

Graph in Graph (GiG): A novel graph AI framework for integrating and interpreting medical and omics data

Heming Zhang¹, Yifei Lu¹, Kaiwen Fang¹, Zixi Xu¹, Vaha Akabry Moghaddam², Ping An², Shiow Jin², Mary Wojczynski², Michael Province^{2#}, Fuhai Li^{1,3#}

¹Institute for Informatics (I2), ²Division of Statistical Genomics, Department of Genetics,

³Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis School of Medicine, St. Louis, MO 63110, USA.

#Correspondence: mprovince@wustl.edu; fuhai.li@wustl.edu

Abstract: Medical records and omics data are rapidly becoming standard in healthcare settings, which characterize the whole-person from dysfunctional molecules to phenotypes, and thus offer potential for precise disease diagnosis and target discovery. Whereas, it remains an open problem to systematically integrate and interpret medical record and omics data of individual patients. In this study, for the first time, we propose a novel graph AI model framework, Graph in Graph (GiG), to integrate and interpret the whole-person medical and omic datasets. Specifically, the medical record data is modeled using a person-phenotype graph, followed by omics signaling graphs of individual patients, which enables the integration of information learned from omic-signaling graph and medical phenotype features to characterize individual patients and to prioritize important omic biomarkers and phenotypes. As an exploratory study, we applied and evaluated the GiG model to study the type 2 diabetes (T2D) and pre-T2D vs healthy using the Long Life Family Study (LLFS) cohort, which enrolls families with exceptional longevity to uncover biological mechanisms of healthy aging with medical and omics data. The evaluation results showed that GiG not only achieve a high prediction but also can interpret the prediction by ranking the essential clinical and omic biomarkers. The GiG framework can be applied to other studies by effectively integrating and interpreting medical and omic datasets for disease diagnosis and pathogenesis discovery.

1. Introduction

The availability of both medical records and omics data is rapidly becoming standard in healthcare settings, offering immense potential for precise disease diagnosis and target discovery. However, integrating these distinct modalities presents unique computational hurdles. Medical phenotypic data are typically sparse and consist of mixed data types, making it difficult to align patients or accurately estimate phenotypic similarities across diverse clinical conditions. Conversely, omics data are high-dimensional, capturing complex signaling pathways and molecular functions. While integration of these datasets can holistically characterize individual patients to identify key phenotypic factors and genetic biomarkers, systematic integration and interpretation remain in their early stages and an open problem^{1,2}. Thus, novel AI models are required to mitigate these new and unique data challenges^{1,2}.

The National Institution of Aging (NIA) Long Life Family Study (LLFS)³ project was established in 2005 to study EL and systematically identify the genetic and environmental factors that contribute to exceptional longevity. LLFS is an expansive study spanning multiple centers and nations, focusing on the genetic and environmental factors contributing to extraordinary longevity and healthy aging. LLFS recruited families with exceptional longevity, providing a unique resource to identify protective factors for healthy aging and diverse chronic diseases, like type 2 diabetes (T2D), pre-T2D and alzheimer's diseases. In LLFS, comprehensive data are collected including both medical phenotypes, like demographic information, physical measurements, cognitive abilities, daily activities, and various health markers, and multi-omics data from blood samples. In the LLFS project, one challenge is to integrate and interpret the phenotypic (demographic and clinical) data and omic data to identify the key factors and biomarkers associated with EL.

Artificial intelligence (AI) models have been widely used in the healthcare field^{4–16}. Especially, the Graph Neural Networks (GNNs)^{17–23}, have revolutionized the field of learning with graph-structured data and empirically achieved the current state-of-art performance in various graph learning tasks, including node classification, link prediction, and graph classification²⁴. Broadly, GNNs follow a recursive neighborhood aggregation scheme where the node features from the neighborhood of each node are aggregated to update the node's feature. Such framework allows GNNs, as one of the most prominent AI models, to be more interpretable compared with other non-GNN AI models^{25–28}. The GNNs have been widely used in modeling signaling pathways with omics^{14,29–35}, based on three advantages: 1) They can naturally convert graph-

structured data into a numerical space, adeptly handling the sparse and mixed data types of phenotypic data; 2) They model omics data in a biologically meaningful way by mapping this data onto signaling pathways or protein-protein interactions; 3) They facilitate the identification of crucial protective factors, which can be framed as the interpretable target-ranking and subnetwork discovery challenges tackled by GNNs through attention analysis. These capabilities are further manifested in their applications in biomedicine, where GNNs are reshaping our comprehension of diseases by elucidating the intricate connections among genes and cells.

In this study, for the first time, we presented a novel deep learning framework, ***Graph in Graph (GiG phenotype-genotype network)***, to resolve the aforementioned data analysis challenges in integrating and interpreting medical phenotype and omics datasets. Specifically, we first convert data frame format phenotypic data into a phenotype graph/network, which links individual T2D patients to phenotypes. Also, each patient will contain an embedded genotype graph in the constructed phenotype network. Thus, all phenotypic nodes and genotypes for each patient can be used as the input of an attention-based model (GAT¹⁹ and Graph Transformer¹⁸) to predict the types of individual patients. The advantage of using an attention-based model is that the attention scores of the phenotypes can indicate the importance of phenotypes and gene biomarkers, which makes the model interpretable and facilitates the clinical decision-making of longevity. In the following sections, the data analysis results and details of the methodology are presented.

2. Results

2.1 GiG architecture and analysis workflow overview

Figure 1a illustrates the architecture of the GiG framework, which employs a dual-graph transformer mechanism to integrate gene-level and patient-level features for classification (Section 4.3). As shown in Figure 1b, we further established a downstream validation pipeline to link the top 142 prioritized genes to pharmacological evidence. By cross-referencing these genes with ChEMBL and conducting citation-consistency checks, we generated a curated ledger of candidate target-drug pairs, which was then overlaid onto the GiG core network (Section 4.5).

