GRAPH MIXING ADDITIVE NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world temporal data often consists of multiple signal types recorded at irregular, asynchronous intervals. For instance, in the medical domain, different types of blood tests can be measured at different times and frequencies, resulting in fragmented and unevenly scattered temporal data. Similar issues of irregular sampling occur in other domains, such as the monitoring of large systems using event log files. Effectively learning from such data requires handling sets of temporally sparse and heterogeneous signals. In this work, we propose Graph Mixing Additive Networks (GMAN), a novel and interpretable-by-design framework for learning directly from sets of graphs that represent such signals. GMAN provides diverse interpretability capabilities, including node-level, graph-level, and subset-level importance, and enables practitioners to trade finer-grained interpretability for greater expressivity when domain priors are available. GMAN achieves state-of-the-art performance in real-world high-stakes tasks, including predicting Crohn's disease onset and hospital length of stay from routine blood test measurements and detecting fake news. Furthermore, we demonstrate how GMAN's interpretability properties assist in revealing disease development phase transitions and provide crucial insights in the healthcare domain.

1 INTRODUCTION

Modern clinical data consist of diverse types of signals, often collected at irregular and asynchronous time intervals. For example, a patient's medical record may include 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 viewed as a set of sparse temporal signals, each with its own fragmented temporal structure. Similar patterns occur in other domains. For instance, the spread of news articles in social media networks often follow asynchronous tree-like patterns of dissemination. Another example is system event logs, which typically include different types of events occurring at varying times and rates.

A common approach for learning from such data is to align the signals to a fixed-size time grid, thereby enforcing a shared timeline. This is typically achieved by trimming or aggregating signals and filling in missing values through interpolation or learned imputation models (Cao et al., 2018; Tashiro et al., 2021; Wu et al., 2022; Du et al., 2023). However, these procedures can lead to substantial information loss and ignore the informative patterns in the irregularity itself, such as the varying time intervals between different measurement types.

In this work, we introduce Graph Mixing Additive Networks (GMAN), a novel framework designed to learn directly from sets of graphs. GMAN can learn over sparse, irregular, temporal heterogeneous signals, without information loss or imputation, by representing them as sets of graphs. For example, in blood test data, each biomarker is modelled as a graph whose nodes correspond to individual measurements. These graphs may differ in structure, size, and feature space, reflecting the diversity of real-world signals.

GMAN provides multiple interpretability capabilities, including node-level, graph-level, and subset-level importance. It also enables practitioners to trade fine-grained interpretability for greater expressivity when domain priors are available. Specifically, signal graphs can be grouped into subsets, shifting interpretability from individual nodes or graphs to the subset level. GMAN builds on Graph Neural Additive Networks (GNAN) (Bechler-Speicher et al., 2024), an interpretable class of GNNs that operate on individual nodes or graphs. Rather than applying GNAN directly, GMAN employs an extended and more flexible variant, we denote as ExtGNAN. In analogy to signal grouping, ExtGNAN supports grouping features into subsets, thereby replacing feature-level interpretability with subset-level interpretability, such as importance scores. We prove that grouping signals or features strictly increases expressivity, and that GMAN is strictly more expressive than GNAN.

We demonstrate the effectiveness of GMAN on real-world high-stakes medical datasets, where it achieves state-of-the-art (SoTA) performance while also providing valuable clinical and biological insights through its interpretability. In addition, we show that GMAN attains SoTA performance in fake news detection, highlighting its flexibility in operating on sets of graphs with arbitrary structure. This contrasts with existing approaches, which are restricted to sets of path-like signals and offer no interpretability. Finally, we demonstrate how GMAN's interpretability reveals phase transitions and provides crucial insights in healthcare.

Our main contributions are as follows:

1. We introduce GMAN, a novel framework for learning directly from sets of sparse, irregular temporal signals, without information loss or imputation.
2. We allow practitioners to integrate domain priors, when available, by grouping features or signal types into subsets. This shifts interpretability from fine-grained (node- or feature-level) to subset-level, while strictly improving expressivity. This capability is particularly valuable in the medical domain, where such priors are common.
3. We provide a theoretical analysis proving that groupings of features and signals make GMAN strictly more expressive.
4. We demonstrate the effectiveness of GMAN on real-world high-stakes medical datasets and fake news detection, achieving state-of-the-art performance in both domains.
5. We show that GMAN provides valuable interpretability capabilities, including node-level, graph-level, and subset-level importance, which yield meaningful clinical and biological insights.

2 RELATED WORK

Graph Neural Additive Networks Graph Neural Networks (GNNs) (Kipf & Welling, 2016; Gilmer et al., 2017; Velickovic et al., 2017; Xu et al., 2018) have become the dominant framework for learning over graph-structured data, enabling flexible representation learning across diverse domains such as healthcare (Paul et al., 2024; Ochoa & Mustafa, 2022; Peng et al., 2023), chemistry (Reiser et al., 2022; Jumper et al., 2021) and social networks (Li et al., 2023; Sharma et al., 2024), amongst others. GNNs leverage both the graph topology and node features to compute learned representations for individual nodes or for entire graphs. Recently, Bechler-Speicher et al. (2024) introduced Graph Neural Additive Networks (GNANs) a novel interpretable-by-design graph learning framework inspired by generalized additive models (GAMs) (Hastie & Tibshirani, 1986; 1987). GNAN applies univariate neural networks to each feature of the nodes separately, and then linearly combines their outputs across nodes to produce node-level and graph-level representations. As features are not mixed non-linearly, GNAN is fully interpretable, and provides feature-level and node-level interpretability which shows exactly how each feature and each node contribute to the final target variable.

Learning from sparse data In many real-world settings, data often presents missing values from irregular sampling and variable feature availability. Recurrent models (Cao et al., 2018) treat missing data as latent variables, while attention-based methods (Du et al., 2023; Wu et al., 2022; Tipirneni & Reddy, 2022; Labach et al., 2023) reconstruct them via contextual masking and temporal blocks. Diffusion models (Tashiro et al., 2021; Alcaraz & Strodthoff, 2022; Senane et al., 2024; Dai et al., 2024) learn conditional distributions over missing values using score-based processes. Graph-based approaches use GNNs to model feature dependencies through bipartite graphs, adaptive message passing, or spatio-temporal attention (You et al., 2020; Cini et al., 2022; Marisca et al., 2022; Ye et al., 2021; Chen et al., 2024). These imputation methods often distort dynamics and may not improve prediction (Qian et al., 2025). An alternative is to model sparsity directly. Neural and Latent ODEs (Chen et al., 2018; Rubanova et al., 2019) address irregular gaps via continuous dynamics but are compute-intensive and rely on missingness encodings. Recent models (Zhang et al., 2021) have used graph representations to capture sparsity without imputation. Importantly, none of the aforementioned methods offer built-in multi-grain interpretability in the way that GMAN does.

