip.

To ensure comparability between discovery and validation datasets, batch effects resulting from differences between the two recruitment centers were corrected using the ComBat algorithm [136] as implemented in the *sva* R package [137]. This empirical Bayes method adjusts for systematic technical variation while preserving biological variation of interest, enabling valid cross-dataset comparisons.

Identification of SEMs, computation of EML, and annotation of epimutated CpG sites on the ComBat-corrected dataset were performed using the same analytical pipeline used in the discovery dataset. SEMs were subsequently annotated and tested for enrichment within the genes and TFBSs prioritized in the discovery phase using SKAT-O. Pathway enrichment analysis on the list of significant genes was performed using the Molecular Signature Database (MSigDB) [138] database via the *enrichR* R package [139].

External validation of transcriptomic signature

To validate the transcriptomic findings, we analyzed two independent datasets from the GEO repository (GSE26112 and GSE79970), both including transcriptomic profiles from JIA patients. Unlike our primary dataset, these validation datasets do not include DNAm data and were derived from PBMCs rather than purified $CD4^+$ T cells.

The GSE26112 dataset included gene expression profiles from 17 children with rheumatoid factor negative (RF^-) polyarticular JIA, measured using Illumina HumanWG-6 v3.0 expression BeadChip. Samples were collected at two time points: baseline (all patients with inactive disease) and follow-up (8 with active disease and 9 with continued inactive disease).

The GSE79970 datasets consisted of RNA sequencing data from PBMC of 85 JIA patients across three clinical categories: inactive treated patients, active treated patients, and active untreated patients.

For both datasets, we performed differential expression analyses using *DESeq2* R package [140], comparing inactive vs. active JIA patients (adjusting for treatment in the GSE79970 dataset). The resulting statistics were combined through random-effects meta-analysis using the *metafor* R package [141]. Differentially expressed genes with consistent direction of fold change compared to our signature have been cross-referenced with the MSigDB to investigate possible enrichment with JIA- and inflammation-related pathways.

2.1.3. Results

Significant association between EML and clinical activity

We investigated differences in EML and 19 epigenetic clocks by clinical status between ID and NO ID patients at both time points (T_0 and T_{end}) using linear regression models adjusted for age and sex. In these analyses, epigenetic biomarkers were considered as dependent variables and the disease activity as the predictor. Effect sizes were standardized (expressed as standard deviation differences between the two groups) to enable comparison across different epigenetic biomarkers, which are originally expressed in varying units of measurement. The results are summarized in **Figure 1.3**.

We observed higher EML values in NO ID compared to ID patients at T_0 ($\beta = 0.37$; 95% CI [-0.12–0.86]; p=0.15), with this difference becoming more pronounced and statistically significant at T_{end} ($\beta = 0.60$; 95% CI [0.13–1.07]; p=0.02).

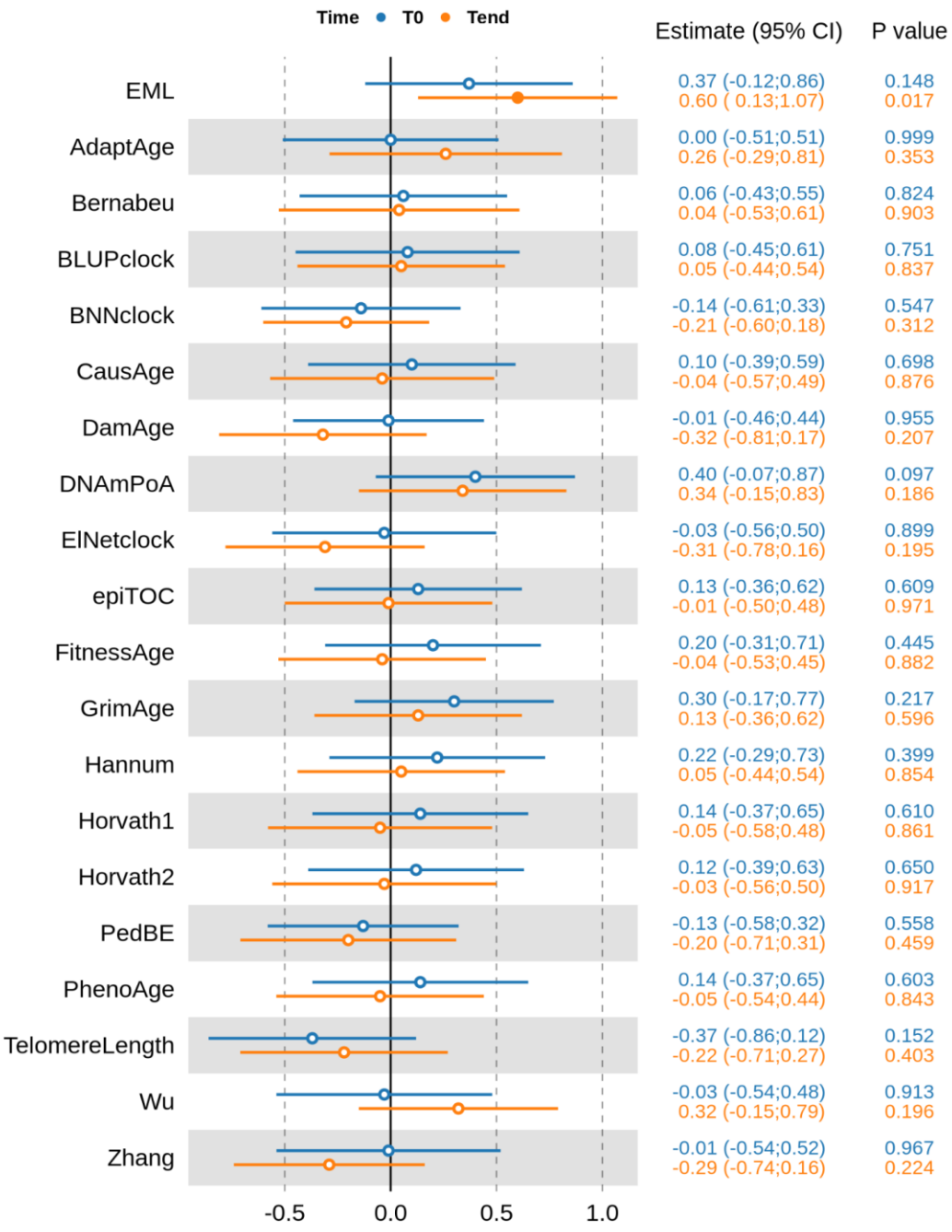

Figure 1.3 – Results from the linear regression analyses presented as a forest plot. Estimates (with 95% confidence intervals) for the differences between ID and NO ID patients across various epigenetic clocks are presented at two time points, T_0 (blue) and T_{end} (orange). EML values are higher in NO ID compared to ID patients at both T_0 and T_{end}, with a statistically significant difference observed at T_{end} (β = 0.60; 95% CI [0.13–1.07]; p=0.02). Instead, no epigenetic clock has been found to be associated with JIA clinical activity.

Sensitivity analysis including flare patients

To assess robustness, we aggregated flare patients with the ID group (ID + Flare vs. NO ID) and repeated the primary models. Linear regressions confirmed higher EML in NO ID at both T_0 $\beta = 0.52$ (95% CI: 0.07–0.97; $p = 0.028$) and T_{end} $\beta = 0.78$ (95% CI: 0.43–1.13; $p < 0.001$). Effect sizes were consistent with the main analysis, with substantial overlap of confidence intervals and concordant directions. For the transcriptomic layer, differential expression on the candidate genes showed strong agreement with the discovery signature across the 104 genes in common (Spearman $\rho = 0.98$, $p = 1.6 \times 10^{-69}$). These findings indicate that excluding flare patients did not materially affect conclusions on EML or the directionality of the epigenetically driven transcriptomic signature.

Multi-omics feature selection identified candidate genes

Based on the above findings, we focused our investigation on EML. Linear regression results indicated a higher number of epimutated CpG sites in NO ID than ID patients across the whole genome. To prioritize the genes showing the most relevant differences between ID and NO ID patients, we employed three distinct feature selection approaches using SKAT-O, Elastic Net Penalized regression, and XGBoost. Features jointly identified by all three algorithms were considered robust candidates for further biological interpretation and therapeutic prioritization.

For this analytical step, we focused exclusively on the epigenetic and transcriptomic data at T_{end}. This allowed us to focus on genes exhibiting increased epimutations likely driven by disease reactivation, which appears to be the predominant mechanism according to the linear regression results. The available sample size did not allow a gene-level comparison between the two groups with enough statistical power. Indeed, we conducted SKAT-O, ElasticNet, and XGBoost analyses, aggregating data at the gene level.

However, this level of aggregation did not sufficiently capture significant differences between the two groups in the SKAT-O analysis, nor did it yield robust classification results with ElasticNet and XGBoost. Therefore, we annotated the epimutated CpG sites into TFBSs according to the ChIP-seq experiments performed by the ENCODE consortium in $CD4^+$ T cells [120], and into KEGG pathways [121]. This approach allowed to aggregate epimutated CpG sites within 159 TFBS and 358 KEGG pathways reducing the number of features and consequently increasing the statistical power.

Using this approach, SKAT-O identified 98 TFBSs with significantly higher numbers of SEMs in NO ID than ID patients with false discovery rate (FDR) adjusted p-value lower than 0.1. No KEGG pathways reached statistical significance at this threshold (see further details in **Supplementary Tables S1** and **Table S2** in the original publication [119]). Elastic Net and XGBoost classification algorithms were trained using a 5-fold cross-validation framework, with hyperparameters optimization through grid search. Feature selection was performed by evaluating the importance metrics unique to each model: non-zero coefficients and the frequency of their use in tree splits, for ElasticNet and XGBoost, respectively. Detailed performance metrics for XGBoost and ElasticNet models, including F1 score, accuracy, precision, recall, and average precision, can be accessed at the original publication [119].

ElasticNet identified 61 TFBSs and 50 KEGG pathways, while XGBoost identified 48 TFBSs and 84 KEGG pathways. Considering the overlap among all three methods, 12 TFBSs were consistently selected (**Figure 1.4**) as the most robust features distinguishing ID from NO ID patients. According to the GSE107011 dataset, 10 out of the 12 TFs targeting the selected TFBSs were indeed expressed in $CD4^+$ T cells. No KEGG pathway was identified by all three ML models.

The 12 TFBSs were further used to identify hub genes, defined as those mapping to multiple selected TFBSs. In **Supplementary Table S9** in the original publication [119], we listed the top 100 genes associated with the 12 TFBSs identified by machine learning, ranked by their frequency in each annotation. By using the frequency of genes annotated to multiple TFBSs, we observed a robust selection of hub genes with several of them being regulated by more than 80% of the 12 selected TFBSs and the gene *FOXP1* targeted by all 12. Based on the above, we defined the list of candidate genes as those that are targeted by at least 50% of the identified TFs (N=909).

As the last step, we further prioritized genes in which a higher number of SEMs in NO ID than ID led to a significant difference in gene expression profile. For this step, we performed a differential gene expression analysis using transcriptomic data from the 909 candidate genes at T_{end}. We found 157 significant genes (FDR < 0.2), of which 80 were up- and 77 down-regulated in NO ID.

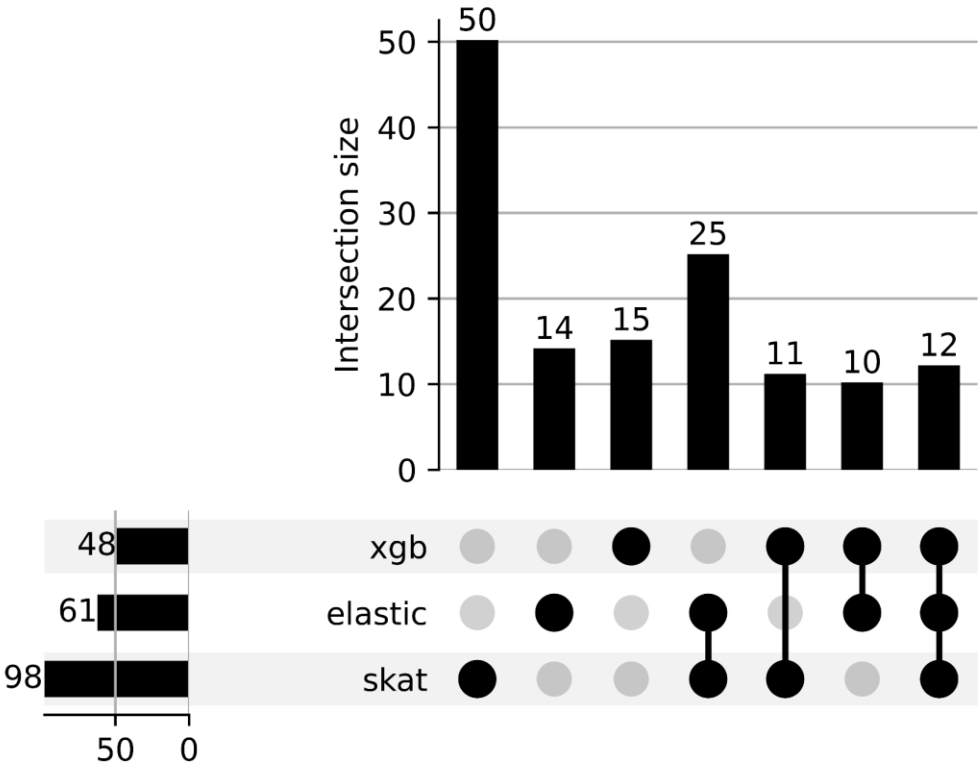

Figure 1.4 - The upSet plot showing the overlap of TFBSs selected by three different methods: XGBoost (xgb), ElasticNet (elastic), and SKAT-O (skat). Bars in the top panels represent the number of intersecting features for each combination of methods, while the bars on the left panels indicate the total number of features selected by each method independently. Black dots connected by lines indicate the specific combinations of methods contributing to the intersection size.

Connectivity MAP identified potentially clinically actionable drugs

We used the CMAP tool [124] to identify a list of potential drugs that counteract the identified signatures associated with clinical reactivation of JIA, and that could potentially be used in the clinical practice. Previously selected genes were used as input for CMAP and it provided a list

of candidate compounds, ranked by the raw connectivity score which is proportional to their ability to reverse the gene expression signatures associated with the disease reactivation.

Only compounds with a FDR-adjusted q-value < 0.05 and negative normalized connectivity score were retained. We focused on pre-selected immune-relevant cell lines, including '*JURKAT*', '*THP1*', '*K562*', '*HL60*', '*BJAB*', '*NALM6*', '*CD34*'. According to these criteria, CMAP identified 12 candidate compounds with potential therapeutic relevance (see **Table 2** of the original publication [119]).

External validation confirmed higher EML in active JIA patients

Application of the ComBat algorithm successfully harmonized DNAm distributions between discovery and validation datasets. Linear regression analyses, adjusted for age and sex, confirmed significantly higher EML levels in active JIA patients from the validation dataset compared to inactive JIA patients from the discovery set ($p = 0.04$), although the effect size was smaller than that observed in the discovery dataset. No significant difference was found between healthy controls in the validation dataset and inactive JIA patients in the discovery dataset.

Furthermore, SKAT-O analysis revealed a significantly higher SEM burden in active JIA patients compared to healthy controls in 7 out of 12 TFBSs and in 16 out of 157 genes prioritized in the discovery set. The direction and magnitude of the SEM burden across the 12 TFBSs were overall confirmed in the validation dataset, with a positive correlation in $\log_2$FC values ($\rho = 0.46$). Pathway enrichment analysis of genes showing higher SEM burden in JIA patients than healthy controls revealed enrichment in key inflammatory pathways, including '*TNF-alpha signaling via NF-kB'* and '*TGF-beta signaling'*, which has been previously implicated in JIA and rheumatoid arthritis [142–144], further supporting the biological relevance of our findings.

External validation refined JIA transcriptomic signature revealing inflammation pathways

Differential gene expression analyses on independent external datasets were performed comparing inactive vs. active JIA patients and combining the results through meta-analysis.

Among the genes in our signature, 22 showed significant differential expression (FDR-adjusted p-value < 0.2) with consistent direction of fold change compared to our original analysis (15 up-regulated; 7 down-regulated genes). Gene set enrichment analysis on the reduced gene list against the MSigDB revealed significant enrichment in inflammatory and immune related pathways like '*TNF-alpha signaling via NF-kB*', '*KRAS signaling*', and '*TGF-beta signaling*', consistent with results of the enrichment analysis on differentially epimutated genes, demonstrating convergent evidence between epigenetic and transcriptomic signatures across different validation datasets.

2.1.4. Discussion

We reanalyzed a publicly available dataset including DNAm and gene expression profiles from 44 JIA patients at two time points: withdrawal of anti-TNF therapy (T_0) and eight months later (T_{end}), when 14 patients experienced disease reactivation.

For our analysis, we focused on comparing patients who maintained inactive disease with those who did not, excluding 24 patients who had acute flares at T_{end}. The decision to exclude "flares" patients from the analysis follows the observation by Spreafico et al. [117] that DNAm and transcriptomic profiles of flare patients closely resemble those of ID rather than NO ID patients, plausibly because pathogenic immune cells rapidly extravasate during acute flare events, leaving minimal and transient epigenetic traces in circulating cells. Based on this rationale, we chose to focus our comparison on ID versus NO ID patients, accepting a reduction in statistical power to avoid overlooking relevant biomarkers potentially masked by the inclusion of flare patients. Nevertheless, our sensitivity analyses demonstrated that including "flares" patients yielded consistent results.

In contrast to the weighted gene co-expression network analysis (WGCNA) [145] conducted in the original study [117], where epigenetic variability in the major histocompatibility complex and T-cell functions emerged as key correlates of clinical activity, we pursued a complementary strategy centered on rare stochastic epigenetic events.

