Communication

A Network-of-Networks Framework for Multi-Omics Data Integration and Analysis via Graph Neural Networks

Pietro Hiram Guzzi *

Posted Date: 18 May 2026

doi: 10.20944/preprints202605.1163.v1

Keywords: multi-omics integration; network of networks; graph neural networks; heterogeneous graphs; bioinformatics infrastructure; precision medicine

 doi:10.20944/preprints202605.1163.v1

Article

A Network-of-Networks Framework for Multi-Omics Data Integration and Analysis via Graph Neural Networks

Pietro Hiram Guzzi

Department of Medical and Surgical Sciences, Magna Graecia University of Catanzaro, Italy; hguzzi@unicz.it

Abstract

The integration of heterogeneous multi-omics data — spanning genomics, transcriptomics, proteomics, epigenomics, and metabolomics — remains one of the central open challenges in computational biology. Existing approaches either flatten omics layers into feature matrices, losing relational structure, or adopt multilayer network formalisms that treat layers as independent graphs coupled only by alignment edges. In this position paper we propose a fundamentally different data model: a *Network of Networks* (NoN), in which each node of a top-level graph is itself a complete graph, defined recursively. This recursive structure naturally encodes the hierarchical organisation of biological systems — from molecular interactions within an omics layer, through pathway-level modules, up to patient-level similarity networks — without collapsing any level of resolution. We formalise the NoN model with a rigorous recursive graph definition, describe a bioinformatics infrastructure built on top of it, and outline how heterogeneous Graph Neural Networks (GNNs) can operate across all levels of the hierarchy simultaneously. We argue that the NoN paradigm offers a principled, scalable, and biologically interpretable foundation for next-generation multi-omics analysis platforms, and we identify key research directions and open challenges that must be addressed to realise this vision.

Keywords: multi-omics integration; network of networks; graph neural networks; heterogeneous graphs; bioinformatics infrastructure; precision medicine

1. Introduction

High-throughput sequencing and mass-spectrometry technologies now routinely produce molecular profiles at multiple biological scales within the same cohort or even the same cell [5,18]. The resulting *multi-omics* datasets encode complementary views of cellular state: DNA sequence and copy-number variation (genomics), gene expression (transcriptomics), protein abundance and post-translational modification (proteomics), chromatin accessibility and methylation (epigenomics), and small-molecule concentrations (metabolomics). Integrating these layers promises a systems-level understanding of disease mechanisms that no single omics can provide [1,10].

Despite rapid methodological progress, most integration frameworks still treat omics data as *feature matrices* — rows are samples, columns are molecular features — and concatenate or project them into a shared latent space [5]. This representation discards the rich relational structure that exists *within* each omics layer (e.g., protein–protein interactions, gene co-expression modules, metabolic reaction networks) and *between* layers (e.g., transcription-factor regulation, enzymatic catalysis, methylation-expression quantitative trait loci).

Graph-based methods have emerged as a natural remedy. Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and Heterogeneous Graph Transformers (HGT) have been applied to multi-omics classification [4,13,20], biomarker discovery [8,9,17,22], and drug-target identification [11]. However, these methods typically construct a *single* heterogeneous graph in which omics layers are represented as disjoint node sets connected by inter-layer edges — a *multilayer* model that does not capture the internal structure of each layer as a first-class graph object.

 doi:10.20944/preprints202605.1163.v1

We argue that the correct abstraction is a *Network of Networks* (NoN): a graph $\mathcal{G}$ whose nodes are themselves graphs, defined recursively. This model has been studied in the context of infrastructure resilience and social networks [12], but has not been systematically applied to multi-omics bioinformatics. The NoN paradigm offers three key advantages over flat multilayer models:

1. **Recursive resolution**: molecular, pathway, cellular, and patient levels are all first-class graph objects, not collapsed into node features.
2. **Compositional GNN operators**: message passing can be applied at each level independently and then composed across levels, enabling hierarchical representation learning.
3. **Biological fidelity**: the recursive structure mirrors the actual hierarchical organisation of living systems.

This paper makes the following contributions:

- A formal recursive definition of the NoN data model for multi-omics data (Section 3).
- A reference bioinformatics infrastructure built on the NoN model, covering data ingestion, graph construction, storage, and GNN analysis (Section 4).
- A discussion of open challenges and a research agenda (Section 7).

2. Background and Related Work

2.1. Multi-Omics Integration

Multi-omics integration methods can be broadly classified into three families [1,5]: *early fusion* (concatenate feature matrices before modelling), *intermediate fusion* (learn a shared latent representation jointly), and *late fusion* (train modality-specific models and combine predictions). Each family has well-known trade-offs: early fusion ignores modality heterogeneity; late fusion may miss cross-omics interactions; intermediate fusion is computationally demanding and sensitive to missing data.

Similarity Network Fusion (SNF) [1] constructs a patient-similarity network per omics layer and fuses them iteratively, but operates only at the sample level and discards molecular-level topology. MOFA+ [1] decomposes multi-omics matrices into shared latent factors but again treats each layer as a flat matrix.

2.2. Graph Neural Networks for Multi-Omics

GNNs have been applied to multi-omics data with increasing sophistication. MOGONET [20] builds a patient-similarity graph per omics layer using GCN, then fuses predictions via a View Correlation Discovery Network. MoGCN [13] extends this to cancer subtype classification. MOHGCN [21] introduces specificity-aware heterogeneous GCN to model high-order interactions. MO-GCAN [7] combines GCN with attention for subtyping. GNNRAI [17] integrates biological knowledge graphs as structural priors into GNN message passing, with gradient-based explainability. Amogel [16] uses associative GNNs with multiple prior knowledge sources. BioNeuralNet [15] provides an end-to-end GNN pipeline specifically designed for network-structured multi-omics data.