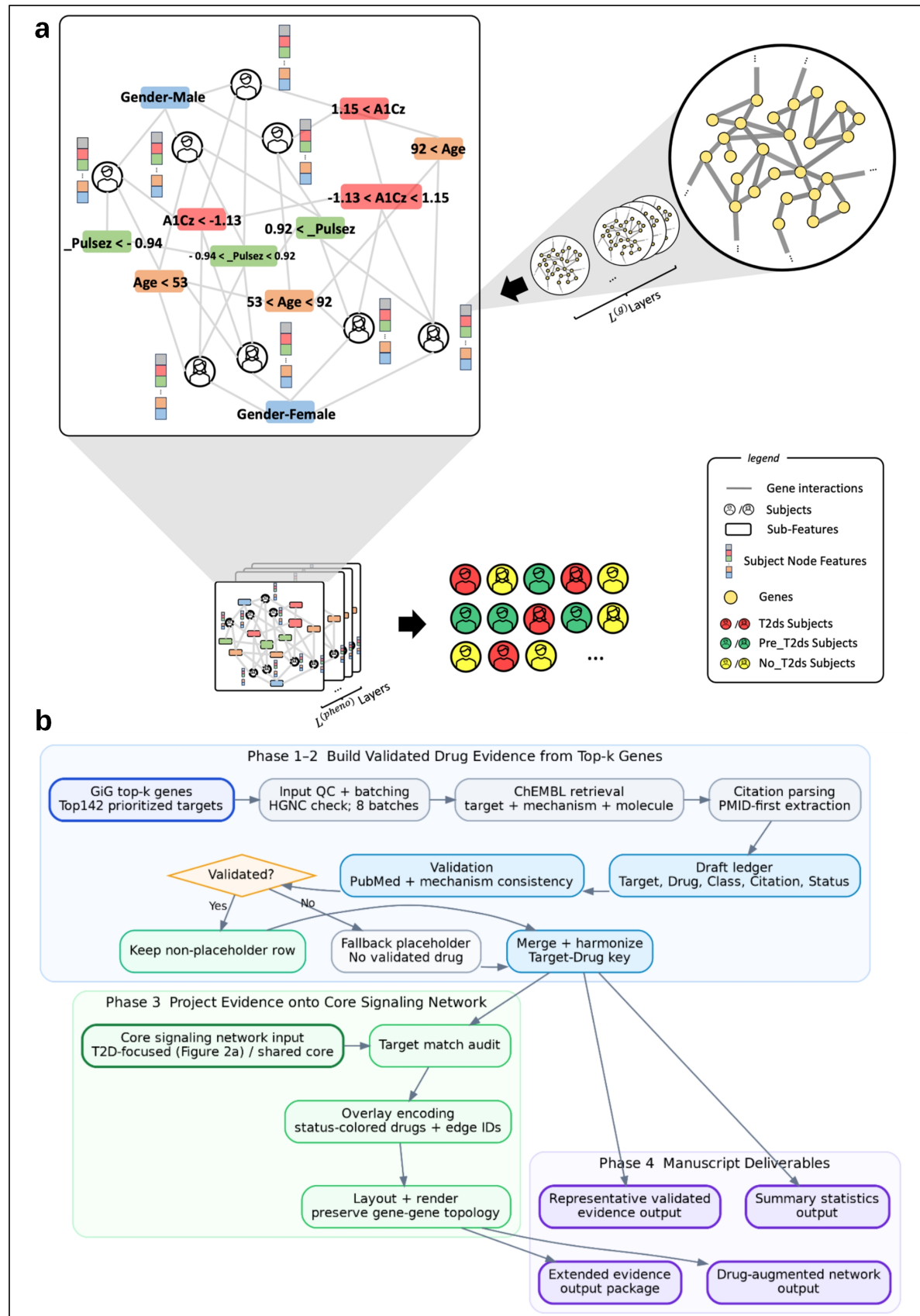


Figure 1. a. Overview of the GiG Model Architecture for T2D Risk Stratification. **b.** Workflow for post-GiG drug-evidence augmentation. GiG top-k genes undergo ChEMBL retrieval, PMID-first parsing, and PubMed/mechanism validation, then are harmonized into a target-drug evidence key, projected onto the core signaling network, and reported as summary statistics, representative validated evidence, a drug-augmented network, and an extended evidence package.

2.2 Identification of Key Biomarkers for Pre-T2D and T2D via the GiG Framework

This study collected 813 patients' electronic records with 122 sub-features from 41 characteristic groups (Table 1). Based on the classification, patients have been divided into three categories: 104 pre_t2ds, 68 t2ds, and 641 no_t2ds.

Table 1. 41 Characteristic Groups

Sex	Age	HAI
FEV1 (LK)	FEV1	FEV6
FEV1/FVC	FVC	PP-FEV1
PP-FEV6	BMI	Waist Circ.
Body Weight	Arm Span	Pulse Rate
SBP	DBP	Pulse Pressure
ABI	MMSE	Animal Fluency
Digit Forward	Digit Backward	Digit Symbol
Logical Mem. (Imm.)	Logical Mem. (Del.)	Total Score (SPPB)
Gait Speed	5x Stand Test	Grip Strength
Albumin	Cholesterol	Creatinine
Cystatin C	hsCRP	NT-BNP
Transferrin	Inorg. T (TSadj)	TSadj BP
D2	D3	

