Thesis Report
To be defended on: 05-06-2025

Age-Independent Prediction of Allergy using Graph Learning from DNA Methylation

Annabelle Vletter *TU Delft*, *Rewire*
Prof. M. Prendergast, *M. van Breugel*

Abstract

Differential DNA methylation patterns can serve as biomarkers for allergic diseases such as pediatric asthma and rhinitis, but age-dependent variability in epigenetic profiles undermines the reliability of predictive models. This thesis addresses that challenge by introducing a graph-based deep learning approach for **age-independent** allergy prediction from DNA methylation data. Each subject's DNA methylation profile is represented as an individualized graph constructed via an extended Weighted Gene Co-expression Network Analysis (WGCNA) that captures global co-methylation structure and subject-specific patterns, thus balancing population-level relationships with individual epigenetic heterogeneity. Edges between CpG sites are assigned weights using a Gaussian kernel on methylation values, ensuring the graph reflects personalized similarity while maintaining biologically meaningful connections. A Graph Neural Network (GNN) with an Edge Convolution (EdgeConv) architecture is then trained on these subject-specific graphs to predict allergy outcomes. We evaluated this framework on DNA methylation data from three harmonized pediatric cohorts (PIAMA, MAKI, COPSAC) processed with the MEFFIL pipeline for cross-cohort normalization and quality control. An Epigenome-Wide Association Study (EWAS) identified key CpG features associated with asthma, rhinitis and IgE, which were used to guide feature selection for model training. Our graph-based model outperformed conventional methods like ElasticNet and XGBoost in certain cohorts and maintained robust predictive accuracy between the ages of 6 and 16, demonstrating a certain resilience to age-related methylation differences. Furthermore, we applied gradient-based saliency analysis to the trained GNN to highlight influential methylation features, providing interpretability and revealing plausible epigenetic markers of allergy. The proposed pipeline is **scalable** and **interpretable**, and its ability to deliver reliable, age-invariant risk predictions from early-life epigenetic data underscores its potential clinical utility for early allergy diagnostics in children.

I. Introduction

Asthma remains one of the most prevalent chronic respiratory conditions among children, significantly impacting quality of life and posing a considerable healthcare burden globally [1]. Early and accurate diagnosis is essential for timely intervention, yet traditional diagnostic approaches often struggle due to symptom variability and the complexity of disease presentation across different populations especially young children [1]. Although artificial intelligence (AI) techniques have shown promise in enhancing predictive capabilities for complex diseases [2], their effectiveness is frequently limited by challenges in generalizing across diverse patient cohorts [3].

DNA methylation (DNAm) has emerged as a powerful early-detection biomarker: it integrates genetic predisposition with environmental and lifestyle exposures, offering a read-out of the underlying biological state [4]. However, most conventional prediction pipelines—typically linear models or other single-site analyses—cannot capture the intricate, non-linear and multi-locus interactions that govern the methylome [5]. Their apparent success within a single cohort often masks a lack of robustness when faced with new data [3].

We posit that deep-learning frameworks, with their capacity to learn hierarchical patterns, are uniquely suited to model the inherent three-dimensional and long-range structure of the genome relationships that classical models struggle to represent. By encoding these complex DNA-wide interactions, deep learning should yield more informative features and, ultimately, stronger predictive power [6].

This thesis specifically, investigates advanced graph-based approaches to solve this age-independent prediction issue, specifically integrating Weighted Gene Co-expression Network Analysis (WGCNA) [7] [8] for graph building and EdgeConv for feature extraction and classification [9], to construct individualized methylation networks for each patient and subsequently use those to predict disease. In doing so, it aims to address critical gaps in prediction accuracy and cross-cohort generalizability, ultimately advancing personalized and clinically actionable diagnostic strategies for asthma and other allergic diseases.

Our goal, therefore, is to learn DNAm signatures that remain predictive across the full paediatric age spectrum, even when the model is trained on only a few small, age-skewed cohorts. Achieving such age-robust performance

would make DNAm-based screening a practical clinical tool, particularly for the youngest children, whose symptoms are most ambiguous and who are hardest to recruit for studies [3].

This research is non-trivial as the methylation dataset is very large with 450,000 points per patient, and allergic disease is not the only biomarker impacting methylation sites. Additionally research cohorts are expensive to obtain, meaning we have relatively few patients, however we have been provided access to three different medical cohorts making the use of Deep Learning models, while still difficult, possible. The use of three harmonized cohorts in combination with deep learning makes this work unique in the research space for DNA methylation for allergic disease.

The guiding research questions are as follows:

1. Does combining and harmonizing data from diverse cohorts enhance the robustness of allergy prediction across age groups?
2. Can deep learning and graph-search techniques be used to incorporate the global connections of DNA methylation into the feature selection process?
3. Does using deep learning and graph-search techniques lead to more age-independent prediction of allergies?

II. Background

A. *Epigenetics and DNA Methylation in Disease Prediction*

Epigenetics refers to heritable shifts in gene activity that occur without changing the underlying DNA sequence [4]. Of the many epigenetic marks, DNA methylation (DNAm) is especially prominent because of its strong influence on transcriptional regulation. By binding methyl groups to cytosines within CpG dinucleotides, DNAm compacts chromatin and restricts access of transcription factors and other enzymes to the underlying DNA [10](Figure I). In this way, methylation acts as an epigenetic switch, governing which genes are expressed and thereby shaping the full spectrum of physiological functions in the human body [10].

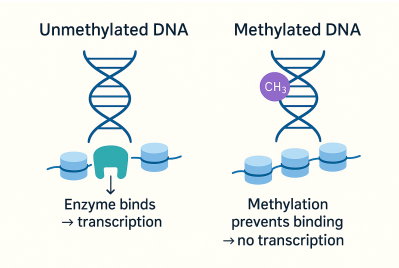

Figure I: Visualization of the DNA-methylation process: a methyl group blocks enzyme binding and compacts chromatin, so the DNA cannot be transcribed.

DNA methylation patterns can be significantly influenced by environmental factors, age, lifestyle, and various stressors, adding layers of complexity and variability to the methylome [10]. However, these dynamic do interactions mean that the methylome provides a valuable biomarker for early disease detection, potentially even before the clinical manifestation of symptoms, making DNA methylation a promising target for developing predictive models for personalized medicine and early intervention strategies.

DNA methylation has proven its predictive ability across various different diseases. A 30-CpG nasal signature identifies children who will go on to develop atopic asthma with an AUC of 0.93 [11], and a compact 3-site model [3] keeps similar performance while eliminating over-fit. In oncology, which is typically slightly easier as one compares tumor cells to healthy ones, a nested-genetic-algorithm pipeline classifies TCGA tumours with 99.1 % accuracy for malignant lesions and 93.9 % for benign ones [12]. Even early, subtle disease is detectable: the DeepMeth auto-encoder [13] spots lung cancer in circulating cfDNA with an AUC of 0.81—well before clinical presentation [13]. These successes are possible because methylation shifts often precede symptoms [7] and because profiling is relatively affordable and minimally invasive [3]. Taken together, consistently high accuracy/AUC, pre-symptomatic signal, and easy sampling make DNA-methylation a uniquely powerful biomarker for precision prediction.

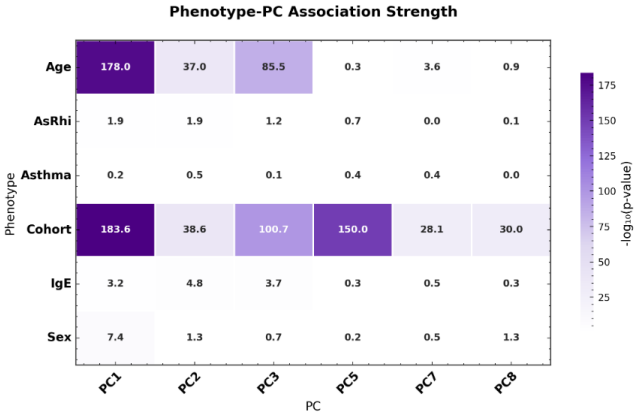

Figure II: Principal-component analysis of the harmonised cohorts, showing how each phenotype drives variance along its most influential principal components.

However, an examination of the principal component analysis (PCA) results by phenotype for allergic disease (Figure II) highlights why cross-cohort prediction is challenging and why linear dimensionality reduction techniques like PCA may be suboptimal for disease classification. Asthma and rhinitis have but subtle associated methylation signatures, especially when compared to stronger influences like age and cohort. These dominant sources of variation overshadow the disease-related signals, making them harder to detect.

This challenge is particularly evident in Figure III, which presents the principal components accounting for the most variance in relation to the combined Asthma and Rhinitis (AsRhi) phenotype. The lack of clear separation between affected and unaffected groups underscores the limited discriminatory power of linear methods in capturing the complex and nuanced epigenetic patterns associated with allergic disease.

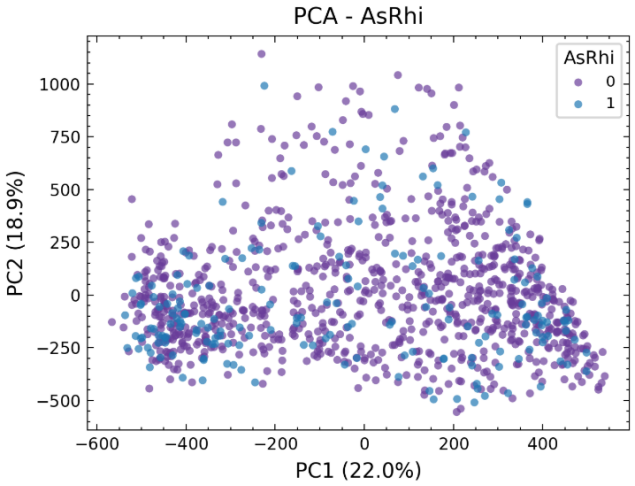

Figure III: Principal component illustrating the dominant variance axis for the combined Asthma + Rhinitis phenotype, highlighting how little variance there is in the methylation of the two groups.

B. *Limitations of Current Methylation Analysis Approaches*

DNA methylation analyses often use single-site methods like Epigenome-Wide Association Studies (EWAS) [14], which identify CpGs with significant methylation differences between cases and controls. While simple and widely used, these approaches are highly sensitive to outliers and prone to errors from measurement noise or biological variability [15, 16].

Consequently, predictors that are trained on a single cohort often under-perform when applied to new cohorts, undermining their clinical usefulness where robust, age-independent accuracy is essential. Maas et al. already illustrated this problem: when their single-CpG models were tested outside the discovery set, AUCs swung from 0.60 to 0.84 [17]. Yet the change in performance depends on which external cohort is chosen. A study by **van Breugel et al.** used an Elastic-Net classifier trained on nasal DNA-methylation data from the PIAMA cohort. They distilled the signal to just three CpG sites and obtained a PR-AUC of 0.502 on held-out PIAMA subjects. Performance rose to 0.837 in the age-matched EVA-PR cohort (children aged 7–20) but fell when the same three-CpG model was applied to younger populations—PR-AUC 0.348 in MAKI and 0.372 in COPSAC, both cohorts of 6-year-olds [3]. These age-dependent swings underscore that single-CpG models currently lack the robustness required for broad clinical deployment. Note that this study was trained and tested on the same data as the current research and thus will make a good comparison for the results.