Despite this progress, all existing methods construct a *single-level* graph — either a patient-similarity graph, a molecular interaction graph, or a heterogeneous graph mixing both — and do not exploit the recursive hierarchical structure of biological networks.

2.3. Network of Networks

The concept of a Network of Networks (or multilayer network) has been formalised in the complex-systems literature [12]. A multilayer network is defined as a set of graphs (layers) coupled by inter-layer edges. However, in the standard multilayer formalism, nodes across layers are aligned (the same entity appears in multiple layers), and the internal structure of each layer is not itself a graph object — it is simply a set of edges. Our NoN model differs fundamentally: each node *is* a graph, and this nesting is recursive, enabling arbitrary depth of hierarchical representation.

 doi:10.20944/preprints202605.1163.v1

3. The Recursive Network-of-Networks Model

3.1. Formal Definition

We begin with a standard graph definition and build the NoN recursively.

Definition 1 (Graph). *A* graph *is a triple $G = (\mathcal{V}, \mathcal{E}, \mathcal{F})$ where $\mathcal{V}$ is a finite set of nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges, and $\mathcal{F} : \mathcal{V} \cup \mathcal{E} \to \mathbb{R}^d$ is a feature function assigning d-dimensional real-valued attributes to nodes and edges.*

Definition 2 (Network of Networks). *A* Network of Networks *of depth $k \geq 0$ is defined recursively as follows.*

- ***Base case*** *($k = 0$): A* NoN *of depth 0 is a standard graph $G^{(0)} = (\mathcal{V}^{(0)}, \mathcal{E}^{(0)}, \mathcal{F}^{(0)})$ as in Definition 1.*
- ***Recursive case*** *($k \geq 1$): A* NoN *of depth k is a tuple*

$$G^{(k)} = \big(\mathcal{V}^{(k)},\ \mathcal{E}^{(k)},\ \{G_v^{(k-1)}\}_{v \in \mathcal{V}^{(k)}},\ \mathcal{F}^{(k)}\big),$$

where

  - *$\mathcal{V}^{(k)}$ is a finite set of* meta-nodes*;*
  - *$\mathcal{E}^{(k)} \subseteq \mathcal{V}^{(k)} \times \mathcal{V}^{(k)}$ is a set of* inter-network edges*;*
  - *for each meta-node $v \in \mathcal{V}^{(k)}$, $G_v^{(k-1)}$ is a* NoN *of depth $k-1$ called the* inner network *of v; and*
  - *$\mathcal{F}^{(k)}$ assigns features to meta-nodes and inter-network edges, where the feature of a meta-node v may be derived from a summary (readout) of $G_v^{(k-1)}$.*

Remark 1. *Definition 2 subsumes the standard multilayer network as a special case with $k = 1$, where each meta-node contains a single-layer graph and inter-network edges encode inter-layer coupling. The key distinction is that in our model the inner networks are* first-class objects *that participate in GNN message passing, not merely sets of node features.*

3.2. Instantiation for Multi-Omics Data

We instantiate the NoN model at three levels of biological organisation (Figure 1).

Level 0 — Molecular graphs ($k = 0$).

Each omics layer $\ell \in \{\text{genome, transcriptome, proteome, epigenome,metabolome}\}$ is represented as a molecular graph $G_\ell^{(0)} = (\mathcal{V}_\ell^{(0)}, \mathcal{E}_\ell^{(0)}, \mathcal{F}_\ell^{(0)})$. Node sets and edge semantics are layer-specific:

- *Transcriptome*: nodes are genes; edges encode co-expression (Pearson $r > \theta$) or regulatory interactions (TF→target from ChIP-Atlas).
- *Proteome*: nodes are proteins; edges encode physical interactions (STRING, BioGRID).
- *Metabolome*: nodes are metabolites; edges encode enzymatic reactions (KEGG).
- *Epigenome*: nodes are CpG sites or chromatin regions; edges encode co-methylation or chromatin loop contacts (Hi-C).
- *Genome*: nodes are genomic loci or genes; edges encode linkage disequilibrium or copy-number co-variation.

Node features $\mathcal{F}^{(0)}$ encode measured omics values (expression levels, methylation β-values, protein abundances, etc.).

Level 1 — Pathway/module graphs ($k = 1$).

Each biological pathway or functional module p (from KEGG, Reactome, or MSigDB) is a meta-node whose inner network $G_p^{(0)}$ is the subgraph of the molecular graph induced by the genes/proteins in that pathway. Inter-network edges at level 1 connect pathways that share member genes or that are

 doi:10.20944/preprints202605.1163.v1

linked by known regulatory crosstalk. The feature of meta-node p is a readout (e.g., mean pooling or attention pooling) of the node embeddings in $G_p^{(0)}$.

Level 2 — Patient/sample graphs ($k = 2$).

Each patient or biological sample s is a meta-node whose inner network $G_s^{(1)}$ is the pathway graph constructed from that patient's omics profiles. Inter-network edges at level 2 connect patients by phenotypic or molecular similarity (e.g., cosine similarity of pathway embeddings, clinical covariates). This level supports patient stratification, subtype discovery, and survival prediction.