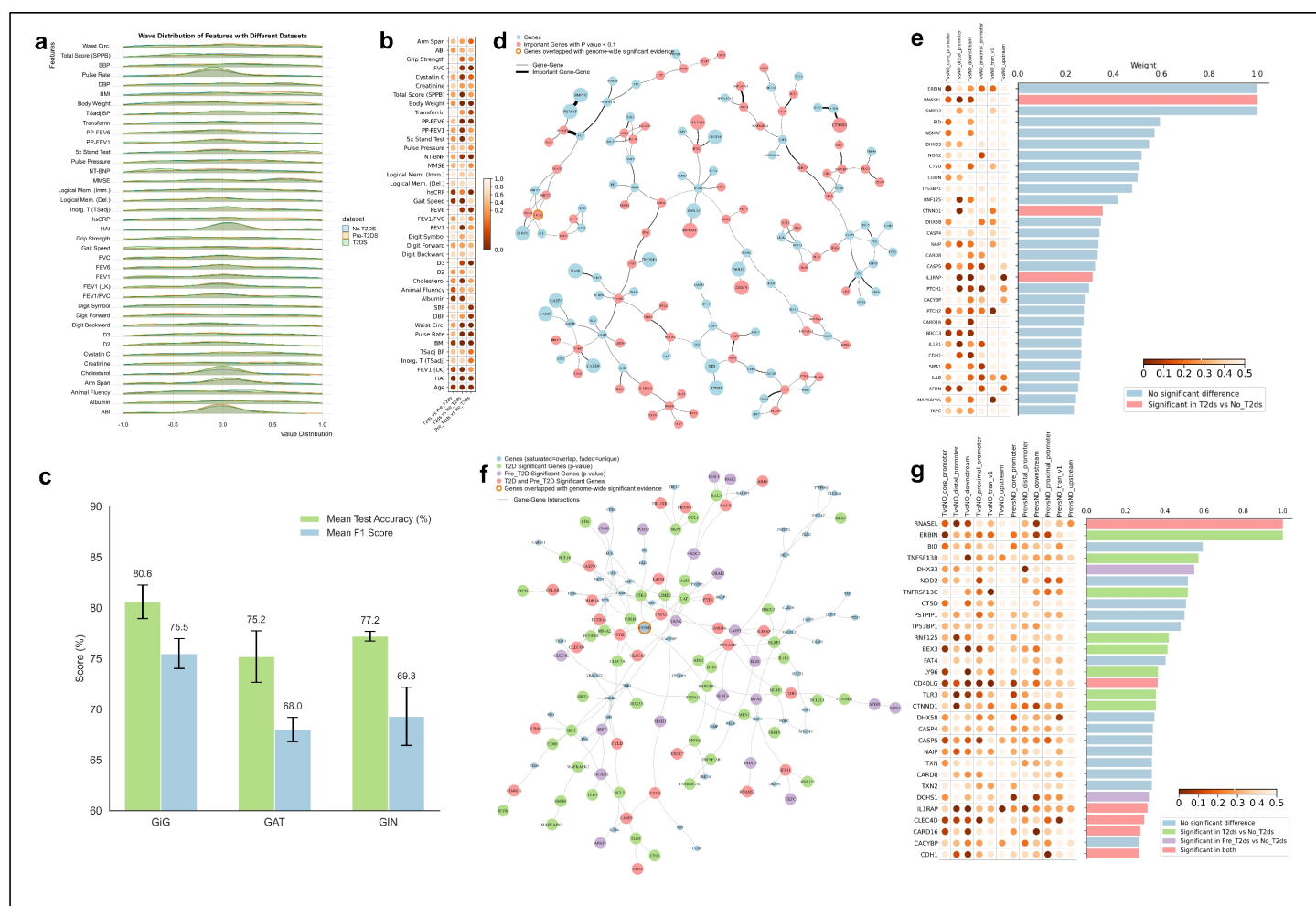


Figure 2. Statistical analysis of clinical phenotypes. **a.** illustrative of wave distribution of features in each clinical phenotypes for each type of patients (T2ds, Pre_t2ds, No_t2ds). **b.** the two-sample Kolmogorov-Smirnov test was demonstrated by comparing the three different types of patients in each clinical phenotype. **c.** Performance comparison between the full GiG model and baseline GNN architectures. Mean test accuracy and mean F1 score are shown for the full GiG model using both phenotype and genotype data and the baseline GAT and GIN models. Error bars represent the standard deviation across cross-validation folds. **d.** the visualization of the top selected nodes by filtering out of the edge weight lower than 0.05 in t2ds patient cohort. The red nodes represent the nodes with any p-value lower than 0.05 when comparing the gene expression and methylation in t2ds cohort and no_t2ds cohort. **e.** the bar plot of the node weights for all of top selected genes (red bar are the genes with any p-value lower than 0.05 by comparing the gene expression and methylation in t2ds cohort and no_t2ds cohort) and the left panel of the bar plot shows the p-value in each type of comparison. **f.** the visualization of the top selected genes combining all of the genes filtered out of the edge weight lower than 0.13 in t2ds patient cohort and the genes filtered out of the edge weight lower than 0.13 in pre_t2ds patient cohort. The green, blue nodes represent the nodes with any p-value lower than 0.05 when comparing the gene expression and methylation in t2ds cohort and no_t2ds cohort and the nodes with any p-value lower than 0.01 when comparing the gene expression and methylation in pre_t2ds cohort and no_t2ds cohort respectively, and the red nodes are the overlapped genes for above 2 types of comparisons. **g.** the bar plot demonstrates the nodes importance in different type of nodes (the colors align with the visualization in the network), and the bottom panel shows the p-value in each type of comparison

We applied GiG to analyze the medical phenotype and omics data of Pre_T2D and T2D. With the trained parameters of GiG in the transformer, we can obtain the edge weight for the genetic network. We also developed a tool for visualizing the critical genes and components in different groups of patients. By selecting the edge weight threshold, we can filter out the edges and genes to shrink the network to the core graph, indicating the critical signaling pathways and genes (check **Figure 2**).