Epigenetic mutation load as a biomarker of disease reactivation

Guided by evidence implicating SEMs in autoimmunity [107,146,147], we evaluated EML and epigenetic clocks as candidate biomarkers of JIA reactivation. EML was higher in NO ID than ID patients at both T_0 and T_{end}, reaching statistical significance at T_{end}. These results suggest that elevated EML could both contribute to and result from disease reactivation, highlighting its potential as a predictive biomarker for monitoring disease status. The differences between the two groups were more pronounced at T_{end}, leading us to interpret the EML increase as likely a consequence of JIA clinical reactivation.

Although some previous studies have already examined SEM accumulation in pediatric populations [148], to our knowledge, none have investigated its longitudinal trajectory to estimate the rate of accumulation under physiological conditions. This gap limits the ability to contextualize our findings against a reference healthy population. However, our results indicate that EML increases over time in JIA patients, with a significantly faster rate observed in NO ID compared to ID. Worthy of note, the presence of slightly higher EML in NO ID compared to ID patients at T_0 (when all patients have inactive disease), despite not statistically significant, suggests that epigenetic changes might precede clinical symptoms. If confirmed in future studies, this would have important implications for early disease monitoring and a deeper understanding of disease etiology. By contrast, we found no evidence of association between epigenetic clocks and JIA clinical activity.

Machine learning-driven gene prioritization

Building on these results, we leveraged ML methods to prioritize molecular features associated with a higher EML in NO ID patients. Given the limited power for genome-wide gene-level testing (14 vs. 30 patients), we aggregated epimutated CpGs by TFBSs and KEGG pathways. The use of three distinct feature selection methods significantly enhances the robustness of our findings. We focused on features extracted by all three approaches to reduce false positive rate and increase the reliability of the selected biomarkers. Importantly, all three methods effectively addressed the class imbalance between ID and NO ID groups, avoiding possible bias in the results. Furthermore, hyperparameters for XGBoost and ElasticNet were properly tuned to mitigate overfitting issues related to the limited sample size. Our results show that the three algorithms demonstrated greater concordance in feature ranking when CpGs were annotated by TFBSs compared to grouping by KEGG pathways. Specifically, 12 different TF

proteins were selected by all three ML methods, 10 of which are expressed in $CD4^+$ T cells according with reference dataset [149].

A crucial step in our gene prioritization process was the integration of epigenetic and transcriptomic data, enabling us to definitively prioritize genes where variations in the total number of SEMs corresponded to significant changes in gene expression. This multi-omics approach strengthens the biological relevance of the identified up- and down-regulated genes, allowing us to identify an epigenetically-driven gene expression signature of JIA clinical activity.

Exploratory clinical insights from CMAP

We used CMAP to identify potential drugs or compounds that could reverse our (epigenetically-derived) transcriptomic signature. Twelve statistically significant compounds showing a counteracting effect were found, including the *avicin-g* (BRD-A33746814), a *nuclear factor kappa B (NF-kB)* inhibitor. The *NF-kB signaling pathway* plays a key role in JIA by regulating the expression of pro-inflammatory genes [142]. Another significant compound is *methyl-fasudil* (BRD-K12683773), a derivative of *fasudil*, a well-known Rho-kinase inhibitor. Although *fasudil* is primarly used for treating cardiovascular conditions, evidence suggests it also modulates *NF-kB signaling* and shows beneficial effects on rats with RA [150]. *Cobimetinib* (BRD-K03390685) is *MEK* inhibitor. Although *MEK* inhibitors are mainly used for cancer treatment, they have been shown to be candidate agents for the treatment of systemic JIA [151]. Among the other compounds, fostamatinib is a *SYK inhibitor* involved in B-cell and innate immune cell activation, and has been tested in RA clinical trials with positive outcomes [152]. Additionally, *aprepitant* (BRD-K65170927) has shown anti-inflammatory effects in RA models [153,154].

External validation

External validation of our findings in an independent dataset of Australian JIA patients from the Royal Children's Hospital in Melbourne confirmed higher EML levels in active patients compared with both control and inactive subjects, in line with the results obtained in the discovery dataset.

Furthermore, SKAT-O analysis of SEMs within candidate TFs replicated 7 out of 12 identified targets, providing strong independent confirmation of our epigenetic signature. Pathway enrichment analysis of differentially epimutated genes revealed significant overrepresentation of key inflammatory pathways, including *TNF-alpha signaling via NF-κB* and *TGF-beta signaling* which has been previously implicated in JIA and rheumatoid arthritis [142–144].

For validating the epigenetically-driven transcriptomic signature, we analyzed gene expression profiles from PBMCs in two independent datasets comparing active and inactive JIA patients. Despite cross-tissue differences between PBMC and $CD4^+$ T cells, 22 genes from our signature showed significant differential expression with consistent direction of change relative to our discovery dataset. Enrichment analysis of these validated genes revealed significant overrepresentation of inflammatory pathways, again including *TNF-alpha signaling via NF-κB* and *TGF-beta signaling*, indicating remarkable consistency across datasets and omic layers.

This convergence of pathway-level results across independent datasets and complementary omics strongly supports the biological robustness of our findings and highlights these inflammatory cascades as promising therapeutic targets for JIA.

Strength and limitations

We acknowledge some limitations in this study. The sample size in the discovery dataset is small for multi-omics analyses, with only 14 patients in the NO ID class, reflecting the rarity of the disease. However, we employed a hypothesis-driven approach, guided by preliminary observations of higher epimutation load in NO ID compared to ID patients, rather than an agnostic genome-wide investigation prone to type II errors after multiple-testing correction. A key strength of the dataset is the availability of paired DNAm and transciptomic data at two time points, allowing longitudinal multi-omics analysis.

The limited sample size also posed challenges for machine learning analyses and feature prioritization. To mitigate this, we used three complementary ML methods and focused on features consistently identified across all of them, thereby reducing false positives and increasing robustness. External validation of prioritized TFs and genes in independent datasets further confirmed the reproducibility of our findings.

We did not apply multiple-testing correction when identifying differentially expressed genes in the discovery phase because this analysis was restricted to a candidate gene list derived from DNAm-based ML results. Although this exploratory strategy warrants cautious interpretation, the consistency with previous literature and replication across independent datasets support the validity of the findings. We also note that some genes exhibited modest fold changes, which could increase the risk of false positives; however, key results were independently confirmed in two external transcriptomic datasets, and consistent enrichment of immune-related pathways across methylation and expression data strengthens confidence in their biological relevance.

A major strength of this study is the reanalysis of a public dataset using updated analytical and integrative multi-omics methods, which revealed new insights into the biomolecular mechanisms of JIA clinical activity. The approach introduced here, characterizing epimutated CpGs and annotating them to TFBSs, can be generalized to other complex diseases with epigenetic dysregulation to uncover epigenetically driven alterations in gene expression. Furthermore, validation of our transcriptomic signature across independent datasets yielded highly consistent pathway enrichments.

While immediate clinical translation lies beyond the scope of this work, the convergence of evidence across omics layers highlights *NF-κB signaling* pathway as a promising therapeutic target deserving further investigation. Finally, although functional validation was not performed, the convergence of evidence from validation epigenetic and transcriptomic datasets provides strong support for the biological relevance of our findings.

2.1.5. Concluding remarks

This study demonstrates a significant association of epigenetic mutations with JIA clinical activity, suggesting a critical role of epigenetic mutations in the reactivation of JIA. We describe molecular mechanisms potentially implicated in JIA etiology and clinical reactivation of the disease. Future studies are required to accumulate epidemiological evidence and to assess the efficacy of the identified compounds.

Beyond its biological findings, this study presents a novel multi-omics methodological approach to identify epigenetically-driven gene expression modifications moving beyond

traditional EWAS approaches, potentially useful to investigate other multifactorial diseases. Our approach integrates the annotation of epimutated CpG sites to the TFs they are targeted, providing a more comprehensive understanding of the molecular mechanisms underlying disease activity.

From a broader perspective, this project exemplifies how rare and complex diseases can be investigated through integrative multi-omics strategies combined with ML. The approach adopted here fits within the hierarchical integration class described by Picard et al. [155], in which molecular layers are analyzed in a biologically informed sequence, linking DNA methylation changes to transcriptional regulation and, ultimately, to functional pathways. This hierarchical design reflects the causal flow of molecular information and sets the conceptual foundation for the subsequent projects in this thesis, where progressively larger and more complex datasets are analyzed using increasingly data-driven frameworks.

2.2. Project 2. Machine Learning-Based Prediction of Cell-type Resolved Brain eQTLs Enhances Discovery of Variants Explaining Alzheimer's Disease Heritability.

This project was conducted during my research period abroad, between the Department of Neurology at Columbia University Medical Center and the New York Genome Center, under the supervision of Prof. Gao Wang, Assistant Professor of Neurological Sciences at Columbia University and head of the Statistical Functional Genomics Lab at the Center for Statistical Genetics, and Prof. David Knowles, Core Faculty Member at the New York Genome Center and Associate Professor of Computer Science at Columbia University.

Unlike the first project presented in this thesis, which focused on a rare autoimmune disease and relied on a limited number of samples, this study addressed a common and complex neurodegenerative disorder, Alzheimer's disease (AD), using large-scale datasets from the FunGen-xQTL consortium. Access to this extensive resource made it possible to investigate molecular regulation in the human brain at single-cell resolution, across thousands of individuals and multiple brain cell types.

The goal of this work was to bridge the gap between genetic predisposition and the molecular mechanisms of gene regulation. Most genetic variants associated with AD are located in noncoding regions of the genome, where they are believed to modulate gene expression through cell type-specific regulatory processes, particularly in microglia. To capture these effects, we developed predictive ML models that integrate genomic, epigenomic, and transcriptomic features to infer cell type-specific eQTLs. The resulting framework, named single-cell Enhanced Expression Modifier Scores (sc-EEMS), represents a step forward in the prediction of regulatory variant effects at the single-cell level.

In contrast to the hierarchical multi-omics strategy applied in the JIA study, this work adopted a transformation-based (mixed) integration approach, combining biologically informed features derived from genomic, epigenomic, and transcriptomic annotations, into a unified predictive framework. This design enabled the model to capture complementary regulatory signals while enhancing both predictive performance and interpretability.

My contribution focused on the computational development of the framework. I designed and implemented the preprocessing pipelines, curated the multi-omic annotations for model training, and trained the initial ML models for single-cell eQTL prediction. Working with such large datasets also allowed me to gain experience in managing large-scale analyses on high-performance computing clusters, optimizing parallel jobs and computational resources for model training.

After my research stay, I continued contributing remotely to model training and methodological refinement. Dr. Chirag Lakhani subsequently led the final stages of the project, including gene-level association analyses, and external validation. The manuscript derived from this work has now been submitted to *Nature Genetics* and is currently under peer review. A preprint version of the study is publicly available on *medRxiv* [194].

Overall, this project expanded my experience from small-scale, hypothesis-driven epigenetic studies to large-scale, data-driven functional genomics. It also enabled me to gain hands-on experience in QTL analyses as a strategy to bridge genetic findings with their functional consequences, deepening my understanding of the regulatory architecture underlying complex diseases. In this context, the integration of ML and DL methods with single-cell QTL data proved particularly powerful in refining disease-gene mapping and uncovering cell type-specific mechanisms underlying genetic risk.

2.2.1. Introduction

Alzheimer's disease (AD) is a multifactorial neurodegenerative disorder and the most common form of dementia [156]. Genetic factors play a crucial role in AD pathogenesis, ranging from rare, highly penetrant mutations in genes such as *APP*, *PSEN1*, and *PSEN2* that cause early-onset familial AD, to common variants identified through GWAS that collectively contribute to the genetic liability for late-onset sporadic disease. Recent large-scale GWAS [157,158] have identified over 70 genomic loci associated with AD risk, significantly advancing our understanding of the genetic architecture underlying this complex disorder.

Despite these advances, a major challenge in post-GWAS analysis lies in elucidating the functional consequences of disease-associated variants, the majority of which are located in non-coding regions of the genome. It is widely believed that many of these non-coding variants modulate disease risk by influencing gene regulatory mechanisms, including transcription, RNA splicing, and chromatin accessibility, thereby affecting the expression and function of causal genes.

Large-scale initiatives such as GTEx [159] have catalogued eQTLs that mediate gene expression in bulk tissue, but across traits, only ~11% of disease heritability is explained by these signals [160]. This discrepancy, known as the "missing regulation" problem [161], reflects the fact that disease-associated variants are enriched in distal regulatory regions that are often invisible to bulk-tissue analyses. Indeed, eQTLs tend to cluster near promoters and affect genes under weaker selection, whereas GWAS loci are typically enriched in enhancer regions and near genes under stronger selection [162,163].

Recent efforts to address these limitations have employed context-specific approaches, including single-cell eQTL [164,165] and cell type-specific molecular QTL studies [166]. Moreover, investigations of intermediate molecular QTLs such as chromatin accessibility QTLs (caQTLs) have provided critical insights into the cell-type specific regulatory mechanisms underlying AD pathogenesis. However, it is challenging to generate population-scale data for all possible molecular phenotypes in all possible cellular contexts.

Genomic deep learning models [167–170], which predict epigenomic or transcriptomic profiles as a function of genomic sequence, can help overcome this limitation by providing assay-specific variant prediction scores for any genetic variant. These models are typically trained either on a single genomic assay or in a multi-task fashion using thousands of functional genomics assays from the Roadmap [171] or ENCODE [172] consortia.

Building upon the Expression Modifier Score (EMS) [173], we propose single-cell Enhanced Expression Modifier Score (sc-EEMS) for the prediction of causal single-cell eQTLs as a function of distance to transcription start site (TSS), ABC scores [174,175], cell type and non cell type-specific features, deep learning based variant effect prediction (DL-VEP) scores, and gene features. We developed a framework for training such models in the context of challenges arising in single-cell eQTL mapping, such as reduced power for rare cell types. We use the CatBoost model [176] to predict causal eQTL status as a function of 4,839 multi-omic features for 6 brain cell types profiled by single-nucleus post-mortem brain RNA-seq

from the ROSMAP [177–179] cohort. Single-cell eQTL summary statistics were generated as part of the FunGen-xQTL resource [180,181] within the Alzheimer's Disease Sequencing Project [182].

We use sc-EEMS to identify putative causal cell type eQTLs, especially in enhancer regions, which are likely to be more disease relevant. We used SHapley Additive exPlanations (SHAP) [183] model interpretation to calculate Shapley values for all high probability predicted eQTLs. On average, the largest contributing set of features were the Enformer [168] DL-VEP scores.

We also compared the contribution of fine-mapped versus predicted eQTLs to the Bellenguez AD GWAS [158]. While fine-mapped eQTLs had higher enrichment of heritability in the GWAS, the predicted eQTLs explained more trait heritability while still maintaining high enrichment of heritability.

The contribution of sc-EEMS-predicted and fine-mapped eQTLs to AD GWAS heritability revealed that the former accounted for a larger proportion of AD heritability while maintaining comparable levels of heritability enrichment, particularly in microglia.

Finally, sc-EEMS predictions were integrated GWAS summary statistics using the eMAGMA framework [184] to nominate cell type-specific AD risk genes.

2.2.2. Methods

Dataset composition and preprocessing

The data used in this study were obtained from the FunGen-xQTL consortium, which integrated three independent single-cell brain atlases to generate the harmonized *MEGA* cohort: Columbia University Irving Medical Center - Batch 1 (CUIMC1) [179],

Massachusetts Institute of Technology (MIT) [178], and Columbia University Irving Medical Center - Batch 2 (CUIMC2) [185].

All data were derived from post-mortem brain tissue samples from participants in the Religious Orders Study and the Rush Memory and Aging Project (ROSMAP) [177], for whom paired whole-genome sequencing and single-nucleus RNA-sequencing data were available for 788 individuals.

Based on previous analysis [186], cells were classified into six major types: astrocytes (Ast), excitatory neurons (Exc), inhibitory neurons (Inh), microglia (Mic), oligodendrocytes (Oli), and oligodendrocyte precursor cells (OPC).

Raw single-cell count matrices were aggregated into pseudo-bulk expression profiles per individual and cell type. Individuals with fewer than ten cells of a given cell type were excluded to ensure reliability of the averaged expression signal.

Expression values were filtered using the *filterByExpr* function from the *edgeR* [187] R package, normalized with the Trimmed Mean of M-values (TMM) method, and transformed using voom to obtain $\log_2$ counts per million ($\log_2$CPM) values with precision weights.

Batch effects among the three datasets were removed using the *removeBatchEffect* function from *limma* [187], and quantile normalization was applied to harmonize expression distributions across samples in the final integrated dataset (**Figure 2.1a**).