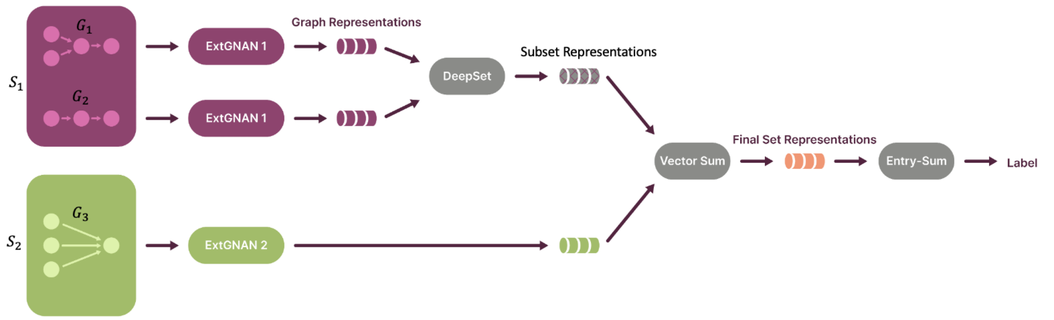


Figure 1: In this example, the input is a set of three graphs, G_1, G_2, G_3, grouped into two subsets S_1 and S_2. Within each subset, the same ExtGNAN instance is applied to all graphs to produce their graph-level representations. For subsets containing multiple graphs, a DeepSets module aggregates these graph representations into a single subset representation. For subsets of size one, the subset representation is simply the graph representation itself. The final set representation is then obtained by summing the subset representations, and the final label prediction is produced by summing the entries of this set representation.

3 GRAPH MIXING ADDITIVE NETWORKS

In this section, we present GMAN, an interpretable and flexible method for learning over sets of arbitrary graphs.

Preliminaries GMAN acts on a set of m graphs $S = \{G_1, \dots, G_m\}$, which can be directed or undirected. Each node $v \in G_i$, $1 \leq i \leq m$ is associated with a feature vector $x_v \in \mathbb{R}^d$ and a time-stamp t_v. GMAN utilizes the distances between each pair of nodes in each graph, denoted by Δ_{uv}, where

$$\Delta_{uv} = \begin{cases} t_u - t_v, & \text{if there exists a path from } u \text{ to } v \\ 0, & \text{otherwise.} \end{cases}$$

Graphs may be provided directly through an explicit layout, e.g., the news propagation graphs we evaluate GMAN over in section 4. More commonly, however, sparse temporal heterogeneous data does not come with a predefined graph structure. In such cases, we construct directed path-graphs for each signal type based on its measurement time stamps, as demonstrated for the medical dataset in section 4. For instance, in patient blood test records, each graph corresponds to a specific biomarker. Nodes in each graph represent individual measurements of that biomarker, annotated with the observed test value as a feature and the associated time stamp. We mark vectors with bold, and 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 $[\mathbf{h}]_S$.

Signal Grouping To incorporate domain priors, the graphs in S can be partitioned into k disjoint subsets $S_1, \dots, S_k$ with $\bigcup_{i=1}^{k} S_i = S$. If at least one subset S' contains multiple graphs, the model's expressivity increases, as we prove in Theorem 3.2, at the cost of shifting interpretability from individual nodes and graphs in S' to the subset level. In practice, this means we can attribute importance to S' as a whole, but not to its individual components. This trade-off—enhanced expressivity at reduced granularity of interpretability- is especially valuable in domains such as medicine, where priors often suggest natural groupings of signals and interpretability is only needed at the subset level. This is demonstrated in Section 4.

GMAN linearly aggregates representations of the subsets of S to form a final set representation, and then assigns a single label to S.

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 produce 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 Extended GNAN (EXTGNAN), as described in Section 3.1. For subsets containing multiple graphs, a featuregroupgnan is applied to each graph, followed by a DeepSet aggregation (Zaheer et al., 2018) 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 (NN) $f : \mathbb{R}^d \to \mathbb{R}^d$ for each vector in the set $\{h_l\}_{G_l \in S_i}$, sums the results, and then applies another NN $g : \mathbb{R}^d \to \mathbb{R}^d$.

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

Here, g and f are NNs of arbitrary depth and width. A high-level visual overview of GMAN is presented in Figure 1. We now turn to define EXTGNAN.

3.1 EXTGNAN

In GNAN, univariate NNs 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, or feature-level interpretability is not required for all features, it may result in sub-par performance. Therefore, EXTGNAN extends GNAN by allowing multivariate NNs to operate on groups of features to gain accuracy at the cost of reducing the feature-level interpretability only for features that are grouped together, and obtaining interpretability for their subset as a whole instead.

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 NN for all the features in the subset together, instead of a univariate NN 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 NN 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:

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

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

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

This concludes the description of EXTGNAN, which computes graph-level representations. We provide a complexity analysis of GMAN in the appendix. Next, we describe how GMAN combines these representations across sets and enables multi-level interpretability.

3.2 NODE, GRAPH AND SUBSET IMPORTANCE

GMAN retains all interpretability properties of GNAN, 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 obtain importance scores. A key property of GMAN's interpretability is that its importance scores directly reflect the contribution of each node, graph, or subset to the predicted label, since these terms are combined additively to produce the final output. Section 4 illustrates how node-level importances yield insightful real-world insights.

We can extract the total contribution of each node j to the prediction by summing the contributions of the node across all feature subsets. 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, we provide instead the total contribution of the set to the final prediction

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

3.3 EXPRESSIVITY PROPERTIES

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

Theorem 3.1. GMAN *is strictly more expressive than GNAN.*

The following theorem shows that a GMAN which is applied to subsets of graphs of size at least two, is more expressive than a GMAN that is applied to only subsets of size one:

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 $\{S_i\}_{i=1}^{k}$. Then a* GMAN *trained over S_2 is strictly more expressive than a* GMAN *trained over S_1.*

4 EMPIRICAL EVALUATION

In this section, we evaluate GMAN on real-world tasks, and demonstrate its interpretability properties [1].

---

[1] Code is provided in Supplementary Materials

4.1 MEDICAL PREDICTIONS

We evaluate GMAN on two high-impact clinical prediction tasks: LoS of intensive care patients and onset of CD. In both settings, each individual is represented as a set of time-stamped biomarker trajectories, where each trajectory forms a directed path graph with nodes corresponding to test results and edges encoding the time elapsed between measurements. This representation preserves the temporal structure of each biomarker independently while enabling joint reasoning across biomarkers during learning.