Of the top 50 highest-weighted genes identified by the model, 48 have documented associations with T2D or closely related metabolic and inflammatory processes in the published literature, providing broad empirical support for the biological relevance of the model's output. The most prominent functional cluster, comprising top-weighted genes such as DHX33, CARD8, BRCC3, CASP4/CASP5, and APAF1, centers on pyroptosis and NLRP3 inflammasome signaling, a pathway with well-established mechanistic links to T2D pathogenesis³⁶. DHX33 has been shown to sense cytosolic RNA and activate the NLRP3 inflammasome³⁷, while CARD8 harbors genetic variants that modulate inflammasome-associated diabetes risk³⁸. BRCC3 regulates NLRP3 activity through deubiquitination³⁹, and CASP4/CASP5 are activated by saturated fatty acids to trigger downstream IL-1 β and IL-18 release⁴⁰. APAF1 has further been implicated in pyroptosis-related retinal pathology in diabetic contexts⁴¹. Taken together, these genes converge on a coherent inflammatory axis that is mechanistically consistent with β -cell dysfunction and systemic insulin resistance in T2D.

The second functional cluster couples β -cell and insulin signaling with innate immune activation. ERBIN suppresses NOD2-dependent NF- κ B activation and inflammasome-mediated innate immune signaling, and is downregulated in diabetic tissue⁴², and IL1RAP, as the obligate co-receptor of the IL-1 receptor complex, mediates IL-1 β -driven NF- κ B activation and β -cell dysfunction in pancreatic islets⁴³, while TXNIP methylation has been independently associated with both inflammation and diabetes⁴⁴. These findings are complemented by innate immune genes with established T2D relevance: NOD2 deficiency aggravates T2D-associated metainflammation⁴⁵, TLR3 displays inflammatory imbalance in diabetic patients⁴⁶, STING1 has been proposed as a therapeutic target in T2D⁴⁷, and IFIH1 variants associate with both autoimmune and type 2 forms of diabetes⁴⁸.

Among the 50 genes, includes high-confidence hits such as TNFRSF13C⁴⁹ and CD40LG⁵⁰. The convergence between model-assigned weights and independently reported genetic and

functional evidence indicates that the node scoring scheme captures biologically meaningful disease signal rather than statistical noise.

2.3 Pre_T2D and T2D stage-specific targets and pathways

To further characterize the biological relevance of the GiG network, we examined differential RNA-seq expression and disease-stage specificity among all 142 high-weight genes (the full set from which the top 50 were selected) identified by the model. Figure 2a presents the disease-stage specificity landscape of these genes, plotting the significance of regulatory-region methylation differences (Mann–Whitney U test; minimum raw p-value across six regulatory regions per gene) in T2D versus No-T2D (x-axis) against Pre-T2D versus No-T2D (y-axis), with bubble size proportional to network edge weight. The genes partition into three distinct groups: 25 genes reach significance in both comparisons (both significant), reflecting dysregulation shared across early and established disease; 45 genes are T2D-specific, achieving significance only against the T2D versus No-T2D axis; and 19 genes are Pre-T2D-specific, elevated in significance on the Pre-T2D axis alone. Among the most prominent both-significant genes, CARD16, CDH1, and IFIH1 occupy the upper-right quadrant with high network weights, suggesting that these genes are not only strongly co-expressed within the GiG network but also consistently dysregulated from the pre-diabetic stage onward. IL1B and SIPA1 likewise fall in this quadrant with moderate network weights, consistent with their established roles in inflammasome signaling⁵¹ and immune cell migration⁵². Conversely, T2D-specific genes such as IL1RAP, RNASEL, and LTB, which carry some of the highest x-axis significance scores, may reflect molecular changes that emerge only in frank disease rather than during the metabolic transition preceding it.

