Interpretable Graph Learning on Irregular Clinical Time Series

Andrea Zerio[2] **Maya Bechler-Speicher**[1] **Maor Huri**[7]
Marie Vibeke Vestergaard[2] **Ran Gilad-Bachrach**[6]
Tine Jess[2,3] **Samir Bhatt**[4,5] **Aleksejs Sazonovs**[2]
[1]Meta
[2]Center of Excellence for Molecular Prediction of IBD, PREDICT
Department of Clinical Medicine, Aalborg University
[3]Department of Gastroenterology & Hepatology, Aalborg University Hospital
[4]University of Copenhagen [5]Imperial College London
[6]Department of Bio-Medical Engineering, Tel-Aviv University
[7]Sagol School of Neuroscience, Tel-Aviv University

Abstract

Real-world medical data often include measurements collected at irregular, asynchronous intervals across different tests and frequencies. For example, routine blood tests are taken at different times, producing fragmented, unevenly sampled temporal data. Effectively learning from such data requires models that can handle temporally sparse, heterogeneous measurements. In this paper, we propose Graph Mixing Additive Networks (GMAN), an interpretable-by-design model for learning from irregular temporal measurements. Our method achieves strong performance on a real-world medical task: predicting the onset of Crohn's disease (CD) from routine biomarkers, comparable to black-box models, both graph and sequential architectures. Its interpretable design also yields clinically meaningful insights, such as dysregulation patterns in the pre-diagnostic phase of complex diseases.

1 Introduction

Modern clinical data consist of diverse types of measurements, often collected at irregular and asynchronous time intervals. For example, a patient's medical record often includes a set of blood tests taken over their lifetime, where each type of test is performed at its own frequency. As a result, the data can be effectively viewed as a set of sparse temporal trajectories. To learn predictive tasks over such data, conventional approaches often rely on imputation [4, 11, 14, 6] to flatten or regularize temporally-irregular data into fixed-size representations. A common strategy is to align observations to a uniform time grid and fill in missing values using interpolation or learned imputation models. Such preprocessing imposes constraints that can potentially lose informative content, risks losing the conditional dependence structure of previous events, and ignore the natural set-like structure of the data, where each example contains a collection of related but non-synchronized temporal sequences. In this paper, we introduce Graph Mixing Additive Networks (GMAN), a novel framework designed to learn directly from sets of sparse, irregular temporal trajectories. We handle irregular sampling by representing each trajectory as a directed path graph and modelling the entire dataset as a set of such graphs.

Our approach extends Graph Neural Additive Networks (GNAN) [2], which are interpretable by design but constrained by their restriction on non-linear feature interactions, limiting expressive power. In contrast, GMAN offers a flexible trade-off between interpretability and expressive power,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

overcoming GNAN's constraints. Additionally, while GNAN operates only on single graphs, GMAN supports learning from sets of graphs, enabling further interpretability through analysis of the relative importance of individual graphs within the set. We demonstrate the effectiveness of GMAN on a real-world medical task: predicting the onset of CD from routine biomarkers. We show that GMAN achieves strong performance while also providing valuable clinical and biological insights through its interpretability features. Finally, we provide a theoretical analysis proving that GMAN is strictly more expressive than GNAN, and that the proposed mechanism for balancing interpretability and expressivity leads to strictly higher expressivity compared to models that enforce full interpretability.

2 Graph Mixing Additive Networks

Let $S = \{G_1, \ldots, G_m\}$ be a *set* of m graphs. Each node v is associated with a feature vector $x_v \in \mathbb{R}^d$ and a time-stamp t_v. In the current case of a patient's blood test data, each graph corresponds to a specific biomarker, and each node within the graph represents an individual measurement of that biomarker, annotated with its feature vector and time of collection. We denote the entry c of a vector $\mathbf{h}$ by $[\mathbf{h}]_c$, and the set of entries corresponding to a set of features S by $[\mathbf{h}]_S$. We assume that the graphs in S are partitioned into $k, 1 \leq k \leq m$ disjoint subsets S_k such that $\bigcup_{i=1}^{k} S_i = S$.