Data Next, we describe the two medical datasets used in our evaluation.

P12 (ICU Length of Stay): The PhysioNet2012 (P12) dataset, introduced by Goldberger et al. (2000), contains records from 11,988 intensive-care unit (ICU) patients, following the exclusion of 12 samples deemed inappropriate according to the criteria in Horn et al. (2020). For each patient, longitudinal measurements from 36 physiological signals were recorded over the initial 48 hours of ICU admission. Additionally, each patient has a static profile comprising 9 features, including demographic and clinical attributes such as age and gender. The dataset is labelled for a binary classification task: predicting whether or not the total LoS in the ICU exceeded 72 hours.

Crohn's Disease (CD Onset): The Danish health registries are comprehensive, nationwide databases covering healthcare interactions for over 9.5 million individuals (Pedersen, 2011). 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 (Arendt et al., 2020). From this data, we constructed a cohort of 8,567 individuals later diagnosed with CD and 8,567 age-matched controls. For each person, we extracted temporal trajectories of 17 routinely measured biomarkers, reflecting key physiological processes. The complete list and descriptions of these biomarkers are provided in the Appendix. The task is binary classification: predicting future CD onset from pre-diagnostic medical histories.

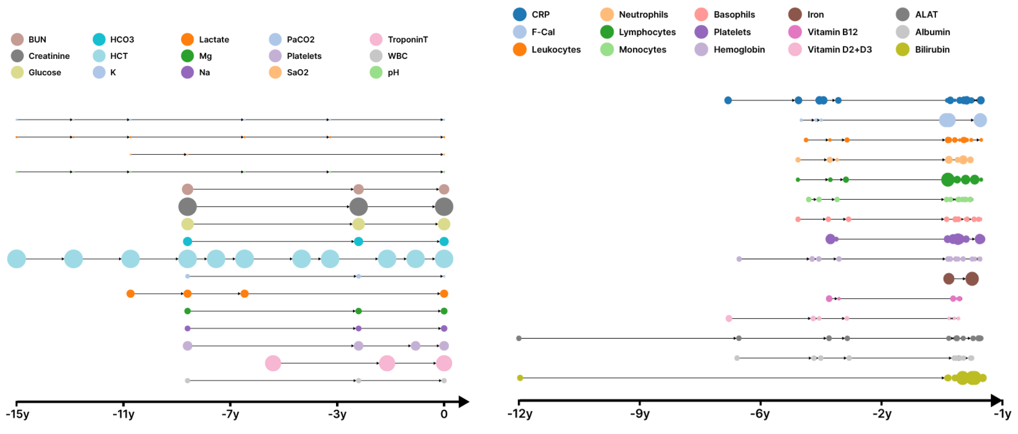
Setup We compare GMAN to 8 baselines from Zhang et al. (2021), comprising both sequential and graph-based models, including: *Transformer* (Vaswani et al., 2017), *Trans-mean*, *GRU-D* (Che et al., 2016), *SeFT* (Horn et al., 2020), *mTAND* (Shukla & Marlin, 2021), DGM^2(Wu et al., 2021) MTGNN (Wu et al., 2020) and Raindrop (Zhang et al., 2021). For the $P12$ we used the splits as in (Zhang et al., 2021). For CD, we randomly split the data into train (80%), validation (10%), and test (10%) sets. We conducted a grid search by training on the training set and evaluating on the validation set. We then selected the best performing model over the validation set and report results over the test set. We define each individual biomarker as a unique signal type and group them according to common biomarker domain knowledge. We tune the biomarker subsets over no grouping at all (i.e, full interpretability), and 5 additional subset groupings motivated by public common clinical knowledge.

Table 1: Evaluation of GMAN on two real-world medical tasks. The metric reported in the mean AUPRC with standard deviation, calculated over 3 random seeds.

Methods	**LoS in ICU**	**CD onset**
Transformer	96.06 ± 0.32	75.60 ± 0.52
Trans-mean	96.44 ± 0.17	75.96 ± 0.92
GRU-D	95.91 ± 2.10	83.36 ± 0.40
SeFT	95.89 ± 0.08	71.22 ± 2.30
mTAND	93.02 ± 1.04	83.17 ± 0.67
DGM^2	97.00 ± 0.40	83.02 ± 0.56
MTGNN	96.20 ± 0.78	75.26 ± 3.04
RAINDROP	96.32 ± 0.13	82.60 ± 0.82
GMAN	**97.41 ± 0.38**	**83.93 ± 0.27**

To account for class imbalance, and as commonly done for these datasets, we report the average AUPRC score and standard-deviation of the selected configuration with 3 random seeds. Additional details on the datasets, experimental setup, subset groupings, and hyperparameter configurations are provided in the Appendix.

Results The results in Table 1 show that GMAN achieves the highest AUPRC across both medical prediction tasks. On Length of Stay in ICU, it improves upon the best baseline with a relative uplift of about 0.41 points, while on Crohn's Disease onset, it outperforms the best baseline by 0.57 points.



(a) P12 (b) CD

Figure 2: Node-level importances for two individuals from (a) the P12 ICU LoS and (b) CD onset datasets. Node size indicates the exact node (measurement) contribution to the prediction.

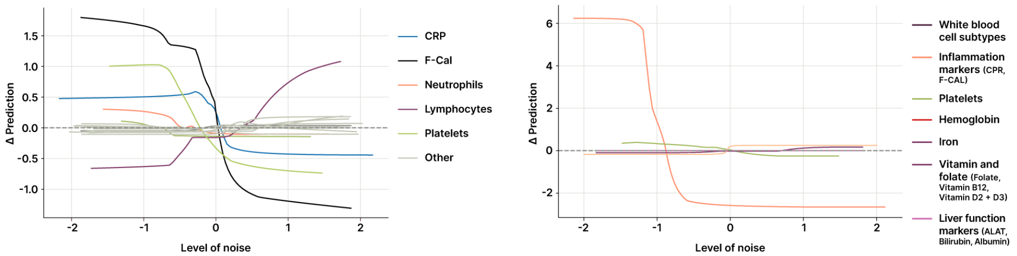
Notably, these improvements are achieved while GMAN also provides interpretability properties, as shown in detail in the next subsection.

4.1.1 CLINICAL INSIGHTS THROUGH INTERPRETABILITY

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 that go beyond global feature importance, 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 outcome 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 highlight the interpretability of GMAN, we conduct attribution analyses on both the CD and P12 datasets, examining how the model assigns importance across time and signals.