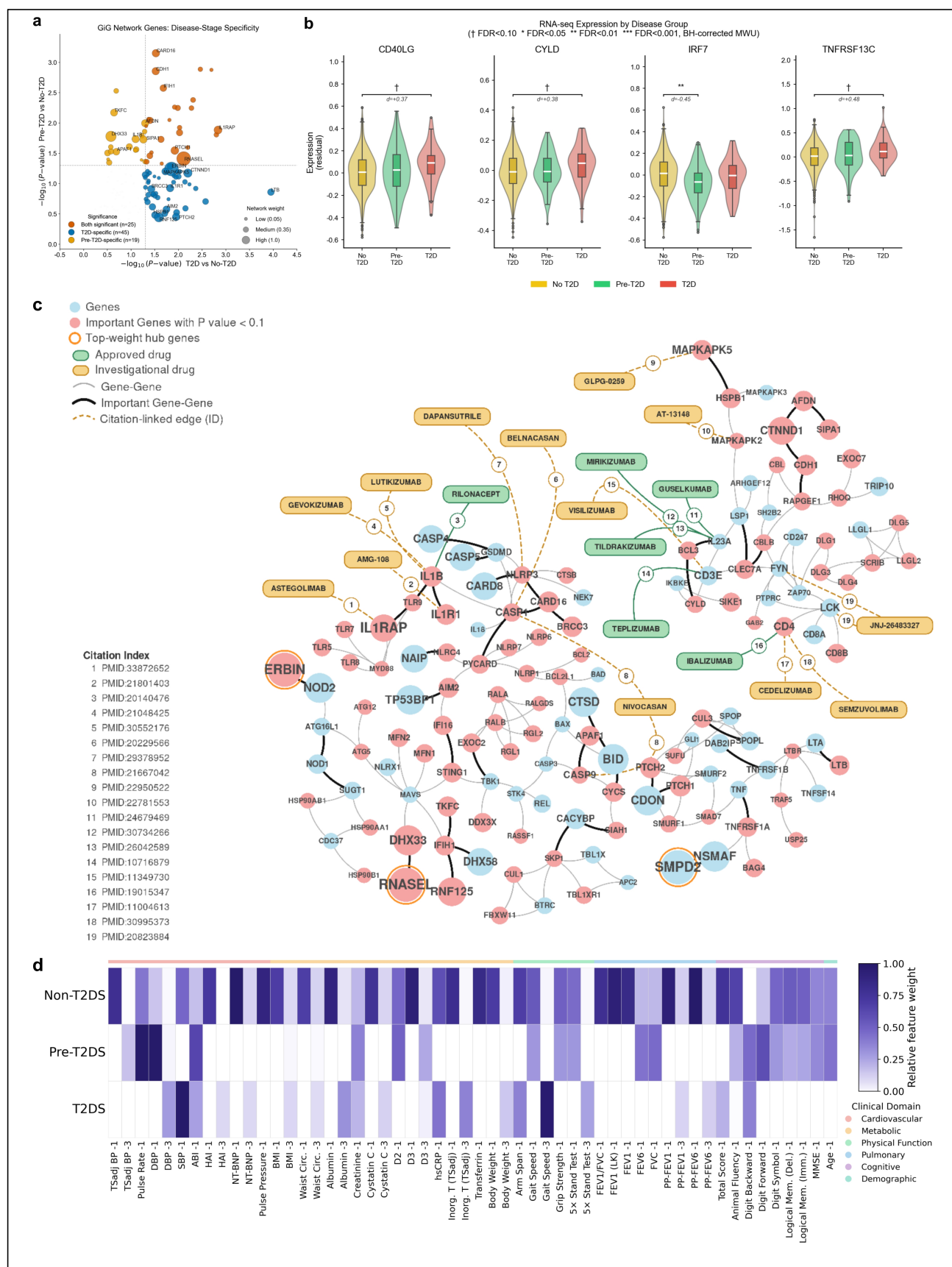


Figure 3. a. Disease-stage specificity of GiG network genes across T2D progression, stratified by expression significance. **b.** Residual RNA-seq expression distributions of four representative high-weight genes across no T2D, pre-T2D, and T2D groups. **c.** Drug-augmented GiG core network of T2D-focused overlay. Validated target-drug evidence from the final drug-evidence ledger is projected onto the T2D-focused GiG core network derived from Figure 2a. The original gene-gene topology and node-significance encoding are preserved, while drug nodes and target-drug links are added as an overlay. Citation-linked edges are shown as dashed lines, and edge ID labels map to PubMed citations in the in-figure Citation Index. The final overlay contains 21 target-drug edges across 13 core-network targets and 19 drug nodes, with approved and investigational drugs displayed using distinct node styles. **d.** Top phenotypic biomarkers across t2ds, pre_t2ds, and no_t2ds subgroups, where the suffixes -1, -2, and -3 denote the bottom 5%, middle 90%, and top 5% percentile ranges of each feature group, respectively.

Panel A of Figure 3b strengthens these network-level findings at the individual gene level by displaying residual RNA-seq expression distributions for four representative high-weight genes across the three disease groups. IRF7 exhibits a significant and progressive decline from No T2D to T2D (FDR < 0.01, Cohen's $d = -0.45$), indicating that interferon regulatory factor 7-mediated

innate immune signaling is suppressed as disease advances⁵³. CD40LG, CYLD, and TNFRSF13C each show a trend toward higher expression in T2D relative to No T2D ($d = +0.37$, $+0.38$, and $+0.48$, respectively; FDR < 0.10), pointing to modest but consistent upregulation of co-stimulatory⁵⁴, deubiquitinase⁵⁵, and B-cell survival signaling in the T2D state. Taken together, these expression patterns corroborate the network weights assigned by the GiG model, reinforcing the interpretation that the top-ranked genes capture genuine disease-associated transcriptional signals rather than statistical artifacts.

2.4 Drugs targeting on the key signaling targets

To evaluate whether GiG-prioritized genes connect to actionable pharmacologic space, we apply the drug-evidence augmentation workflow to all 142 prioritized targets and construct a validated, citation-auditable evidence ledger (Figure 3c). The final merged ledger contains 159 rows across 142 targets, including 42 validated non-placeholder target-drug rows and 117 standardized placeholder rows indicating that no direct validated evidence is retained under current retrieval and validation rules. At the target level, 25 of 142 genes show at least one validated non-placeholder linkage. A compact summary of retrieval coverage, validation outcomes, and development-status distribution is provided in Table 2a.

Among validated non-placeholder rows, development status includes 7 approved and 35 investigational rows, with no retained preclinical-only rows after confidence filtering. Validation

accounting further reflects the conservative behavior of the pipeline: 42 candidate rows are retained as validated, 4 rows are removed during citation-consistency checks, and 117 targets are represented by fallback placeholder rows. This design emphasizes traceable support and reduces the risk of overstating druggability for genes without direct mechanism-linked evidence.

Approved-linked evidence is concentrated in immuno-inflammatory targets, including CD3E, CD4, IL1B, IL23A, and SYK. The broader investigational landscape additionally includes CASP1, CASP8/9, CD40/CD40LG, IL1R1/IL1RAP, NLRP3, TLR3, TNFRSF13C, TNFSF13B, and TRAF3IP2. This profile is consistent with the inflammatory and innate-immune emphasis of the GiG-derived network, indicating that high-weight network biology is partially connected to ongoing therapeutic programs.

We then project validated target-drug evidence onto the T2D-focused core network (Figure 2f) to generate a drug-augmented overlay (Figure 3c). In this overlay, 21 target-drug edges are mapped across 13 core-network targets and 19 drug nodes, enabling direct visual inspection of where translationally supported targets reside in the core signaling architecture. Together with Table 2, this figure provides a concise translational readout that links mechanistic prioritization to auditable intervention hypotheses.

Interpretation remains conservative. Target-level evidence mapping does not establish efficacy in the T2D/pre-T2D indication context, and development status reflects global drug-development context rather than T2D-specific approval. In addition, absence of validated direct evidence in this ledger should not be interpreted as absence of therapeutic potential, because it may reflect curation density, citation granularity, or strict confidence filters. Overall, this section extends GiG from mechanistic prioritization to a translationally annotated network output that is auditable and suitable for downstream therapeutic triage.