The partition $\{S_i\}_{i=1}^{k}$ provides a flexible trade-off between expressivity and interpretability. GMAN linearly aggregates representations of the subsets of S to form a final set representation, and then assigns a single label to S. The level of interpretability that GMAN provides for each graph depends on the size of the subset it belongs to. When a subset contains a single graph, GMAN offers fine-grained, node-level importance scores. In contrast, for larger subsets, it provides only subset-level importance scores, trading interpretability for improved expressivity.

First, GMAN applies a function Φ_i to each subset S_i to obtain a representation of the subset S_i, denoted as $\mathbf{h}_i \in \mathbb{R}^d$.

$$\mathbf{h}_i = \Phi_i(S_i),$$

Then, it produces a representation for the whole set, $\mathbf{h}_S$ by summing the subsets' $\mathbf{h}_S = \sum_{i=1}^{k} h_i$. Finally, to produce the label, it sums over the d entries of $\mathbf{h}_S$. Overall:

$$\text{GMAN}(S) = \sum_{c=1}^{d} \sum_{i=1}^{k} [\Phi_i(S_i)]_c \tag{1}$$

Where $\Phi_i(S_i) = \mathbf{h}_{S_i}$ is a representation of the subset S_i.

For subsets of size one, $\Phi_i(S_i)$ applies an ExtGNAN, as described in Section 2.1. For subsets containing multiple graphs, a ExtGNAN is applied to each graph, followed by a DeepSet aggregation [17] over the resulting vectors. Importantly, each subset is assigned its own ExtGNAN, and all graphs within a subset share the same one. A DeepSet first applies a neural network $f : \mathbb{R}^d \to \mathbb{R}^d$ for each vector in the set $\{h_l\}_{G_l \in S_i}$, sums the results, and then applies another neural network $g : \mathbb{R}^d \to \mathbb{R}^d$.

$$g\left(\sum_{i \in S_2} f(h_i)\right)$$

Here, g and f are neural networks of arbitrary depth and width. We now turn to define ExtGNAN.

2.1 ExtGNAN

In GNAN, univariate neural networks are applied to each feature of each node in isolation, to learn a representation for a graph. This has the benefit of generating interpretable models as features do not mix non-linearly. Nonetheless, when interactions between features are crucial for the task, it may result in sub-par performance. Therefore, ExtGNAN extends GNAN by allowing multivariate neural networks to operate on groups of features to gain accuracy at the cost of reducing the interpretability of the model.

Assume that the features are partitioned into K subsets $\{F_l\}_{l=1}^{K}$. For any subset of features greater than one, ExtGNAN applies a multivariate neural network for all the features in the subset together, instead of a univariate neural network for each one separately. To learn a representation of a graph G, ExtGNAN first computes representations for the nodes of G as follows. ExtGNAN learns a distance function $\rho(x;\theta) : \mathbb{R} \to \mathbb{R}$ and a set of feature shape functions $\{\psi_l\}_{l=1}^{K}, \psi_l(X;\theta_k) : \mathbb{R}^{|F_l|} \to \mathbb{R}^{|F_l|}$. Each of these functions is a neural network of arbitrary depth. For brevity, we omit the parameterization θ and θ_k for the remainder of this section. The entries of the representation of node j corresponding to the indices of the features in F_l, denoted as $[\mathbf{h}_j]_{F_l}$, is computed by summing the contributions of the features in the subset F_l from all nodes in the graph:

$$[\mathbf{h}_j]_{F_l} = \sum_{w \in V} \rho\left(\Delta(w,j)\right) \cdot \psi_l\left([\mathbf{X}_w]_{F_l}\right),$$

where $\Delta(w,j) = t_w - t_j$ and $[\mathbf{X}_w]_{F_l}$ are the features of node w corresponding to the subset F_l. Overall, the full representation of node j can be written as:

$$\mathbf{h}_j = \big([\mathbf{h}_j]_{F_1}, [\mathbf{h}_j]_{F_2}, \ldots, [\mathbf{h}_j]_{F_K}\big).$$

Then ExtGNAN produces a graph representation by summing the node representations,

$$\mathbf{h}_G = \sum_{i \in V} \mathbf{h}_i. \tag{2}$$

Further details on how node- and feature-level interpretability are achieved are included in the Appendix

2.2 Expressivity properties

In this section we provide a theoretical analysis of the expressiveness of GMAN. Proofs are included in the Appendix.