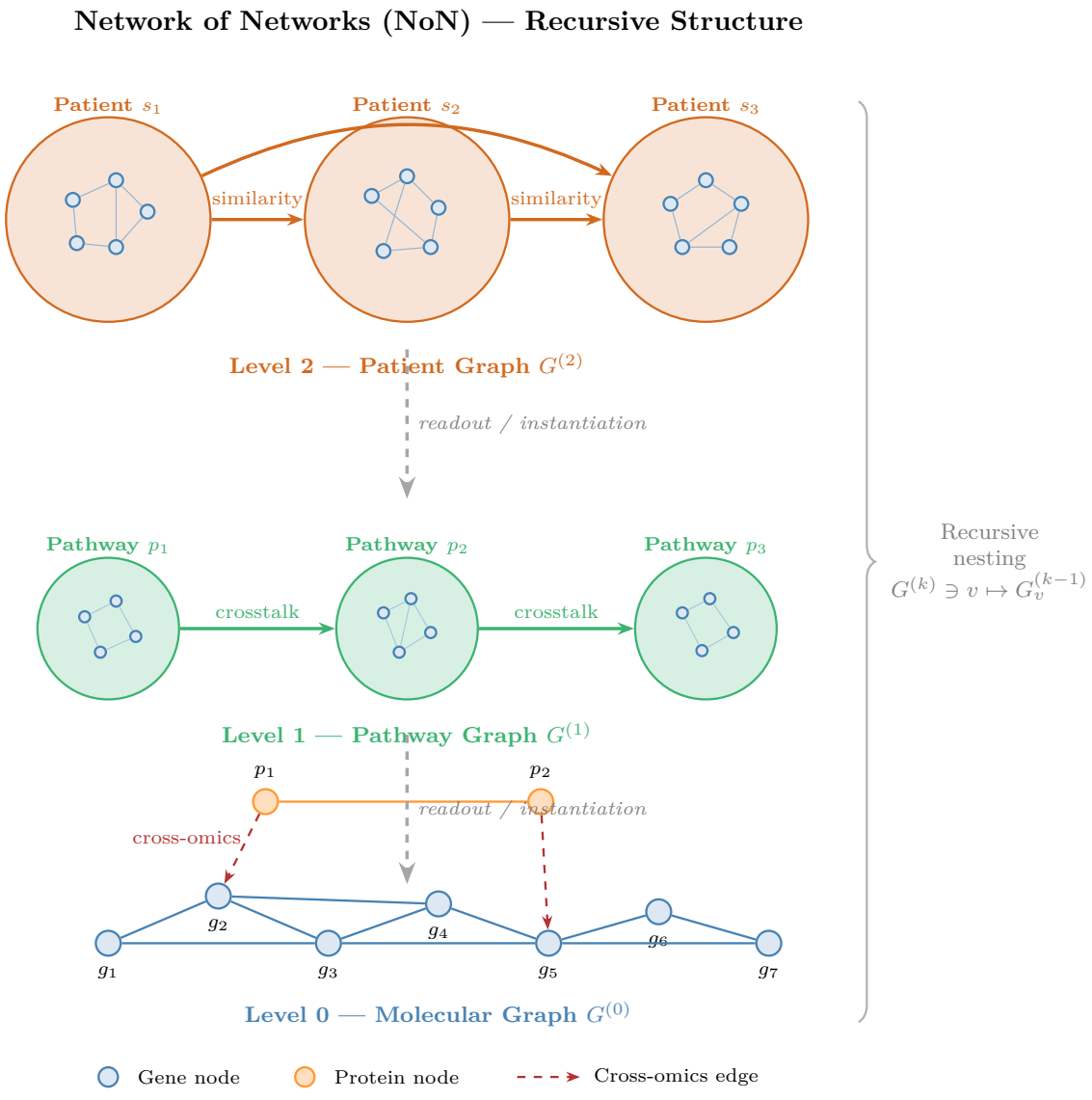

Figure 1. Recursive Network of Networks (NoN) structure for multi-omics data. *Bottom (Level 0)*: Molecular graphs $G^{(0)}$ encode intra-layer relationships (gene co-expression, PPI) and cross-omics regulatory edges (dashed red arrows). *Middle (Level 1)*: Pathway meta-nodes $G^{(1)}$ whose inner networks are induced subgraphs of the molecular layer; inter-pathway crosstalk edges connect functionally related modules. *Top (Level 2)*: Patient meta-nodes $G^{(2)}$ whose inner networks are the pathway graphs of each individual; patient-similarity edges form the population-level graph. Vertical dashed arrows indicate the readout/instantiation operations that propagate information across levels. The right-hand brace formalises the recursive mapping $G^{(k)} \ni v \mapsto G_v^{(k-1)}$.

 doi:10.20944/preprints202605.1163.v1

Proposition 1. *The* NON *of depth 2 described above contains, as special cases, the patient-similarity networks used by MOGONET [20], the pathway-level graphs used by the multilevel GNN of Yan et al. [23], and the heterogeneous molecular graphs used by MOHGCN [21].*

3.3. Cross-Level Edges

In addition to intra-level edges, the NON supports *cross-level edges* that connect nodes at different depths:

- *Molecule → Pathway*: a gene node at level 0 is linked to the pathway meta-node at level 1 that contains it (membership edge).
- *Pathway → Patient*: a pathway meta-node at level 1 is linked to the patient meta-node at level 2 whose inner network contains it (instantiation edge).
- *Cross-omics*: edges between molecular nodes of different omics layers (e.g., TF gene → target protein, CpG → gene expression QTL).

Cross-level and cross-omics edges are typed, enabling heterogeneous GNN operators to apply type-specific transformations.

4. Proposed Infrastructure

Figure 2 illustrates the overall infrastructure, which comprises four layers: data ingestion, graph construction, storage, and GNN analysis.