Although EWAS is often treated as a complete classification strategy, it can also serve as a pragmatic dimensionality-reduction filter, ranking CpGs on variance and trimming the $\sim$ 450,000 loci assayed on the Illumina 450K array down to a tractable subset before more complex modeling techniques thereby mitigating the curse of dimensionality in the small-sample studies. This approach gains empirical support from studies demonstrating that EWAS-filtered CpG subsets achieve clinical-grade predictive performance, while outperforming conventional dimension reduction methods like PCA in computational efficiency and biological interpretability [18].

To overcome the limitations of single-CpG analyses, region-based methods like Bump Hunting [19], Probe Lasso [20], and DMRcate [5] group adjacent CpGs into Differentially Methylated Regions (DMRs), improving interpretability, dimensionality reduction [6], and noise robustness [21]. However, these methods focus on local patterns and often miss broader, trans-chromosomal interactions. Additionally, their reliance on linear models and correlation-based clustering limits their ability to capture the complex, non-linear methylation landscape, and their sensitivity to cohort differences hampers generalizability.

These challenges highlight the need for advanced approaches capable of modeling global methylation interactions and integrating biological context, such as genomic annotations and regulatory elements [6]. Graph-based and deep learning methods offer a promising solution, enabling the representation of intricate, long-range dependencies in a scalable and interpretable framework suitable for robust disease prediction

C. *Global Connectivity and Graph-Based Methods*

Global connectivity methods attempt to address these limitations by integrating information across the entire genome [22], recognizing both cis- (within chromosome) and trans-chromosomal (across chromosome) interactions. Methods such as Autoencoders [23, 13, 24] and Variational Autoencoders (VAEs) [25, 26] have been proposed to extract latent biological patterns from high-dimensional DNA methylation data. These approaches, despite showing improved prediction capabilities, face interpretability challenges that complicate clinical translation.

Graph-based machine learning methods could represent a particularly promising advancement in this domain [27, 28, 29]. By modeling methylation data as a network (graph), where nodes correspond to CpG sites and edges represent their interactions, graph-based approaches could be used to capture both local and global methylation dynamics. These interactions include biologically

relevant non-linear and non-local relationships, thus providing deeper biological insights.

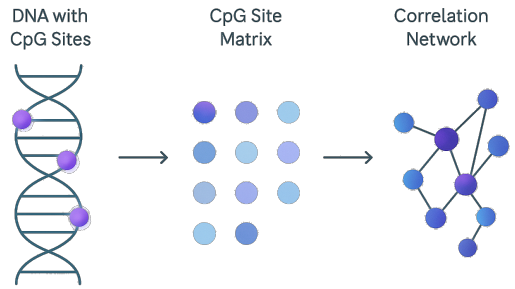

Figure IV: (Left) A DNA double helix with marked methylated CpG sites. (Center) Each CpG site becomes a node in a feature matrix whose entries are the sites' methylation percentages. (Right) Edges are inferred with various similarity measures—most often Pearson correlation [7], Gaussian-kernel similarity [30], or representation-learning methods such as DeepWalk [31]—yielding a CpG network that captures coordinated methylation patterns.

Recent work illustrates that graph learning could provide capabilities that other methylation models lack. Jiang *et al.* who completed the only other study using graph learning on methylation data, converting methylation data into a graph as in figure IV, transformed pairwise CpG correlations into a weighted interaction graph similar to WGCNA, and passed it through a Self-Attention GCN, achieving an AUC of 0.9987 across 32 tumour types—far above convolutional or multilayer-perceptron baselines that ignore graph topology [7] At the unsupervised end of the spectrum, Li *et al.* applied DeepWalk to a miRNA–disease bipartite graph: random-walk embeddings preserved higher-order connectivity and lifted association prediction to AUC $\approx$ 0.90 compared with attribute-only or matrix-factorisation approaches [32].

These studies show that graph learning captures long-range dependencies missed by single-CpG or local methods, is data-efficient via sparse edge sharing, and offers interpretable outputs like attention weights or node importance. These strengths make it well suited for clinically robust methylation-based prediction.

Interpretability is crucial in medical models, where incorrect predictions can have serious consequences. For graph-based methylation models, methods like gradient saliency help identify the CpG sub-networks driving predictions. Gradient-based approaches are scalable to large graphs and maintain biological relevance by computing first-order importance scores without retraining. More advanced post-hoc methods—such as GNNExplainer [33], and PGM-Explainer [34]—offer deeper insight into model decisions. Combining these tools with GNNs ensures both predictive performance and transparency, critical for clinical deployment in age-independent allergy prediction.

1. *WGCNA.*

Among hand-crafted network methods, *Weighted Gene/ CpG Co-expression Network Analysis* (WGCNA) is commonly used for uncovering structure in biological data like gene co-expression [35], and was the method selected by Jiang et al. for their graph building model [7]. WGCNA assumes a scale-free topology (figure V) an expected property of epigenetic networks [36, 37] and yields interpretable modules whose hub CpGs often map to key regulatory genes [7].

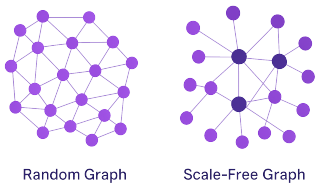

Figure V: WGCNA creates scale-free graphs which are more accurate towards epigenetic systems compared to random graphs [37]

2. *EdgeConv*

Edge Convolution (EdgeConv) was introduced by Wang et al. in the Dynamic Graph CNN (DGCNN) architecture for 3-D point-cloud analysis [9]. Their key idea was to impose a k-nearest-neighbour graph on an otherwise unordered point set and learn "edge features" that encode the geometric relation between each point and its neighbours. By recomputing the graph after every layer, DGCNN lets these edge features evolve with the learned representation, so information diffuses non-locally while the network remains permutation-invariant and lightweight. This design closed the performance gap with grid-based CNNs, achieving state-of-the-art results on ModelNet40 classification, ShapeNetPart segmentation, and S3DIS indoor-scene parsing benchmarks.

In our approach, the CpG methylation data is treated as the "point cloud," with each CpG site represented by its methylation level. WGCNA structures this point cloud into a biologically informed, scale-free graph where edges capture co-methylation strength. EdgeConv is then applied to this graph, using the WGCNA-defined neighborhoods to guide feature aggregation. Unlike traditional GCNs, EdgeConv learns edge-aware representations by modeling both the relative differences between CpGs and their interaction strengths. This enables the capture of complex, non-linear dependencies across the methylome. By recomputing edge features through multiple layers, EdgeConv supports multiscale information flow, making it particularly effective for high-dimensional, low-sample methylation data. Its ability to combine biological priors with deep feature extraction makes it a powerful tool for epigenetic analysis.

III. Methods

We developed a biologically-informed pipeline for disease classification from DNA methylation data across three pediatric cohorts (MAKI, PIAMA, COPSAC). After harmonizing the data using the MEFFIL framework, we performed an EWAS to select the 50 most discriminative CpG sites. These features were used to construct subject-specific co-methylation graphs via an extended WGCNA approach that incorporates both global correlation structure and individual methylation profiles. We then trained a weighted EdgeConv graph neural network on these graphs, using contrastive learning and focal loss to address class imbalance. To benchmark performance, we compared the GNN against Elastic-Net logistic regression and XGBoost, both trained on the same CpG features. Evaluation was conducted using cohort-stratified cross-validation and standard classification metrics.

A. Cohorts

Throughout this study we analyse three paediatric cohorts—MAKI, PIAMA and COPSAC. Immunoglobulin E (IgE) is the allergen-specific antibody that initiates mast-cell degranulation and thus serves as a sensitive marker of atopic sensitisation; yet sensitisation alone is often asymptomatic. Following the MeDALL criteria developed by Pinart et al [38], we label a participant as "diseased" only when IgE positivity co-occurs with physician-diagnosed asthma or rhinitis. A summary of the cohorts is provided in Figure VI.

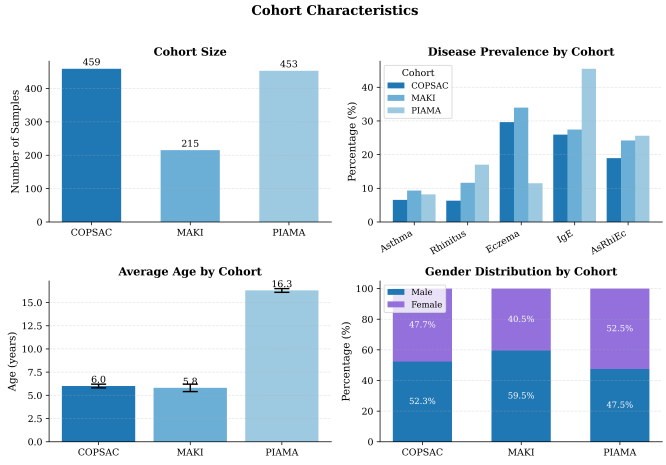

Figure VI: Summary of the three cohorts used in training and testing of the model. In this figure Eczema data is included though it is not used to classify disease status, see Appendix B.

B. MEFFIL

Prior to any downstream analysis, DNA-methylation data from the three cohorts were pre-processed and harmonized with the `meffil` R package [39]. The workflow—quality control, probe filtering, and functional normalization—removes technical variation (batch, array slide, and other platform effects) and dampens cohort-specific age effects, making the cohorts more comparable, despite their inherant technical variation. Residual variation was assessed with principal-component analysis (PCA); PCs associated with technical variation were regressed out before the harmonized β-values were passed to all subsequent analyses, including WGCNA module detection and graph-neural-network modeling [39]. Note that figure II and III were calculated post-MEFFIL meaning that there are still biological and cohort effects to be found in the data.

C. Epigenome-Wide Association Study (EWAS)

To make the high-dimensional research plausible it was necessary to reduce the number of CpG sites included in the network process down to 50, this is to mitigate the impact of the inherent imbalance in the dataset with only 1100 samples as opposed to 450,000 CpG sites.

Therefore, after pre-processing with MEFFIL an EWAS was conducted on the harmonized methylation data from the COPSAC, MAKI, and PIAMA cohorts. This was done to reduce the size of the dataset, as well as to identify CpG sites associated with key phenotypes. Phenotypic data, including derived outcomes such as `IgE_AsRhi` (co-occurrence of IgE positivity and asthma or rhinitis), were cleaned, encoded, and merged with methylation matrices via sample IDs.

For each CpG site, logistic regression was performed, adjusting for covariates such as age, sex, and batch. The output is a list of the 50 CpGs that show the highest variance across diseased/healthy patients. This list is then used to filter the relevant CpGs in downstream analyses.