Theorem 2.1. GMAN *is strictly more expressive than GNAN.*

Theorem 2.2. *Let S be a set of graphs $\{G_i\}_{j=1}^{m}$. Let $S_1 = \{S_i\}_{i=1}^{m}$ be a partition of S such that $|S_i| = 1$. Let $S_2 = \{S_i\}_{i=1}^{k}$ such that there exists k with $|S_k| > 1$. with a subset partition $\{S_i\}_{i=1}^{k}$. Then a* GMAN *trained over S_2 is strictly more expressive than a* GMAN *trained over S_1.*

3 Empirical evaluation

We evaluate GMAN on a real-world clinical task: predicting Crohn's Disease (CD) onset from 17 routinely measured blood and stool biomarkers. Our cohort is derived from the Danish health registries and the Registry of Laboratory Results for Research (RLRR), which records routine laboratory tests from hospitals and general practitioners since 2015 [9, 1]. We identify 8,567 individuals with incident CD (first recorded CD diagnosis) and treat them as cases. To form a comparable reference group, we randomly sample individuals from the same registries and downsample by age to mirror the expected prediagnostic testing frequency and the typical young-adult onset of CD; from this pool we select 8,567 age-matched controls, yielding two balanced classes.

Table 1: Preliminary results on CD onset prediction.

Methods	**AUROC**
Transformer	66.17 ± 0.3
Trans-mean	70.78 ± 2.4
SeFT	73.27 ± 1.70
MTGNN	76.47 ± 0.77
DGM^2-O	82.57 ± 0.90
GRU-D	82.29 ± 0.95
Raindrop	80.62 ± 0.98
Graph Transformer	81.21 ± 1.10
GraphSAGE	75.67 ± 4.91
GATv2	78.43 ± 4.88
GMAN	**83.64 ± 0.9**

Each person is modelled as a set of time-stamped biomarker trajectories, where each trajectory is a directed path graph with nodes as test results and edges encoding elapsed time; this preserves the original irregular sampling of the data, while enabling joint reasoning across biomarkers. Full details are included in the Appendix.

We group the biomarkers into seven physiology-based subsets reflecting distinct clinical processes: (i) white blood cell subtypes capture cellular immune activation patterns; (ii) inflammation markers summarize systemic and gut-specific inflammatory burden; (iii) platelets provide an independent hematologic signal of chronic inflammation; (iv) hemoglobin/anemia tracks pre-diagnostic anemia dynamics, which are common in CD due to chronic intestinal inflammation; (v) iron status reflects iron metabolism and malabsorption that often precede or accompany CD-related anemia; (vi) vitamin/folate status captures joint nutrient-deficiency signatures linked to intestinal malabsorption and disease location; and (vii) liver function markers reflect hepatic response and systemic effects of chronic inflammation. We train GMAN on this partition to distinguish incident CD from controls. Full subset definitions, including biomarker membership are provided in the Appendix.

The subset partition is a hyperparameter of GMAN: finer partitions (up to single-biomarker subsets) maximize interpretability, while coarser physiologically motivated subsets allow richer cross-biomarker interactions. In this work we use clinically motivated subsets to balance these two objectives. Conceptually, subset design plays a role similar to structured feature selection: it specifies which biomarkers should be interpreted independently versus as a coupled signal. In many settings, grouping is clinically sensible because key biomarkers are routinely interpreted through their non-linear combinations rather than in isolation.