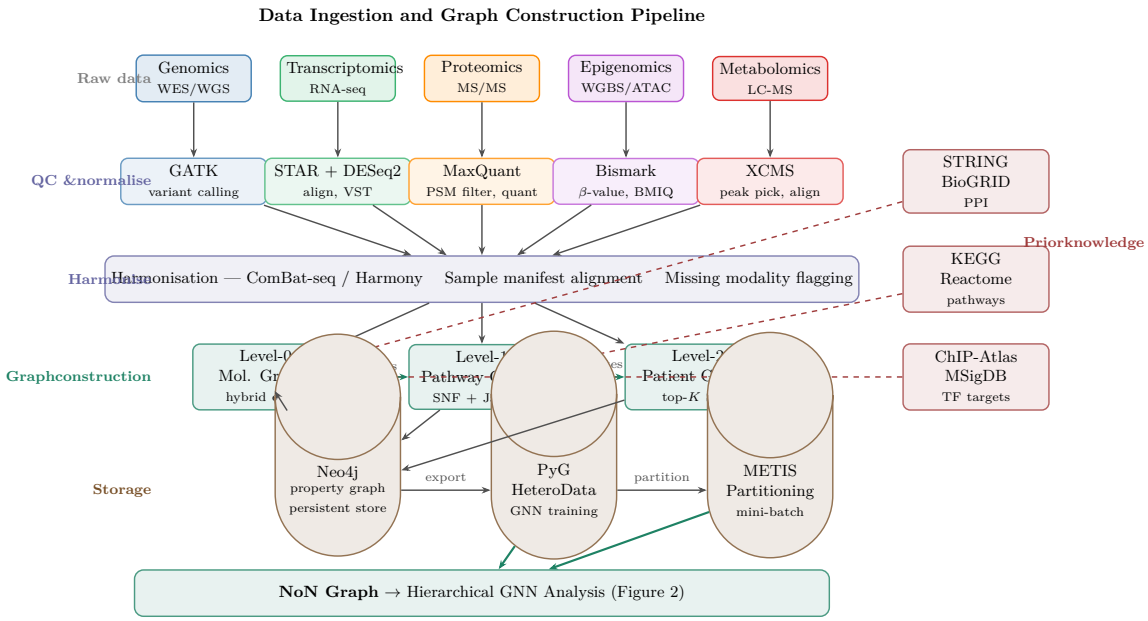

Figure 2. Data ingestion and graph construction pipeline. Raw omics data from five modalities undergo modality-specific QC and normalisation, followed by cross-modality harmonisation (ComBat-seq / Harmony). Graph construction proceeds bottom-up: level-0 molecular graphs use hybrid data-driven and knowledge-driven edges (STRING, ChIP-Atlas, KEGG); level-1 pathway graphs are induced from MSigDB/Reactome membership; level-2 patient graphs are built by SNF over pathway embeddings. The resulting NON is stored in Neo4j and exported to PyG `HeteroData` for GNN training.

4.1. Data Ingestion and Harmonisation

Raw omics data arrive from heterogeneous platforms and must be harmonised before graph construction. Each omics modality undergoes modality-specific quality control and normalisation:

- *RNA-seq*: read alignment (STAR/HISAT2), count normalisation (DESeq2 variance-stabilising transformation or TMM), batch correction (ComBat-seq).

 doi:10.20944/preprints202605.1163.v1

- *Proteomics*: peptide-spectrum match filtering, protein inference, quantile normalisation, missing-value imputation (k-NN or MinProb).
- *DNA methylation*: bisulfite conversion efficiency filtering, β-value computation, BMIQ normalisation.
- *Metabolomics*: peak picking, retention-time alignment, log-transformation, Pareto scaling.
- *WES/WGS*: variant calling (GATK), annotation (VEP), MAF matrix construction.

Sample identifiers are harmonised across modalities using a master sample manifest, and missing modalities are flagged for downstream imputation.

4.2. Graph Construction

Given harmonised omics matrices, the graph construction module builds the NoN bottom-up:

Level-0 molecular graphs.

Intra-layer edges are constructed by combining data-driven and knowledge-driven approaches [17, 22]:

- *Data-driven*: Pearson or Spearman correlation above a threshold θ (tuned by permutation testing); mutual information; ARACNE for regulatory networks.
- *Knowledge-driven*: curated interactions from STRING (PPI), BioGRID (PPI), ChIP-Atlas (TF→target), KEGG (metabolic reactions), and Hi-C contact maps (chromatin topology).

Hybrid edges are assigned weights combining data-driven scores and knowledge-based confidence scores.

Level-1 pathway graphs.

Pathway membership is retrieved from MSigDB (C2, C5 collections) and Reactome. Each pathway meta-node is instantiated as the induced subgraph of the molecular graph. Inter-pathway edges are added when two pathways share $\geq \delta$ member genes (Jaccard similarity) or are linked in the Reactome hierarchy [2,3].

Level-2 patient graphs.

Patient-similarity edges are constructed using Similarity Network Fusion (SNF) [1] applied to pathway-level embeddings, yielding a fused patient similarity network. Edge weights are thresholded to retain the top-K neighbours per patient.

4.3. Storage and Query Layer

The NoN is stored in a **Neo4j** property graph database, which natively supports typed nodes and edges and enables Cypher queries for subgraph retrieval. For GNN training, subgraphs are exported to **PyTorch Geometric** `HeteroData` objects, which support heterogeneous node and edge types with type-specific feature tensors. Large-scale graphs are partitioned using METIS-based graph partitioning for mini-batch training.