Critical phase detection through node-level importance In the GMAN framework, nodes represent individual signals within a biomarker trajectory. As such, highly influential nodes can highlight critical phases where specific signals most strongly impact the prediction. We use Equation (3) to quantify node-level contributions across biomarkers in both the CD and P12 datasets. To comply with privacy and data protection legislation, CD data is anonymised via noise and temporal shifting. Figure 2 displays node-level importance across biomarker trajectories for two randomly selected individuals from each dataset, with node size reflecting the magnitude of each node's contribution to the model's prediction. In both clinical tasks, the model's attributions appear consistent with established biomedical knowledge. For CD prediction, GMAN highlights key inflammatory and immune markers, such as F-Cal, platelets, and lymphocytes, as primary contributors. All of these are known to play central roles in disease onset (Vestergaard et al., 2023). In the P12 example, the model assigns high importance to markers of renal function, liver injury, cardiac stress, and metabolic imbalance, aligning well with clinical predictors of severity in intensive care settings.

Total biomarker contribution In addition to node-level importances, GMAN provides subset-level importance scores, quantifying the contribution of entire biomarker groups to the model's prediction. This enables flexible, system-level interpretability, allowing users to assess the collective influence of physiologically related biomarkers on the target variable. We conduct a subset-level importance analysis on the CD dataset using two clinically motivated grouping strategies. In the first, we assess individual biomarkers to determine whether the model prioritizes features known to be associated with CD onset, allowing us to verify its alignment with established biomedical knowledge. In the second, we group biomarkers into clinically coherent subsets based on physiological function, such as immune response; inflammation; oxygen transport; and liver function; to examine whether



(a) Single-biomarker groups GMAN (b) Best-performing GMAN groups

Figure 3: Subset-level contribution curves for Crohn's Disease prediction. Each curve shows how the GMAN's output changes as increasing noise is added to the latent representation of a biomarker group. (a) uses individual biomarkers; (b) uses physiologically coherent groups.

the model captures system-level patterns consistent with disease progression. Full details on the clinical significance of these groupings are provided in the Appendix. For each grouping strategy, we trained a separate instance of GMAN. To quantify the importance of each biomarker subset, we use Equation (4), measuring how the model's prediction changes when increasing the noise added to the features of nodes within that subset. To introduce noise in a structured way, we apply PCA (Abdi & Williams, 2010) to the feature vectors of all nodes corresponding to biomarkers in the group, and progressively perturb the input along the 1st principal component. This procedure is performed independently for each subset and allows us to assess the model's sensitivity to perturbations in biologically meaningful groupings, offering insight into the relative predictive weight of each group. Importantly, GMAN is interpretable by design, which makes perturbation analysis directly reflect its internal computation. Unlike post-hoc attribution, we do not estimate influence indirectly but observe how the additive contribution changes. Adding structured noise (via PCA) and measuring the resulting prediction shift faithfully traces the model's internal mechanism rather than relying on an external approximation. Thus, perturbation effects align exactly with feature importance, making the experiment both natural and principled. Figure 3 presents the results of the subset-level attribution analysis. In the single-biomarker setting (Figure 3(a)), F-Cal, platelets, and lymphocytes show strong directional effects on model output. F-Cal and platelets are positively associated with CD risk, while lymphocytes have an inverse effect, findings that are both biologically grounded and consistent with prior work (Vestergaard et al., 2023). In the clinically-coherent group setting (Figure 3(b)), the inflammation subset, emerges as the most influential, with a pronounced non-linear effect on predictions, aligning with established diagnostic relevance in CD (Vestergaard et al., 2023).

4.2 FAKE-NEWS DETECTION

Table 2: Evaluation of GMAN on the Gossipcop (GOS) fake news detection dataset.

Methods	Accuracy
GATv2	96.10 ± 0.3
GraphConv	96.77 ± 0.1
GraphSage	94.45 ± 1.5
GCNFN	96.52 ± 0.2
GMAN	**97.34 ± 0.2**

Data GossipCop (GOD) is a dataset of news articles annotated by professional journalists and fact-checkers, containing both content-based labels and social context information verified through the GossipCop fact-checking platform. It is composed of $5,464$ tree-structured graphs based on sharing information, where the news article is the root node and sharing users are subsequent nodes in the cascade, with edges signifying sharing relationships. Each node in these graphs is associated with 4 features: 768-dimensional embeddings generated using a pretrained BERT model on user historical posts, 300-dimensional embeddings from a pretrained word2vec (Mikolov et al., 2013) on the same historical posts, 10-dimensional features extracted from user profiles, and a 310-dimensional "content" feature that combines the 300-dimensional embedding of user comments with the 10-dimensional profile features. Each graph is labelled to indicate whether it originates from a fake news post or not. We decomposed the tree into a set of directed graphs rooted at the origin, each representing a distinct path of news propagation.

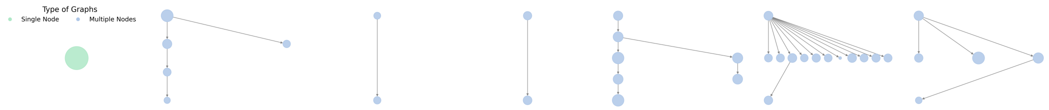


Figure 4: Node importance for fake-news spread graphs, over the GOS dataset. The node size corresponds to its importance learned by tGMAN, according to Equation (3). All graphs with a single node are grouped into one subset. Therefore, the importance is provided on the subset level rather than the node level (green node).

Setup While GMAN can be applied to sets of graphs with any structure, the baselines used in the CD and P12 experiments are limited to path-like graphs and are unable to act on more complex graphs as in this dataset. Therefore, we instead evaluate GMAN against 4 GNNs, including GATv2 (Brody et al., 2022), GraphConv (Morris et al., 2021), GraphSAGE (Hamilton et al., 2018) and GCNFN (Monti et al., 2019). We use random splits of train (80%), validation (10%), and test (10%) sets over the data and selected the best performing model over the validation set. We then report the average Accuracy score and std of the selected configuration with three seeds. Since many subgraphs reduce to a single node after decomposition, we group all size-one graphs into a shared subset and combine them non-linearly. For the remaining graphs, whose identities are not uniquely distinguishable, we apply a shared EXTGNAN. In total, we use two distinct EXTGNAN instances for this experiment.

Results and node-level importance Results are provided in Table 2. GMAN outperforms all baselines, with an uplift of 0.57 accuracy points. Figure 4 presents the node importance of a random sample from the datasets, where the size of a node corresponds to its importance score according to Equation (3).