Table 2a. Summary of Drug-Evidence Augmentation Outcomes for GiG-Prioritized Targets

Category	Metric	Value
Coverage	Input targets	142
	Targets with validated evidence	25 (17.61%)
Ledger composition	Total rows	159
	Validated non-placeholder rows	42
	Placeholder rows	117

Validation filtering	Removed rows	4
Development status	Approved	7
	Investigational	35

Table 2b. Representative Sample of Validated Drug Evidence for GiG-Prioritized Targets

Target	Candidate Drug	Drug Class	Development Status	Evidence Citation
CD3E	TEPLIZUMAB	other	approved	PMID:10716879; DOI:10.1006/cimm.2000.1617
CD4	IBALIZUMAB	inhibitor	approved	PMID:19015347; DOI:10.1128/AAC.00942-08
IL23A	GUSELKUMAB	inhibitor	approved	PMID:24679469; DOI:10.1016/j.jaci.2014.01.025
IL23A	MIRIKIZUMAB	inhibitor	approved	PMID:30734266; DOI:10.1111/bjd.17628
SYK	FOSTAMATINIB	inhibitor	approved	PMID:19965662; DOI:10.1182/blood-2009-08-236471
IL1B	GEVOKIZUMAB	negative allosteric modulator	investigational	PMID:21048425; DOI:10.4161/mabs.3.1.13989
CASP1	BELNACASAN	inhibitor	investigational	PMID:20229566; DOI:10.1002/cmdc.200900531
NLRP3	DAPANSUTRILE	inhibitor	investigational	PMID:29378952; DOI:10.1073/pnas.1716095115
TLR3	RINTATOLIMOD	agonist	investigational	PMID:11607032; DOI:10.1038/35099560
TNFSF13B	ATACICEPT	inhibitor	investigational	PMID:18046446; DOI:10.1038/sj.leu.2405048
CD40LG	DAPIROLIZUMAB PEGOL	inhibitor	investigational	PMID:25784719; DOI:10.1177/0961203315574558
IL1R1	AMG-108	antagonist	investigational	PMID:21801403; DOI:10.1186/ar3430
TRAF3IP2	IGURATIMOD	inhibitor	investigational	PMID:23116200; DOI:10.1186/ar4071

Table 2. Drug-evidence augmentation results for GiG-prioritized targets. (a) summarizes coverage, validation filtering, development-status distribution, and T2D-focused network mapping of the post-GiG drug-evidence workflow. (b) provides representative validated target-drug-citation examples from the final non-placeholder evidence ledger.

2.5 Key phenotypic biomarkers

By analyzing the group features with suffixes, we identified distinct phenotypic biomarker profiles for the three patient groups. The group features with suffixes indicate the range of subfeatures in each domain. For example, the group feature ‘_Pulsez_v1-1’ with suffix 1 means that this subfeature falls into the lowest five percentile of the whole feature. Suffix 2 implies that this subfeature falls into the range of the 5 to 95 percentiles of the entire feature group. Thus, the suffix 3 means that his subfeatures are over 95 percentiles of the whole feature group. Figure 3d displays the top subfeatures contributing most to the t2ds, pre_t2ds, and no_t2ds patient groups. Tables 3-5 detail the top 20 subfeatures for each of these categories.

Table 3. Top 20 features for t2ds patients (barplot)

Rank	Group Name	Subgroup Name	Weight
1	SBP	SBP-1	0.0667
2	Gait Speed	Gait Speed-3	0.0645
3	hsCRP	hsCRP-1	0.0286
4	Arm Span	Arm Span-1	0.0250
5	Digit Backward	Digit Backward-1	0.0250
6	Inorg. T (TSadj)	Inorg. T (TSadj)-3	0.0247
7	Albumin	Albumin-3	0.0212
8	DBP	DBP-3	0.0191
9	ABI	ABI-1	0.0180
10	5x Stand Test	5x Stand Test-3	0.0173
11	Total Score (SPPB)	Total Score (SPPB)-1	0.0143
12	Creatinine	Creatinine-1	0.0133
13	Body Weight	Body Weight-3	0.0100
14	PP-FEV6	PP-FEV6-3	0.0084
15	Cystatin C	Cystatin C-3	0.0078
16	PP-FEV1	PP-FEV1-3	0.0072
17	BMI	BMI-3	0.0066
18	HAI	HAI-3	0.0062
19	NT-BNP	NT-BNP-3	0.0061
20	Waist Circ.	Waist Circ.-3	0.0055

Table 4. Top 20 features for pre_t2ds patients

Rank	Group Name	Subgroup Name	Weight
------	------------	---------------	--------

1	Pulse Rate	Pulse Rate-1	0.1333
2	DBP	DBP-1	0.1333
3	ABI	ABI-1	0.0857
4	Digit Forward	Digit Forward-1	0.0800
5	D2	D2-1	0.0667
6	Digit Backward	Digit Backward-1	0.0615
7	FVC	FVC-1	0.0615
8	Age	Age-1	0.0571
9	FEV6	FEV6-1	0.0571
10	Creatinine	Creatinine-1	0.0500
11	MMSE	MMSE-1	0.0462
12	Gait Speed	Gait Speed-1	0.0401
13	Digit Symbol	Digit Symbol-1	0.0400
14	Grip Strength	Grip Strength-1	0.0400
15	5x Stand Test	5x Stand Test-1	0.0399
16	Logical Mem. (Imm.)	Logical Mem. (Imm.)-1	0.0333
17	D3	D3-3	0.0324
18	Logical Mem. (Del.)	Logical Mem. (Del.)-1	0.0308
19	Animal Fluency	Animal Fluency-1	0.0286
20	TSadj BP	TSadj BP-3	0.0220