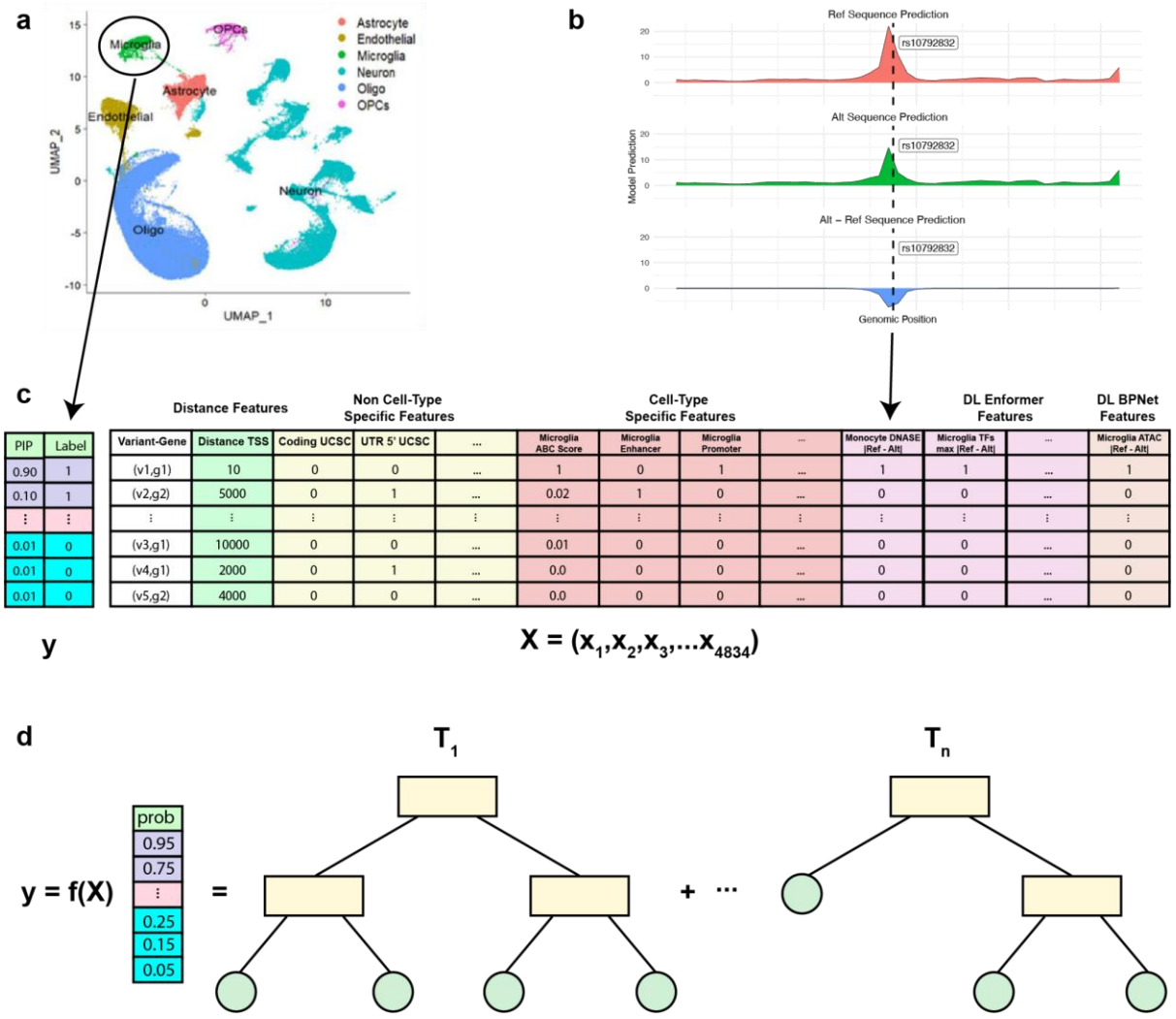

Figure 2.1 – sc-EEMS framework for predicting single-cell eQTLs. a) UMAP plot representing neuronal cell types which have single-cell eQTL data. b) Enformer predictions of *PU.1* TF binding for reference and alternate sequences as well as the difference in predictions. The difference in predictions is used to create DL-VEP scores. c) Feature matrix construction showing 4,839 features per variant-gene pair, including distance to TSS (log-transformed), ABC scores, non cell type-specific features, cell type-specific features, DL-VEP scores, and gene-level features. d) CatBoost model architecture for predicting eQTL probabilities as a function of the 4,839 features.

eQTL mapping and fine-mapping

Following data processing and normalization, genes were classified based on whether they met the mean $\log_2$CPM ≥ 2.0 filtering criterion after combining the three datasets. Genes passing this threshold were designated as *MEGA* genes and analyzed jointly to maximize

statistical power, whereas genes not meeting the threshold were analyzed within their original constituent datasets, here referred to as *Other* genes.

For both gene categories, eQTL mapping was conducted for all cis-variants within the corresponding dataset, and the resulting summary statistics were fine-mapped using SuSiE (Sum of Single Effects) [188]. This approach estimates posterior inclusion probabilities (PIPs) for each variant and constructs credible sets capturing those most likely to be causal within each locus. PIPs represent the posterior probability that a given variant drives the observed association signal, conditional on the data. By explicitly accounting for LD structure, this method improves causal variant localization and reduces confounding from correlated variants.

Feature annotation

Fine-mapped results were used to generate two training datasets for model development:

- *MEGA*: variants from genes analyzed in the combined cohort;
- *MEGA+Other*: variants from both *MEGA* and *Other* genes, offering broader coverage but higher heterogeneity.

Each dataset included variant-gene pairs annotated with 4,839 features (**Figure 2.1c**), grouped as follows:

- Distance to transcription start site (TSS) (log-transformed);
- Activity-by-Contact (ABC) scores computed from H3K27ac and ATAC-seq data in four brain cell type: astrocytes, microglia, neurons, and oligodendrocytes [175];
- Non-cell type-specific annotations, including 82 baseline annotations from PolyFun (e.g., UTRs, introns, MAF from gnomAD [189]);
- Cell type-specific annotations based on chromatin marks (H3K27ac, H3K4me1, ATAC-seq) in astrocytes, microglia, neurons, and oligodendrocytes;
- Deep learning-derived variant effect prediction (DL-VEP) scores, including predictions from Enformer [168] (**Figure 2.1b**), BPNet [168] and ChromBPNet [170] models;

- Gene-level features, such as the GeneBayes constraint score and expression levels [190].

Model training and evaluation

Several classification algorithms were benchmarked to predict whether a variant is a causal eQTL for a given gene. Among them, CatBoost and XGBoost achieved the best overall performance. CatBoost was ultimately selected as the final model, as it provided slightly higher predictive accuracy and handled categorical features more efficiently.

Cell type-specific binary classifiers were then trained using CatBoost (**Figure 2.1d**), to predict whether a variant represented a causal eQTL for a given gene.

Given the substantially higher number of low-PIP variants compared to high-PIP ones, variants with PIP > 0.50 (a lower threshold than the 0.9 used for EMS [173]) were labeled as positives, while those with PIP < 0.01 were labeled as negatives, maintaining a 1:10 positive-to-negative ratio. Positive variants were weighted proportionally to their PIP to emphasize high-confidence signals and mitigate class imbalance.

To obtain a conservative estimate of model performance, a stricter testing set was used, requiring PIP > 0.90 for positives and PIP < 0.01 for negatives. Evaluation followed a leave-one-chromosome-out (LOCO) cross-validation scheme, and predictive performance was summarized using the area under the precision-recall curve (AUPRC) averaged across folds.

Model validation on external microglia data

Model generalization was evaluated on an independent dataset from the *MIGA* study [191], which includes both genotyping array and bulk cell-sorted microglia RNA-seq data from four brain regions in 100 individuals.

Fine-mapping was then performed on this dataset to identify high-confidence microglia eQTLs, defined as variants with PIP > 0.25 or PIP > 0.50 in at least two regions.

Evaluating the contribution of DL-VEP features

To assess whether upweighting DL-VEP features could improve the model's ability to distinguish causal variants from those in high LD, we trained an alternative model configuration in which all DL-VEP features were upweighted tenfold relative to all other features. This weighting increased the influence of DL-VEP features during training keeping all other parameters identical to the standard (unweighted) model.

The rationale for this approach was to prioritize sequence-based functional features, which are less correlated with simple genomic proximity and may better capture regulatory effects relevant to eQTL activity (further discussed in the Discussion section).

Both model configurations were evaluated using LOCO cross-validation within the *MEGA* cohort and further validated on the *MIGA* [191] independent dataset. Predicted probabilities from the upweighted and unweighted models were compared on these variants to evaluate which configuration more accurately prioritized true causal signals.

In addition, we examined how well the two models captured the AD polygenic signal using stratified LD score regression (SLDSC) [192]. For each model, we constructed binary functional annotations where each variant was assigned a value of 1 if it was predicted as an eQTL with probability ≥ p for any gene, and 0 otherwise. We generated 20 annotations per model using probability thresholds ranging from p = 0.80 to 0.99 (increment 0.01), and for each annotation we calculated both the proportion of heritability explained (p_{h^2}) and the enrichment of heritability (e_{h^2}) in the AD GWAS.

As a baseline comparison, additional models were trained using only the log-transformed distance to the transcription start site (TSS) as predictor, since this feature is known to be a strong individual determinant of eQTL status [173]. Model performance across all cell types was evaluated by AUPRC under LOCO cross-validation.

Predicted vs. fine-mapped eQTLs: thresholds and heritability comparison

A key step of the analysis was to define a meaningful prediction probability threshold to determine which variant-gene pairs should be classified as predicted eQTLs. For each cell

type, similarly to the DL-VEP weighting analysis, the sc-EEMS models were used to generate multiple binary annotations across a range of probability thresholds (p = 0.80-0.99).

For each threshold, SLDSC was applied to estimate the proportion (p_{h^2}) and enrichment (e_{h^2}) of Bellenguez AD GWAS [158] heritability, selecting the cutoff that best balanced the two metrics. These optimal thresholds were then used to define the final predicted eQTL sets for each cell type and to compare their heritability contribution with fine-mapped eQTLs.

Feature importance analysis

To interpret the relative contribution of each annotation type to model predictions, SHapley Additive exPlanations (SHAP) values were computed for all features.

Features were then grouped into five categories: (i) distance to TSS, (ii) Activity-by-Contact (ABC) scores, (iii) non cell type-specific features, (iv) cell type-specific features, and (v) DL-VEP features derived from Enformer and ChromBPNet models. The mean absolute SHAP value within each category was used to summarize its overall importance in predicting eQTL probability.

Nomination of putative AD risk genes

Predicted and fine-mapped eQTLs were integrated with AD GWAS summary statistics from Bellenguez et al. [158] using eMAGMA to derive cell type-specific gene-level associations and nominate candidate AD risk genes.

2.2.3. Results

Upweighting deep learning features improves the localization of AD polygenic signal

We compared models trained with and without upweighting of DL-VEP features and both achieved similar cross-validation performance (average AUPRC ~= 0.67 across cell types).

However, when evaluated on the external *MIGA* [191] dataset, the upweighted model better discriminated true causal variants: among variants with PIP > 0.25 in at least two brain regions (n=46), 41.3% were predicted with probabilities > 0.25 in the upweighted model compared with 36.9% in the unweighted one; for PIP > 0.50, the proportions were 66.6% vs 55.5%, respectively.

Moreover, when comparing the proportion of heritability (p_{h^2}) vs. enrichment of heritability (e_{h^2}), the binary functional annotations based on the DL-VEP upweighted model outperformed the ones from the DL-VEP unweighted model (**Figure 2.2**). That is, at every value of p_{h^2}, there was a DL-VEP upweighted model threshold with higher e_{h^2} compared to the DL-VEP unweighted model, thus better localizing the polygenic signal from the AD GWAS. Given that the DL-VEP upweighted model better identifies true eQTLs in the *MIGA* dataset and achieves superior polygenic localization in the AD GWAS, we use this model for all subsequent analyses and refer to it simply as sc-EEMS.

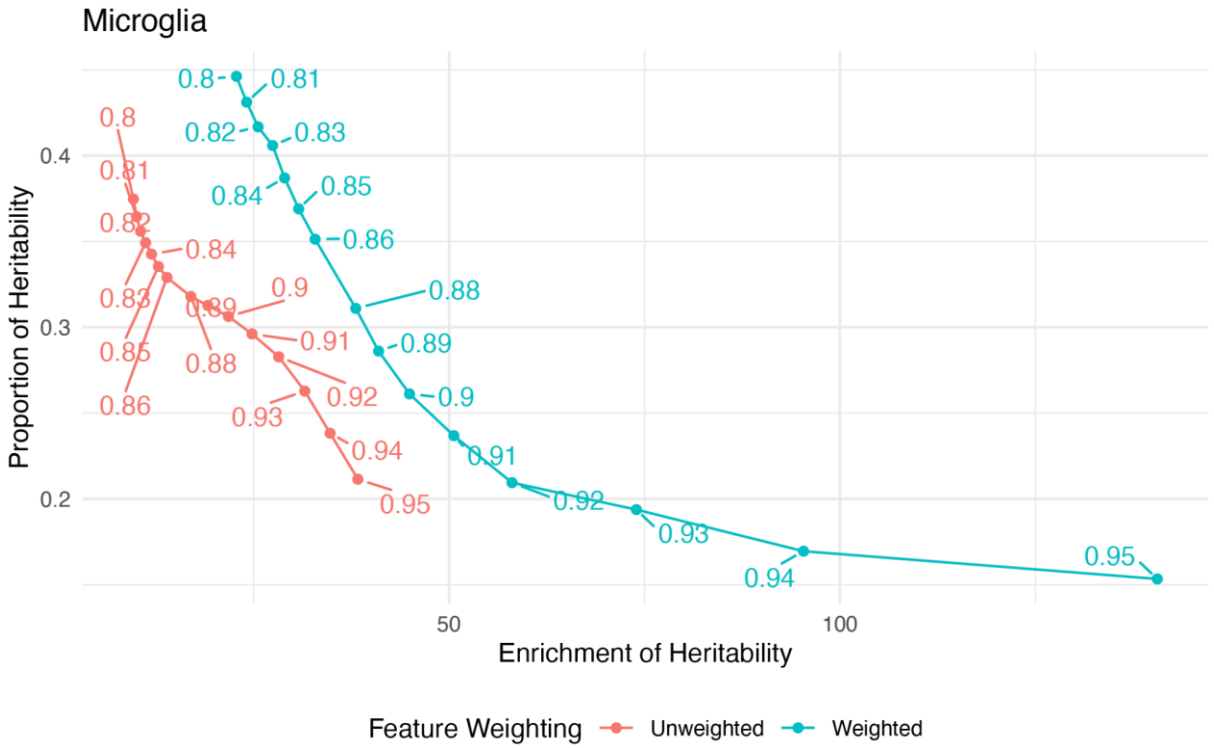

Figure 2.2 – AD GWAS heritability comparisons for DL-VEP unweighted and upweighted models. Enrichment of heritability (x-axis) versus proportion of heritability explained (y-axis) for binary annotations created using model predictions of DL-VEP unweighted (red) and upweighted (light blue) models in microglia. Binary annotations consist of variants with predicted probability above a certain numerical threshold for each model.

sc-EEMS outperforms Distance to TSS in predicting cell type-specific eQTLs

The number of variant-gene pairs available for model training varied substantially across cell types, reflecting the underlying abundance of each cell type in the single-cell datasets. Training set sizes ranged from 3,810 positive variants across 801 genes in microglia to 32,090 positive variants across 5,843 genes in excitatory neurons (**Figure 2.3a**). The more stringent selection criteria applied to the test set (PIP > 0.90 for positive variants) resulted in correspondingly smaller test sets, with 173 positive variants from 157 genes in microglia and 1,990 positive variants from 1,691 genes in excitatory neurons (**Figure 2.3b**).

To evaluate model performance, we obtained predicted probabilities for each test variant using the model trained with its corresponding chromosome held out. We then aggregated predictions across all test variants to calculate a single AUPRC for each cell type.

sc-EEMS performance was compared with a baseline model using only the log-transformed distance to TSS as predictor. Models trained using *MEGA* generally performed similarly to using *MEGA+Other*. Both models substantially outperformed the baseline model using only Distance to TSS (**Figure 2.3c**), achieving higher AUPRC values in all cell types (average AUPRC on *MEGA* = 0.67 vs. 0.53 for baseline models). The best performance was observed for oligodendrocytes (AUPRC = 0.74), while inhibitory neurons yielded the lowest (AUPRC = 0.62).

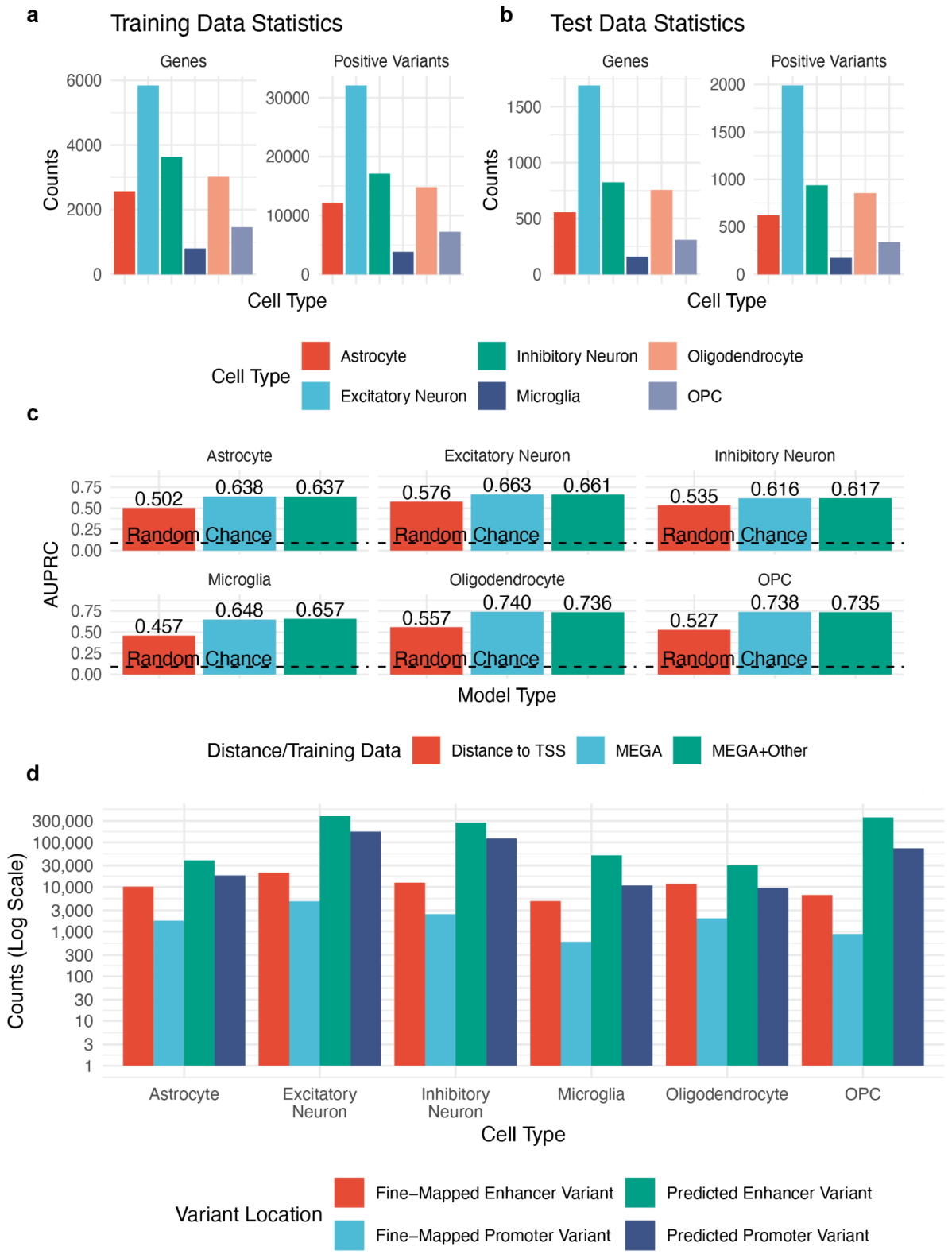

Figure 2.3 – Training statistics across brain cell types. a) Number of eQTLs and eGenes used as training data for each cell type model. b) Number of eQTLs and eGenes used as testing data for each cell type model. c) AUPRC for cell-specific prediction of eQTLs using just distance to TSS (log-transformed), sc-EEMS trained on *MEGA* genes only, and sc-EEMS

trained on combined *MEGA+Other* genes. The dashed line represents the AUPRC based on random chance. d) Number of fine-mapped eQTLs (PIP > 0.10) and predicted eQTLs (at cell-type specific thresholds) identified using sc-EEMS. Stratified by whether the variant is considered a promoter (within 10kb of TSS) or enhancer.