4.2.1 ABLATION STUDY

Finally, we carry out an ablation study on the CD dataset, to isolate the impact of key components on performance. Specifically, we take the best-performing configuration from the training-data grid search and re-train it with individual components ablated to assess their effect on performance. We test: (i) replacing DeepSet with mean pooling to assess the value of learned non-linear aggregation; (ii) replacing the distance function NN (ρ) with a constant 1 value. This tests the usefulness of structural information; (iii) substituting ExtGNAN with a node-wise MLP to test the importance of graph inductive bias; (iv) replacing ExtGNAN with an identity mapping as a lower bound without feature learning; and (v) using standard GNAN (no multivariate feature groups) to evaluate the benefit of grouped feature processing. We report the performance degradation (AUPRC difference) in Table 3. The ablation results demonstrate that the core components of GMAN are critical to its effectiveness, as their removal consistently leads to notable performance degradation.

Table 3: Ablation study of GMAN components.

Ablation	**AUPRC drop**
(i) DeepSet→mean pooling	*- 19.98% ± .28 %*
(ii) ρ→1	*- 12.39% ± 1.39%*
(iii) ExtGNAN→MLP	*- 15.00% ± 2.09 %*
(iv) ExtGNAN→Identity	*- 17.70% ± 0.15%*
(v) ExtGNAN→GNAN	*- 4.38 % ± 2.85%*

5 CONCLUSION

We introduced Graph Mixing Additive Networks (GMAN), a framework for learning from sets of graphs that represent irregular and asynchronous temporal signals. GMAN was designed to handle real-world scenarios where multiple signal types are collected at uneven intervals, such as medical records with heterogeneous blood tests or event logs in complex systems. The framework combined strong predictive performance with built-in interpretability, offering importance scores at the node, graph, and subset levels. GMAN allows practitioners to integrate domain priors when available, trading fine-grained interpretability for greater expressivity. Across experiments on real-world high-stakes tasks, GMAN achieved state-of-the-art results. Beyond predictive accuracy, GMAN's interpretability capabilities proved particularly valuable in domains like healthcare, where uncovering phase transitions and providing actionable insights is critical.

REPRODUCIBILITY STATEMENT

The full code is provided in the Supplementary Material. We will release our code upon acceptance, including all training and evaluation scripts. The README file in the code provides all necessary details to ensure our results and evaluations can be reproduced faithfully.

REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399*, 2022.

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*, pp. 469–475, 2020.

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.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks?, 2022. URL `https://arxiv.org/abs/2105.14491`.

Peter Buneman. A note on the metric properties of trees. *J. Combin. Theory Ser. B*, 17(1):48–50, 1974.

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.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values, 2016. URL `https://arxiv.org/abs/1606.01865`.

Qianyi Chen, Jiannong Cao, Yu Yang, Wanyu Lin, Sumei Wang, and Youwu Wang. Multistage graph convolutional network with spatial attention for multivariate time series imputation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the g_ap_s: Multivariate time series imputation by graph neural networks, 2022. URL `https://arxiv.org/abs/2108.00298`.

Zongyu Dai, Emily Getzen, and Qi Long. Sadi: Similarity-aware diffusion model-based imputation for incomplete temporal ehr data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4195–4203. PMLR, 2024.

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

Manik Garg, Marcin Karpinski, Dorota Matelska, Lawrence Middleton, Oliver S Burren, Fengyuan Hu, Eleanor Wheeler, Katherine R Smith, Margarete A Fabre, Jonathan Mitchell, et al. Disease prediction with multi-omics and biomarkers empowers case–control genetic discoveries in the uk biobank. *Nature Genetics*, 56(9):1821–1831, 2024.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018. URL `https://arxiv.org/abs/1706.02216`.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical science*, 1(3):297–310, 1986.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

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

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G Krishnan. Duett: dual event time transformer for electronic health records. In *Machine Learning for Healthcare Conference*, pp. 403–422. PMLR, 2023.

Xiao Li, Li Sun, Mengjie Ling, and Yan Peng. A survey of graph neural network based recommendation in social networks. *Neurocomputing*, 549:126441, 2023.

Ivan Marisca, Andrea Cini, and Cesare Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. *Advances in neural information processing systems*, 35: 32069–32082, 2022.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. Fake news detection on social media using geometric deep learning, 2019. URL `https://arxiv.org/abs/1902.06673`.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks, 2021.

Juan G Diaz Ochoa and Faizan E Mustafa. Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses. *Artificial Intelligence in Medicine*, 131:102359, 2022.

Showmick Guha Paul, Arpa Saha, Md Zahid Hasan, Sheak Rashed Haider Noori, and Ahmed Moustafa. A systematic review of graph neural network in healthcare-based applications: Recent advances, trends, and future directions. *IEEE Access*, 12:15145–15170, 2024.

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

Wei Peng, Tielin Chen, Hancheng Liu, Wei Dai, Ning Yu, and Wei Lan. Improving drug response prediction based on two-space graph convolution. *Computers in Biology and Medicine*, 158: 106859, 2023.

Linglong Qian, Yiyuan Yang, Wenjie Du, Jun Wang, Richard Dobsoni, and Zina Ibrahim. Beyond random missingness: Clinically rethinking for healthcare time series imputation, 2025. URL `https://arxiv.org/abs/2405.17508`.

Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, et al. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93, 2022.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Zineb Senane, Lele Cao, Valentin Leonhard Buchner, Yusuke Tashiro, Lei You, Pawel Andrzej Herman, Mats Nordahl, Ruibo Tu, and Vilhelm Von Ehrenheim. Self-supervised learning of time series representation via diffusion process and imputation-interpolation-forecasting mask. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2560–2571, 2024.

Kartik Sharma, Yeon-Chang Lee, Sivagami Nambi, Aditya Salian, Shlok Shah, Sang-Wook Kim, and Srijan Kumar. A survey of graph neural networks for social recommender systems. *ACM Computing Surveys*, 56(10):1–34, 2024.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series, 2021. URL `https://arxiv.org/abs/2101.10318`.

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.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

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.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

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.

Jakob Woerner, Thomas Westbrook, Seokho Jeong, Manu Shivakumar, Allison R Greenplate, Sokratis A Apostolidis, Seunggeun Lee, Yonghyun Nam, and Dokyoon Kim. Plasma protein-based and polygenic risk scores serve complementary roles in predicting inflammatory bowel disease. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 30, pp. 522, 2025.

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.

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`.

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`.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yongchao Ye, Shiyao Zhang, and James JQ Yu. Spatial-temporal traffic data imputation via graph attention convolutional network. In *International Conference on artificial neural networks*, pp. 241–252. Springer, 2021.

Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33: 19075–19087, 2020.

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

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

A APPENDIX

B THEORETICAL FRAMEWORK

B.1 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.

B.2 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 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 :contentReference[oaicite:0]index=0:contentReference[oaicite:1]index=1. 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.