D. Genomically aware WGCNA

Weighted Gene Co-*methylation* Network Analysis (WGCNA) is the backbone of our graph construction, however this methodology is extended on three fronts: *i)* every child receives an *individual* network rather than sharing a single cohort graph, this is to ensure clinical applicability; *ii)* edges encode methylation similarity as edge weights and *iii)* the final output is a sparse PyTorch-Geometric object that preserves biological context while remaining GPU-friendly. A full visualization and summary of the methodology can be seen in figure VII.

1. Per-CpG normalisation and imputation

After MEFFIL harmonisation and EWAS feature selection, each retained CpG column is z-scored, $x_i \leftarrow (x_i - \mu_i)/\sigma_i$, and any residual missing values are set to 0. This stabilises probe variance and guarantees numerically robust correlations.

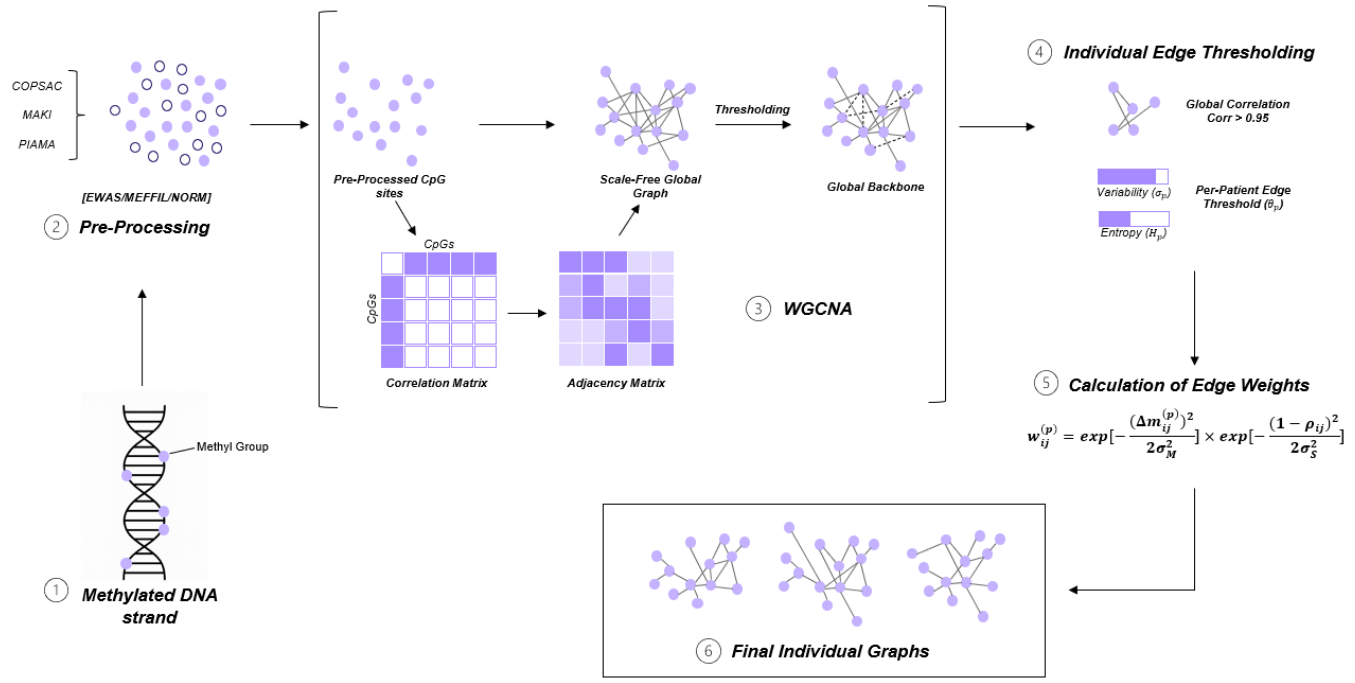

FIGURE VII: BREAKDOWN OF WGCNA METHODOLOGY: 1. THE ORIGINAL METHYLATED DNA STRAND, 2. THE METHYLATION AT THE RELEVANT LOCATIONS IS ANALYZED, THE TOP 50 SITES ARE CALCULATED BASED ON AN EWAS, AND THE THREE MEDICAL COHORTS ARE HARMONIZED 3. WGCNA IS PERFORMED OVER THE TRAIN SET TO CALCULATE A GLOBAL CORRELATION BACKBONE THAT IS LATER USED AS A BLUEPRINT FOR WHICH SITES TO ANALYZE 4. TO CREATE INDIVIDUAL GRAPHS, ASIDE FROM THE TOP 5 % OF EDGES FROM WGCNA, ADDITIONAL EDGES ARE CHOSEN BASED ON THE θ_p (EQ. 2) 5. EDGE WEIGHTS ARE CALCULATED USING GAUSSIAN SIMILARITY 6. THIS PROCESS IS REPEATED TO CREATE GRAPHS FOR EACH PATIENT

2. Global backbone construction

Across the training cohort we compute an *absolute* Pearson correlation matrix $\rho_{ij} = |\operatorname{cor}(x_i, x_j)|$. Following classic WGCNA we raise it to the power $\beta = 4$ and zero out weak entries (< 0.75), yielding a sparse template $A^{\text{global}} \in \mathbb{R}^{N \times N}$. This backbone is reused when building every subject graph, ensuring all networks share a common "scaffolding" of population-supported links.

3. Subject-specific edge selection

For each child p the normalised methylation profile is the vector $\mathbf{m}^{(p)} = (m_1^{(p)}, \ldots, m_N^{(p)})$. The goal is to decide which CpG pairs (i, j) to connect.

1. Core links. We first lock in a *cohort-wide* skeleton: every pair whose global correlation already satisfies $A_{ij}^{\text{global}} \geq 0.95$ (95th percentile of all absolute correlations) is included in *every* graph. These edges are so consistently strong that excluding them could erase biology shared by all subjects.

2. Adaptive fringe. Remaining sites from the global correlation matrix enter a second gate that adapts to each child's own methylome. Define

$$\sigma_p = \operatorname{sd}(\mathbf{m}^{(p)}) \quad \text{and} \quad H_p = -\sum_k p_k \log_2 p_k, \quad (1)$$

where p_k is the fraction of sites whose z-score falls into histogram bin k. A small standard deviation (σ_p) or entropy (H_p) means the subject's methylation levels cluster tightly; a large value signals dispersion.

We translate these two statistics into a personalised edge-weight cut-off based on Gaussian similarity

$$\theta_p = \operatorname{clip}\Big(0.80\,[1-\alpha\sigma_p/0.30]\,[1+\beta H_p/4],\ 0.60,\ 0.95\Big), \quad (2)$$

with $\alpha = \beta = 1$. Hence: if $\sigma_p > 0.30$ (above-average variation) $\Rightarrow$ the product $[1 - \alpha\sigma_p/0.30]$ dips below 1, *lowering* θ_p and admitting more edges so that the graph is not starved of information.
Alternatively, $H_p > 4$ (high entropy) $\Rightarrow$ the term $[1 + \beta H_p/4]$ exceeds 1, *raising* θ_p to suppress noisy links.

A hard clip confines θ_p to $[0.60, 0.95]$, preventing the threshold from drifting to extremes.

4. Edge weighting

Each retained edge carries the following weight value calculated as follows:

$$w_{ij}^{(p)} = \exp\Big[-\frac{(\Delta m_{ij}^{(p)})^2}{2\sigma_M^2}\Big] \times \exp\Big[-\frac{(1-\rho_{ij})^2}{2\sigma_S^2}\Big] \tag{3}$$

$\Delta m_{ij}^{(p)}$ Absolute methylation difference $\big|m_i^{(p)} - m_j^{(p)}\big|$. Small differences yield values near 1, large ones near 0.

ρ_{ij} Absolute Pearson correlation across the training cohort— a proxy for long-term co-regulation.

Both Gaussian kernels share bandwidth $\sigma_M = \sigma_S = 0.10$, chosen so that a half-sigma deviation shrinks the factor to ≈ 0.78.

5. Adaptive sparsification

Dense graphs bias message-passing toward a few highly connected CpGs. As we have already created highly connected graphs with WGCNA, we want to cap this effect, and therefore we restrict each node's out-degree:

$$d_i^{(p)} = \min\Big\{15,\ \max\big(5,\ \lceil 0.01\,|E^{(p)}|\rceil\big)\Big\}. \tag{4}$$

Thus every node keeps ≥ 5 edges (avoiding isolates), ≤ 15 edges (avoiding hubs). If a node exceeds its budget, only its top-weighted connections survive. The resulting network is small enough to learn from with relatively few samples, while it is also large enough to represent global structure, and balanced so that no single CpG dominates information flow.

6. Final graph representation

Each child p is represented by a PyG-ready object

$$G^{(p)} = \big(V,\ E^{(p)},\ \mathbf{x}^{(p)},\ \mathbf{e}^{(p)},\ y_p\big), \tag{5}$$

- **Nodes V:** one vertex per CpG (fixed order). Node feature $x_i^{(p)}$ is the subject's z-scored methylation value.
- **Edges $E^{(p)}$:** selected via the described logic
- **Edge attr. $\mathbf{e}_{ij}^{(p)}$:** $[\,w_{ij}^{(p)}]$, containing the edge weight.
- **Graph label y_p:** binary outcome (0 =control, 1 =case), used only during GNN training.

This *genomically aware WGCNA* thus weaves together subject-level co-methylation, as well as population evidence, yielding compact yet information-rich graphs ready for downstream deep-learning analysis.

E. Graph Feature Extraction and Model Training

After constructing biologically-informed graphs using WGCNA and subject-specific adaptation, the next stage involved training a graph neural network (GNN) to classify phenotypic labels. This section outlines the full computational pipeline from feature preprocessing to model evaluation, emphasizing patient-level graph normalization, model architecture, loss functions, and evaluation procedures. This process is also visualized in figure VIII.

1. EdgeConv GNN Architecture

Our model is a three-block *weighted* EdgeConv graph neural network. First, each node feature vector $\mathbf{x}_i \in \mathbb{R}^{d_{\text{in}}}$ is passed through a linear layer, batch normalisation, and a ReLU to obtain an encoded representation $\mathbf{h}_i^{(0)} \in \mathbb{R}^{d_{\text{hid}}}$.

For every edge (i, j) with scalar weight w_{ij} (met hylation correlation), an EdgeConv block computes the message

$$\mathbf{m}_{ij} = \text{MLP}\big[\mathbf{h}_i^{(\ell)},\ \mathbf{h}_j^{(\ell)} - \mathbf{h}_i^{(\ell)},\ w_{ij}\big], \tag{6}$$

followed by mean aggregation over the neighbourhood $\mathcal{N}(i)$:

$$\mathbf{h}_i^{(\ell+1)} = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij}. \tag{7}$$

We use three such blocks ($\ell = 0, 1, 2$); blocks 2 and 3 include residual connections: $\mathbf{h}_i^{(2)} \leftarrow \mathbf{h}_i^{(2)} + \mathbf{h}_i^{(1)}$ and $\mathbf{h}_i^{(3)} \leftarrow \mathbf{h}_i^{(3)} + \mathbf{h}_i^{(2)}$. After the third block, node features are pooled with global mean pooling to yield a graph embedding $\mathbf{z} \in \mathbb{R}^{d_{\text{hid}}}$, which is fed through a two-layer MLP (ReLU + dropout) to produce the final logits $\mathbf{o} \in \mathbb{R}^2$.