With the aim of determining the best probability threshold for classifying variant-gene pairs as causal or non-causal eQTLs for each cell type, we used SLDSC to maximize the per-standardized-annotation effect size (tau*) and get the best tradeoff between the proportion of AD GWAS heritability explained (p_{h^2}) and the corresponding enrichment (e_{h^2}).

Based on this optimization, we obtained distinct thresholds across the six cell types (**Figure 2.4**).

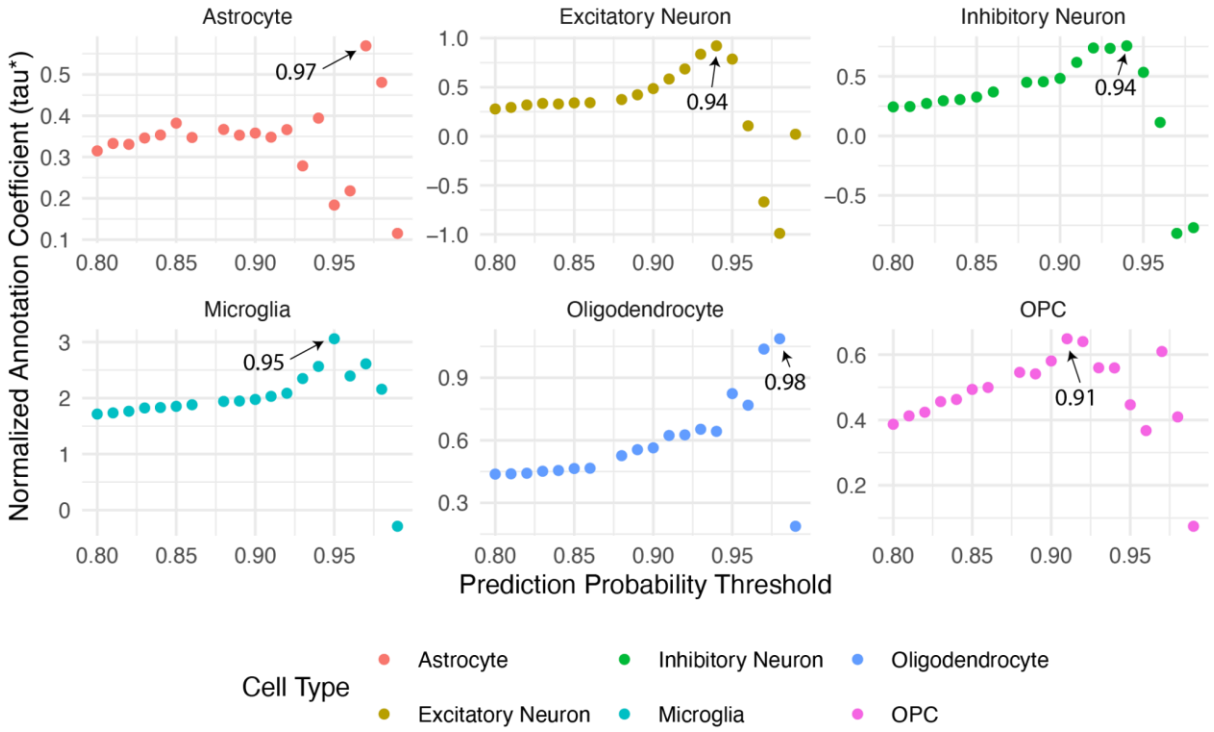

Figure 2.4 – Cell type-specific tau* values (y-axis) across prediction probability thresholds (x-axis). The x-axis represents the binary annotation generated by including all variants with predicted probability above a given threshold. The labeled point denotes the prediction probability threshold with the highest tau* value for each cell type.

Applying these thresholds resulted in an average 20.9-fold increase in the number of predicted eQTLs relative to fine-mapped eQTLs.

Following Mostafavi et al. [162], variants within 10 kb of a TSS were considered promoter-like, and those beyond 10 kb enhancer-like.

Across all cell types, we observed an average 33.6-fold increase in promoter-like predicted eQTLs and an 18.7-fold increase in enhancer-like variants compared to fine-mapped eQTLs (**Figure 2.3d**).

DL-derived features collectively contribute more to model predictions than distance to TSS

Model interpretation using SHAP confirmed that distance to TSS remained a strong individual predictor. However, the overall contribution of DL features, particularly Enformer-based scores, exceeded the distance to TSS across all cell types (**Figure 2.5a**).

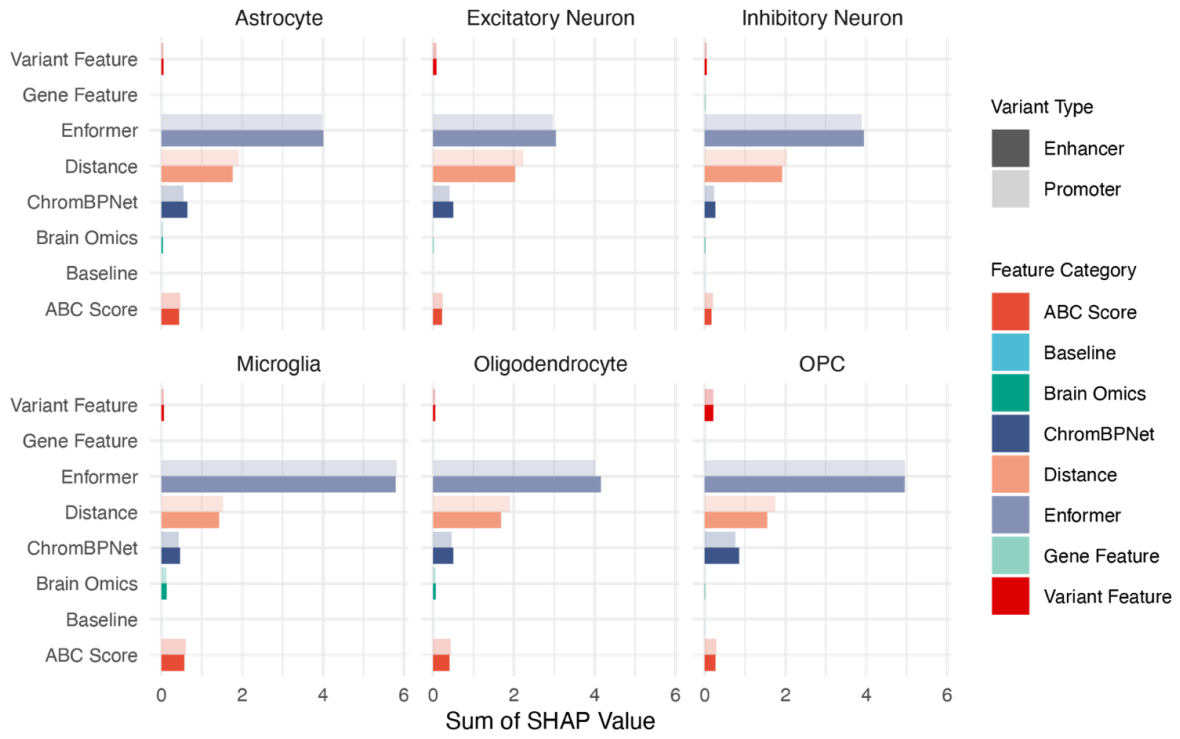

Figure 2.5 – Feature importance. SHAP value analysis of feature contributions. Sum of absolute SHAP values for feature categories across six brain cell types, separated by promoter and enhancer variants. Stratified LD score regression (SLDSC) statistics for binary

annotations based on fine-mapped eQTLs (PIP > 0.10) and predicted eQTLs (cell-specific thresholds).

Predicted eQTLs better capture AD heritability and enhance gene discovery

The contribution of predicted and fine-mapped eQTLs to AD GWAS heritability was evaluated using SLDSC. Across all brain cell types, sc-EEMS–predicted eQTLs explained a larger proportion of AD heritability than original fine-mapped eQTLs, while maintaining comparable enrichment levels, with the strongest improvement observed in microglia, where the proportion of heritability explained increased approximately threefold (from $p_{h^2} \simeq 0.05$ to $p_{h^2} \simeq 0.15$).

To extend the functional interpretation of these variants, predicted eQTLs were integrated with AD GWAS summary statistics using the eMAGMA framework. This analysis identified 215 significant cell type-gene pairs corresponding to 111 unique genes, compared with 76 pairs (55 unique genes) obtained from fine-mapped eQTLs.

These findings demonstrate that incorporating predicted variant-gene relationships substantially improves the power of gene-level association analyses, providing a more comprehensive view of the regulatory mechanisms contributing to AD susceptibility.

2.2.4. Discussion

In this study, we developed single-cell Enhanced Expression Modifier Scores (sc-EEMS), a ML framework for predicting causal eQTLs across major brain cell types by integrating summary multi-omic features. This model extends the original EMS [173] framework to the single-cell context, enabling the identification of regulatory variants that are difficult to capture with standard fine-mapping analyses.

A key methodological advance of sc-EEMS is its ability to exploit low-powered single-cell eQTL datasets through a training strategy that incorporates weaker fine-mapping signals while upweighting high-confidence variants according to their PIPs. This design allows the

model to learn from both the strength and uncertainty of fine-mapping results, enhancing its generalization across cell types with varying data availability.

Standard supervised frameworks assume that training labels are correct, yet even state-of-the-art fine-mapping methods such as SuSiE may assign high PIPs to non-causal variants in regions of complex LD. Consequently, learning directly from these labels can propagate LD-driven correlations. Genomic annotations such as distance to TSS or ABC scores, are also spatially autocorrelated, which can further reinforce these dependencies. In contrast, DL-VEP features capture molecular regulatory effects directly from sequence context and exhibit lower spatial redundancy. By explicitly upweighting these features during training, sc-EEMS was encouraged to prioritize sequence-based regulatory information, thereby improving its ability to disentangle causal from correlated variants and to better localize the polygenic signal observed in AD GWAS.

Biological interpretation and implications for AD genetics

The sc-EEMS model outperformed baseline predictors based solely on genomic proximity, with DL-VEP features emerging as the most informative predictors across all brain cell types. This emphasizes that regulatory information encoded in DNA sequence context can meaningfully complement traditional annotations in identifying causal eQTLs.

Across cell types, predicted eQTLs captured a larger proportion of AD heritability than experimentally fine-mapped eQTLs, particularly in microglia, where the heritability explained was approximately threefold higher. This result reinforces the pivotal role of microglia in AD pathogenesis, consistent with prior studies linking immune activation and myeloid regulatory programs to disease risk [158,193]. Notably, most predicted eQTLs mapped to enhancer-like regions, supporting the concept that AD risk is primarily mediated through distal, cell type-specific regulatory mechanisms rather than promoter-proximal variation.

Integration of sc-EEMS predictions with eMAGMA further demonstrated the biological relevance of this framework. The model nominated over twice as many AD risk genes as fine-mapped eQTLs alone, revealing cell type-specific regulatory connections that were not detected by standard approaches.

Limitations and methodological considerations

A key aspect influencing model performance is the choice of PIP thresholds used to define positive and negative training examples. In sc-EEMS, a lower PIP threshold for positives (0.50) than the one used in the EMS framework (0.90) was adopted to include a broader range of candidate eQTLs, thereby increasing the diversity of training examples. While this strategy improves sensitivity and allows the model to capture weaker but biologically relevant signals, it may also introduce additional noise due to imperfect fine-mapping.

The use of PIP-weighted labels and the upweighting of DL-VEP features mitigated this limitation by emphasizing high-confidence examples and prioritizing sequence-derived information, which is less prone to LD-driven artifacts. However, residual uncertainty from correlated variants cannot be entirely excluded, and the resulting predictions should be interpreted probabilistically rather than as definitive causal assignments.

Incorporating additional cell type-specific regulatory annotations could improve the predictive performance and interpretability of the model. Beyond model optimization, extending the framework to other molecular QTL modalities (e.g., caQTLs, sQTLs) may reveal complementary regulatory layers and deepen our understanding of the molecular mechanisms contributing to disease risk.

2.2.5. Concluding remarks

Overall, sc-EEMS advances the original EMS framework by integrating a more comprehensive and tissue-specific set of genomic, epigenomic, and transcriptomic features to predict causal eQTLs at single-cell resolution. This expanded annotation landscape enhances both predictive accuracy and biological interpretability, allowing the model to capture cell type-specific regulatory variants that are difficult to identify through standard fine-mapping. By linking these predicted eQTLs to downstream gene-level effects through integration with GWAS data, sc-EEMS bridges the gap between noncoding genetic variation and functional mechanisms, facilitating the identification of risk genes and pathways relevant to Alzheimer's disease.

More broadly, this project highlights the pivotal role of ML methods in biomedical research and reflects the central theme of this thesis: using data-driven, integrative approaches to uncover the molecular basis of complex diseases. Here, machine learning has proven essential not only for deriving meaningful features from high-dimensional multi-omic data but also for combining them across molecular layers into biologically interpretable frameworks. Through approaches like sc-EEMS, predictive modeling becomes a powerful means to translate heterogeneous omic data into biologically actionable insights, paving the way toward a more mechanistic and personalized understanding of human disease.

2.3. Project 3. Genome-wide association study and xQTL analysis in Syndrome of Undifferentiated Recurrent Fever (SURF): a multi-omics approach for identifying novel candidate biomarkers and molecular pathophysiological mechanisms.

This work was conducted within the framework of the Personalised medicine for Systemic AutoInflammatory Diseases (PerSAIDs) project, a large multicenter initiative that connects some of the major European registries of systemic autoinflammatory diseases with the goal of elucidating their genetic and molecular mechanisms through multi-omics data analysis. The consortium encompasses a wide spectrum of disorders, ranging from well-characterized monogenic diseases such as Familial Mediterranean Fever (FMF), Mevalonate Kinase Deficiency (MKD), and TNF Receptor-Associated Periodic Syndrome (TRAPS), to more complex and heterogeneous conditions including systemic-onset Juvenile Idiopathic Arthritis (sJIA), Periodic Fever, Aphthous stomatitis, Pharyngitis, and Adenitis syndrome (PFAPA), and the Syndrome of Undifferentiated Recurrent Fever (SURF), which was the focus of this study.

SURF represents a heterogeneous group of autoinflammatory conditions characterized by recurrent episodes of fever and systemic inflammation in the absence of identifiable monogenic or autoimmune causes. Unlike classical autoinflammatory syndromes, SURF is believed to result from the interplay of multiple genetic and regulatory factors, making its molecular basis elusive to conventional diagnostic approaches.

To investigate the molecular complexity of SURF, we implemented a genome-wide analytical framework that combined association testing of both common and rare variants with downstream multi-omics xQTL analyses across multiple omic layers, including proteomics, immunomics, cytokine profiling, transcriptomics, and epigenomics.

I conceived this study with Dr. Giovanni Fiorito and carried it out with support from other PerSAIDs consortium members. The project is ongoing, with current efforts focused on dataset expansion, framework refinement, and result validation.

Building on the knowledge gained from the analysis of QTL data during my research stay abroad (Project 2), this work represented an opportunity to extend those concepts to a context where omic layers are integrated without any prior transformations. This approach presents several analytical challenges but also provides unique opportunities to jointly explore complementary molecular dimensions of complex autoinflammatory diseases.

2.3.1. Introduction

Autoinflammatory diseases are rare disorders characterized by recurrent or chronic inflammation due to dysregulation of the innate immune system, in the absence of infection or autoimmunity. They include monogenic syndromes, such as Familial Mediterranean Fever (FMF), Mevalonate Kinase Deficiency (MKD), and TNF Receptor-Associated Periodic Syndrome (TRAPS), as well as polygenic forms, including Periodic Fever, Aphthous stomatitis, Pharyngitis, and Adenitis (PFAPA) and systemic-onset juvenile idiopathic arthritis (SoJIA).

However, a subset of patients does not meet the diagnostic criteria for PFAPA [195] or other well-recognized autoinflammatory diseases. These cases are classified as Syndrome of Undifferentiated Recurrent Fever (SURF), a term first introduced by Broderick [196] and later consolidated in subsequent clinical studies [197–199]. SURF represents a heterogeneous clinical entity characterized by self-limiting inflammatory episodes in the absence of a confirmed molecular diagnosis.

Recent evidence has highlighted distinctive clinical features of SURF, including multi-organ involvement and, in a relevant subset of patients, a complete or partial response to colchicine. This therapeutic responsiveness, rarely observed in PFAPA but reminiscent of FMF, suggests a potential mechanistic link to the cytoskeleton and pyrin inflammasome pathway [200]. Altogether, these findings indicate that SURF may not be a single disease, but rather a spectrum of autoinflammatory conditions whose genetic and molecular bases remain largely unexplored.

In this context, integrating genomic and multi-omics data represents a promising strategy to dissect the biology of SURF, refine patient stratification, and identify molecular drivers of susceptibility and progression [197].