4.4. GNN Analysis Layer

The GNN analysis layer operates on the NoN through a hierarchical message-passing scheme (Figure 3).

 doi:10.20944/preprints202605.1163.v1

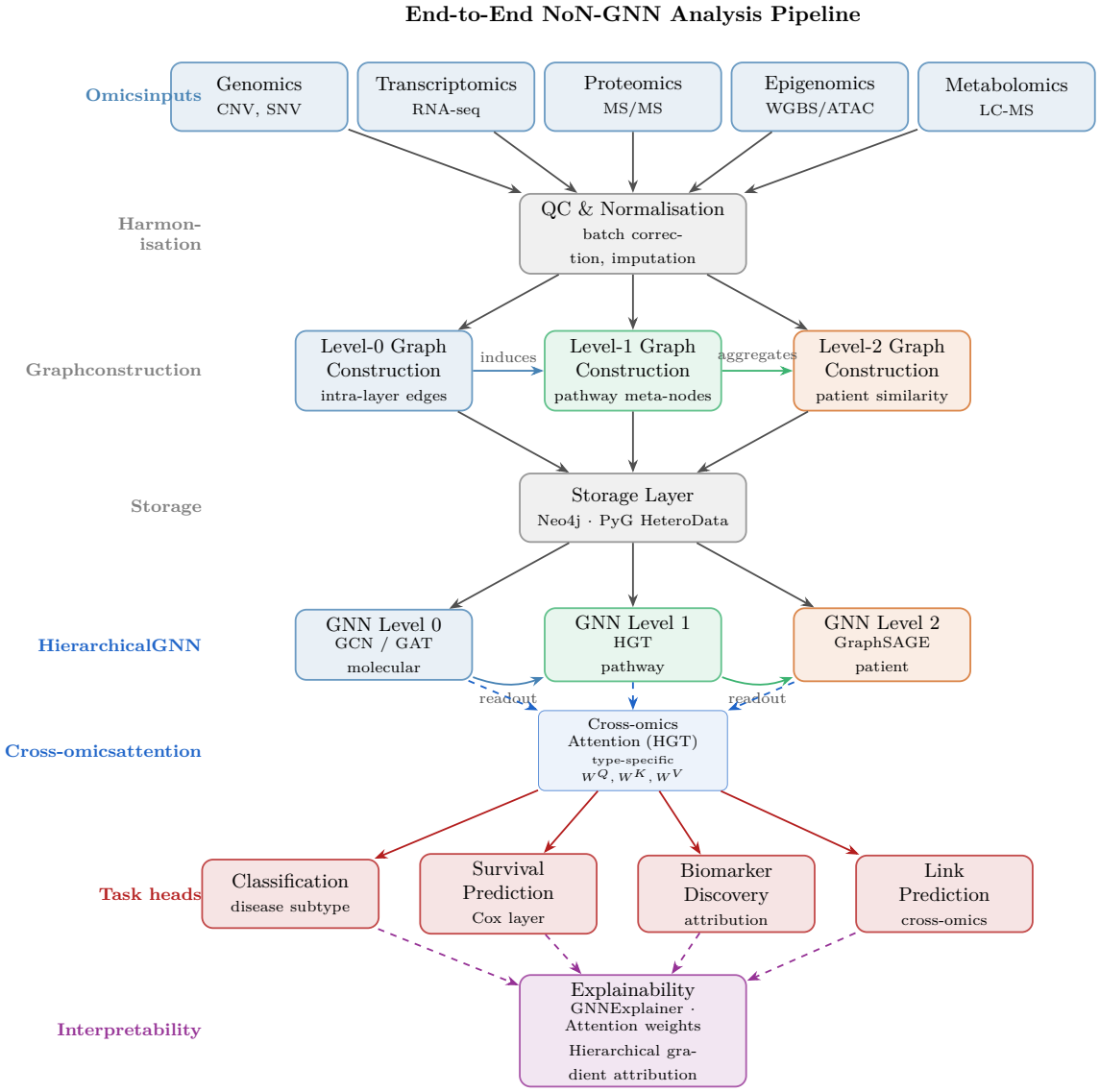

Figure 3. End-to-end NoN-GNN analysis pipeline. Harmonised omics data feed three parallel graph construction modules (levels 0–2). Hierarchical GNNs (GCN/GAT at level 0, HGT at level 1, GraphSAGE at level 2) perform message passing within each level; readout functions propagate embeddings upward. A cross-omics attention module (HGT) integrates information across omics types. Task-specific heads support classification, survival prediction, biomarker discovery, and link prediction. Explainability is provided by GNNExplainer, attention weight extraction, and hierarchical gradient attribution.

Level-0 message passing.

Within each molecular graph $G_\ell^{(0)}$, a layer-specific GNN (GCN, GAT, or GraphSAGE) computes node embeddings:

$$\mathbf{h}_{v,\ell}^{(0)} = \mathrm{GNN}_\ell\big(\mathbf{x}_{v,\ell},\ \{\mathbf{x}_{u,\ell} : u \in \mathcal{N}(v)\}\big),$$

where $\mathbf{x}_{v,\ell}$ is the feature vector of node v in layer ℓ and $\mathcal{N}(v)$ denotes its neighbours.

Level-1 readout and message passing.

A differentiable readout function (attention pooling or hierarchical pooling) aggregates level-0 embeddings into pathway meta-node representations [6]:

$$\mathbf{h}_p^{(1)} = \mathrm{READOUT}\Big(\{\mathbf{h}_{v,\ell}^{(0)} : v \in G_p^{(0)}\}\Big).$$

 doi:10.20944/preprints202605.1163.v1

A second GNN then propagates information across the pathway graph:

$$\mathbf{h}_p^{(1)\prime} = \mathrm{GNN}_{\mathrm{path}}\Big(\mathbf{h}_p^{(1)},\ \{\mathbf{h}_q^{(1)} : q \in \mathcal{N}(p)\}\Big).$$

Level-2 readout and message passing.