What each EdgeConv layer learns. An EdgeConv layer can be regarded as a local, edge-aware "convolution" on graphs that proceeds in three steps:

1. **Edge-conditioned message.** For every neighbour $j \in \mathcal{N}(i)$ the layer concatenates the *centre feature* $\mathbf{h}_i^{(\ell)}$, the *relative feature* $\mathbf{h}_j^{(\ell)} - \mathbf{h}_i^{(\ell)}$, and the *edge attribute* w_{ij}. This lets the kernel adapt both to direction and to methylation-correlation strength, unlike a vanilla GCN.
2. **Non-linear filtering.** The shared MLP transforms this concatenated vector, extracting a higher-level, task-specific representation of the edge.
3. **Permutation-equivariant aggregation.** The messages for all neighbours are averaged, giving an update that is invariant to the ordering of $\mathcal{N}(i)$ yet sensitive to their content. The output $\mathbf{h}_i^{(\ell+1)}$ therefore embeds both the node's own state and an edge-weighted summary of its local context.

Stacking three EdgeConv blocks allows information to propagate up to three hops, or three steps away, so the fi-

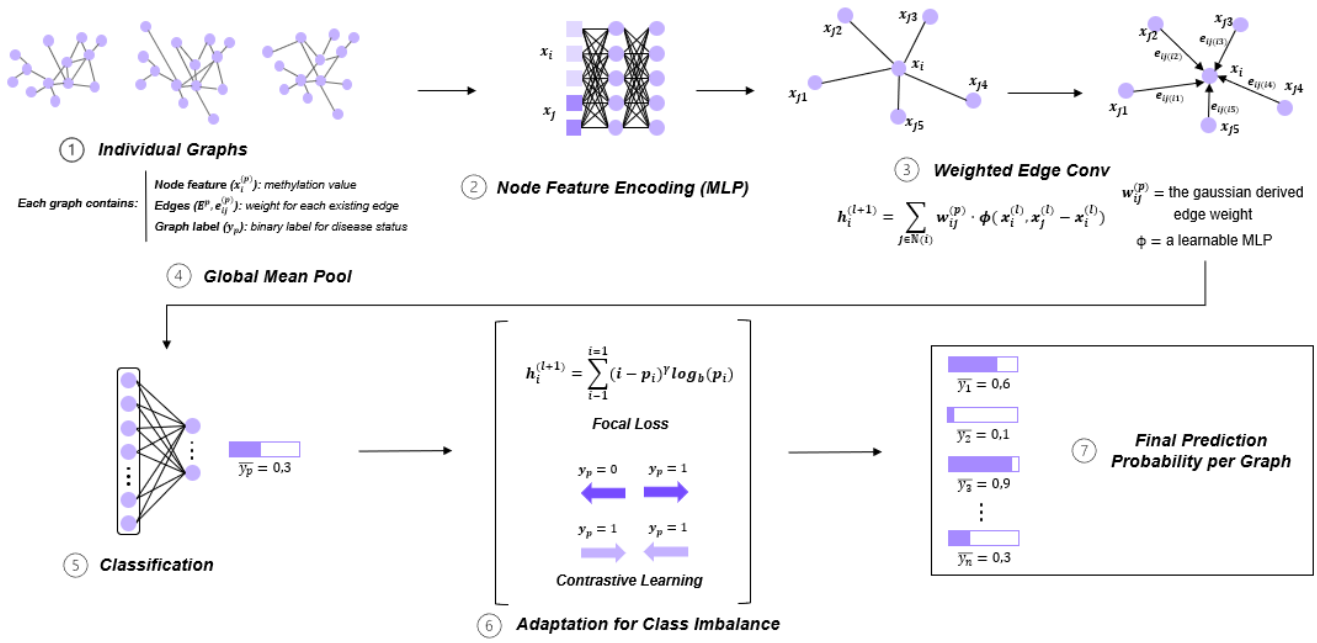

FIGURE VIII: BREAKDOWN OF EDGECONV METHODOLOGY: 1. GRAPHS CREATED USING GENOMIC AWARE WGCNA, 2. NODE FEATURE ENCODING, TRANSFORMING RAW NODE INPUT INTO RICHER REPRESENTATIONS 3. EDGECONV METHODOLOGY WITH MESSAGE PASSING [9] 4. POOLING OF RESULTS 5. INITAL DISEASE CLASSIFICATION 6. USE OF FOCAL LOSS AND CONTRASTIVE LEARNING TO ADAPT FOR THE CLASS IMBALANCE 7. FINAL CLASSIFICATION PREDICTION

nal embedding captures not only local promoter–TSS links but also distal enhancer–gene and even inter-chromosomal co-methylation patterns. Residual shortcuts ensure stable gradients through the deep, edge-conditioned stack.

2. *Loss Function*

Training minimises a composite objective

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{CL}} \mathcal{L}_{\text{contrast}} \tag{8}$$

where (i) $\mathcal{L}_{\text{focal}}$ is class-balanced focal loss ($\gamma = 2.0$, $\alpha = 0.35$); (ii) $\mathcal{L}_{\text{contrast}}$ is a supervised graph-level contrastive loss that pulls together embeddings of graphs with the same label and pushes apart different classes, using temperature-scaled cosine similarity.

3. *Training and Evaluation*

We perform 5-fold cross-validation. Each fold is trained for up to 200 epochs with Adam ($lr = 3 \times 10^{-4}$, weight-decay $= 10^{-4}$), using a one-cycle cosine learning-rate schedule with warm-up. Early stopping monitors validation PR-AUC with a patience of 150 epochs. At the end of training we select the checkpoint with the best validation PR-AUC, tune an optimal probability threshold on the validation data, and then report metrics on the hold-out test set: ROC-AUC, PR-AUC, F1, and Accuracy.

F. *Baseline Methods and Evaluation Metrics*

1. *ElasticNet Logistic Regression*

As a linear-model baseline we trained an *ElasticNet* logistic regression on the same tabular input that is output after MEFFIL and EWAS: the top–50 differentially methylated CpG features per sample. Prior to modelling, each feature vector was cleaned for NaN/ $\pm\infty$ entries, imputed with the training-set mean, and scaled with a `RobustScaler`. To mitigate label imbalance we supplied inverse-frequency class weights to the loss. Hyper-parameters were tuned by a five-fold grid search over the inverse regularisation strength $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and ℓ_1 mixing coefficient $\alpha_{\text{EN}} = 1 - \ell_2$ ratio l_1-ratio $\in \{0.1, 0.5, 0.7, 0.9\}$, using the `saga` optimiser and an ElasticNet penalty. The resulting sparse coefficient vector provides an interpretable ranking of CpGs: the absolute weight $|\beta_j|$ of each feature directly reflects its influence on the log-odds of the class label.

2. *Gradient-Boosted Trees (XGBoost)*

For a non-linear baseline we employed *XGBoost*—an ensemble of gradient-boosted decision trees—trained on the same CpG feature matrix. After robust scaling, an `XGBClassifier` was fit with 100 trees, maximum depth 4, learning rate 0.1, subsample ratio 0.8, and column-subsample ratio 0.8; early stopping halted boosting when the validation objective ceased improving. Class imbal-

ance was addressed via the built-in `scale_pos_weight` parameter, computed from the positive/negative sample counts in the training set. The model outputs probability estimates through logistic leaves, and its built-in gain–based feature importance yields a ranked list of CpGs whose splits contribute most to the boosted ensemble's predictions. Together, Elastic-Net and XGBoost serve as strong tabular-data baselines against which to benchmark the EdgeConv graph network.

3. *Classification Performance*

To evaluate the classification performance there are two common metrics Receiver-Operator Curve and Precision-Recall. The components of which are as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{10}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{11}$$

A *ROC* curve plots TPR against FPR; a *PR* curve plots TPR against PPV. ROC-AUC is robust but can look overly optimistic when the negative class vastly outnumbers the positive—as in methylation studies—because FPR is diluted by the many true negatives. PR-AUC avoids this bias: every extra false positive lowers precision, and its baseline equals the positive-class prevalence, making it the preferred metric for highly imbalanced data [3]. The performance of Precision-Recall is relative to the class imbalance, i.e. if the cohort has 20% positive samples, then a PR-score of 0.2 would mean that the model is guessing randomly.

4. *Ability of the model to capture global structure*

To identify the network elements that drive the asthma prediction we applied a lightweight *gradient–saliency* explainer. For each subject-graph we back-propagated the positive-class logit y to every **input** tensor—node features $\mathbf{x}$ and scalar edge weights w_{ij}—and took the absolute gradient $|\partial y/\partial x|$ as a first-order *importance* score: the larger the magnitude, the more a small perturbation of that CpG site or edge would change the output. Node scores were obtained directly; an edge score was defined as the sum of the two endpoint scores. Gradients were computed for the first 50 training graphs (one backward pass each), then averaged across subjects within a cohort to yield cohort-level importance maps. This saliency approach requires no additional training or perturbation cycles, making it scalable to thousands-of-node methylation graphs while still offering biologically interpretable rankings for downstream analysis. This method could however be overly positive as to the relative importance of the edges.

5. *Representation-learning advantage via dimensionality reduction*

To assess how message passing augments the learned representations, we compared our full GNN (node encoder + three WeightedEdgeConv layers) against an *MLP-only* baseline that shares the same node encoder but omits all edge convolutions. For each sample graph, we extracted a fixed-length embedding by global mean-pooling the final node features. These per-graph vectors were then projected into two dimensions using both t-SNE and PCA, yielding four scatterplots (GNN vs. MLP under t-SNE and PCA). No additional training or parameter tuning was required beyond a single forward pass per graph and standard dimensionality-reduction steps.

IV. Results

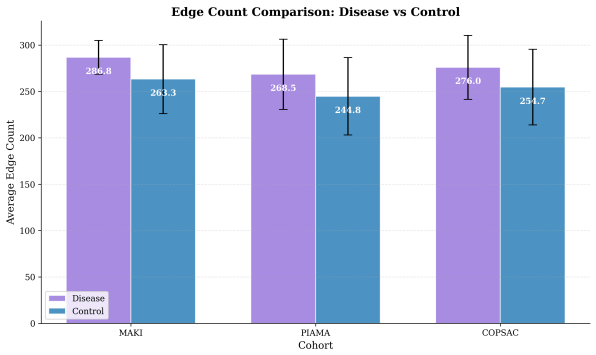

Figure IX: Comparison of the number of edges for all three cohorts.