Table 5. Top 20 features for no_t2ds patients

Rank	Group Name	Subgroup Name	Weight
1	NT-BNP	NT-BNP-1	0.0185
2	PP-FEV6	PP-FEV6-1	0.0185
3	FEV1 (LK)	FEV1 (LK)-1	0.0172
4	D3	D3-1	0.0172
5	PP-FEV1	PP-FEV1-1	0.0172
6	Albumin	Albumin-1	0.0164
7	Pulse Pressure	Pulse Pressure-1	0.0164
8	FEV1	FEV1-1	0.0156
9	Transferrin	Transferrin-1	0.0156
10	Inorg. T (TSadj)	Inorg. T (TSadj)-1	0.0149
11	FEV1/FVC	FEV1/FVC-1	0.0147
12	TSadj BP	TSadj BP-1	0.0147
13	Cystatin C	Cystatin C-1	0.0143
14	Total Score (SPPB)	Total Score (SPPB)-1	0.0140
15	HAI	HAI-1	0.0135

16	Body Weight	Body Weight-1	0.0130
17	BMI	BMI-1	0.0128
18	Arm Span	Arm Span-1	0.0127
19	Waist Circ.	Waist Circ.-1	0.0127
20	Animal Fluency	Animal Fluency-1	0.0123

2.6 Pre-T2D and T2D pathway analysis

These categorizations are derived from the high-level classifications provided in sources such as KEGG PATHWAY database⁵⁶ and Gene Ontology (GO) Biological Process (BP)^{57,58}. which broadly group pathways into categories like metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases, and drug development. The above categorizations are structured with a more specific focus on the functional or disease-related aspects of the pathways (Table 6-7).

Table 6. KEGG Pathway enrichment analysis

Term	Genes	PValue	FDR
NOD-like receptor signaling pathway	RNASEL, NLRX1, HSP90AB1, NOD1, BRCC3, NLRC4, NOD2, TNF, PYCARD, TBK1, CASP5, DHX33, IFI16, NLRP6, NLRP7, CASP4, CASP1, MFN1, MFN2, NLRP3, NLRP1, TP53BP1, IKBKE, ATG5, CTSB, GSDMD, HSP90AA1, CARD8, NEK7, IL18, ERBIN, ATG12, MAVS, AIM2, STING1, ATG16L1, IL1B, TRAF5, BCL2, CARD16, SUGT1, MYD88, NAIP, BCL2L1	6.30E-40	6.62E-38
RIG-I-like receptor signaling pathway	NLRX1, DDX3X, TNF, ATG12, IFIH1, CYLD, MAVS, SIKE1, TBK1, TKFC, RNF125, STING1, DHX58, IKBKE, ATG5	2.30E-12	3.01E-11
NF-kappa B signaling pathway	TNFSF14, IL1R1, TNF, TNFRSF1A, CYLD, ZAP70, LCK, IL1B, TRAF5, LTA, BCL2, LTBR, MYD88, BCL2L1	4.62E-10	3.73E-09
Cytosolic DNA-sensing pathway	GSDMD, IL18, PYCARD, MAVS, AIM2, TBK1, IFI16, STING1, IL1B, CASP3, CASP1, NLRP3, IKBKE	3.32E-09	2.49E-08

Apoptosis	APAF1, BAD, DAB2IP, TNF, TNFRSF1A, CASP9, CASP3, BCL2, BAX, CYCS, BID, CTSD, CTSB, BCL2L1	1.28E-07	6.73E-07
-----------	---	----------	----------

Table 7. GO Pathway enrichment analysis

Term	Genes	PValue	FDR
positive regulation of interleukin-1 beta production	GSDMD, CARD8, NEK7, HSPB1, NOD1, NLRC4, NOD2, TNF, PYCARD, AIM2, IFI16, NLRP6, CLEC7A, CASP1, TLR8, NLRP3, NLRP1, CARD16, MYD88, NAIP	3.61E-25	5.48E-22
positive regulation of canonical NF-kappaB signal transduction	CUL1, NOD1, NOD2, TNF, PYCARD, TBK1, CLEC7A, CASP1, IKBKE, TNFSF14, IL1R1, FBXW11, DAB2IP, TNFRSF1A, CYLD, CD4, MAVS, IL1B, TRAF5, REL, LTA, TLR9, TLR8, TLR7, CARD16, LTBR, MYD88	1.10E-24	8.34E-22
apoptotic process	RALB, DDX3X, CUL3, CUL1, NOD1, NLRC4, STK4, CASP9, PYCARD, CASP5, IFI16, CASP3, CASP4, CASP1, MFN2, NLRP3, NLRP1, BID, ATG5, SKP1, TNFSF14, APAF1, BAD, SIAH1, DAB2IP, TNFRSF1B, TNFRSF1A, IL1B, TRAF5, BCL3, LTA, BCL2, BAX, CYCS, CARD16, LTBR, MYD88, NAIP, BCL2L1	8.36E-21	2.12E-18
pyroptotic inflammatory response	GSDMD, CARD8, NEK7, NLRC4, PYCARD, AIM2, NLRP6, CASP3, CASP4, CASP1, NLRP3, NLRP1, CARD16, NAIP	1.47E-18	2.89E-16
innate immune response	NLRX1, DDX3X, NOD1, IL1RAP, NLRC4, NOD2, TNF, IFIH1, PYCARD, TBK1, IFI16, NLRP6, CLEC7A, CASP4, DHX58, NLRP3, NLRP1, ATG5, GSDMD, CARD8, DAB2IP, CYLD, MAVS, AIM2, STING1, IL23A, REL, TLR9, TLR8, TLR7, TLR5, MYD88, NAIP	1.52E-18	2.89E-16