In this study, we analyzed multi-omics data collected through the PerSAIDs project, a European collaborative initiative that connects some of the largest registries of systemic autoinflammatory diseases with the goal of developing standardized protocols and bioinformatics pipelines for data management, collecting and integrating multi-omics datasets, and ensuring data reusability. Data were collected for a broad spectrum of autoinflammatory diseases, but for this study, we specifically focused on SURF.

A hierarchical multi-omics integration strategy was adopted. First, we performed genome-wide analyses to identify candidate SNPs and genes. We then conducted downstream QTL analyses of these candidates across multiple omic layers, including proteomics, immunomics, cytokine profiling, transcriptomics and epigenomics, in order to identify potentially relevant omic biomarkers linked to genetic variation. This approach was designed to move beyond purely clinical characterization and provide further molecular insights into the pathogenesis of SURF.

2.3.2. Methods

Dataset and workflow description

We assembled a dataset comprising 76 patients with SURF and 266 controls.

WGS data were obtained partly from the Giannina Gaslini Institute's internal registries, which provided all controls and 45 cases, and partly from the PerSAIDs project, which provided 31 additional cases.

The downstream multi-omics analysis was conducted on data collected from the PerSAIDs project and included proteomics, immunomics, cytokine profiling, transcriptomics, and epigenomics.

The overall study workflow is summarized in **Figure 3.1**. Starting from WGS data, we applied stringent quality control and ancestry correction. Variants were then stratified by frequency: common variants (MAF > 1%) were tested using GWAS [6], while rare variants (MAF < 1%) were assessed at the gene level using SKAT-O [33]. Candidate SNPs and genes emerging from these analyses were subsequently interrogated through multi-omics QTL (xQTL) analyses, integrating genetic variation with multiple molecular layers available in a subset of SURF

patients. This hierarchical strategy enabled us to move from genetic associations to functional hypotheses on disease-relevant biological mechanisms.

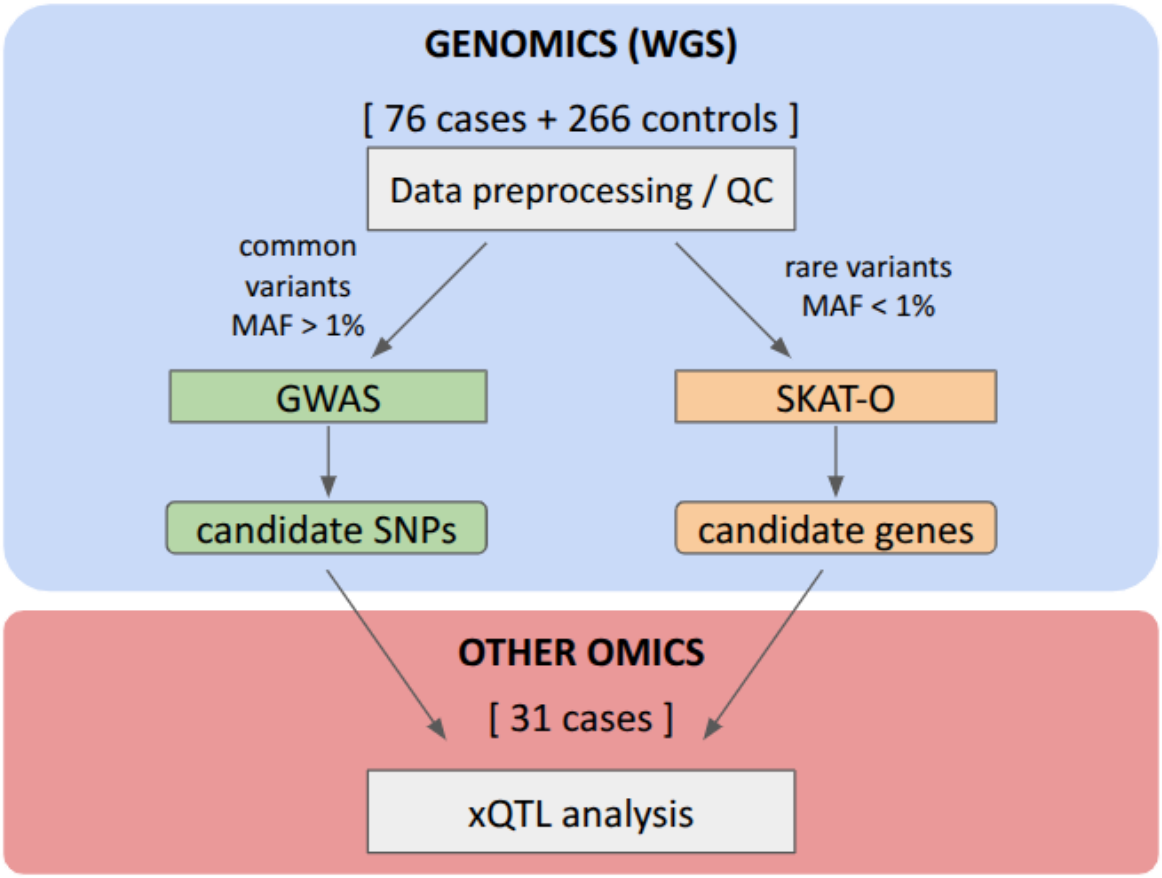

Figure 3.1 – Overall workflow of the study. Whole-genome sequencing data from 76 SURF cases and 266 controls underwent stringent preprocessing and quality control. Common variants (MAF > 1%) were analyzed using GWAS to identify candidate SNPs, while rare variants (MAF < 1%) were analyzed using SKAT-O to highlight candidate genes. These genetic candidates were subsequently integrated with multi-omics data (proteomics, immunomics, cytokine profiling, transcriptomics, and epigenomics) from 31 SURF patients through xQTL analyses to explore potential molecular biomarkers linked to genetic variation.

Data preprocessing and quality control

Genetic variants underwent a standard QC pipeline, including removal of sex chromosomes, filtering based on genotype quality and allelic balance, and exclusion of variants with >5% missingness. At the sample level, individuals with excessive missing data, extreme

heterozygosity, or close relatedness were excluded. PCA with HapMap reference populations was used to check samples' ancestry, leading to the removal of non-European outliers.

Common variants analysis

Common variants (MAF > 1%) were tested using a GWAS based on an additive logistic regression model with sex as covariate. Analyses were performed in PLINK [201], and results were visualized with Manhattan and QQ plots to evaluate association signals and potential population stratification.

Rare variants analysis

Rare variants (MAF < 1%) were aggregated at the gene level and analyzed using SKAT-O [33], which tests for enrichment of potentially pathogenic alleles in cases compared to controls. QQ plots were used to evaluate genomic inflation and overall model fit.

Multi-omics QTL analyses

To explore the functional consequences of genetic variation, we performed QTL analyses on candidate SNPs and genes identified by GWAS and SKAT-O, respectively. Molecular features were modeled as a function of genotype or gene-level burden, adjusting for sex, disease activity, and treatment status.

The following omics layers were included:

- **Proteomics** (inflammation-related panel, Olink PEA): 92 ComBat-corrected Normalized Protein eXpression (NPX) measurements derived from retrospective plasma samples;
- **Immunomics** (antibody profiling, 16k-protein microarrays): 15,750 ComBat-corrected antibody intensity measurements.
- **Cytokine profiling** (LegendPlex bead-based immunoassays): inflammatory cytokines and chemokines quantified with multiplex flow cytometry panels.

- **Targeted expression profiling** (NanoString nCounter platform): 76 gene expression measurements obtained from RNA extracted from patient samples, quantified using a multiplex hybridization-based assay.
- **Epigenomics** (Illumina Infinium MethylationEPIC v2 BeadChip): DNA methylation measurements of 913,893 CpG obtained from blood samples, preprocessed, batch-corrected, and quality-controlled for downstream analyses including differential methylation and pathway enrichment.

Comparison with PFAPA patients

To assess whether the associations identified in SURF patients reflected disease-specific mechanisms rather than general features of recurrent inflammatory conditions, we compared the observed molecular trends with those detected in patients affected by PFAPA. This comparative analysis allowed us to distinguish signals potentially unique to SURF from those shared with other autoinflammatory phenotypes, thereby refining the list of candidate biomarkers for biological interpretation.

2.3.3. Results

Quality control and dataset composition

The preprocessing pipeline progressively excluded samples and variants that did not meet quality standards. After ancestry correction with PCA, five cases (three Italian and two

Turkish) and thirteen controls were identified as outliers and removed (**Figure 3.2**). The final dataset consisted of 69 SURF cases and 197 controls.

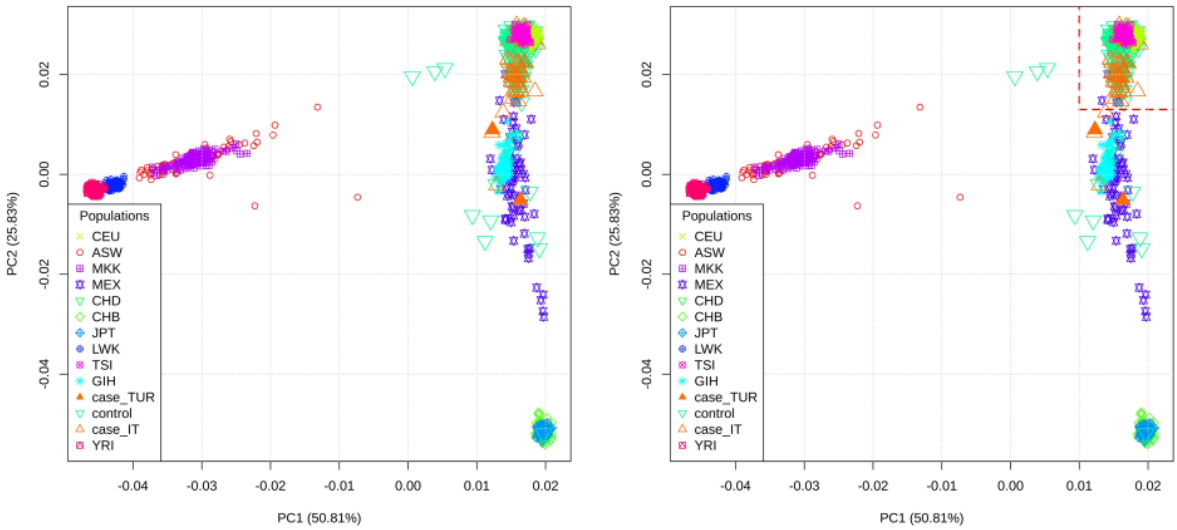

Figure 3.2 – Principal Component Analysis (PCA) of study samples and HapMap reference populations. The plots show the first two principal components of the PCA. Study cases (orange upward triangles: empty for Italian subjects, filled for Turkish) and controls (green downward triangles) are projected together with HapMap populations. **Left panel**: several cases and controls cluster outside of the European reference groups (TSI, CEU), overlapping instead with other HapMap populations, indicating divergent genomic backgrounds. **Right panel**: after removal of these outlier samples, the distribution of cases and controls aligns more consistently with the expected European ancestry clusters.

Genome-wide analyses results

In the GWAS of common variants, we adopted the classical genome-wide significance threshold of $p < 5 \times 10^{-8}$, together with a suggestive threshold of $p < 1 \times 10^{-5}$. No variants surpassed the genome-wide threshold, but several loci reached the suggestive level (**Figure**

3.3). The QQ plot indicated good control of population stratification, with a genomic inflation factor close to 1 ($\lambda = 1.022$).

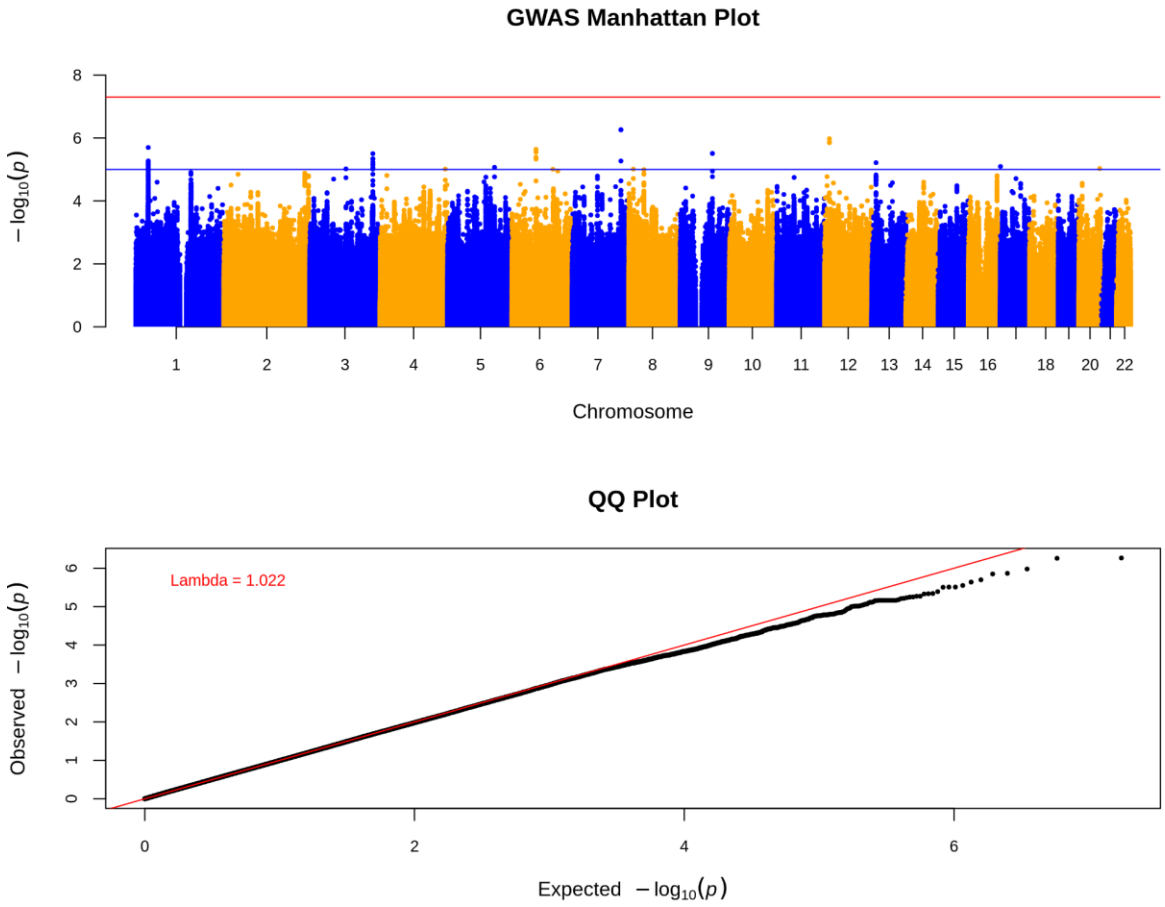

Figure 3.3 – GWAS results are shown in the Manhattan plot (above) and QQ plot (below). No SNPs reached genome-wide significance (red line), although several reached the suggestive threshold (blue line).

Rare variants were analyzed at the gene level using SKAT-O. Here we applied a Bonferroni-corrected threshold for exome-wide significance ($p < 2.5 \times 10^{-6}$) and a suggestive threshold of $p < 5 \times 10^{-4}$. Although no genes reached the stringent Bonferroni threshold, several candidates showed suggestive evidence of association and were retained for downstream analyses (**Figure 3.4**). The QQ plot indicated some degree of inflation ($\lambda = 1.371$), but still within acceptable limits for exploratory investigations.

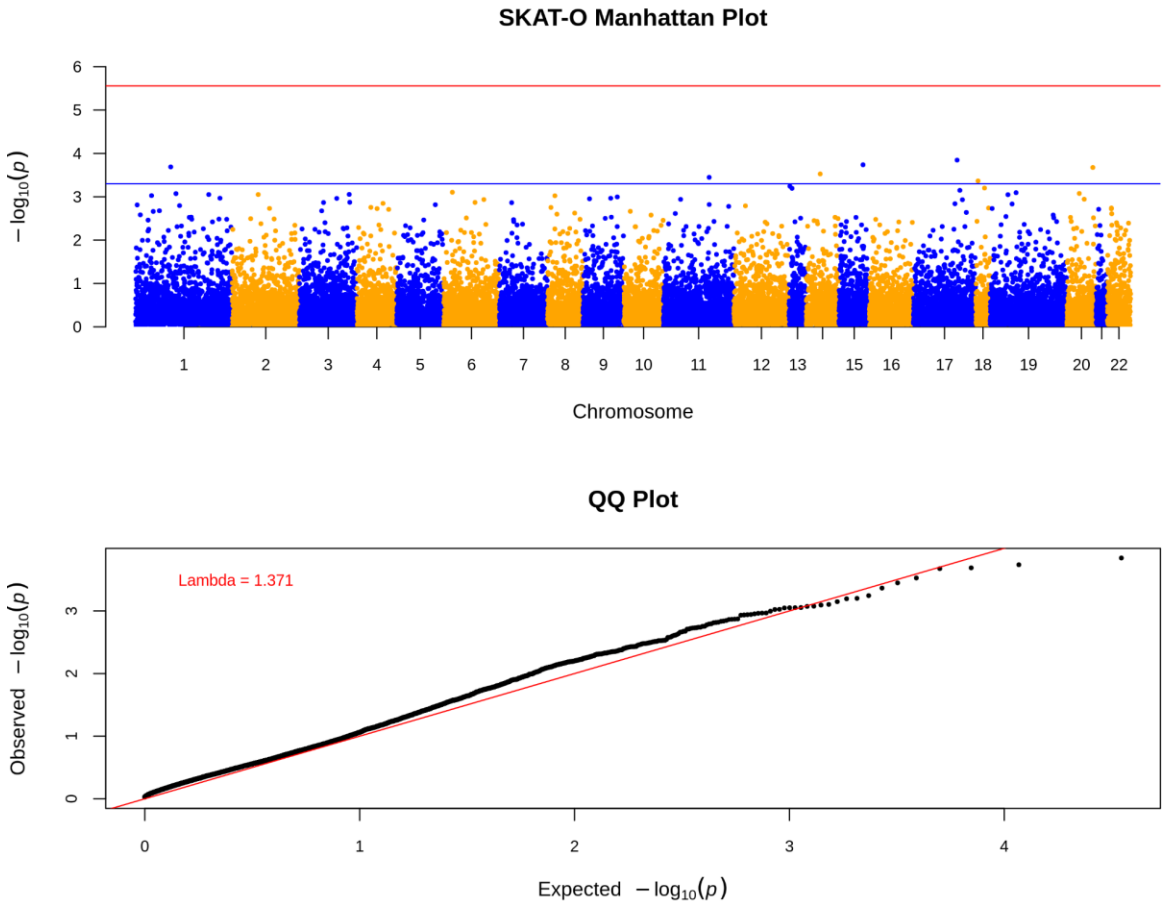

Figure 3.4 – SKAT-O results are shown in the Manhattan plot (above) and QQ plot (below). No genes reached genome-wide significance (red line), although several reached the suggestive threshold (blue line).