B.3 FOUR-POINT CONDITION AND RECOVERABILITY

We now turn to structural identifiability. We prove that when input graphs are connected, acyclic, and positively weighted (i.e., trees), the pairwise distance matrix learned by GMAN encodes the full structure of the graph, up to isomorphism. This provides theoretical justification for the model's ability to reason over temporal structure without needing explicit graph supervision.

The following theorem shows that if the graph satisfies the four-point condition (Buneman, 1974), GMAN can reconstruct the original graph from the transformed distance matrix that is fed to GMAN as the graph input:

Theorem B.1. *Let G be a graph represented by an adjacency matrix A, and D be the transformed distance-matrix for* GMAN. *Then if D satisfies the four-point condition,* GMAN *can learn ρ such that $\rho(D) = A$.*

Proof. Let $G = (V, E, w)$ be a positively weighted path graph, i.e.

$$V = \{v_1, \dots, v_n\}, \qquad E = \big\{\{v_i, v_{i+1}\} \mid i = 1, \dots, n-1\big\}, \qquad w(\{v_i, v_{i+1}\}) > 0.$$

Define the pair-wise distance matrix $D \in \mathbb{R}^{n \times n}$ by

$$D_{uv} \ = \sum_{e \in P_G(u,v)} w(e) \ ,$$

where $P_G(u, v)$ is the unique u–v path in G. Then:

(a) Tree-metric property. D satisfies the four-point condition of Buneman (Buneman, 1974); hence (V, D) is a *tree metric*.

(b) Uniqueness (*no information loss*). By Buneman's theorem the tree that realises D is unique up to isomorphism. For a path graph the only automorphism is the reversal $(v_1, \dots, v_n) \mapsto (v_n, \dots, v_1)$, so D determines G completely except for left–right orientation.

(c) Efficient reconstruction. G can be reconstructed from D in $O(n^2)$ time:

   i. Choose an endpoint $s = \arg\max_{v \in V} \max_{u \in V} D_{vu}$.
   ii. Order the vertices $v_1 = s,\ v_2, \dots, v_n$ so that $D_{sv_1} < D_{sv_2} < \dots < D_{sv_n}$.
   iii. Set edge weights $w(\{v_i, v_{i+1}\}) = D_{sv_{i+1}} - D_{sv_i}$ for $i = 1, \dots, n-1$.

□

C COMPUTATIONAL COMPLEXITY AND EFFICIENT IMPLEMENTATION

In this section, we provide a big-O analysis of the time complexity of GMAN. We also provide an efficient approach to implement GMAN.

Computational Complexity The computational complexity of GMAN is as follows:

- **Scales linearly with the number of graphs m:** Each graph is processed independently or in small subsets, so total cost is $\mathcal{O}(m)$ assuming fixed per-graph cost.
- **Scales quadratically with the number of nodes per graph n:** Due to the dense aggregation over all node pairs in ExtGNAN, the per-graph cost is $\mathcal{O}(n^2)$.
- **Overall complexity** for a set of graphs is:

$$\mathcal{O}\big(m \cdot K \cdot n^2 \cdot d_\psi\big)$$

where K is the number of feature groups and d_ψ is the cost of evaluating the multivariate neural networks.

Efficient Implementations In the main paper, we present GMAN with the objective of maximal clarity, e.g., by presenting vector entry-wise operations. Nonetheless, the operations of GMAN can be done in an optimized fashion for GPU, through tensor operations.

D DATASET DETAILS

D.1 PHYSIONET P12

We provide the full list of the 36 physiological signals and 3 static patient features used in our experiments.

1. Alkaline phosphatase (ALP): A liver- and bone-derived enzyme; elevations suggest cholestasis, bone disease, or hepatic injury.

2. Alanine transaminase (ALT): Hepatocellular enzyme; increased values mark acute or chronic liver cell damage.

3. Aspartate transaminase (AST): Enzyme in liver, heart, and muscle; rises indicate hepatocellular or muscular injury.

4. Albumin: Major plasma protein maintaining oncotic pressure and transport; low levels reflect inflammation, malnutrition, or liver dysfunction.

5. Blood urea nitrogen (BUN): End-product of protein catabolism cleared by the kidneys; elevation signals renal impairment or high catabolic state.

6. Bilirubin: Hemoglobin breakdown product processed by the liver; accumulation indicates hepatobiliary disease or hemolysis.

7. Cholesterol: Circulating lipid essential for membranes and hormones; dysregulation is linked to cardiovascular risk.

8. Creatinine: Waste from muscle metabolism filtered by the kidneys; higher levels imply reduced glomerular filtration.

9. Invasive diastolic arterial blood pressure (DiasABP): Pressure during ventricular relaxation; low readings may reflect vasodilation or hypovolemia.

10. Fraction of inspired oxygen (FiO$_2$): Proportion of oxygen delivered; values above ambient air denote supplemental therapy.

11. Glasgow Coma Score (GCS): Composite neurologic score for eye, verbal, and motor responses; scores ≤ 8 indicate severe impairment.

12. Glucose: Principal blood sugar; hypo- or hyper-glycemia can cause neurologic compromise and metabolic instability.

13. Serum bicarbonate (HCO_3): Key extracellular buffer; low levels signal metabolic acidosis, high levels metabolic alkalosis or compensation.

14. Hematocrit (HCT): Percentage of blood volume occupied by red cells; reduced values denote anemia, elevated values hemoconcentration.

15. Heart rate (HR): Beats per minute reflecting cardiac demand; tachycardia indicates stress or shock, bradycardia conduction disorders.

16. Serum potassium (K): Crucial intracellular cation; deviations predispose to dangerous arrhythmias.

17. Lactate: By-product of anaerobic metabolism; elevation marks tissue hypoxia and shock severity.

18. Invasive mean arterial blood pressure (MAP): Time-weighted average arterial pressure; low values threaten organ perfusion.

19. Mechanical ventilation flag (MechVent): Binary indicator of ventilatory support; presence denotes respiratory failure or peri-operative care.

20. Serum magnesium (Mg): Cofactor for numerous enzymatic reactions; abnormalities contribute to arrhythmias and neuromuscular instability.

21. Non-invasive diastolic arterial blood pressure (NIDiasABP): Cuff-derived diastolic pressure; trends mirror vascular tone without an arterial line.

22. Non-invasive mean arterial blood pressure (NIMAP): Cuff-based mean pressure; used when invasive monitoring is unavailable.

23. Non-invasive systolic arterial blood pressure (NISysABP): Cuff-derived systolic pressure; elevations suggest hypertension or pain response.

24. Serum sodium (Na): Principal extracellular cation governing osmolality; dysnatremias cause neurologic symptoms and fluid shifts.

25. Partial pressure of arterial carbon dioxide ($PaCO_2$): Indicator of ventilatory status; hypercapnia implies hypoventilation, hypocapnia hyperventilation.