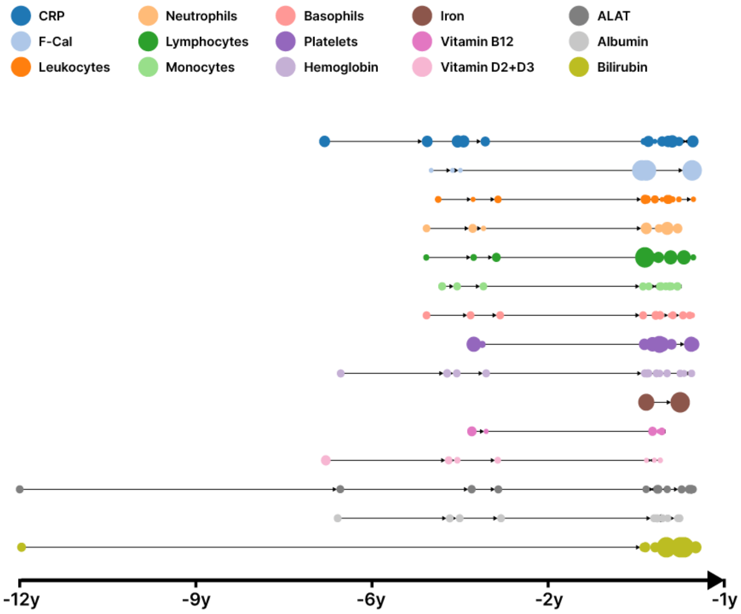
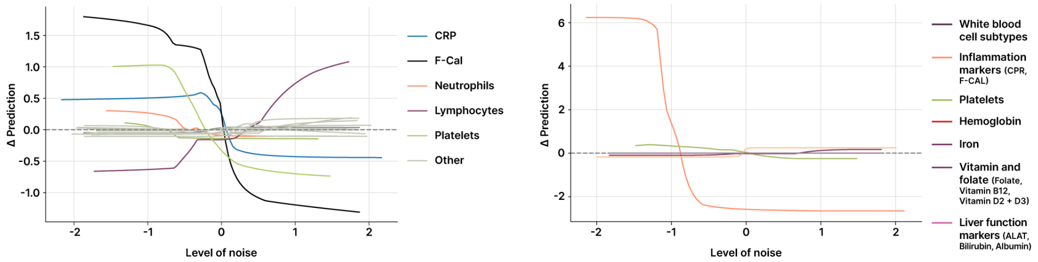


Figure 1: Node-level attributions for an individual CD patient. Node size indicates each measurement's contribution to the prediction. The time dimension represents distance to the diagnosis date. Data are privacy-protected via noise and temporal shifting.

For example, clinically established composites such as the neutrophil-to-lymphocyte ratio (NLR) reflect a non-additive immune-balance effect, and the albumin–bilirubin score (ALBI) combines albumin and bilirubin through a non-linear, log-weighted relationship rather than independent additive contributions. Grouping such biomarkers therefore encourages GMAN to model these clinically meaningful interaction patterns directly, while singleton subsets retain fully granular node-level attributions when that level of detail is required. Importantly, GMAN does not require a priori medical groupings: when domain knowledge is unavailable, one can default to singleton subsets for full interpretability or use data-driven group proposals (e.g., correlation-based clustering), which we leave to future work.

We compare GMAN against 10 other baselines [10, 7, 3, 18, 12, 8, 16, 15, 5] and report the results in Table 1. Hyperparameters were tuned via grid search: models were trained on the training split and selected by validation AUROC. We fixed the best configuration and report test AUROC as mean ± std over three random seeds. The complete subsets information is provided in the Appendix, alongside the full experimental setup.



(a) Single-biomarker groups GMAN

(b) Best-performing GMAN groups

Figure 2: Subset-level contribution curves for CD prediction. Each curve shows how the model's output changes as increasing noise is added to the latent representation of a biomarker group. (a) considers individual biomarkers, showing F-Cal, platelets, and CRP increase risk, whereas lymphocytes exhibit an opposing (protective) effect. (b) uses physiologically coherent groups, where inflammation markers (CRP, F-Cal) dominate the prediction.

Beyond predictive performance, a central strength of GMAN is that interpretability is built into the model by design, rather than added post-hoc. This allows for fine-grained, temporally resolved explanations, enabling users to understand how and when specific biomarkers influence the model's predictions. In high-stakes domains such as healthcare, this level of transparency is crucial. Clinical decision-making often depends not only on the accuracy of a prediction but on a clear understanding of the reasoning behind it.