Taken together, the genome-wide analyses did not reveal clear and definitive associations. Nevertheless, both approaches highlighted a subset of candidate variants and genes that warrant further investigation. To this end, we performed multi-omics xQTL analyses to test whether these genetic signals were linked to molecular phenotypes.

Multi-omics QTL analysis

We investigated candidate variants from GWAS and SKAT-O through multi-omics xQTL analyses. For each molecular feature, we applied multiple linear regression using the following model:

Omics variable (Y) ~ Genetics (X) + Sex + Age + Disease activity + Treatment status

being the independent variable *X* the genotype of a candidate SNP in the case of GWAS-based common variants analysis, or the aggregated rare variants load of a candidate gene in the case of the SKAT-O-based rare variants analysis. The dependent variable *Y* represents each omics feature in turn, spanning proteomics, immunomics, cytokine profiling, expression profiling, and epigenomics.

This analysis revealed several statistically significant molecular mechanisms after correction for multiple testing (FDR < 0.05). For instance, immunomic profiling revealed that antibody levels against *NCOA3* (Nuclear Receptor Coactivator 3) were higher in patients carrying the alternative allele of SNP rs2268126 ($\beta = 0.401$, $p = 4.32 \times 10^{-7}$, **Figure 3.5a**). This aligns with the biological interpretation for which *NCOA3* is involved in the activation of the *NF-kB* pathway [202], a well-known inflammatory pathway.

QTL analysis on transcriptomics showed that *CXCL9* expression, a pro-inflammatory chemokine [203], followed a similar trend with respect to the same SNP ($\beta = 0.851$, $p = 1.17 \times 10^{-3}$).

Furthermore, analysis of rare variants load highlighted further candidate signals. For instance, a higher rare variants load in *TOP1* (Topoisomerase 1) gene was positively associated with increased immunomic signatures against *DNMT1* ($\beta = 0.483$, $p = 3.47 \times 10^{-8}$, **Figure 3.5b**), the enzyme responsible for DNA methylation [204]. Although *TOP1* has primarily been implicated in autoimmune diseases, it also plays a crucial role in activating pathogen-induced and inflammatory genes [205]. This correlation suggests that genetic variability in TOP1 may modulate immune recognition of epigenetic regulators such as *DNMT1*.

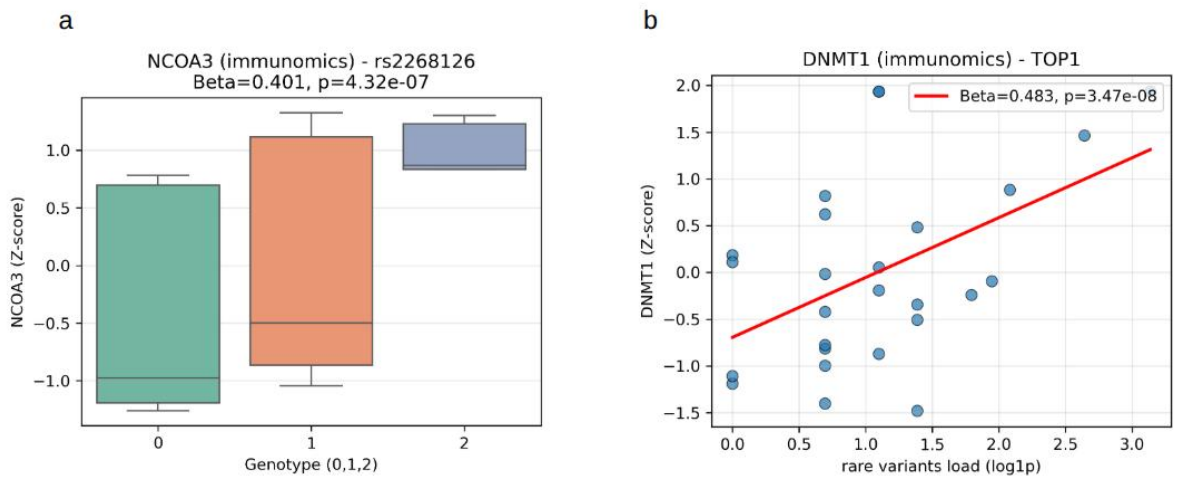

Figure 3.5 – Examples of significant xQTL associations in immunomic profiles of SURF patients.

a) Boxplot shows a positive association ($\beta = 0.401$, $p = 4.32 \times 10^{-7}$) between antibody levels against NCOA3 (Z-score standardized) and the rs2268126 genotype (0 = homozygous reference, 1 = heterozygous, 2 = homozygous alternative).

b) Scatterplot shows a positive correlation ($\beta = 0.483$, $p = 3.47 \times 10^{-8}$) between antibody levels against *DNMT1* and the gene-level rare variant load of *TOP1* (log1p-transformed).

Comparison with PFAPA patients

To evaluate the disease specificity of the signals identified in SURF, we compared their molecular trends with those observed in PFAPA patients. This analysis confirmed that some of the most relevant associations appeared to be restricted to SURF. In particular, both *NCOA3* antibody levels (**Figure 3.6a**) and *CXCL9* transcript expression showed a consistent correlation with the alternative allele of rs2268126 in SURF patients, whereas no comparable trend was observed in PFAPA (**Figure 3.6b**). These findings suggest that such features may represent SURF-specific biomarkers, distinguishing this condition from other autoinflammatory syndromes with overlapping clinical presentations.

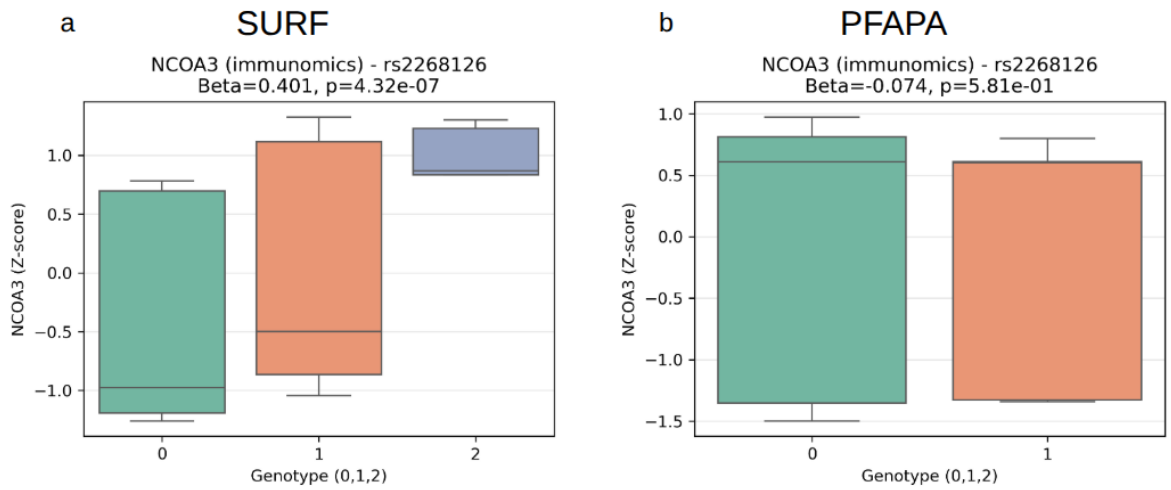

Figure 3.6 – Comparison of xQTL associations between SURF and PFAPA patients.

(a) Boxplot shows the positive association ($\beta = 0.401$, $p = 4.32 \times 10^{-7}$) between antibody levels against NCOA3 (Z-score standardized) and the rs2268126 genotype in SURF patients.

(b) In contrast, no significant association was observed in PFAPA patients ($\beta = -0.074$, $p = 5.81 \times 10^{-1}$).

2.3.4. Discussion

To the best of our knowledge, this study represents one of the first attempts to provide a genome-wide overview of SURF by jointly analyzing both common and rare variants. By adopting a hierarchical multi-omics integration strategy that starts from genetic variation and progressively explores downstream molecular effects across proteomic, immunomic, cytokine, transcriptomic, and epigenomic layers, this work aims to deepen the understanding of the biological mechanisms underlying the disease.

Biological interpretation

Genome-wide analyses did not identify associations surpassing conventional significance thresholds, either for common or rare variants. This outcome may reflect the limited sample size (76 cases and 266 controls), which restricts statistical power to detect variants with small

to moderate effects, or alternatively, a genuine absence of strong genetic determinants underlying the disease. While no definitive associations were observed, these findings do not exclude a polygenic or regulatory contribution to SURF. The ongoing collection of new WGS data within the PerSAIDs project is expected to nearly double the current number of SURF samples, substantially improving statistical power and the ability to validate preliminary findings.

Despite these constraints, the downstream multi-omics xQTL analyses provided more biologically informative results. Several omic features showed significant associations with specific genotypes or rare variant burdens, suggesting potential molecular mechanisms through which genetic variation may contribute to disease activity. The comparison with PFAPA patients was particularly informative, as it helped discriminate SURF-specific molecular signatures from general inflammatory responses. For instance, the association between rs2268126 and increased antibody levels against *NCOA3*, together with a parallel increase in *CXCL9* transcript levels, suggests a SURF-specific perturbation involving transcriptional coactivators and pro-inflammatory chemokines [202,203].

From a biological perspective, these results support the involvement of regulatory mechanisms controlling inflammation and transcriptional activation. *NCOA3* acts as a coactivator of *NF-kB*, linking transcriptional regulation to inflammatory signaling, while the association between the rare variant load in *TOP1* and immune reactivity against *DNMT1* highlights a potential interaction between DNA topology, epigenetic regulation, and immune recognition [204,205]. Although exploratory, these findings provide a valuable foundation for future functional studies aimed at validating the identified mechanisms.

Strengths, limitations, and future perspectives

However, some limitations must be acknowledged. First, the number of available samples remains limited, particularly for the multi-omics analyses, where only 31 patients had a complete multi-omics profile. Second, the hierarchical design of the study inherently depends on the robustness of upstream genomic findings: if the genetic associations identified at the initial stage are incomplete or contain false positives, downstream omic associations may reflect indirect or spurious effects. These limitations emphasize the importance of expanding the cohort and replicating results in independent datasets. Ongoing recruitment and

sequencing efforts within the PerSAIDs consortium are expected to address these issues, enabling more reliable detection of both genetic and functional signals.

Additional omic layers, including metabolomic and lipidomic profiles, are currently being collected and will soon be integrated into the analytical framework, enabling the generation of more comprehensive multi-omic profiles within a larger dataset. Although machine learning did not play a central role in the present phase of the study, the upcoming expansion of available omic layers will allow the implementation of integrative multi-omics approaches such as MOFA [90] and DIABLO [92]. Hopefully, these frameworks will support the identification of cross-omic biomarkers and molecular pathways, improving the prioritization of candidate features and offering a more comprehensive view of the biological processes underlying SURF.

Overall, this study demonstrates the feasibility and promise of a hierarchical multi-omics strategy for the investigation of rare and heterogeneous autoinflammatory diseases. Despite the small sample size, the integrative approach revealed biologically coherent molecular patterns that provide new hypotheses on the pathophysiology of SURF. The ongoing expansion of the dataset will allow refinement of these analyses, strengthening the link between genetic variation, molecular function, and clinical phenotype.

2.3.5. Concluding remarks

Overall, this study represents one of the first efforts to investigate the molecular basis of SURF through a genetics-driven multi-omics approach. By combining whole-genome sequencing with downstream molecular profiling across proteomic, immunomic, cytokine, transcriptomic, and epigenomic layers, the project provides an initial framework for linking genetic variation to molecular phenotypes in a condition that remains poorly characterized at the biological level. Although limited in sample size, this hierarchical integration strategy enabled the identification of biologically consistent patterns, revealing molecular signatures that may contribute to disease susceptibility and immune dysregulation.

The analytical approach adopted here demonstrates the feasibility of applying hierarchical multi-omics integration to rare and clinically heterogeneous autoinflammatory diseases. The

ongoing expansion of the SURF dataset within the PerSAIDs consortium, together with the inclusion of additional omic layers such as metabolomics and lipidomics, will further strengthen the capacity to uncover molecular mechanisms underlying disease activity. In future stages, integrative ML frameworks will be applied to jointly analyze these datasets, enabling the identification of cross-omic biomarkers and the reconstruction of relevant biological pathways.

More broadly, this project reflects the progressive methodological trajectory of this thesis: from small-scale, hypothesis-driven investigations to increasingly comprehensive, data-driven analyses. Across the three studies presented, the combined use of multi-omics data and machine learning has proven instrumental for elucidating the complex molecular architecture of human diseases. Together, these approaches provide a powerful means to integrate diverse layers of biological information, transform high-dimensional data into interpretable patterns, and bridge the gap between molecular mechanisms and clinical phenotypes. This integrated perspective represents a crucial step toward a more mechanistic and personalized understanding of complex disorders.

Conclusions

The work presented in this thesis illustrates how the integration of multi-omics data and machine learning can advance the study of complex diseases. Across distinct disease contexts, each project contributed to a unified objective: using data-driven and biologically informed frameworks to move from molecular complexity toward actionable biomedical insight.

From molecular features to mechanistic understanding

The first project, focused on JIA, introduced a hierarchical multi-omic framework that combined DNA methylation and transcriptomic data to investigate the role of stochastic epigenetic mutations (SEMs) in disease reactivation. The observation of an increased epigenetic mutation load (EML) in patients with active disease suggested that widespread stochastic methylation alterations are a downstream consequence of inflammatory reactivation, but may also serve as early molecular indicators of impending relapse. By integrating multiple omic layers and applying ML-based feature prioritization, the study identified transcription factors and signaling pathways as central mediators of immune dysregulation.

The second project, centered on Alzheimer's disease (AD), marked a shift toward large-scale predictive modeling and methodological innovation. The single-cell Enhanced Expression Modifier Score (sc-EEMS) framework extended the concept of expression modifier scores to the single-cell context by integrating multi-omic features to predict causal eQTLs across brain cell types. This approach improved the localization of heritability in AD GWAS data and revealed new candidate risk genes. Beyond its biological findings, the project underscored the value of combining deep learning feature representations with classical machine learning to bridge the gap between noncoding genetic variation and functional interpretation.

Finally, the third project explored Syndrome of Undifferentiated Recurrent Fever (SURF) within the PerSAIDs consortium, applying a hierarchical multi-omics integration strategy to investigate a clinically heterogeneous autoinflammatory condition. Starting from genome-wide and rare variant analyses, genetic findings were linked to molecular traits across proteomic,

immunomic, transcriptomic, and epigenomic layers. Although no genome-wide significant loci were identified, downstream analyses revealed several candidate omic biomarkers. This work provided the first genome-wide molecular overview of SURF, highlighting the feasibility of applying integrative multi-omics to rare and poorly characterized disorders.

Broader implications and future perspectives

Together, these projects delineate a methodological trajectory: from small-scale, hypothesis-driven analyses to increasingly data-driven, large-scale, and integrative frameworks. They demonstrate how multi-omics integration, supported by machine learning and robust statistical design, can uncover molecular mechanisms that remain hidden when each omic is studied in isolation.

A consistent lesson across projects is that data quality and harmonization are as crucial as algorithmic sophistication, particularly in rare and multicenter studies where technical or population heterogeneity can obscure biological signals. Equally important is ensuring biological interpretability, so that computational results translate into mechanistic understanding and, ultimately, clinical insight.

In summary, this thesis contributes to the growing evidence that multi-omics integration combined with machine learning can bridge the gap between large-scale data generation and clinical utility. Whether in biomarker discovery, mechanistic insight, or diagnostic support, these approaches provide a scalable and interpretable framework to navigate biological complexity, paving the way for a more data-driven and personalized medicine.

Acknowledgments

I would like to express my deepest gratitude to Dr. Paolo Uva, my supervisor, for his constant guidance, trust, and scientific mentorship throughout this PhD. His vision and expertise have been a continuous source of inspiration and motivation.

My sincere thanks go to Dr. Giovanni Fiorito, whose scientific insight and enthusiasm have profoundly shaped my analytical and critical thinking, and who played a key role in the conceptualization of several projects presented in this work.

I am also grateful to Davide Cangelosi, Marta Rusmini, Michele Iacomino, Matteo Vergani, and all my colleagues in Clinical Bioinformatics Unit, who have been invaluable colleagues and mentors throughout this journey. Their support, patience, and technical expertise have been essential to my professional and personal growth.

Finally, I thank all collaborators and clinicians from the Giannina Gaslini Institute and partner institutions who contributed to the datasets, ideas, and discussions that made this work possible.

This thesis represents not only the result of my research but also the outcome of a collective effort built on collaboration, curiosity, and shared passion for science.