26. Partial pressure of arterial oxygen (PaO_2): Measure of oxygenation efficiency; low values denote hypoxemia.

27. Arterial pH: Measure of hydrogen-ion concentration; deviations from normal reflect systemic acid–base disorders.

28. Platelet count (Platelets): Thrombocyte concentration essential for hemostasis; low counts increase bleeding risk, high counts thrombosis risk.

29. Respiration rate (RespRate): Breaths per minute; tachypnea signals metabolic acidosis or hypoxia, bradypnea central depression.
30. Hemoglobin oxygen saturation (SaO_2): Percentage of hemoglobin bound to oxygen; values below normal indicate significant hypoxemia.
31. Invasive systolic arterial blood pressure (SysABP): Peak pressure during ventricular ejection; extremes compromise end-organ perfusion.
32. Body temperature: Core temperature; fever suggests infection, hypothermia exposure or metabolic dysfunction.
33. Troponin I: Cardiac-specific regulatory protein; elevation confirms myocardial injury.
34. Troponin T: Isoform of cardiac troponin complex; rise parallels Troponin I in detecting myocardial necrosis.
35. Urine: Hourly urine volume as a gauge of renal perfusion; oliguria signals kidney hypoperfusion or failure.
36. White blood cell count (WBC): Reflects immune activity; leukocytosis suggests infection or stress, leukopenia marrow suppression or severe sepsis.

Static patient features: Age; Gender; *ICUType* – categorical code for the admitting intensive care unit (1 = Coronary Care, 2 = Cardiac Surgery Recovery, 3 = Medical ICU, 4 = Surgical ICU), capturing differences in case mix and treatment environment.

D.2 CROHN'S DISEASE PREDICTION

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.

D.2.1 CLINICAL CONTEXT AND RELATED WORK FOR PREDICTING CD ONSET

Research on predicting the onset of CD has explored a range of approaches, including the use of routinely measured blood-based biomarkers and more complex biological data derived from multi-omics technologies.

Several studies have assessed the predictive potential of standard clinical blood tests. For example, (Vestergaard et al., 2023) analyzed six routine biomarkers from 1,186 Danish patients eventually diagnosed with CD, achieving moderate predictive performance (AUROC of 0.74) approximately six months before clinical diagnosis. Larger-scale analyses, such as those using UK Biobank data, combined multiple standard biomarkers and basic demographic information, reporting similar predictive performances (AUROCs typically between 0.70–0.75). These analyses generally utilized methods like logistic regression, random forests, or gradient-boosted trees, favored for structured clinical datasets.

Other studies have integrated advanced biochemical data, known as multi-omics, including large-scale protein measurements (proteomics), metabolites (metabolomics), or genomic markers. (Garg et al., 2024) for instance, combined 67 blood biomarkers with approximately 2,900 plasma proteins from the UK Biobank, achieving an AUROC of 0.786. (Woerner et al., 2025) combined genetic risk scores with extensive proteomic data, achieving an AUROC of 0.76 for CD prediction up to five years prior to diagnosis. Similar multi-omics approaches employing microbiome profiling, immune signaling molecules (cytokines), or lipid molecules typically achieve AUROCs between 0.75 and 0.80 but often involve significant cost, specialized laboratory analyses, and reduced consistency across diverse patient cohorts.

Overall, routine blood tests provide meaningful predictive signals for CD onset, while integrating complex biochemical measurements can improve predictive accuracy, albeit at greater cost, complexity, and variability across clinical populations.

E BIOMARKER SUBSET GROUPINGS

E.1 P12

In the PhysioNet P12 task, we grouped the 36 physiological signals into one multivariate subset and 29 singleton subsets. Domain knowledge showed that only the respiratory and gas-exchange variables shared sufficiently strong, coherent dynamics to benefit from joint modeling. All other signals were physiologically diverse, so they were left as singletons to retain their unique predictive information.

1. **Arterial blood gas profile**
   [pH, $PaCO_2$, PaO_2, SaO_2, HCO_3, FiO_2]
   This group captures systemic acid–base status (pH, HCO_3), carbon dioxide clearance ($PaCO_2$), oxygenation (PaO_2, SaO_2), and inspired oxygen fraction (FiO_2). Together they form the canonical arterial blood gas panel, enabling the model to detect respiratory derangements such as hypoxemia, hypercapnia, or metabolic compensation.
2. **Complete blood count**
   [WBC, HCT, Platelets]
   This cluster summarizes hematologic composition by measuring leukocyte-mediated immune response (WBC), oxygen-carrying capacity (HCT), and clotting potential (Platelets). Joint modeling supports recognition of systemic inflammation, anemia, and coagulopathy.
3. **Comprehensive metabolic panel**
   [Glucose, Na, K, Mg, BUN, Creatinine]
   These biomarkers represent key substrates and electrolytes (Glucose, Na, K, Mg) and renal waste products (BUN, Creatinine). Grouping them provides a unified view of metabolic balance, electrolyte homeostasis, and kidney function.
4. **Liver function tests**
   [ALT, AST, ALP, Albumin, Bilirubin]
   These biomarkers assess hepatocellular injury (ALT, AST), cholestasis (ALP, Bilirubin), and hepatic synthetic function (Albumin). Their combined interpretation reflects multiple dimensions of liver health.
5. **Lipid and cardiac markers**
   [Cholesterol, TroponinI, TroponinT, HR]
   This group integrates lipid metabolism (Cholesterol), cardiac injury markers (Troponin I, Troponin T), and heart rate (HR). Together, they provide insight into cardiovascular stress, myocardial injury, and metabolic risk.
6. **Blood pressure profiles**
   [SysABP, DiasABP, MAP, NISysABP, NIDiasABP, NIMAP]
   These variables capture invasive and non-invasive arterial blood pressures, reflecting systemic hemodynamics. Grouping them enables the model to learn coherent pressure dynamics rather than treating each measurement in isolation.
7. **Ventilation mechanics**
   [RespRate, MechVent]

This group reflects mechanical and physiological components of ventilation. Their joint dynamics provide context for interpreting respiratory compensation and ventilatory support.

8. **Tissue perfusion**
   [Lactate, Urine]
   Elevated lactate indicates anaerobic metabolism, while urine output tracks renal perfusion. Together they provide complementary signals of global tissue perfusion and shock severity.
9. **Global status indicators**
   [GCS, Temp]
   These variables capture overall neurologic responsiveness (GCS) and systemic temperature regulation (Temp), providing global context on patient stability and severity of illness.

E.2 CD

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.