Patient embeddings are obtained by aggregating pathway embeddings:

$$\mathbf{h}_s^{(2)} = \mathrm{READOUT}\Big(\{\mathbf{h}_p^{(1)\prime} : p \in G_s^{(1)}\}\Big),$$

followed by a patient-level GNN operating on the patient similarity network.

Cross-omics attention.

Cross-omics edges are handled by a Heterogeneous Graph Transformer (HGT) [14] that applies type-specific attention:

$$\mathrm{Attention}(v, e, u) = \frac{\left(\mathbf{W}^{Q}_{\tau(v)}\,\mathbf{h}_v\right)^{\top}\mathbf{W}^{K}_{\phi(e)}\left(\mathbf{W}^{V}_{\tau(u)}\,\mathbf{h}_u\right)}{\sqrt{d}},$$

where $\tau(v)$ and $\phi(e)$ denote the type of node v and edge e, respectively.

Task heads.

The final patient embeddings $\mathbf{h}_s^{(2)}$ are fed into task-specific heads:

- *Classification*: softmax over disease subtypes or clinical outcomes.
- *Survival prediction*: Cox proportional hazards layer.
- *Biomarker discovery*: gradient-based attribution (GNNExplainer, Integrated Gradients) to identify molecular nodes most influential for the prediction.
- *Link prediction*: inner product of node embeddings to discover novel cross-omics regulatory edges.

4.5. Explainability

Biological interpretability is a first-class requirement. The framework supports three complementary explainability mechanisms:

1. **Attention weights**: HGT attention scores identify which cross-omics edges and pathway connections drive each prediction.
2. **GNNExplainer**: identifies the minimal subgraph and node feature subset that maximally explains a prediction [17].
3. **Hierarchical attribution**: gradients are back-propagated through all three levels of the NoN, yielding attribution scores at the molecular, pathway, and patient levels simultaneously.

5. Comparison with Existing Methods

Table 1 compares the proposed NoN framework with representative existing multi-omics GNN methods across six dimensions.

 doi:10.20944/preprints202605.1163.v1

Table 1. Comparison of multi-omics GNN integration frameworks. **Graph model**: type of graph representation used. **Levels**: number of biological abstraction levels modelled. **Prior knowledge**: whether curated biological knowledge is incorporated as graph structure. **Interpretability**: mechanism for biological explanation. **Scalability**: approach to handling large graphs. **Missing omics**: strategy for incomplete modality data. ✓= supported; ○ = partial; × = not supported.

Method	Graph model	Levels	Prior knowledge	Interpretability	Scalability	Missing omics
MOGONET [20]	Patient similarity (per omics)	1	×	×	Mini-batch	○
MoGCN [13]	Patient similarity	1	×	×	Full graph	×
MOHGCN [21]	Heterogeneous	1	○	○	Mini-batch	○
MO-GCAN [7]	Heterogeneous + attention	1	×	Attention	Full graph	×
GNNRAI [17]	Knowledge graph	1	✓	GNNExplainer	Mini-batch	○
Amogel [16]	Associative graph	1	✓	Attention	Full graph	×
BioNeuralNet [15]	Network-structured	1	✓	Gradient	Mini-batch	○
Multilevel GNN [23]	Hierarchical (2 levels)	2	✓	Gradient	Mini-batch	○
NoN (proposed)	**Recursive NoN**	**3+**	✓	**Multi-level**	**Hierarchical**	✓

6. Use Case: Cancer Subtype Discovery

To illustrate the NoN framework in a concrete biomedical scenario, we describe its application to cancer molecular subtype discovery using TCGA multi-omics data (Figure 4).