References

1. Schork NJ. Genetics of complex disease: approaches, problems, and solutions. Am J Respir Crit Care Med. 1997;156:S103-109. https://doi.org/10.1164/ajrccm.156.4.12-tac-5

2. Mitchell KJ. What is complex about complex disorders? Genome Biol. 2012;13:237. https://doi.org/10.1186/gb-2012-13-1-237

3. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010;363:166–76. https://doi.org/10.1056/NEJMra0905980

4. Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med. 2009;360:1759–68. https://doi.org/10.1056/NEJMra0808700

5. Gibson G. Decanalization and the origin of complex disease. Nat Rev Genet. 2009;10:134–40. https://doi.org/10.1038/nrg2502

6. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9:356–69. https://doi.org/10.1038/nrg2344

7. Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nat Rev Methods Primer. 2021;1:59. https://doi.org/10.1038/s43586-021-00056-9

8. The Wellcome Trust Case Control Consortium, Management Committee, Burton PR, Clayton DG, Cardon LR, Craddock N, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–78. https://doi.org/10.1038/nature05911

9. Huang J, McLean GR, Franke A. Twenty years of genome-wide association studies: Health translation challenges and AI opportunities. Eur J Hum Genet [Internet]. 2025 [cited 2025 Oct 29]; https://doi.org/10.1038/s41431-025-01951-5

10. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science. 2005;308:385–9. https://doi.org/10.1126/science.1109557

11. Siminovitch KA. PTPN22 and autoimmune disease. Nat Genet. 2004;36:1248–9. https://doi.org/10.1038/ng1204-1248

12. Bottini N, Vang T, Cucca F, Mustelin T. Role of PTPN22 in type 1 diabetes and other autoimmune diseases. Semin Immunol. 2006;18:207–13. https://doi.org/10.1016/j.smim.2006.03.008

13. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007;316:889–94. https://doi.org/10.1126/science.1141634

14. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, et al. Diverse Genome-wide Association Studies Associate the IL12/IL23 Pathway with Crohn Disease. Am J Hum Genet. 2009;84:399–405. https://doi.org/10.1016/j.ajhg.2009.01.026

15. Moschen AR, Tilg H, Raine T. IL-12, IL-23 and IL-17 in IBD: immunobiology and therapeutic targeting. Nat Rev Gastroenterol Hepatol. 2019;16:185–96. https://doi.org/10.1038/s41575-018-0084-8

16. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. Am J Hum Genet. 2023;110:179–94. https://doi.org/10.1016/j.ajhg.2022.12.011

17. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5. https://doi.org/10.1126/science.1222794

18. The Wellcome Trust Case Control Consortium, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet. 2012;44:1294–301. https://doi.org/10.1038/ng.2435

19. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008;456:18–21. https://doi.org/10.1038/456018a

20. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11:446–50. https://doi.org/10.1038/nrg2809

21. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010;11:31–46. https://doi.org/10.1038/nrg2626

22. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12:745–55. https://doi.org/10.1038/nrg3031

23. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, et al. Clinical Interpretation and Implications of Whole-Genome Sequencing. JAMA. 2014;311:1035. https://doi.org/10.1001/jama.2014.1717

24. Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am J Hum Genet. 2008;83:311–21. https://doi.org/10.1016/j.ajhg.2008.06.024

25. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34:188–93. https://doi.org/10.1002/gepi.20450

26. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: Testing for an Accumulation of Rare Variants Using Next-Generation Sequencing Data. Hum Hered. 2012;73:84–94. https://doi.org/10.1159/000336982

27. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. Schork NJ, editor. PLoS Genet. 2009;5:e1000384. https://doi.org/10.1371/journal.pgen.1000384

28. Price AL, Kryukov GV, De Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. Am J Hum Genet. 2010;86:832–8. https://doi.org/10.1016/j.ajhg.2010.04.005

29. Liu DJ, Leal SM. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. Gibson G, editor. PLoS Genet. 2010;6:e1001156. https://doi.org/10.1371/journal.pgen.1001156

30. Han F, Pan W. A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. Hum Hered. 2010;70:42–54. https://doi.org/10.1159/000288704

31. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an Unusual Distribution of Rare Variants. Leal SM, editor. PLoS Genet. 2011;7:e1001322. https://doi.org/10.1371/journal.pgen.1001322

32. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

33. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91:224–37. https://doi.org/10.1016/j.ajhg.2012.06.007

34. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016;32:1493–501. https://doi.org/10.1093/bioinformatics/btw018

35. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012;109:1193–8. https://doi.org/10.1073/pnas.1119675109

36. Génin E. Missing heritability of complex diseases: case solved? Hum Genet. 2020;139:103–13. https://doi.org/10.1007/s00439-019-02034-4

37. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet. 2002;3:43–52. https://doi.org/10.1038/nrg703

38. Aguet F, Alasoo K, Li YI, Battle A, Im HK, Montgomery SB, et al. Molecular quantitative trait loci. Nat Rev Methods Primer. 2023;3:4. https://doi.org/10.1038/s43586-022-00188-6

39. Mackay TFC, Anholt RRH. Pleiotropy, epistasis and the genetic architecture of quantitative traits. Nat Rev Genet. 2024;25:639–57. https://doi.org/10.1038/s41576-024-00711-3

40. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. Williams SM, editor. PLoS Genet. 2014;10:e1004383. https://doi.org/10.1371/journal.pgen.1004383

41. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99:1245–60. https://doi.org/10.1016/j.ajhg.2016.10.003

42. Pividori M, Rajagopal PS, Barbeira A, Liang Y, Melia O, Bastarache L, et al. PhenomeXcan: Mapping the genome to the phenome through the transcriptome. Sci Adv. 2020;6:eaba2083. https://doi.org/10.1126/sciadv.aba2083

43. Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, et al. Mendelian randomization. Nat Rev Methods Primer. 2022;2:6. https://doi.org/10.1038/s43586-021-00092-5

44. Jacobs BM, Taylor T, Awad A, Baker D, Giovanonni G, Noyce AJ, et al. Summary-data-based Mendelian randomization prioritizes potential druggable targets for multiple sclerosis. Brain Commun. 2020;2:fcaa119. https://doi.org/10.1093/braincomms/fcaa119

45. Porcu E, Rüeger S, Lepik K, eQTLGen Consortium, Agbessi M, Ahsan H, et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat Commun. 2019;10:3300. https://doi.org/10.1038/s41467-019-10936-0

46. Zhou D, Jiang Y, Zhong X, Cox NJ, Liu C, Gamazon ER. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. Nat Genet. 2020;52:1239–46. https://doi.org/10.1038/s41588-020-0706-2

47. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60. https://doi.org/10.1126/science.1262110

48. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021;53:1300–10. https://doi.org/10.1038/s41588-021-00913-z

49. Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Kasela S, Wucher V, et al. Cell type-specific genetic regulation of gene expression across human tissues. Science. 2020;369:eaaz8528. https://doi.org/10.1126/science.aaz8528

50. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014;343:1246949. https://doi.org/10.1126/science.1246949

51. Wen L, Li G, Huang T, Geng W, Pei H, Yang J, et al. Single-cell technologies: From research to application. The Innovation. 2022;3:100342. https://doi.org/10.1016/j.xinn.2022.100342

52. Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet. 2019;20:257–72. https://doi.org/10.1038/s41576-019-0093-7

53. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8:1826. https://doi.org/10.1038/s41467-017-01261-5

54. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9:1825. https://doi.org/10.1038/s41467-018-03621-1

55. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481–7. https://doi.org/10.1038/ng.3538

56. Gerring ZF, Mina-Vargas A, Gamazon ER, Derks EM. E-MAGMA: an eQTL-informed method to identify risk genes using genome-wide association study summary statistics. Valencia A, editor. Bioinformatics. 2021;37:2245–9. https://doi.org/10.1093/bioinformatics/btab115

57. De Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. Tang H, editor. PLOS Comput Biol. 2015;11:e1004219. https://doi.org/10.1371/journal.pcbi.1004219

58. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:83. https://doi.org/10.1186/s13059-017-1215-1

59. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf Fusion. 2019;50:71–91. https://doi.org/10.1016/j.inffus.2018.09.012

60. Kreitmaier P, Katsoula G, Zeggini E. Insights from multi-omics integration in complex disease primary tissues. Trends Genet. 2023;39:46–58. https://doi.org/10.1016/j.tig.2022.08.005

61. Miao Z, Humphreys BD, McMahon AP, Kim J. Multi-omics integration in the age of million single-cell data. Nat Rev Nephrol. 2021;17:710–24. https://doi.org/10.1038/s41581-021-00463-x

62. Mohr AE, Ortega-Santos CP, Whisner CM, Klein-Seetharaman J, Jasbi P. Navigating Challenges and Opportunities in Multi-Omics Integration for Personalized Healthcare. Biomedicines. 2024;12:1496. https://doi.org/10.3390/biomedicines12071496

63. Feldner-Busztin D, Firbas Nisantzis P, Edmunds SJ, Boza G, Racimo F, Gopalakrishnan S, et al. Dealing with dimensionality: the application of machine learning to multi-omics data. Wren J, editor. Bioinformatics. 2023;39:btad021. https://doi.org/10.1093/bioinformatics/btad021

64. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17 Suppl 2:15. https://doi.org/10.1186/s12859-015-0857-9

65. Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46. https://doi.org/10.1016/j.csbj.2021.06.030

66. Wosiak A, Zakrzewska D. Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis. Czarnowski I, editor. Complexity. 2018;2018:2520706. https://doi.org/10.1155/2018/2520706

67. Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: Bergadano F, Raedt L, editors. Mach Learn ECML-94 [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 1994 [cited 2025 Oct 29]. p. 171–82. https://doi.org/10.1007/3-540-57868-4_57

68. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn. 2002;46:389–422. https://doi.org/10.1023/A:1012487302797

69. Scornet E. Trees, forests, and impurity-based variable importance [Internet]. arXiv; 2020 [cited 2025 Oct 29]. https://doi.org/10.48550/ARXIV.2001.04295

70. Ogutu JO, Schulz-Streeck T, Piepho H-P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc. 2012;6:S10. https://doi.org/10.1186/1753-6561-6-S2-S10

71. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26:303–4. https://doi.org/10.1038/nbt0308-303

72. Schölkopf B, Smola A, Müller K-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Comput. 1998;10:1299–319. https://doi.org/10.1162/089976698300017467

73. Nounou MN, Bakshi BR, Goel PK, Shen X. Bayesian principal component analysis. J Chemom. 2002;16:576–95. https://doi.org/10.1002/cem.759

74. Beh EJ. Simple Correspondence Analysis: A Bibliographic Review. Int Stat Rev. 2004;72:257–84. https://doi.org/10.1111/j.1751-5823.2004.tb00236.x

75. Sompairac N, Nazarov PV, Czerwinska U, Cantini L, Biton A, Molkenov A, et al. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. Int J Mol Sci. 2019;20:4414. https://doi.org/10.3390/ijms20184414

76. Zou H, Hastie T, Tibshirani R. Sparse Principal Component Analysis. J Comput Graph Stat. 2006;15:265–86. https://doi.org/10.1198/106186006X113430

77. Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. Mach Learn. 2011;83:331–53. https://doi.org/10.1007/s10994-010-5222-7

78. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16:85–97. https://doi.org/10.1038/nrg3868

79. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018;24:1248–59. https://doi.org/10.1158/1078-0432.CCR-17-0853

80. Xie G, Dong C, Kong Y, Zhong JF, Li M, Wang K. Group Lasso Regularized Deep Learning for Cancer Prognosis from Multi-Omics and Clinical Features. Genes. 2019;10:240. https://doi.org/10.3390/genes10030240

81. Wilson CM, Li K, Yu X, Kuan P-F, Wang X. Multiple-kernel learning for genomic data mining and prediction. BMC Bioinformatics. 2019;20:426. https://doi.org/10.1186/s12859-019-2992-1

82. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333–7. https://doi.org/10.1038/nmeth.2810

83. Lee B, Zhang S, Poleksic A, Xie L. Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. Front Genet. 2020;10:1381. https://doi.org/10.3389/fgene.2019.01381

84. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. IEEE Trans Neural Netw Learn Syst. 2021;32:4–24. https://doi.org/10.1109/TNNLS.2020.2978386

85. Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, et al. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. Front Genet. 2018;9:477. https://doi.org/10.3389/fgene.2018.00477

86. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401:788–91. https://doi.org/10.1038/44565

87. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 2012;40:9379–91. https://doi.org/10.1093/nar/gks725

88. Luo Y, Tao D, Wen Y, Ramamohanarao K, Xu C. Tensor Canonical Correlation Analysis for Multi-view Dimension Reduction [Internet]. arXiv; 2015 [cited 2025 Oct 30]. https://doi.org/10.48550/ARXIV.1502.02330

89. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics. 2014;15:162. https://doi.org/10.1186/1471-2105-15-162

90. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14:e8124. https://doi.org/10.15252/msb.20178124

91. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21:111. https://doi.org/10.1186/s13059-020-02015-1

92. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Birol I, editor. Bioinformatics. 2019;35:3055–62. https://doi.org/10.1093/bioinformatics/bty1054

93. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25:2906–12. https://doi.org/10.1093/bioinformatics/btp543

94. Sun D, Wang M, Li A. A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. IEEE/ACM Trans Comput Biol Bioinform. 2019;16:841–50. https://doi.org/10.1109/TCBB.2018.2806438

95. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12:3445. https://doi.org/10.1038/s41467-021-23774-w

96. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do K-A. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics. 2013;29:149–59. https://doi.org/10.1093/bioinformatics/bts655

97. Wu C, Zhang Q, Jiang Y, Ma S. Robust network-based analysis of the associations between (epi)genetic measurements. J Multivar Anal. 2018;168:119–30. https://doi.org/10.1016/j.jmva.2018.06.009

98. Zarayeneh N, Ko E, Oh JH, Suh S, Liu C, Gao J, et al. Integration of multi-omics data for integrative gene regulatory network inference. Int J Data Min Bioinforma. 2017;18:223–39. https://doi.org/10.1504/IJDMB.2017.10008266

99. Fortelny N, Bock C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biol. 2020;21:190. https://doi.org/10.1186/s13059-020-02100-5

100. Ravelli A, Martini A. Juvenile idiopathic arthritis. Lancet Lond Engl. 2007;369:767–78. https://doi.org/10.1016/S0140-6736(07)60363-8

101. Martini A, Lovell DJ, Albani S, Brunner HI, Hyrich KL, Thompson SD, et al. Juvenile idiopathic arthritis. Nat Rev Dis Primer. 2022;8:5. https://doi.org/10.1038/s41572-021-00332-8

102. Koker O, Aliyeva A, Sahin S, Adrovic A, Yildiz M, Haslak F, et al. An overview of the relationship between juvenile idiopathic arthritis and potential environmental risk factors: Do early childhood habits or habitat play a role in the affair? Int J Rheum Dis. 2022;25:1376–85. https://doi.org/10.1111/1756-185X.14431

103. Meyer B, Chavez RA, Munro JE, Chiaroni-Clarke RC, Akikusa JD, Allen RC, et al. DNA methylation at IL32 in juvenile idiopathic arthritis. Sci Rep. 2015;5:11063. https://doi.org/10.1038/srep11063

104. Li J, Li L, Wang Y, Huang G, Li X, Xie Z, et al. Insights Into the Role of DNA Methylation in Immune Cell Development and Autoimmune Disease. Front Cell Dev Biol. 2021;9:757318. https://doi.org/10.3389/fcell.2021.757318

105. Lovell DJ, Giannini EH, Reiff A, Cawkwell GD, Silverman ED, Nocton JJ, et al. Etanercept in Children with Polyarticular Juvenile Rheumatoid Arthritis. N Engl J Med. 2000;342:763–9. https://doi.org/10.1056/NEJM200003163421103

106. Gentilini D, Garagnani P, Pisoni S, Bacalini MG, Calzari L, Mari D, et al. Stochastic epigenetic mutations (DNA methylation) increase exponentially in human aging and correlate with X chromosome inactivation skewing in females. Aging. 2015;7:568–78. https://doi.org/10.18632/aging.100792

107. Yan Q, Paul KC, Lu AT, Kusters C, Binder AM, Horvath S, et al. Epigenetic mutation load is weakly correlated with epigenetic age acceleration. Aging. 2020;12:17863–94. https://doi.org/10.18632/aging.103950

108. Fiorito G, Caini S, Palli D, Bendinelli B, Saieva C, Ermini I, et al. DNA methylation-based biomarkers of aging were slowed down in a two-year diet and physical activity intervention trial: the DAMA study. Aging Cell. 2021;20:e13439. https://doi.org/10.1111/acel.13439

109. Fiorito G, McCrory C, Robinson O, Carmeli C, Ochoa-Rosales C, Zhang Y, et al. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. Aging. 2019;11:2045–70. https://doi.org/10.18632/aging.101900

110. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115. https://doi.org/10.1186/gb-2013-14-10-r115

111. Duan R, Fu Q, Sun Y, Li Q. Epigenetic clock: A promising biomarker and practical tool in aging. Ageing Res Rev. 2022;81:101743. https://doi.org/10.1016/j.arr.2022.101743

112. Margiotti K, Monaco F, Fabiani M, Mesoraca A, Giorlandino C. Epigenetic Clocks: In Aging-Related and Complex Diseases. Cytogenet Genome Res. 2023;163:247–56. https://doi.org/10.1159/000534561

113. Chervova O, Panteleeva K, Chernysheva E, Widayati TA, Baronik ŽF, Hrbková N, et al. Breaking new ground on human health and well-being with epigenetic clocks: A systematic review and meta-analysis of epigenetic age acceleration associations. Ageing Res Rev. 2024;102:102552. https://doi.org/10.1016/j.arr.2024.102552

114. Mukherjee N, Harrison TC. Epigenetic Aging and Rheumatoid Arthritis. J Gerontol A Biol Sci Med Sci. 2024;79:glad213. https://doi.org/10.1093/gerona/glad213

115. Dammering F, Martins J, Dittrich K, Czamara D, Rex-Haffner M, Overfeld J, et al. The pediatric buccal epigenetic clock identifies significant ageing acceleration in children with internalizing disorder and maltreatment exposure. Neurobiol Stress. 2021;15:100394. https://doi.org/10.1016/j.ynstr.2021.100394

116. Gomaa N, Konwar C, Gladish N, Au-Young SH, Guo T, Sheng M, et al. Association of Pediatric Buccal Epigenetic Age Acceleration With Adverse Neonatal Brain Growth and Neurodevelopmental Outcomes Among Children Born Very Preterm With a Neonatal Infection. JAMA Netw Open. 2022;5:e2239796. https://doi.org/10.1001/jamanetworkopen.2022.39796