F EXPERIMENTAL SETUP, HYPERPARAMETER CHOICES AND GROUPING CONFIGURATIONS

This section outlines key implementation choices and model settings used in our experiments, including the manually tuned biomarker grouping configurations that served as an important hyperparameter for performance and interpretability.

F.1 HYPER-PARAMETERS

We trained all models with 100 epochs using the Adam optimizer with weight decay 1e-5. 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

We trained all models with batch size of range {16, 32}, dropout rate in {0.1, 0.2}, number of layers in the {3, 4, 5} range, hidden channels in the {32, 64} range.

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.

F.2 GROUPING CONFIGURATIONS FOR CLINICAL TASKS

In both clinical tasks, the configuration of input feature subsets (i.e., how we grouped input biomarkers into multivariate trajectories) was treated as a manually tuned hyperparameter. These groupings determine how GMAN combines individual graph representations prior to final prediction, and they affect both the expressivity and interpretability of the model.

In-Hospital Mortality (P12). We compared GMAN's performance under the following grouping strategies:

- **No grouping**: Each biomarker is in its own size one subset.
- **Respiratory** - one group of the biomarkers: *[FiO$_2$, PaO$_2$, PaCO$_2$, SaO$_2$, RespRate, pH, MechVent]* and the rest are singletons.
- **Metabolic Electrolytes** - one group of the biomarkers: *[Na, K, Mg, HCO3, Lactate, Glucose]* and the rest are singletons.
- **Liver Panel** - one group of the biomarkers: *[ALT, AST, ALP, Albumin, Bilirubin, Cholesterol]* and the rest are singletons.
- **Pathway-Based Grouping**: Biomarkers are organised based on their molecular or mechanistic roles, grouping them by their function in metabolism, homeostasis, or cellular composition: Energy metabolism [Glucose, Cholesterol, Lactate]; nitrogen waste clearance [BUN, Creatinine, Urine]; protein synthesis and enzymes [Albumin, ALT, AST]; liver function and cholestasis [ALP, Bilirubin]; acid–base balance [pH, HCO$_3$]; gas transport [PaO$_2$, PaCO$_2$, SaO$_2$, FiO$_2$]; mineral homeostasis [Na, K, Mg]; hematologic composition [HCT, Platelets, WBC]; cardiovascular dynamics [SysABP, DiasABP, MAP, NISysABP, NIDiasABP, NIMAP, HR]; respiratory mechanics [RespRate, MechVent]; and cardiac injury [TroponinI, TroponinT]. Global status indicators are grouped separately as [Temp, GCS].
- **Organ-System Grouping** Biomarkers are organized around organ systems and clinical monitoring domains (respiratory support, cardiovascular dynamics etc: oxygenation support [SaO$_2$, PaO$_2$, FiO$_2$, MechVent]; ventilation and acid–base balance [PaCO$_2$, RespRate,

pH, HCO_3]; cardiovascular dynamics [HR, MAP, SysABP, DiasABP, NIMAP, NISysABP, NIDiasABP]; perfusion and renal function [Urine, Lactate, Creatinine, BUN]; hepatobiliary function [Bilirubin, ALT, AST, ALP, Albumin]; inflammation and coagulation [WBC, Platelets]; electrolyte and oxygen-carrying capacity [Na, K, Mg, HCT]; metabolic reserve [Glucose, Cholesterol]; myocardial injury [TroponinI, TroponinT]; and global status indicators [Temp, GCS].

Crohn's Disease Onset. We evaluated several grouping configurations:

- **No grouping**: Each biomarker is in its own size one subset.
- **Biologically driven grouping** (see Appendix E.2): Biomarkers are grouped into 7 clinically coherent subsets (e.g., inflammation markers, immune cell subtypes, liver function).
- **Diagnostic Panel Grouping**: A clinically motivated grouping that mirrors standard blood test panels used in routine diagnostics.
- **Data-driven grouping**: Groupings are derived from clustering biomarkers based on the complementary signal in their attribution curves (see Section 4.1.1).
- **Merged coarse groupings**: Broad categories such as inflammation, haematology, and micronutrients.
- **Minimal Pairwise Interaction** : Emphasizes minimal yet informative combinations that capture key axes of immune, inflammatory, and metabolic variational proximity.

We report the AUPRC of different biomarker groupings for predicting the onset of CD in Table 4

G LLM USAGE

We relied on Large Language Models solely for grammar and spelling checks, without using them to generate or modify the scientific content.

Table 4: Performance using different biomarker subsets for CD prediction.

Grouping Strategy	**Biomarker Groups**	**AUPRC**
Flat grouping	Each biomarker is modeled independently: CRP, F-Cal, Leukocytes, Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils, Platelets, Hemoglobin, Iron, Folate, Vitamin B12, Vitamin D2+D3, ALAT, Albumin, Bilirubin.	79.66 ± 0.96
Biologically driven grouping	• White blood cell subtypes: Leukocytes, Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils • Inflammatory markers: CRP, F-Cal • Platelets • Haemoglobin • Iron • Vitamin and folate status: Folate, Vitamin B12, Vitamin D2 + D3 • Liver function markers: ALAT, Albumin, Bilirubin	83.93 ± 0.27
Diagnostic Panel Grouping	• Inflammatory markers: CRP, F-Cal • WBC count (main): Leukocytes, Neutrophils, Lymphocytes • WBC rare subtypes: Monocytes, Eosinophils, Basophils • Platelets + Hemoglobin: Platelets, Hemoglobin • Nutrient panel: Iron, Folate, Vitamin B12, Vitamin D2+D3 • Liver function test panel: ALAT, Albumin, Bilirubin	79.06 ± 3.4
Data-driven grouping	• Acute Inflammation Markers: CRP, F-Cal, Platelets • Immune–Iron Axis: Lymphocytes, Iron • Neutrophils • Monocytes • Eosinophils • Basophils • Hemoglobin • Folate • Vitamin B12 • Vitamin D • ALAT • Bilirubin • Albumin • Total Leukocytes	79.04 ± 7.8
Merged coarse groupings	• F-Cal (local inflammation) • Systemic immune/inflammation: CRP, Leukocytes, Neutrophils, Lymphocytes, Monocytes • Allergy-linked eosinophils/basophils • Hematological status: Platelets, Hemoglobin • Nutritional status: Iron, Folate, Vitamin B12, Vitamin D2+D3 • Hepatic status: ALAT, Albumin, Bilirubin	75.21 ± 2.3
Minimal Pairwise Interaction	• CRP + F-Cal • Leukocytes + Neutrophils • Lymphocytes + Monocytes • Eosinophils + Basophils • Platelets • Hemoglobin • Iron + Folate • Vitamin B12 + Vitamin D2+D3 • ALAT + Albumin • Bilirubin	81.50 ± 1.3