Models that can provide such insights are far more likely to be trusted, audited, and integrated into clinical workflows. To showcase GMAN's fine-grained interpretability we compute both node-level and subset-level attributions. Figure 1 visualizes a CD trajectory set from a single patient, with node size encoding each measurement's effect on the prediction. GMAN correctly recognises inflammation and immune markers as dominant drivers as well as prioritising nodes closer to the diagnosis date, consistent with prior evidence [13]. GMAN can also assign *subset*-level importance via Equation (4), quantifying contributions of user-specified biomarker groups. For each grouping strategy, we train a separate model and estimate group effects by structured perturbations: within each subset we compute a PCA over its node features and progressively perturb inputs along the first principal component, recording the induced change in prediction (i.e., change in importance). Figure 2 summarizes these results; full group definitions are detailed in the Appendix.

4 Conclusion

In this extended abstract, we present GMAN, a work-in-progress framework for learning from sets of temporally sparse graphs, balancing interpretability and expressivity. We apply it to incident CD prediction from routine blood tests, which are ubiquitous across primary, emergency, and inpatient care. These time-stamped tests snapshot inflammation, organ function, and nutrition and, despite irregular, indication-biased ordering, capture the molecular state at each visit. Modelling each biomarker as a directed path graph, GMAN achieves strong performance and yields additive, multi-resolution attributions at the level of individual measurements and biomarker groups, revealing patterns of pre-diagnostic dysregulation.

References

[1] Johan Frederik Håkonsen Arendt, Anette Tarp Hansen, Søren Andreas Ladefoged, Henrik Toft Sørensen, Lars Pedersen, and Kasper Adelborg. Existing data sources in clinical epidemiology: laboratory information system databases in denmark. *Clinical epidemiology*, pages 469–475, 2020.

[2] Maya Bechler-Speicher, Amir Globerson, and Ran Gilad-Bachrach. The intelligible and effective graph neural additive network. *Advances in Neural Information Processing Systems*, 37:90552–90578, 2024.

[3] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[4] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

[5] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[6] Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

[7] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[8] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series, 2020. URL `https://arxiv.org/abs/1909.12064`.

[9] Carsten Bøcker Pedersen. The danish civil registration system. *Scandinavian journal of public health*, 39(7_suppl):22–25, 2011.

[10] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.

[11] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[13] Marie Vibeke Vestergaard, Kristine H Allin, Gry J Poulsen, James C Lee, and Tine Jess. Characterizing the pre-clinical phase of inflammatory bowel disease. *Cell Reports Medicine*, 4 (11), 2023.

[14] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[15] Yinjun Wu, Jingchao Ni, Wei Cheng, Bo Zong, Dongjin Song, Zhengzhang Chen, Yanchi Liu, Xuchao Zhang, Haifeng Chen, and Susan Davidson. Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series, 2021. URL `https://arxiv.org/abs/2103.02164`.

[16] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks, 2020. URL `https://arxiv.org/abs/2005.11650`.

[17] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets, 2018. URL `https://arxiv.org/abs/1703.06114`.

[18] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. *arXiv preprint arXiv:2110.05357*, 2021.

Appendix

Node, graph and set importance

GMAN retains all interpretability properties of GNAN [2], including feature-level and node-level importance. However, it extends beyond GNAN by operating on sets of graphs rather than single graphs, enabling additional forms of interpretability such as graph-level and subset-level importance. Because GMAN allows a flexible trade-off between interpretability and expressivity, permitting non-linear mixing within graph subsets, some adaptations are required to extract meaningful attributions under this more expressive regime. We can extract the total contribution of each node j to the prediction by summing the contributions of the node across all feature sets. This is only valid when the node belongs to a graph that is not combined non-linearly with other graphs, i.e., it belongs to a subset of size one.