117. Spreafico R, Rossetti M, Whitaker JW, Wang W, Lovell DJ, Albani S. Epipolymorphisms associated with the clinical outcome of autoimmune arthritis affect CD4 $^{+}$ T cell activation pathways. Proc Natl Acad Sci. 2016;113:13845–50. https://doi.org/10.1073/pnas.1524056113

118. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv; 2018 [cited 2024 Sept 25]. https://doi.org/10.48550/ARXIV.1802.03426

119. Cavalca G, Vergani M, Cangelosi D, Consolaro A, Gattorno M, Ravelli A, et al. Stochastic epigenetic mutation profiles as biomarkers of clinical activity in juvenile idiopathic arthritis: a multi-omic machine learning approach for gene prioritization. Mol Med. 2025;31:289. https://doi.org/10.1186/s10020-025-01348-6

120. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Author Correction: Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2022;605:E3. https://doi.org/10.1038/s41586-021-04226-3

121. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–61. https://doi.org/10.1093/nar/gkw1092

122. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. J R Stat Soc Ser B Stat Methodol. 2005;67:301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x

123. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min [Internet]. San Francisco California USA: ACM; 2016 [cited 2024 Sept 20]. p. 785–94. https://doi.org/10.1145/2939672.2939785

124. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science. 2006;313:1929–35. https://doi.org/10.1126/science.1132939

125. Higgins-Chen AT, Thrush KL, Wang Y, Minteer CJ, Kuo P-L, Wang M, et al. A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking. Nat Aging. 2022;2:644–61. https://doi.org/10.1038/s43587-022-00248-2

126. Hillary RF, Marioni RE. MethylDetectR: a software for methylation-based health profiling. Wellcome Open Res. 2020;5:283. https://doi.org/10.12688/wellcomeopenres.16458.2

127. Ying K, Liu H, Tarkhov AE, Sadler MC, Lu AT, Moqri M, et al. Causality-enriched epigenetic age uncouples damage and adaptation. Nat Aging. 2024;4:231–46. https://doi.org/10.1038/s43587-023-00557-0

128. Jokai M, Torma F, McGreevy KM, Koltai E, Bori Z, Babszki G, et al. DNA methylation clock DNAmFitAge shows regular exercise is associated with slower aging and systemic adaptation. GeroScience. 2023;45:2805–17. https://doi.org/10.1007/s11357-023-00826-1

129. Teschendorff AE. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. Genome Med. 2020;12:56. https://doi.org/10.1186/s13073-020-00752-3

130. Pelegí-Sisó D, De Prado P, Ronkainen J, Bustamante M, González JR. *methylclock* : a Bioconductor package to estimate DNA methylation age. Robinson P, editor. Bioinformatics. 2021;37:1759–60. https://doi.org/10.1093/bioinformatics/btaa825

131. Malousi A, Kouidou S, Tsagiopoulou M, Papakonstantinou N, Bouras E, Georgiou E, et al. MeinteR: A framework to prioritize DNA methylation aberrations based on conformational and cis-regulatory element enrichment. Sci Rep. 2019;9:19148. https://doi.org/10.1038/s41598-019-55453-8

132. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20:163–72. https://doi.org/10.1038/s41590-018-0276-y

133. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python [Internet]. arXiv; 2018 [cited 2024 Sept 25]. http://arxiv.org/abs/1201.0490. Accessed 25 Sept 2024

134. Defazio A, Bach F, Lacoste-Julien S. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives [Internet]. arXiv; 2014 [cited 2024 Sept 25]. http://arxiv.org/abs/1407.0202. Accessed 25 Sept 2024

135. Chavez-Valencia RA, Chiaroni-Clarke RC, Martino DJ, Munro JE, Allen RC, Akikusa JD, et al. The DNA methylation landscape of CD4+ T cells in oligoarticular juvenile idiopathic arthritis. J Autoimmun. 2018;86:29–38. https://doi.org/10.1016/j.jaut.2017.09.010

136. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27. https://doi.org/10.1093/biostatistics/kxj037

137. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3. https://doi.org/10.1093/bioinformatics/bts034

138. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25. https://doi.org/10.1016/j.cels.2015.12.004

139. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44:W90-97. https://doi.org/10.1093/nar/gkw377

140. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8

141. Viechtbauer W. Conducting Meta-Analyses in *R* with the **metafor** Package. J Stat Softw [Internet]. 2010 [cited 2025 May 30];36. https://doi.org/10.18637/jss.v036.i03

142. Zhang W, Cai Z, Liang D, Han J, Wu P, Shan J, et al. Immune Cell-Related Genes in Juvenile Idiopathic Arthritis Identified Using Transcriptomic and Single-Cell Sequencing Data. Int J Mol Sci. 2023;24:10619. https://doi.org/10.3390/ijms241310619

143. Singh K, Deshpande P, Li G, Yu M, Pryshchep S, Cavanagh M, et al. K-RAS GTPase- and B-RAF kinase-mediated T-cell tolerance defects in rheumatoid arthritis. Proc Natl Acad Sci U S A. 2012;109:E1629-1637. https://doi.org/10.1073/pnas.1117640109

144. Bira Y, Tani K, Nishioka Y, Miyata J, Sato K, Hayashi A, et al. Transforming growth factor beta stimulates rheumatoid synovial fibroblasts via the type II receptor. Mod Rheumatol. 2005;15:108–13. https://doi.org/10.1007/s10165-004-0378-2

145. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. https://doi.org/10.1186/1471-2105-9-559

146. Gao X, Huang X, Wang Y, Sun S, Chen T, Gao Y, et al. Global research hotspots and frontier trends of epigenetic modifications in autoimmune diseases: a bibliometric analysis from 2012 to 2022 [Internet]. Allergy and Immunology; 2023 [cited 2025 May 30]. https://doi.org/10.1101/2023.07.05.23292268

147. Surace AEA, Hedrich CM. The Role of Epigenetics in Autoimmune/Inflammatory Disease. Front Immunol. 2019;10:1525. https://doi.org/10.3389/fimmu.2019.01525

148. Spada E, Calzari L, Corsaro L, Fazia T, Mencarelli M, Di Blasio AM, et al. Epigenome Wide Association and Stochastic Epigenetic Mutation Analysis on Cord Blood of Preterm Birth. Int J Mol Sci. 2020;21:5044. https://doi.org/10.3390/ijms21145044

149. Xu W, Monaco G, Wong EH, Tan WLW, Kared H, Simoni Y, et al. Mapping of γ/δ T cells reveals Vδ2+ T cells resistance to senescence. EBioMedicine. 2019;39:44–58. https://doi.org/10.1016/j.ebiom.2018.11.053

150. Okamoto H, Yoshio T, Kaneko H, Yamanaka H. Inhibition of NF-κB signaling by fasudil as a potential therapeutic strategy for rheumatoid arthritis. Arthritis Rheum. 2010;62:82–92. https://doi.org/10.1002/art.25063

151. Zhang M, Dai R, Zhao Q, Zhou L, An Y, Tang X, et al. Identification of Key Biomarkers and Immune Infiltration in Systemic Juvenile Idiopathic Arthritis by Integrated Bioinformatic Analysis. Front Mol Biosci. 2021;8:681526. https://doi.org/10.3389/fmolb.2021.681526

152. Tanaka Y, Millson D, Iwata S, Nakayamada S. Safety and efficacy of fostamatinib in rheumatoid arthritis patients with an inadequate response to methotrexate in phase II OSKIRA-ASIA-1 and OSKIRA-ASIA-1X study. Rheumatology. 2021;60:2884–95. https://doi.org/10.1093/rheumatology/keaa732

153. Pavithra J P, Shila G S, Maheshwaran P M, Staines M J, Sundaram S S. Exploring the Anti-Inflammatory Effects of Aprepitant: An Antagonist of the Neurokinin-1 Receptor. Asian J Biol Life Sci. 2024;12:611–5. https://doi.org/10.5530/ajbls.2023.12.80

154. Liu X, Zhu Y, Zheng W, Qian T, Wang H, Hou X. Antagonism of NK-1R using aprepitant suppresses inflammatory response in rheumatoid arthritis fibroblast-like synoviocytes. Artif Cells Nanomedicine Biotechnol. 2019;47:1628–34. https://doi.org/10.1080/21691401.2019.1573177

155. Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46. https://doi.org/10.1016/j.csbj.2021.06.030

156. Andrews SJ, Renton AE, Fulton-Howard B, Podlesny-Drabiniok A, Marcora E, Goate AM. The complex genetic architecture of Alzheimer's disease: novel insights and future directions. EBioMedicine. 2023;90:104511. https://doi.org/10.1016/j.ebiom.2023.104511

157. Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, Holland D, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet. 2021;53:1276–82. https://doi.org/10.1038/s41588-021-00921-z

158. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet. 2022;54:412–36. https://doi.org/10.1038/s41588-022-01024-z

159. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30. https://doi.org/10.1126/science.aaz1776

160. Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat Genet. 2020;52:626–33. https://doi.org/10.1038/s41588-020-0625-2

161. Connally NJ, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, Chun S, et al. The missing link between genetic association and regulatory function. eLife. 2022;11:e74970. https://doi.org/10.7554/eLife.74970

162. Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Systematic differences in discovery of genetic effects on gene expression and complex traits. Nat Genet. 2023;55:1866–75. https://doi.org/10.1038/s41588-023-01529-1

163. Umans BD, Battle A, Gilad Y. Where Are the Disease-Associated eQTLs? Trends Genet. 2021;37:109–24. https://doi.org/10.1016/j.tig.2020.08.009

164. Bryois J, Calini D, Macnair W, Foo L, Urich E, Ortmann W, et al. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. Nat Neurosci. 2022;25:1104–12. https://doi.org/10.1038/s41593-022-01128-z

165. Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, Andersen S, et al. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. Science. 2022;376:eabf3041. https://doi.org/10.1126/science.abf3041

166. HIPSCI Consortium, Alasoo K, Rodrigues J, Mukhopadhyay S, Knights AJ, Mann AL, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet. 2018;50:424–31. https://doi.org/10.1038/s41588-018-0046-7

167. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015;12:931–4. https://doi.org/10.1038/nmeth.3547

168. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203. https://doi.org/10.1038/s41592-021-01252-x

169. Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nat Genet. 2025;57:949–61. https://doi.org/10.1038/s41588-024-02053-6

170. Pampari A, Shcherbina A, Kvon EZ, Kosicki M, Nair S, Kundu S, et al. ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. BioRxiv Prepr Serv Biol. 2025;2024.12.25.630221. https://doi.org/10.1101/2024.12.25.630221

171. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30. https://doi.org/10.1038/nature14248

172. The ENCODE Project Consortium, Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710. https://doi.org/10.1038/s41586-020-2493-4

173. Wang QS, Kelley DR, Ulirsch J, Kanai M, Sadhuka S, Cui R, et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. Nat Commun. 2021;12:3394. https://doi.org/10.1038/s41467-021-23134-8

174. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9. https://doi.org/10.1038/s41588-019-0538-0

175. Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, et al. Genome-wide enhancer maps link risk variants to disease genes. Nature. 2021;593:238–43. https://doi.org/10.1038/s41586-021-03446-x

176. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features [Internet]. arXiv; 2017 [cited 2025 Oct 30]. https://doi.org/10.48550/ARXIV.1706.09516

177. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis JAD. 2018;64:S161–89. https://doi.org/10.3233/JAD-179939

178. Fujita M, Gao Z, Zeng L, McCabe C, White CC, Ng B, et al. Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. Nat Genet. 2024;56:605–14. https://doi.org/10.1038/s41588-024-01685-y

179. Mathys H, Peng Z, Boix CA, Victor MB, Leary N, Babu S, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. Cell. 2023;186:4365-4385.e27. https://doi.org/10.1016/j.cell.2023.08.039

180. Cao X, Sun H, Feng R, Mazumder R, Buen Abad Najar CF, Li YI, et al. Integrative multi-omics QTL colocalization maps regulatory architecture in aging human brain [Internet]. Neurology; 2025 [cited 2025 Oct 30]. https://doi.org/10.1101/2025.04.17.25326042

181. Leung YY. FunGen-xQTL atlas: a multi-omic atlas of quantitative trait loci in human brain and cerebrospinal fluid for target identification and prioritization in Alzheimer's disease. Alzheimers Dement. 2024;20:e092045. https://doi.org/10.1002/alz.092045

182. Beecham GW, Bis JC, Martin ER, Choi S-H, DeStefano AL, van Duijn CM, et al. The Alzheimer's Disease Sequencing Project: Study design and sample selection. Neurol Genet. 2017;3:e194. https://doi.org/10.1212/NXG.0000000000000194

183. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9

184. Gerring ZF, Mina-Vargas A, Gamazon ER, Derks EM. E-MAGMA: an eQTL-informed method to identify risk genes using genome-wide association study summary statistics. Valencia A, editor. Bioinformatics. 2021;37:2245–9. https://doi.org/10.1093/bioinformatics/btab115

185. Comandante-Lou N, Lama T, Chen K, Zhang J, Hu B, Liu S, et al. PLXNB1 and other signaling drives a pathologic astrocyte state contributing to cognitive decline in Alzheimer's Disease [Internet]. Neuroscience; 2025 [cited 2025 Oct 30]. https://doi.org/10.1101/2025.02.24.639868

186. Green GS, Fujita M, Yang H-S, Taga M, Cain A, McCabe C, et al. Cellular communities reveal trajectories of brain ageing and Alzheimer's disease. Nature. 2024;633:634–45. https://doi.org/10.1038/s41586-024-07871-6

187. Robinson MD, McCarthy DJ, Smyth GK. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40. https://doi.org/10.1093/bioinformatics/btp616

188. Wang G, Sarkar A, Carbonetto P, Stephens M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. J R Stat Soc Ser B Stat Methodol. 2020;82:1273–300. https://doi.org/10.1111/rssb.12388

189. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625:92–100. https://doi.org/10.1038/s41586-023-06045-0

190. Zeng T, Spence JP, Mostafavi H, Pritchard JK. Bayesian estimation of gene constraint from an evolutionary model with gene features. Nat Genet. 2024;56:1632–43. https://doi.org/10.1038/s41588-024-01820-9

191. Lopes KDP, Snijders GJL, Humphrey J, Allan A, Sneeboer MAM, Navarro E, et al. Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. Nat Genet. 2022;54:4–17. https://doi.org/10.1038/s41588-021-00976-y

192. ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47:1228–35. https://doi.org/10.1038/ng.3404

193. Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. Science. 2019;366:1134–9. https://doi.org/10.1126/science.aay0793

194. Lakhani CM, Cavalca G, Liu A, Nidumbur R, Feng R, Raj T, et al. Machine Learning-Based Prediction of Cell-type Resolved Brain eQTLs Enhances Discovery of Variants Explaining Alzheimer's Disease Heritability [Internet]. Genetic and Genomic Medicine; 2025 [cited 2026 Feb 16]. https://doi.org/10.64898/2025.12.03.25341562

195. Hausmann J, Dedeoglu F, Broderick L. Periodic Fever, Aphthous Stomatitis, Pharyngitis, and Adenitis Syndrome and Syndrome of Unexplained Recurrent Fevers in Children and Adults. J Allergy Clin Immunol Pract. 2023;11:1676–87. https://doi.org/10.1016/j.jaip.2023.03.014

196. Broderick L, Hoffman HM. Pediatric recurrent fever and autoinflammation from the perspective of an allergist/immunologist. J Allergy Clin Immunol. 2020;146:960-966.e2. https://doi.org/10.1016/j.jaci.2020.09.019

197. Papa R, Penco F, Volpi S, Sutera D, Caorsi R, Gattorno M. Syndrome of Undifferentiated Recurrent Fever (SURF): An Emerging Group of Autoinflammatory Recurrent Fevers. J Clin Med. 2021;10:1963. https://doi.org/10.3390/jcm10091963

198. Macaraeg M, Baker E, Handorf E, Matt M, Baker EK, Brunner H, et al. Clinical, Immunologic, and Genetic Characteristics in Patients With Syndrome of Undifferentiated Recurrent Fevers. Arthritis Rheumatol Hoboken NJ. 2025;77:596–605. https://doi.org/10.1002/art.43065

199. Palmeri S, Ponzano M, Recchi G, Conti C, La Bella S, Sutera D, et al. Predictive factors for therapeutic response and cluster analysis in syndrome of undifferentiated recurrent fever (SURF). RMD Open. 2025;11:e005874. https://doi.org/10.1136/rmdopen-2025-005874

200. Palmeri S, Penco F, Bertoni A, Bustaffa M, Matucci-Cerinic C, Papa R, et al. Pyrin Inflammasome Activation Defines Colchicine-Responsive SURF Patients from FMF and Other Recurrent Fevers. J Clin Immunol. 2024;44:49. https://doi.org/10.1007/s10875-023-01649-7

201. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81:559–75. https://doi.org/10.1086/519795

202. Sun X, Chen J, Chen X, Ma J, Xiao L, Yao S, et al. Nuclear receptor coactivator 3 transactivates proinflammatory cytokines in collagen-induced arthritis. Cytokine. 2023;161:156074. https://doi.org/10.1016/j.cyto.2022.156074

203. Dillemans L, De Somer L, Neerinckx B, Proost P. A review of the pleiotropic actions of the IFN-inducible CXC chemokine receptor 3 ligands in the synovial microenvironment. Cell Mol Life Sci. 2023;80:78. https://doi.org/10.1007/s00018-023-04715-w

204. Jorgensen BG, Berent RM, Ha SE, Horiguchi K, Sasse KC, Becker LS, et al. DNA methylation, through DNMT1, has an essential role in the development of gastrointestinal smooth muscle cells and disease. Cell Death Dis. 2018;9:474. https://doi.org/10.1038/s41419-018-0495-z

205. Rialdi A, Campisi L, Zhao N, Lagda AC, Pietzsch C, Ho JSY, et al. Topoisomerase 1 inhibition suppresses inflammatory genes and protects from death by inflammation. Science. New York, N.Y.; 2016;352:aad7993. https://doi.org/10.1126/science.aad7993