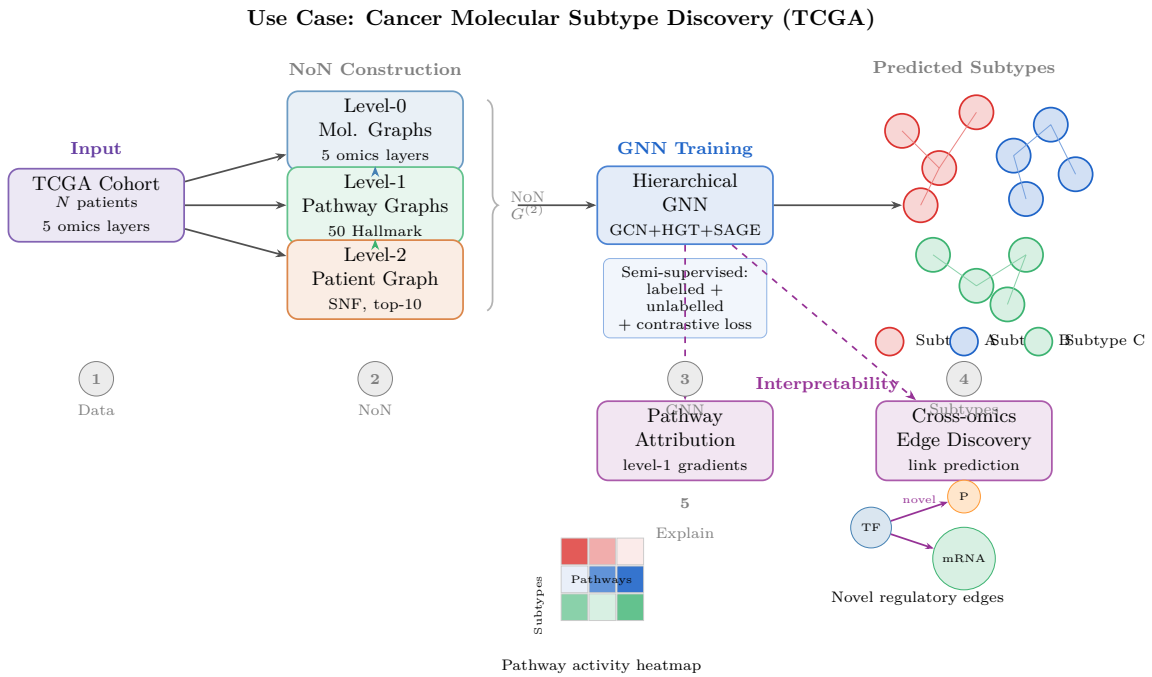

Figure 4. Use case: cancer molecular subtype discovery with NoN-GNN. *Step 1*: A TCGA multi-omics cohort (5 omics layers) is ingested. *Step 2*: The NoN $G^{(2)}$ is constructed bottom-up (molecular → pathway → patient). *Step 3*: A hierarchical GNN is trained in a semi-supervised fashion using labelled and unlabelled patients with a contrastive loss. *Step 4*: Patients are assigned to molecular subtypes (A, B, C) based on their level-2 embeddings. *Step 5*: Pathway attribution (level-1 gradients) and cross-omics link prediction reveal mechanistic signatures and novel regulatory interactions underlying each subtype.

 doi:10.20944/preprints202605.1163.v1

Data.

We consider five omics layers for a cohort of N patients: mRNA expression (RNA-seq), DNA methylation (450K array), copy-number variation (SNP array), miRNA expression, and somatic mutation profiles.

Graph construction.

1. *Level 0*: Five molecular graphs are constructed — one per omics layer — using the hybrid edge construction described in Section 4. Cross-omics edges connect TF genes to their target genes (ChIP-Atlas), CpG sites to their associated genes (methylation QTL), and miRNA nodes to their predicted target mRNAs (miRTarBase).
2. *Level 1*: Pathway meta-nodes are instantiated from the Hallmark gene set collection (MSigDB, 50 pathways). Each pathway meta-node aggregates the molecular nodes of all five omics layers that belong to that pathway.
3. *Level 2*: A patient similarity network is constructed by SNF over pathway-level embeddings, retaining the top-10 neighbours per patient.

GNN training.

The hierarchical GNN is trained in a semi-supervised fashion: a fraction of patients have known subtype labels (from TCGA clinical annotations); the remainder are unlabelled. The model is trained to minimise a cross-entropy loss on labelled patients while simultaneously optimising a graph contrastive loss on the patient similarity network [19].

Expected outcomes.

The NoN framework is expected to:

- Recover established molecular subtypes (e.g., TCGA breast cancer PAM50 subtypes) with accuracy competitive with or exceeding single-level GNN baselines.
- Identify pathway-level signatures (level-1 attribution) that explain subtype differences, providing mechanistic insight beyond gene lists.
- Discover novel cross-omics regulatory edges (link prediction) that may represent subtype-specific regulatory mechanisms.

7. Discussion

7.1. Advantages of the NoN *Paradigm*

The recursive NoN model offers several advantages over existing approaches. First, it preserves relational structure at every biological scale, avoiding the information loss inherent in feature-matrix representations. Second, the hierarchical GNN naturally implements a form of *biological inductive bias*: message passing at the molecular level captures local molecular interactions, while pathway-level aggregation captures functional module activity, and patient-level propagation captures population-level patterns. Third, the recursive definition is *compositional*: new omics layers or biological levels can be added without redesigning the entire model.

7.2. Open Challenges

Scalability.

A full NoN for a genome-wide multi-omics dataset may contain millions of nodes and billions of potential edges. Efficient mini-batch training with neighbour sampling (GraphSAGE-style) is essential, but must be adapted to the hierarchical structure to avoid breaking the recursive aggregation. Hierarchical graph partitioning (e.g., METIS applied level by level) is a promising direction.

 doi:10.20944/preprints202605.1163.v1

Missing modalities.

In practice, not all patients have all omics layers profiled. The NON model must handle missing inner networks gracefully. Possible strategies include: (i) imputing missing molecular graphs from available layers using cross-omics link prediction; (ii) masking missing meta-nodes during training with a modality-dropout regularisation scheme; (iii) learning a prior distribution over inner networks using a variational graph autoencoder.

Graph construction sensitivity.

The topology of level-0 molecular graphs depends on the choice of correlation threshold θ, the set of knowledge databases used, and the normalisation pipeline. Sensitivity analyses and benchmarking against held-out biological annotations (e.g., known disease genes from DisGeNET, drug targets from DGIdb) are necessary to validate graph quality.

Over-smoothing.

Deep GNNs applied to dense biological graphs are prone to over-smoothing, where node embeddings converge to indistinguishable representations. Residual connections, jumping knowledge networks, and limiting depth to 2–3 layers per level are standard mitigations, but their interaction with the hierarchical NON structure requires further study.

Evaluation and benchmarking.

Standard multi-omics benchmarks (TCGA, METABRIC, CPTAC) provide labelled data for classification tasks, but evaluation of biomarker discovery and link prediction requires dedicated biological validation (e.g., CRISPR perturbation screens, proteomics validation of predicted interactions). A community benchmark for NON-based multi-omics methods is needed.

7.3. Research Agenda

We identify the following priority research directions:

1. **Scalable hierarchical mini-batching**: develop sampling strategies that respect the recursive NON structure.
2. **Variational NON autoencoders**: learn generative models of inner networks to handle missing modalities and enable counterfactual reasoning.
3. **Temporal NON**: extend the model to longitudinal multi-omics data by adding a temporal dimension to inter-network edges.
4. **Single-cell NON**: instantiate the model at single-cell resolution, where level-0 graphs are cell-cell communication networks and level-1 graphs are tissue-level cell-type graphs.
5. **Foundation model pre-training**: pre-train a universal NON GNN on large public multi-omics repositories (TCGA, GTEx, ENCODE) and fine-tune on task-specific cohorts.