The WGCNA network contained the top-50 phenotype-enriched CpG sites per chromosome. These 50 sites are those which are expected to be significant for our disease status according to the EWAS. Graph-level analysis revealed a consistent, disease-associated rewiring of the CpG interaction network. Across all three cohorts, graphs from allergic patients had more edges when compared to control patients reflecting a greater overall interaction load as seen in figure IX.

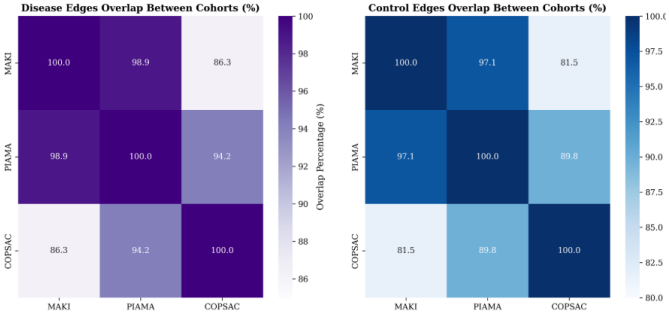

Figure X: Comparison percentage overlap between cohorts diseased v. control. MAKI and PIAMA show clear overlap while COPSAC graphs show clearer differences.

To assess the consistency of graph structure across cohorts, we compared the percentage of shared edges between diseased and control graphs. In figure X, it is visualized that disease graphs consistently exhibited higher edge overlap across cohort pairs than controls, indicating a more conserved co-methylation architecture. This suggests that allergic disease drives a shared reorganization of methylation dynamics in the top-50 sites form the EWAS, reinforcing the potential of graph-based models to generalize predictive patterns across diverse populations.

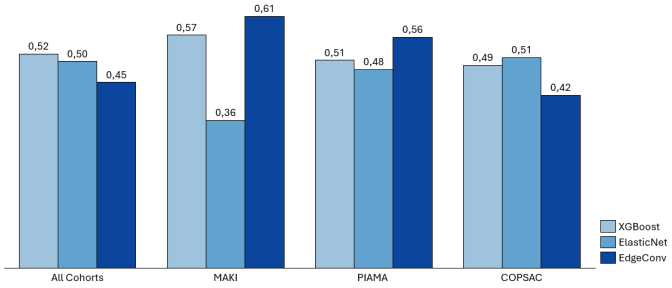

Figure XI: PR-AUC Classification performance per Cohort, across three tested models, XGBoost, ElasticNet, and WGCNA/EdgeConv. PRC results are relative to the prevalance of disease, meaning that for PIAMA the baseline is ≈ 0.23, for MAKI it is ≈ 0.16, and finally for COPSAC it is ≈ 0.12

Figure XI compares the two classical machine-learning models (ElasticNet and XGBoost) with the graph-based approache evaluated in this study. When trained and evaluated on the pooled MAKI–PIAMA–COPSAC dataset, XGBoost achieved the highest aggregate PR-AUC of 0.52. ElasticNet performed slightly worse at 0.50 and EdgeConv performed the worst at 0.45. However, this result is not seen across all cohort. EdgeConv performs better on classification in just MAKI and PIAMA and the relative weaker performance seems to come only from the COPSAC cohort.

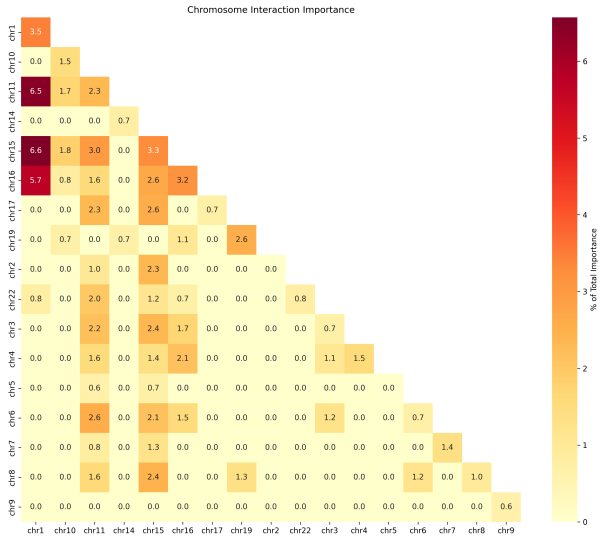

Figure XII: Evaluation of the location of the most important features, comparing the relative importance of trans-chromosomal edges to cis-chromsomal edges

Gradient–saliency analysis was used to identify the CpG features and interactions most influential to the EdgeConv model's predictions. When ranking features (Appendix E), the most important features were often edges—suggesting that the model relies heavily on co-methylation relationships between CpG sites. This emphasis on edges may partly reflect the nature of the gradient-based method, which assigns high importance to an edge if both connected CpGs are influential.

Importantly, many of these top-ranked edges span different chromosomes as seen in figure XII, indicating that EdgeConv captures trans-chromosomal interactions—patterns that are typically missed by methods focused only on local or linear relationships. This highlights the model's capacity to learn biologically meaningful, genome-wide methylation structures.

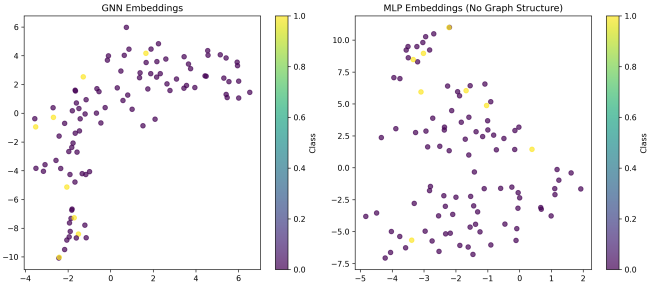

Figure XIII: t-SNE visualization of graph-level embeddings for 100 subjects. Each point represents a pooled embedding of a subject's methylation graph, colored by asthma status. Left: embeddings from the full GNN (node encoder+three WeightedEdgeConv layers) exhibit higher separability. Right: embeddings from an MLP baseline that omits all graph convolutions show no clear separation, demonstrating the representational advantage conferred by leveraging network structure.

The t-SNE and PCA projections in figure XIII reveal a slight improvement in class separability when graph structure is incorporated. In the left panel ("GNN Embeddings"), diseased samples (yellow) cluster more to the left of the curve, indicating that message passing over the methylation network does captures disease-relevant patterns, even still though there are quite a few samples overlapping. By contrast, the right panel ("MLP Embeddings, No Graph Structure") shows that embeddings derived solely from individual CpG features overlap heavily between cases and controls, with no clear cluster boundaries. This comparison indicate that the GNN's ability to aggregate information across correlated CpG sites produces more discriminative representations, which in turn underlie its improved predictive performance in two of the three cohorts.

V. Discussion

This study combined WGCNA–based graph construction with a weighted EdgeConv GNN to predict allergic disease from nasal DNA-methylation profiles across three pediatric cohorts. Results show (i) allergic patients show different methylation graphs when compared to control patiens (Figure IX), and (ii) graph learning offers competitive classification performance, outperforming XGBoost and ElasticNet in MAKI and PIAMA while lagging in COPSAC.

A. Research Questions

To begin we will look back to the original research questions.

1. *Can the combining of variant cohorts lead to a more robust prediction of allergies?*

To look at this results it is best to compare to the research by van Breugel et al [3] (Section II-B) as that study was also trained on PIAMA, and subsequently externally replicated on MAKI and COPSAC, our study was trained and tested on all three. One important caveat to the direct comparison of these results is that the van Breugel et al. study also classified using Eczema data. This reduced the class imbalance present in the dataset (figure VI) and could skew results. The results are summarized in table I.

Study	PIAMA	MAKI	COPSAC
van Breugel *et al.* [3]	0.50	0.35	0.37
ElasticNet	0.48	0.36	0.51
XGBoost	0.51	0.57	0.49
EdgeConv	0.56	0.61	0.42

Table I: PR-AUC comparison to van Breugel et al. [3] across cohorts. Their study included additional phenotype data, which reduced class imbalance. Additionally, they used MAKI and COPSAC for external validation.

We will focus on the results for testing on the PIAMA cohort—the dataset on which the three-CpG model of van Breugel *et al.* was originally trained. In this setting our re-implementation of Elastic-Net (with 50 CpGs) yields a similar PR-AUC (0.48 vs. the baseline 0.50), whereas the non-linear models show clear gains: XGBoost rises slightly from 0.50 to 0.51 and EdgeConv further to 0.56. This suggests that, on the same cohort, supplying a richer 50-CpG feature set in combination with a non-linear model benefits models that can exploit higher-order interactions or graph structure. Additionally, that combining datasets improves prediction, even when that means training on two datasets from different age groups.

External validation on independent cohorts will be needed for the combined dataset to confirm that the combined dataset, larger CpG panel and graph-based learning generalise beyond the data used here. But as an initial result we can see that the harmonized cohort predicts well on both 16-year-olds (PIAMA) as well as 6-year-olds (MAKI).

2. *Can deep learning and graph-search techniques incorporate the global connections of DNA methylation into the feature selection process?*

Figure XIII demonstrates that graph structure does contribute to the predictive ability of the model. This finding supports the hypothesis that graph-based learning frameworks provide a richer and more informative representation of the methylome compared to traditional machine learning models such as ElasticNet, which evaluate CpGs in isolation.

Notably, none of the top-20 most influential edges connect CpG sites located on the same gene (Appendix E). This suggests that global, long-range interactions dominate the feature importance landscape—consistent with our premise that disease-relevant methylation signatures are distributed across the genome rather than being confined to localized regions. However, this conclusion comes with two important caveats:

(i) The findings are influenced by the initial EWAS-based feature selection, which prioritized the top-50 CpG sites by differential signal. If these CpGs are widely dispersed across the genome, the potential for capturing local interactions is inherently limited.

(ii) Although the analysis confirms that many CpGs reside on different chromosomes, the biological implications of these trans-chromosomal links remain speculative. Further validation would require additional data modalities, such as Hi-C, which capture the 3D conformation of the genome.

In Figure XII, the relative importance of *trans*-chromosomal edges does reinforces the global perspective: cross-chromosome interactions dominate the model's learned feature space. While *cis*-chromosomal links do contribute, they appear to play a secondary role. This second analysis is also less impacted by the EWAS preprocessing as there are CpGs in the top-50 on all the plotted chromosomes. Together, these findings underscore the utility of deep graph models in capturing the non-local, system-wide structure of methylation regulation that may underpin allergic disease.