Therefore, the contribution of node j is

$$\text{TotalContribution}(j) = \sum_{l=1}^{K} [\mathbf{h}_j]_{F_k} = \sum_{w \in V} \rho\left(\Delta(w, j)\right) \sum_{l=1}^{K} \psi_k\left([\mathbf{x}_w]_l, l \in F_k\right). \tag{3}$$

The contribution of a graph G is then

$$\text{TotalContribution}(G) = \sum_{v \in G} \text{TotalContribution}(v).$$

For graphs that are mixed non-linearly, i.e., graphs that belong in subsets of size greater than one, interpretability is more limited, and we can only provide the total contribution of the set to the final prediction

$$\text{TotalContribution}(S) = \sum_{l=1}^{K} [\mathbf{S}]_{F_k}. \tag{4}$$

Expressiveness

This section consolidates the theoretical foundations of GMAN by presenting formal proofs of its expressive power and structural recoverability guarantees. We first prove that GMAN is strictly more expressive than GNAN, both in terms of multivariate feature learning and subset level representation. We then establish conditions under which GMAN can recover the latent structure of input graphs purely from learned pairwise distances. Specifically, we show that when graphs form tree-structured trajectories, the model's learned distance matrix implicitly determines the original graph topology, up to isomorphism.

Proof of Theorem 3.1

We will prove that GMAN is strictly more expressive than GNAN. To prove this, we use a ground truth function that is a feature-level XOR. Let a single-node graph be endowed with binary features $x = (x_1, x_2) \in \{0, 1\}^2$ and define the target $f_\oplus(x) = x_1 \oplus x_2$.

First we will show that GNAN cannot express $f_\oplus$. A GNAN scores the graph by $\hat{y} = \sigma\big(\phi_1(x_1) + \phi_2(x_2)\big)$, where each ϕ_i is univariate. Put $a = \phi_1(0)$, $b = \phi_1(1)$, $c = \phi_2(0)$, $d = \phi_2(1)$. To match the XOR truth-table there must exist a threshold τ such that

$$a + c < \tau, \quad b + c > \tau, \quad a + d > \tau, \quad b + d < \tau.$$

Summing the first and last inequalities yields $a + b + c + d < 2\tau$, while the middle pair gives $a + b + c + d > 2\tau$—a contradiction. Thus no GNAN realises $f_\oplus$.

Now we will show that GMAN can express $f_\oplus$. Place the two features in the same subset $F = \{x_1, x_2\}$ and choose the subset-network

$$\psi_F(x_1, x_2) = x_1 + x_2 - 2x_1x_2.$$

For the four binary inputs this mapping returns $(0, 1, 1, 0)$, exactly $f_{\oplus}$. Hence GMAN represents a function unattainable by GNAN, proving that GMAN is strictly more expressive.

Proof of Theorem 3.2

Let S be a set of graphs $\{G_i\}_{j=1}^{m}$. Let $S_1 = \{S_i\}_{i=1}^{m}$ be a partition of S such that $|S_i| = 1$. Let $S_2 = \{S_i\}_{i=1}^{k}$ such that there exists k with $|S_k| > 1$. with a subset partition $\{S_i\}_{i=1}^{k}$. We will prove that a GMAN trained over S_2 is strictly more expressive than a GMAN trained over S_1.

To prove this, we use a ground truth function that is a set-level XOR. Let every graph G_i carry a single binary feature $x_i \in \{0, 1\}$ and let the $\mathsf{ExtGNAN}$ encoder return this feature unchanged, i.e. $h(G_i) = x_i$. Denote a set containing two graphs by $S = \{G_1, G_2\}$ and define the permutation-invariant target

$$f_{\oplus}(S) = x_1 \oplus x_2.$$

Singleton partition (S_1). If each graph is placed in its own subset, GMAN aggregates *additively*: the model output is

$$\hat{y} \;=\; \phi(x_1) + \phi(x_2),$$

because the final GMAN stage simply sums subset scores. Write $a = \phi(0)$ and $b = \phi(1)$. To realise $f_{\oplus}$ via a threshold τ we would need

$$a + a < \tau, \quad b + a > \tau, \quad a + b > \tau, \quad b + b < \tau.$$

Adding the first and last inequalities yields $a + b < \tau$, while the middle pair gives $a + b > \tau$—a contradiction. Hence GMAN_{S_1} cannot represent $f_{\oplus}$.

Paired partition (S_2). Group the two graphs together and use a DeepSet $\Phi(S_2) = g\big(\sum_{i=1}^{2} f(x_i)\big)$ with $f(x) = x$ and $g(s) = s(2 - s)$. Then

$$g\big(x_1 + x_2\big) = \begin{cases} 0 & (x_1, x_2) = (0,0) \text{ or } (1,1), \\ 1 & (x_1, x_2) = (0,1) \text{ or } (1,0), \end{cases}$$

exactly $f_{\oplus}$. The final GMAN sum over feature channels leaves this value unchanged, so GMAN_{S_2} realises $f_{\oplus}$.