8. Conclusions

We have presented a vision for a next-generation multi-omics bioinformatics infrastructure grounded in a formally defined recursive Network of Networks (NON) data model. By treating each biological entity — from individual molecules to pathways to patients — as a first-class graph object, the NON paradigm enables hierarchical Graph Neural Networks to learn representations that are simultaneously molecularly precise, functionally interpretable, and clinically actionable. We have described the key components of the proposed infrastructure, formalised the recursive graph model, outlined a concrete use case in cancer subtype discovery, and identified the principal open challenges that must be addressed to realise this vision. We believe the NON framework represents a principled and timely step towards truly integrative, systems-level computational biology.

Acknowledgments: The author thanks the open-source communities behind PyTorch Geometric, Neo4j, and the TCGA/GTEx data consortia.

 doi:10.20944/preprints202605.1163.v1

References

1. Francis E. Agamah, Jumamurat R. Bayjanov, Anna Niehues, et al. Computational approaches for network-based integrative multi-omics analysis. *Frontiers in Molecular Biosciences*, 9:967205, 2022.
2. Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Cloud4snp: distributed analysis of snp microarray data on the cloud. In *Proceedings of the international conference on bioinformatics, computational biology and biomedical informatics*, pages 468–475, 2013.
3. Giuseppe Agapito, Marianna Milano, Pietro Hiram Guzzi, and Mario Cannataro. Extracting cross-ontology weighted association rules from gene ontology annotations. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(2):197–208, 2015.
4. Fadi Alharbi, Aleksandar Vakanski, Boyu Zhang, et al. Comparative analysis of multi-omics integration using graph neural networks for cancer classification. *IEEE Access*, 2025.
5. Jenna L. Ballard, Zexuan Wang, Wenrui Li, et al. Deep learning-based approaches for multi-omics data integration and analysis. *BioData Mining*, 17:1–28, 2024.
6. Mario Cannataro, Pietro H Guzzi, and Pierangelo Veltri. Impreco: Distributed prediction of protein complexes. *Future Generation Computer Systems*, 26(3):434–440, 2010.
7. Yifan Dou and Golrokh Mirzaei. MO-GCAN: multi-omics integration based on graph convolutional and attention networks. *Bioinformatics*, 2025.
8. Alba Gutiérrez-Sacristán, Arnaud Serret-Larmande, Meghan R Hutch, Carlos Sáez, Bruce J Aronow, Surbhi Bhatnagar, Clara-Lea Bonzel, Tianxi Cai, Batsal Devkota, David A Hanauer, et al. Hospitalizations associated with mental health conditions among adolescents in the us and france during the covid-19 pandemic. *JAMA network open*, 5(12):e2246548, 2022.
9. Pietro H Guzzi and Mario Cannataro. μ-cs: An extension of the tm4 platform to manage affymetrix binary data. *BMC bioinformatics*, 11(1):315, 2010.
10. Pietro Hiram Guzzi, Ugo Lomoio, Barbara Puccio, and Pierangelo Veltri. Structural analysis of sars-cov-2 spike protein variants through graph embedding. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 12(1):3, 2022.
11. Wei Jiang, Weicai Ye, Xiaoming Tan, et al. Network-based multi-omics integrative analysis methods in drug discovery: a systematic review. *BioData Mining*, 2025.
12. Bohyun Lee, Shuo Zhang, Aleksandar Poleksic, et al. Heterogeneous multi-layered network model for omics data integration and analysis. *Frontiers in Genetics*, 10:1381, 2020.
13. Xiao Li, Jie Ma, Ling Leng, et al. MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Frontiers in Genetics*, 13:806842, 2022.
14. Anjun Ma, Xiaoying Wang, Jingxiang Li, et al. Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications*, 14:964, 2023.
15. Vicente Ramos, Sundous Hussein, Mohamed Abdel-Hafiz, et al. BioNeuralNet: A graph neural network based multi-omics network data analysis tool. *arXiv preprint arXiv:2507.20440*, 2025.
16. Chia Yan Tan, H. Ong, Chern Hong Lim, et al. Amogel: a multi-omics classification framework using associative graph neural networks with prior knowledge for biomarker identification. *BMC Bioinformatics*, 2025.
17. Rohit K. Tripathy, Zachary Frohock, Hong Wang, et al. Effective integration of multi-omics with prior knowledge to identify biomarkers via explainable graph neural networks. *NPJ Systems Biology and Applications*, 2025.
18. Nektarios A. Valous, Ferdinand Popp, Inka Zörnig, et al. Graph machine learning for integrated multi-omics analysis. *British Journal of Cancer*, 2024.
19. Jiahui Wang, Nanqing Liao, Xiaofei Du, et al. A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks. *BMC Genomics*, 2024.
20. Tongxin Wang, Wei Shao, Zhi Huang, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12:3445, 2021.
21. Wenhao Wu, Shudong Wang, Yuanyuan Zhang, et al. MOHGCN: A trustworthy multi-omics data integration framework based on specificity-aware heterogeneous graph convolutional neural networks for disease diagnosis. *Expert Systems with Applications*, 2025.
22. Shunxin Xiao, Huibin Lin, Conghao Wang, et al. Graph neural networks with multiple prior knowledge for multi-omics data analysis. *IEEE Journal of Biomedical and Health Informatics*, 2023.

 doi:10.20944/preprints202605.1163.v1

23. Hongxi Yan, Dawei Weng, Dongguo Li, et al. Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration. *Briefings in Bioinformatics*, 2024.