3. *Does using deep learning and graph-search techniques lead to more age-independent prediction of allergies?*

Our results suggest that deep learning and graph-based methods do contribute to more age-independent allergy prediction. EdgeConv performed strongly in both the PIAMA and MAKI cohorts, which together represent ages 16 and 6 respectively. This indicates improved gener-

alizability across age groups. Performance in COPSAC was weaker, likely in part due to the stronger class imbalance in COPSAC. Additionally, according to the dataset owners, asthma in COPSAC was labeled using alternative clinical criteria than is common, introducing potential label noise. Deep Learning models like EdgeConv appear more sensitive to this kind of inconsistency, whereas other methods like XGBoost and ElasticNet remain comparatively robust. This difference was likely not corrected by MEFFIL preprocessing as MEFFIL focuses on technical variation like batch size. While the overall findings are promising, especially across PIAMA and MAKI, confirming true age-independence will require external validation on additional cohorts beyond the scope of this thesis.

B. Biological interpretation and Clinical Outlook

The disease-associated increase in edge count observed in our graphs suggests biologically meaningful changes in the methylome of allergic children. However, to fully confirm the spatial and regulatory significance of these connections, integration with three-dimensional chromatin conformation data (e.g., Hi-C) would be ideal, such datasets are currently expensive and challenging to generate [40]. Clinically, this work demonstrates the feasibility of applying graph-based learning to DNA methylation at the individual-patient level, with a scalable and interpretable framework. Since a patient-specific graph can be constructed from minimally invasive nasal samples and processed through the trained model, this method holds promise for future diagnostic use.

C. Limitations

This study also presents several limitations that should be addressed in future work. First, discrepancies between cohorts impact model performance: for example, the MAKI cohort is relatively small, and the COPSAC cohort uses differing phenotyping criteria, which may reduce classification accuracy even when techniques such as focal loss and class weighting are applied. Second, graph sparsity remains a constraint. While limiting the number of CpG sites to 50 per chromosome helps manage GPU memory, and balance the discrepancy between the number of CpG sites and size of the dataset, this restriction may inadvertently exclude biologically relevant loci. More adaptive pruning strategies—such as attention-based feature selection—could improve information retention. Third, the current use of gradient saliency for model interpretability may oversimplify the true underlying feature interactions. As this method ignores higher-order graph dependencies and can suffer from saturation effects, more expressive explanation techniques, such as GNNExplainer or PGExplainer, are necessary to accurately identify and interpret the subgraphs most influential in driving predictions.

VI. Conclusion and Future Work

This thesis presents a novel graph-based deep learning framework for predicting allergic disease from DNA methylation data in pediatric cohorts. By first using three harmonized cohorts and subsequently combining EWAS-guided feature selection, genomically aware WGCNA graph construction, and the EdgeConv GNN architecture, we model long-range, patient-specific comethylation patterns in a scalable and interpretable way. Our results demonstrate that graph-based learning not only captures biologically meaningful relationships—often spanning chromosomes and aligning with immune-related pathways but can match or outperforms traditional approaches like ElasticNet and XGBoost in certain cohorts.

Despite the promising results, the work is constrained by several limitations. The top-CpG selection process may bias the graph's genomic coverage, and current analyses lack integration with 3D genome data (e.g., Hi-C), which could validate or refine inferred interactions. Furthermore, while the framework is designed to be clinically applicable its predictive accuracy and robustness must still be tested in external, independent datasets.

To that end, three additional cohorts (ATLANTIS, ADEM, and VIVA) have already been identified for external validation. Future work should focus on applying stronger graph-explainer models (e.g., GNNExplainer, PGExplainer) for subgraph attribution and interpretability, and on incorporating multi-omics modalities such as RNA-seq and Hi-C to evaluate functional and spatial coherence of the learned methylation graphs.

References

[1] H. Jones, A. Lawton, and A. Gupta, "Asthma attacks in children—challenges and opportunities," *Indian Journal of Pediatrics*, vol. 89, no. 4, pp. 373–377, 2022. [Online]. Available: https://doi.org/10.1007/s12098-021-04069-w

[2] N. Sheliemina, "The use of artificial intelligence in medical diagnostics: Opportunities, prospects and risks," *Health Economics and Management Review*, vol. 5, no. 2, pp. 104–124, 2024, open Access Article. [Online]. Available: https://doi.org/10.61093/hem.2024.2-07

[3] M. van Breugel, C. Qi, Z. Xu, C.-E. T. Pedersen, I. Petoukhov, J. M. Vonk, U. Gehring, M. Berg, M. Bügel, O. A. Carpaij, E. Forno, A. Morin, A. U. Eliasen, Y. Jiang, M. van den Berge, M. C. Nawijn, Y. Li, W. Chen, L. J. Bont, K. Bønnelykke, J. C. Celedón, G. H. Koppelman, and C.-J. Xu, "Nasal dna methylation at three cpg sites predicts childhood allergic disease," *Nature Communications*, vol. 13, no. 1, p. 7415, 2022. [Online]. Available: https://doi.org/10.1038/s41467-022-35088-6

[4] M. Neidhart, *DNA Methylation and Complex Human Disease*, 1st ed. Elsevier, 2015.

[5] T. J. Peters, M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, R. V. Lord, S. J. Clark, and P. L. Molloy, "De novo identification of differentially methylated regions in the human genome," *Epigenetics & Chromatin*, vol. 8, no. 1, p. 6, 2015. [Online]. Available: https://doi.org/10.1186/1756-8935-8-6

[6] E. Gatev, N. Gladish, S. Mostafavi, and M. S. Kobor, "CoMeBack: DNA methylation array data analysis for co-methylated regions," *Bioinformatics*, vol. 36, no. 9, pp. 2675–2683, 01 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa049

[7] X. Jiang, Z. Li, A. Mehmood, H. Wang, Q. Wang, Y. Chu, X. Mao, J. Zhao, M. Jiang, B. Zhao, G. Lin, E. Wang, and D. Wei, "A self-attention graph convolutional network for precision multi-tumor early diagnostics with dna methylation data," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 15, no. 3, pp. 405–418, September 2023. [Online]. Available: https://doi.org/10.1007/s12539-023-00563-1

[8] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008. [Online]. Available: https://doi.org/10.1186/1471-2105-9-559

[9] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, oct 2019. [Online]. Available: https://doi.org/10.1145/3326362

[10] J. N.-G. P. F. A. Toraño E. Garcia, M. Fernández-Morera, "The impact of external factors on the epigenome: In utero and over lifetime," *BioMed Research International*, vol. 2016, 2016.

[11] E. Forno, T. Wang, C. Qi, Q. Yan, C.-J. Xu, N. Boutaoui, Y.-Y. Han, and et al., "Dna methylation in nasal epithelium, atopy, and atopic asthma in children: a genome-wide study," *The Lancet Respiratory Medicine*, vol. 7, no. 4, pp. 336–346, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213260018304661

[12] N. S. Eissa, U. Khairuddin, and R. Yusof, "A hybrid metaheuristic-deep learning technique for the pan-classification of cancer based on dna methylation," *BMC Bioinformatics*, vol. 23, no. 1, p. 273, 2022. [Online]. Available: https://doi.org/10.1186/s12859-022-04815-7

[13] X. Cai, J. Tao, S. Wang, Z. Wang, J. Wang, M. Li, H. Wang, X. Tu, H. Yang, J.-B. Fan, and H. Ji, "Noninvasive lung cancer early detection via deep methylation representation learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 11 828–11 836, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21439

[14] S. Wei, J. Tao, J. Xu, X. Chen, Z. Wang, N. Zhang, L. Zuo, Z. Jia, H. Chen, H. Sun, Y. Yan, M. Zhang, H. Lv, F. Kong, L. Duan, Y. Ma, M. Liao, L. Xu, R. Feng, G. Liu, T. E. Project, and Y. Jiang, "Ten years of ewas," *Advanced Science*, vol. 8, no. 20, p. 2100727, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.202100727

[15] C. Wu, X. Mou, and H. Zhang, "Gbdmr: identifying differentially methylated cpg regions in the human genome via generalized beta regressions," *BMC Bioinformatics*, vol. 25, no. 1, p. 97, 2024. [Online]. Available: https://doi.org/10.1186/s12859-024-05711-y

[16] S. Li, F. E. Garrett-Bakelman, A. Akalin, P. Zumbo, R. Levine, B. L. To, I. D. Lewis, A. L. Brown, R. J. D'Andrea, A. Melnick, and C. E. Mason, "An optimized algorithm for detecting and annotating regional differential methylation," *BMC Bioinformatics*, vol. 14, no. 5, p. S10, 2013. [Online]. Available: https://doi.org/10.1186/1471-2105-14-S5-S10

[17] S. Maas, A. Vidaki, A. Teumer, R. Costeira, R. Wilson, and J. e. a. van Dongen, "Validating biomarkers and models for epigenetic inference of alcohol consumption from blood," *Clinical Epigenetics*, vol. 13, no. 1, p. 198, 2021.

[18] Y. Cheng, C. Gieger, A. Campbell, A. M. McIntosh, M. Waldenberger, D. L. McCartney, R. E. Marioni, and C. A. Vallejos, "Feature pre-selection for the development of epigenetic biomarkers," *medRxiv*, 2024. [Online]. Available: https://www.medrxiv.org/content/early/2024/02/15/2024.02.14.24302694

[19] D. Li, Z. Xie, M. L. Pape, and T. Dye, "An evaluation of statistical methods for dna methylation microarray data analysis," *BMC Bioinformatics*, vol. 16, no. 1, p. 217, July 2015. [Online]. Available: https://doi.org/10.1186/s12859-015-0641-x

[20] L. M. Butcher and S. Beck, "Probe lasso: A novel method to rope in differentially methylated regions with 450k dna methylation data," *Methods*, vol. 72, pp. 21–28, 2015, (Epi)Genomics approaches and their applications. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1046202314003697

[21] L. Gomez, G. J. Odom, J. I. Young, E. R. Martin, L. Liu, X. Chen, A. J. Griswold, Z. Gao, L. Zhang, and L. Wang, "comethdmr: accurate identification of co-methylated and differentially methylated regions in epigenome-wide association studies with continuous phenotypes," *Nucleic Acids Research*, vol. 47, no. 17, p. e98, September 2019.

[22] A. Jajoo, O. Hirschi, K. Schulze, Y. Guan, and N. A. Hanchard, "A first-generation genome-wide map of

correlated dna methylation demonstrates highly coordinated and tissue-independent clustering across regulatory regions," *Research Square [Preprint]*, pp. rs.3.rs–2 852 818, 2023.

[23] Z. Si, H. Yu, and Z. Ma, "Learning deep features for dna methylation data analysis," *IEEE Access*, vol. 4, pp. 2732–2737, 2016.