Strict separation. Because $f_{\oplus}$ is representable by GMAN_{S_2} but not by GMAN_{S_1}, the former is strictly more expressive.

Dataset Details

The CD dataset is derived from the Danish health registries, a comprehensive, nationwide database covering healthcare interactions for over 9.5 million individuals [9]. A key resource is the Registry of Laboratory Results for Research (RLRR), which has collected laboratory test results from hospitals and general practitioners since 2015 [1]. From this data, we constructed a cohort of 8,567 individuals later diagnosed with Crohn's Disease (CD) and 8,567 age-matched controls from the same registry. For each person, we extracted temporal trajectories of 17 routinely measured biomarkers, reflecting key physiological processes such as immune response, inflammation, organ function, and nutritional status. The only exception is faecal calprotectin (F-Cal), a stool-based biomarker specifically used to detect intestinal inflammation in conditions like CD. The task is binary classification: predicting future CD onset from pre-diagnostic medical histories.

Here we detail the full list of the 17 biomarkers extracted from the Danish health registries.

1. C-reactive protein (CRP): A protein produced by the liver in response to inflammation. Elevated CRP indicates active inflammation, often associated with inflammatory diseases like CD.

2. Faecal Calprotectin (F-Cal): A protein released from neutrophils into the intestinal lumen, detectable in stool samples. Elevated levels indicate gastrointestinal inflammation and are commonly used to detect and monitor inflammatory bowel disease.

3. Leukocytes (White Blood Cells): Cells that are central to the body's immune response. Elevated leukocyte counts typically suggest infection or inflammation, including flare-ups in CD.

4. Neutrophils: A type of leukocyte involved, among other things, in fighting bacterial infections. High neutrophil counts often indicate acute inflammation or infection, including intestinal inflammation in CD.

5. Lymphocytes: A group of white blood cells that form the core of the adaptive immune system, including T cells, B cells, and natural killer (NK) cells. They are responsible for antigen-specific immune responses. Abnormal levels can signal immune dysregulation, often implicated in autoimmune and chronic inflammatory diseases such as CD.

6. Monocytes: A type of white blood cell that circulates in the blood and differentiates into macrophages or dendritic cells upon entering tissues. These cells are essential for phagocytosis, antigen presentation, and regulation of inflammation. Elevated levels may reflect immune activation or tissue damage.

7. Eosinophils: Immune cells involved primarily in allergic reactions and parasitic infections. Elevated eosinophil counts might reflect allergic responses or gastrointestinal inflammation.

8. Basophils: The least common type of leukocyte, involved in allergic and inflammatory responses. Their elevation is uncommon but may accompany certain inflammatory or allergic conditions.

9. Platelets: Cell fragments critical for blood clotting and also involved in inflammatory responses. High platelet counts (thrombocytosis) are commonly seen during active inflammation in conditions like CD.

10. Hemoglobin (Hb): The protein in red blood cells responsible for oxygen transport. Low hemoglobin (anemia) is frequently observed in chronic inflammatory conditions such as CD due to blood loss or nutrient deficiencies.

11. Iron: An essential mineral for red blood cell production. Low iron levels often indicate chronic blood loss or malabsorption, both common in CD due to intestinal inflammation.

12. Folate (Vitamin B9): A vitamin necessary for red blood cell production and DNA synthesis. Deficiency may result from impaired absorption in inflamed intestinal tissue.

13. Vitamin B12 (Cobalamin): Required for red blood cell production and neurological function. Deficiencies are common in CD, especially when the ileum is affected.

14. Vitamin D2+D3 (Ergocalciferol + Cholecalciferol): Vitamins essential for bone health and immune regulation. Low levels are often seen in CD due to malabsorption and systemic inflammation.

15. ALAT (Alanine Aminotransferase): An enzyme indicating liver function. Elevated levels may reflect liver inflammation, medication effects, or co-occurring autoimmune liver disease.