[24] S. Katz, V. A. Martins dos Santos, E. Saccenti, and G. V. Roshchupkin, "methae: an explainable autoencoder for methylation data," *bioRxiv*, 2024. [Online]. Available: https://www.biorxiv.org/content/early/2024/01/19/2023.07.18.549496

[25] C. Doersch, "Tutorial on variational autoencoders," 2016, august 16, 2016, with very minor revisions on January 3, 2021. [Online]. Available: https://arxiv.org/abs/1606.05908

[26] J. J. Levy, A. J. Titus, C. L. Petersen, Y. Chen, L. A. Salas, and B. C. Christensen, "Methylnet: an automated and modular deep learning approach for dna methylation analysis," *BMC Bioinformatics*, vol. 21, no. 1, p. 108, 2020. [Online]. Available: https://doi.org/10.1186/s12859-020-3443-8

[27] N. A. Valous, F. Popp, I. Zörnig, D. Jäger, and P. Charoentong, "Graph machine learning for integrated multi-omics analysis," *British Journal of Cancer*, 2024, published on 2024/05/10. [Online]. Available: https://doi.org/10.1038/s41416-024-02706-7

[28] O. A. Montesinos-López, G. I. H. Prado, J. C. Montesinos-López, A. Montesinos-López, and J. Crossa, "A graph model for genomic prediction in the context of a linear mixed model framework," *The Plant Genome*, vol. 17, no. 4, p. e20522, 2024. [Online]. Available: https://doi.org/10.1002/tpg2.20522

[29] F. Xia, K. Sun, S. Yu, A. Aziz, L. Wan, S. Pan, and H. Liu, "Graph learning: A survey," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 109–127, 2021.

[30] X. Fan, M. Gong, Y. Wu, Z. Tang, and J. Liu, "Neural gaussian similarity modeling for differential graph structure learning," 2023. [Online]. Available: https://arxiv.org/abs/2312.09498

[31] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 701–710. [Online]. Available: https://doi.org/10.1145/2623330.2623732

[32] G. Li, J. Luo, Q. Xiao, C. Liang, P. Ding, and B. Cao, "Predicting microrna-disease associations using network topological similarity based on deepwalk," *IEEE Access*, vol. 5, pp. 24 032–24 039, 2017.

[33] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," 2019. [Online]. Available: https://arxiv.org/abs/1903.03894

[34] M. N. Vu and M. T. Thai, "Pgm-explainer: Probabilistic graphical model explanations for graph neural networks," 2020. [Online]. Available: https://arxiv.org/abs/2010.05788

[35] B. L. Tremblay, F. Guénard, B. Lamarche, L. Pérusse, and M.-C. Vohl, "Network analysis of the potential role of dna methylation in the relationship between plasma carotenoids and lipid profile," *Nutrients*, vol. 11, no. 6, p. 1265, 2019, epub 2019 Jun 4. [Online]. Available: https://doi.org/10.3390/nu11061265

[36] J. Li, D. Zhou, W. Qiu, Y. Shi, J.-J. Yang, S. Chen, Q. Wang, and H. Pan, "Application of weighted gene co-expression network analysis for data from paired design," *Scientific Reports*, vol. 8, no. 1, p. 622, 2018. [Online]. Available: https://doi.org/10.1038/s41598-017-18705-z

[37] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. Pt 21, pp. 4947–4957, 2005. [Online]. Available: https://doi.org/10.1242/jcs.02714

[38] M. Pinart, M. Benet, I. Annesi-Maesano, A. von Berg, D. Berdel, and K. C. L. e. a. Carlsen, "Comorbidity of eczema, rhinitis, and asthma in ige-sensitised and non-ige-sensitised children in medall: a population-based cohort study," *The Lancet Respiratory Medicine*, vol. 2, no. 2, pp. 131–140, Feb. 2014, epub 2014-01-14.

[39] J. L. Min, G. Hemani, G. Davey Smith, C. Relton, and M. Suderman, "Meffil: efficient normalization and analysis of very large dna methylation datasets," *Bioinformatics*, vol. 34, no. 23, pp. 3983–3989, Dec. 2018.

[40] J.-P. Fortin and K. D. Hansen, "Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data," *Genome Biology*, vol. 16, no. 1, p. 180, 2015.

Supplementary Information

Appendix A: Additional Results for the Final Model

Convergence of the Model

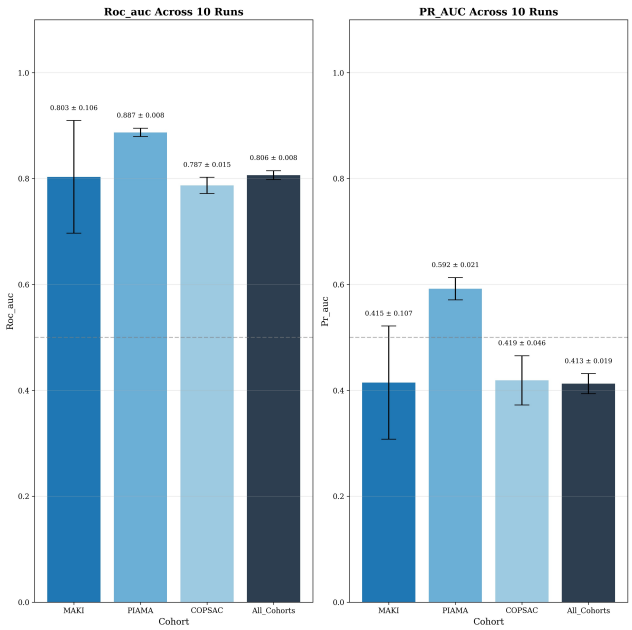

Figure 1: ROC-AUC and PR-AUC across 10 runs of the WGCNA/EdgeConv model, with the addition of error bars to show the variation in the model and thus its ability to converge. Only MAKI shows significant variation, in all likelihood caused by its small cohort size. These results suggest that EdgeConv consistently learns across datasets and provides reliable predictions.

ROC perfomance of the Model

Throughout this study we have used precision-recall to evaluate the model as that is generally less affected by the class imbalance inherent in this dataset. However, ROC is still very commonly used in methylation research therefore we also decided to include this in the supplement.

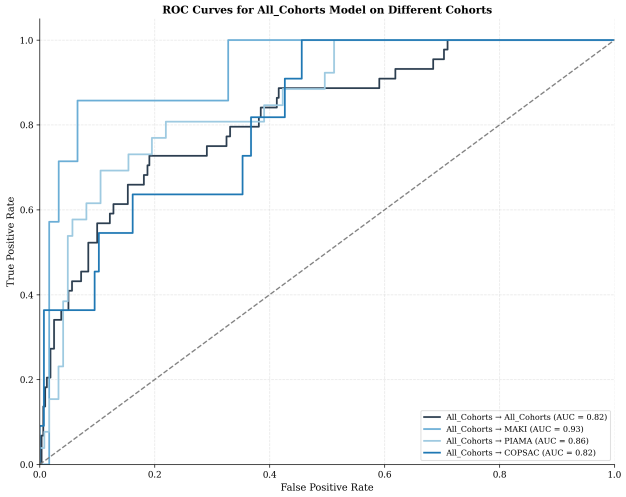

Figure 2: ROC curves for MAKI, PIAMA, COPSAC and the entire dataset for the WGCNA/EdgeConv model showing consistently high performance with COPSAC performing the worst

Appendix B: Performance in a model with eczema

The complete dataset, including the eczema data, was eventually shared for this research however this was very shortly before the completion of this research, so while tests were performed on this data, the model was not optimized to account for the difference, but the results can be found below. There is no significant difference, but optimization of the model to the new situation would likely improve results, as the class imbalance in the new situation is much less extreme.

Study	**PIAMA**	**MAKI**	**COPSAC**
van Breugel *et al*	0.50	0.35	0.37
ElasticNet	0.48	0.36	0.51
XGBoost	0.51	0.57	0.49
EdgeConv	0.59	0.57	0.42

Table 1: PR-AUC comparison to van Breugel et al. across cohorts. They externally validated on MAKI and COPSAC and did not include it in the train data.

Appendix C: SAGCN Architecture

A Self-Attention Graph Convolutional Network (SAGCN) was initially tested as part of this research due to its use in the paper by Jiang et al [1] , which represents the only other use of graph learning for methylation data. Its performance was however found to be unstable and significantly lower than that of the EdgeConv model. Due to its poor generalization across cohorts and high variance between runs, it was ultimately excluded from the final methodology. The Jiang et al. study was a cancer study which had many more samples which is a potential reason that their model was able to converge.

SAGCN Implementation and Evaluation

SAGCN was implemented using a standard two-layer Graph Convolutional Network (GCN) with `SAGPooling`. This approach introduces a learnable self-attention mechanism that prunes the graph hierarchy by retaining only the most important nodes, dynamically adapting the structure during training.

[1] X. Jiang, Z. Li, A. Mehmood, H. Wang, Q. Wang, Y. Chu, X. Mao, J. Zhao, M. Jiang, B. Zhao, G. Lin, E. Wang, and D. Wei, "A Self-attention Graph Convolutional Network for Precision Multi-tumor Early Diagnostics with DNA Methylation Data," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 15, no. 3, pp. 405–418, Sep. 2023. doi: 10.1007/s12539-023-00563-1.

- **Architecture:** The model comprised two GCN layers interleaved with self-attention pooling, followed by a multi-layer perceptron (MLP) classification head.
- **Input:** Subject-specific graphs constructed using WGCNA, where nodes represent CpG sites and edges encode co-methylation patterns.
- **Loss Function:** Focal loss and contrastive learning were used to address class imbalance, and early stopping based on validation AUC was applied.
- **Training Setup:** Models were trained using a cosine-annealing learning rate scheduler, with 10 repeated runs to account for initialization variability.

Performance and Exclusion Justification

Despite its conceptual appeal, and previous success, SAGCN exhibited poor and unstable performance across runs, especially on the COPSAC cohort. Compared to EdgeConv, the SAGCN model showed:

- Lower average ROC-AUC and PR-AUC across all cohorts.
- Higher variance in results
- A tendency to overfit cohort-specific noise due to the learnable adjacency mechanism.

These limitations were most pronounced in cross-cohort generalization, where SAGCN failed to learn robust, transferable representations. Figure 3 illustrates the inconsistency and performance degradation due to the models inability to converge. Based on these findings, SAGCN was not used in the final analysis pipeline.

Cohort	ROC–AUC ($\mu \pm \sigma$)	PR–AUC ($\mu \pm \sigma$)
MAKI	0.70 ± 0.05	0.30 ± 0.06
PIAMA	0.71 ± 0.05	0.28 ± 0.06
COPSAC	0.58 ± 0.07	0.12 ± 0.08

Table 2: Performance metrics (mean ± standard deviation) for SAGCN across cohorts.

Compared to Elastic Net (ROC–AUC: 0.89, PR–AUC: 0.50) and EdgeConv (ROC–AUC: 0.82, PR–AUC: 0.48), SAGCN showed up to 30% lower performance in precision-recall space. The overall SAGCN performance is summarized in table 2.

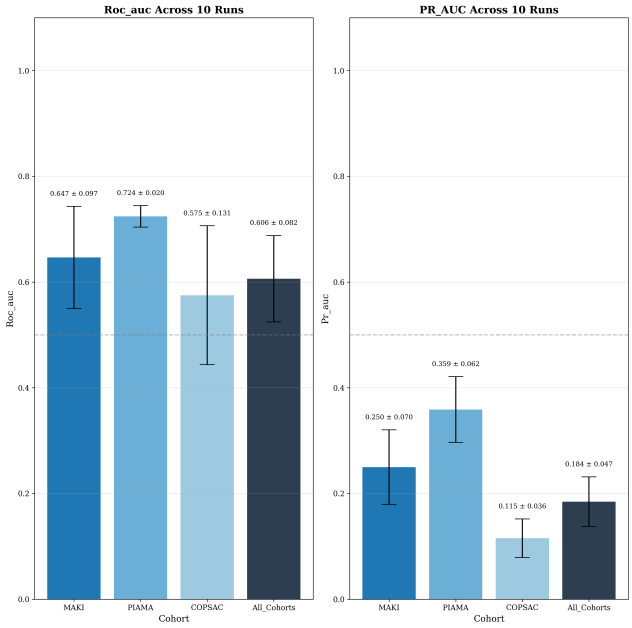

Figure 3: Mean performance over 10 runs for the WGCNA/SAGCN model, error bars repsenting the variation across runs. Model performance for this model was much poorer than the performance seen in EdgeConv

Appendix D: Addition of Genomic Edge Typing

Given the well-established influence of genomic context on methylation function [2] [3] —such as proximity to promoters, gene bodies, and chromosomal co-localization—we incorporated genomic edge typing as a biologically motivated extension to the graph construction process. The goal was to enhance model performance and interpretability by categorizing CpG–CpG edges based on their underlying genomic relationships. Although ultimately this edge typing did not yield measurable improvements in predictive accuracy, and in fact decreased the performance of the model, it was still valuable to test as an addition to the model architecture.

Categorisation Pipeline

The function `add_genomic_edge_types` assigned a five-level categorical label to each CpG–CpG edge based on gene annotations and chromosomal position. The classification pipeline proceeds as follows:

(a) **Annotation lookup:** Chromosome, genomic coordinate, and gene identifiers are retrieved from provided CpG annotation files.

(b) **Same-gene test:** Two CpGs are flagged as belonging to the same gene if they share at least one gene symbol (including overlapping genes).

(c) **Distance test:** Genomic distance is computed for CpGs located on the same chromosome.

(d) **Edge-type assignment:** Each edge is assigned a type (0–4) based on a rule-based decision tree described in Table 3.

[2] Pierce BL et al. (2018). Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nature Communications*, 9, 804. doi:10.1038/s41467-018-03209-9

[3] Zhang X-M et al. (2021). Graph Neural Networks and Their Current Applications in Bioinformatics. *Frontiers in Genetics*, 12, 690049. doi:10.3389/fgene.2021.690049

Edge Type Definitions

Code	Label	Rule	Distance
0	Same gene, proximal	CpGs in the same gene with genomic distance < 5 kb	<5 kb
1	Same gene, distal	CpGs in the same gene but ≥ 5 kb apart	≥ 5 kb
2	Cis-regulatory*	CpGs on the same chromosome, < 50 kb apart, with one in a promoter/TSS region and the other in a gene body or 5'UTR	<50 kb
3	Same chromosome	All other intra-chromosomal pairs	≥ 50 kb
4	Inter-chromosomal	CpGs located on different chromosomes	n/a

Table 3: Edge types and corresponding genomic rules.

* Putative cis-regulatory links where one site is near a transcription start site and the other is in a transcribed region.

Edge Typing Integration into Graph Construction

Genomic edge types were integrated during and after graph assembly via the `create_genomic_aware_graph` function. This graph construction process involved three primary edge classes:

- **Local edges:** High-confidence links with strong methylation similarity (> 0.95), retained without modification.
- **Global edges:** Moderate-similarity links above a relaxed threshold (> 0.8), prioritized if supported by genomic proximity or shared annotation.
- **Genomic edges:** Edges with weaker methylation similarity but retained due to strong genomic context—such as proximity within 100 kb, co-membership in a gene, or putative regulatory roles.

To integrate genomic information into edge inclusion and weighting, the methylation similarity score between each CpG pair was scaled by a multiplicative *genomic boost factor*. This factor was additive and based on biologically motivated criteria:

- +0.3 boost if CpGs belonged to the same gene.
- +0.2 if the edge linked promoter/TSS regions to gene bodies or 5'UTRs (putative cis-regulatory interaction).
- +0.1 to +0.4 based on physical distance (e.g., < 5 kb received the highest boost).

This adjustment allowed edges that were genomically meaningful—but might fall below strict similarity thresholds—to be retained in the final graph. The adjusted similarity score (`combined_similarity`) was then used to determine whether an edge qualified as local, global, or genomic.

Once edges were selected, they were labeled with a categorical genomic edge type using `add_genomic_edge_types`, based on gene co-membership, chromosomal identity, and physical distance. While these edge types were not explicitly used in message passing or model parameters (e.g., via edge-type-specific kernels), they were preserved for downstream stratified analysis and interpretability.

Performance of EdgeConv Model with Edge Typing

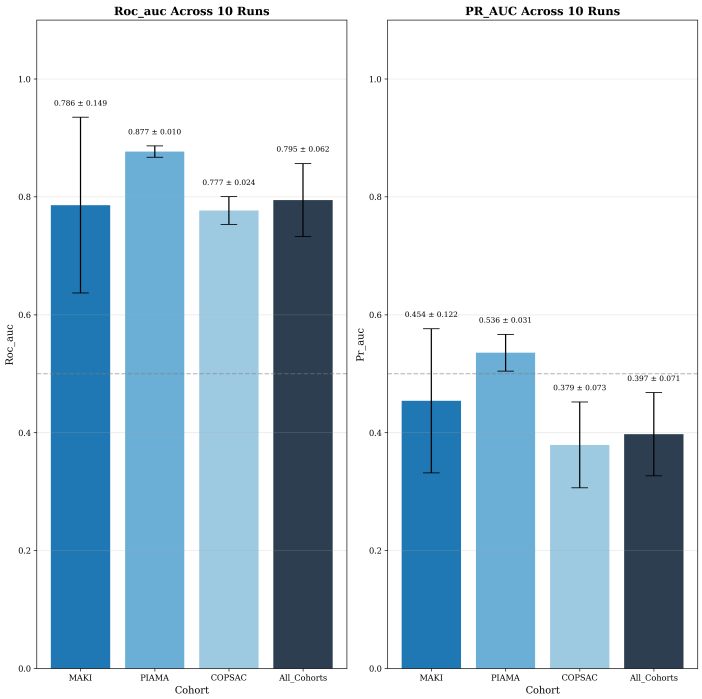

Figure 4: Mean performance over 10 runs for the WGCNA/EdgeConv model with Edge Typing, error bars repsenting the variation across runs. Model performance with this inclusion decreased, and results were less likely to converge

With the addition of genomic edge categories, the performance of the model did not differ significantly from the baseline using methylation similarity alone. These results suggest that although the edge types are biologically reasonable, they did not offer a meaningful inductive bias under the current model structure.

Appendix E: Additional Feature Extraction Information

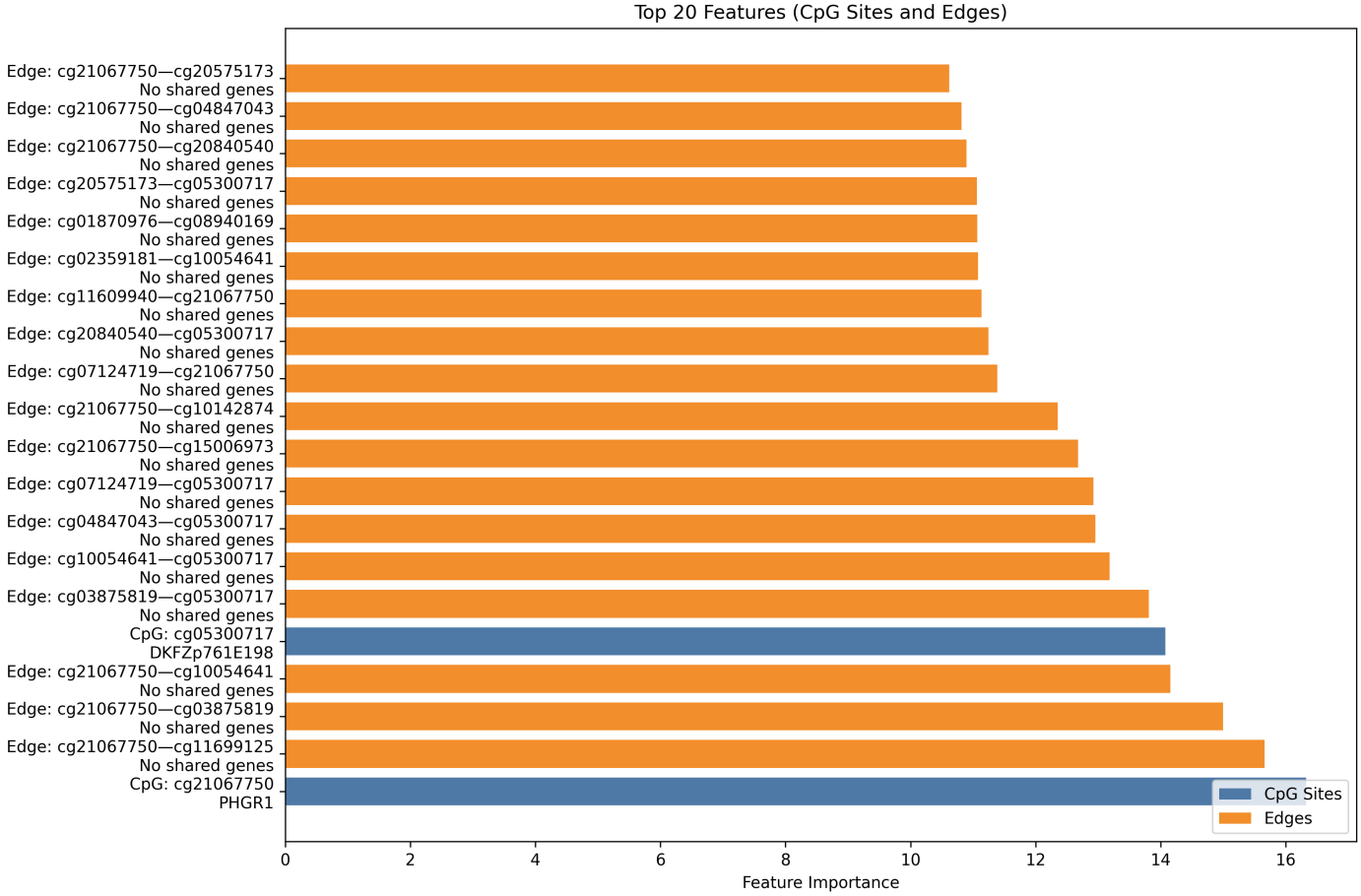

Figure 5: Ranking of the relative importance of the features of the EdgeConv model, both for the nodes as well as the weighted edges. Created using a gradient-saliency explainer.