16. Albumin: A protein produced by the liver that helps maintain blood volume and transport nutrients. Low albumin can reflect chronic inflammation, malnutrition, or protein loss in CD.

17. Bilirubin: A compound produced from red blood cell breakdown. It is filtered by the liver and excreted into the intestine via bile. Elevated levels may indicate liver dysfunction, bile

duct obstruction, or hemolytic anemia.

Biomarker Subset Groupings

In the Crohn's Disease prediction task, we grouped the 17 selected biomarkers into 7 subsets based on shared physiological function, clinical relevance, and correlated patterns observed in exploratory analyses. This configuration produced the most robust and interpretable results, balancing domain knowledge with empirical performance. The grouping is as follows:

1. **White blood cell subtypes**
   [Leukocytes, Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils]
   These biomarkers all represent components of the immune system's cellular response. Grouping them enables the model to learn shared immune activation patterns, which are known to be dysregulated in inflammatory bowel diseases like CD. Combining them in a multivariate subset captures both their relative proportions and total counts, which are clinically relevant for distinguishing inflammation subtypes.

2. **Inflammation markers**
   [CRP, Faecal Calprotectin]
   These are key indicators of systemic and intestinal inflammation, respectively. CRP reflects acute-phase liver response, while F-Cal is specific to intestinal neutrophilic activity. Though mechanistically distinct, both are strongly correlated with inflammatory disease activity and complement each other in modeling CD-specific inflammation signatures.

3. **Platelets**
   [Platelets]
   Thrombocytosis (elevated platelet count) is a well-established marker of chronic inflammation. As platelet behavior is relatively independent from other hematological and nutritional markers, we model it as its own trajectory.

4. **Hemoglobin**
   [Hemoglobin]
   Hemoglobin concentration is a direct measure of anemia, which is prevalent in CD patients due to chronic blood loss and inflammation-induced iron sequestration. Its temporal dynamics often diverge from those of other blood components, warranting a separate representation.

5. **Iron status**
   [Iron]
   Iron metabolism is tightly linked to both hemoglobin levels and systemic inflammation but shows distinct dynamics. Modeling it separately allows the model to learn delayed or decoupled effects (e.g., iron deficiency preceding hemoglobin drop).

6. **Vitamin and folate status**
   [Folate, Vitamin B12, Vitamin D2+D3]
   These nutrients are absorbed in different regions of the gastrointestinal tract (e.g., B12 in the ileum, folate in the jejunum), and their deficiency profiles can be informative of CD location and severity. Grouping them allows the model to detect joint patterns of malabsorption and systemic nutrient depletion.

7. **Liver function markers**
   [ALAT, Bilirubin, Albumin]
   These biomarkers reflect hepatic function and protein synthesis. Abnormal liver enzymes and hypoalbuminemia are frequently observed in CD due to medication effects, chronic inflammation, or comorbid autoimmune liver disease. Combining them supports learning of systemic inflammatory effects beyond the gut.

This grouping reflects known biological relationships, enhances the interpretability of the model's subset-level attributions, and improves performance compared to unstructured or purely univariate representations. It enables GMAN to exploit interactions among related features while maintaining a modular structure that aligns with clinical reasoning.

5 Experimental Setup and Hyperparameter Choices

This section outlines key implementation choices and model settings used in our experiments.

5.1 Experimental Setup

Unless otherwise noted, we trained all CD models for a maximum of 300 epochs using the Adam optimizer with weight decay of 1e-4.

We trained all GMAN models with batch size of range {64, 256}, dropout in the {0.1-0.2} range, n_layers in the {2, 10} range, hidden_channels in the {4, 64} range, num_lab_ids_embed in the {5, 8} range, num_biom_embed in the {3, 5} range, num_units_embed in the {3, 5} range. We used a ReduceLROnPlateau scheduler with a max learning rate in the {1e-2, 1e-4} range, min learning rate in the {1e-7, 1e-8} range, factor in the {0.2-0.9} range and patience=100

Random seeds were fixed for reproducibility, and results are reported across three independent runs. All models were trained on a single NVIDIA Tesla V100-PCIE-16GB GPU.