As summarized in **Figure 4a**, pathway and ontology enrichment analyses across multiple annotation databases consistently converge on innate immune signaling and inflammasome-

mediated cell death as the dominant biological themes. Among the KEGG pathways, the RIG-I-like receptor signaling pathway achieves the second-highest fold enrichment, followed by the NOD-like receptor signaling pathway and the NF- κ B signaling pathway, which rank third and fourth, respectively. This indicates that the prioritized genes are disproportionately concentrated in innate immune sensing and cytokine-driven inflammatory cascades. At the GO Biological Process level, immune system process and innate immune response rank as the most significantly enriched terms, with pattern recognition receptor signaling pathway exhibiting the largest fold enrichment among all GO BP terms. Pyroptotic inflammatory response and positive regulation of interleukin-1 beta production are also prominently enriched, directly implicating the inflammasome-caspase axis⁵⁹ in the biological signal captured by the model. GO Molecular Function analysis reveals strong enrichment for CARD domain binding and cysteine-type endopeptidase activator activity, reflecting the functional specificity of the gene set toward inflammasome assembly interfaces and caspase catalytic regulation. At the cellular compartment level, canonical inflammasome complex and NLRP3 inflammasome complex display the most extreme fold enrichment values, with NLRP1 inflammasome complex and ripoptosome also significantly over-represented, collectively pointing to the intracellular multi-protein death-signaling machinery as the spatial substrate of the enrichment signal.

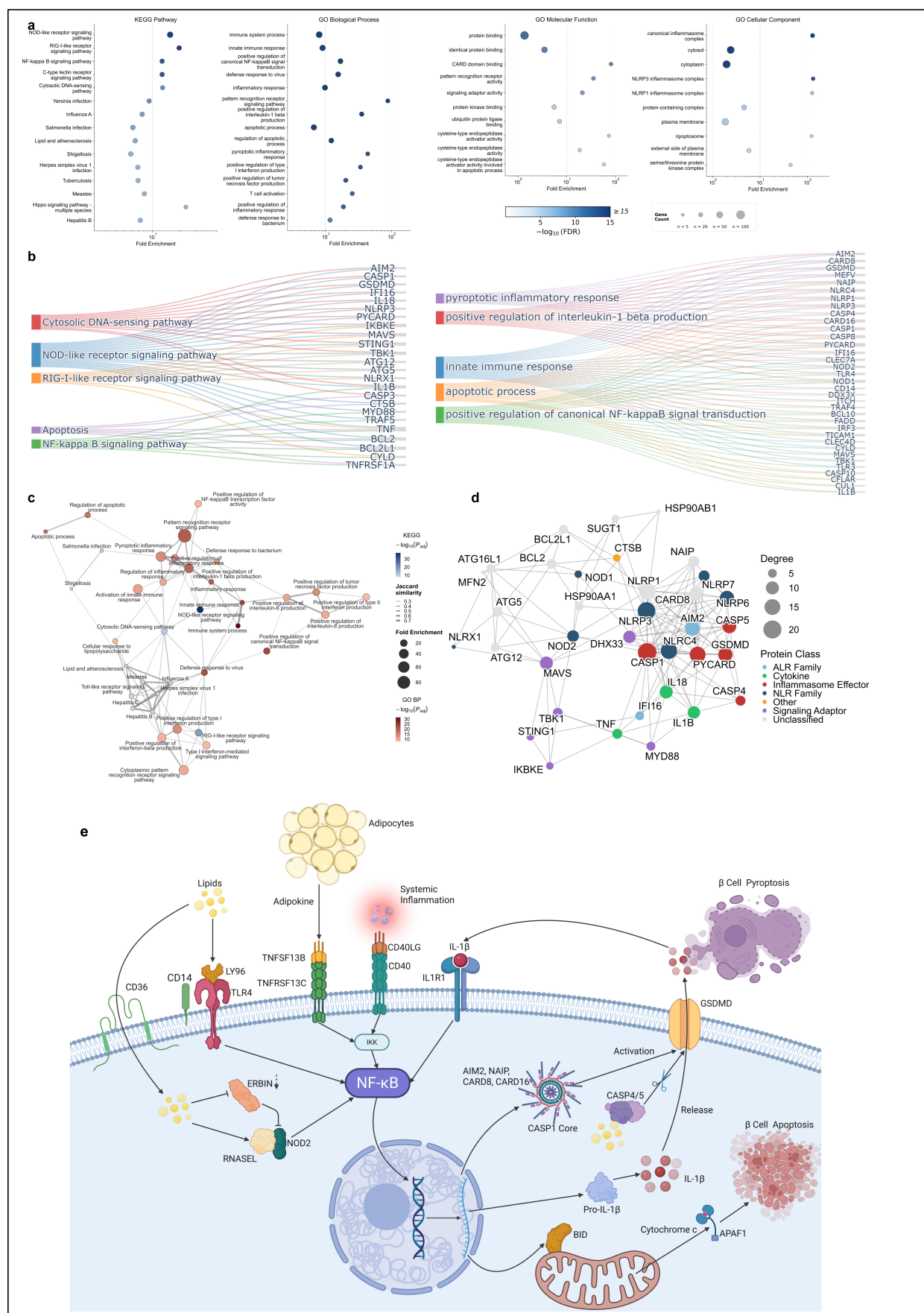


Figure 4. a. Functional enrichment analysis of prioritized genes across KEGG pathways, GO biological process, GO molecular function, and GO cellular component, highlighting prominent over-representation in innate immune signaling, inflammasome assembly, and caspase regulatory pathways. **b.** P-values and FDRs of top five representative GO terms and KEGG pathways. **c.** Enrichment map of pathway co-enrichment network, with nodes representing individual KEGG pathways and GO BP. **d.** Protein-protein interaction network of NOD-related inflammasome components, with node size scaled by interaction degree and color indicating protein class. **e.** Molecular mechanisms of β -cell death via NOD-like receptor–NF- κ B–inflammasome signaling in the diabetic microenvironment.

To delineate the gene-level composition of the most enriched pathways, we constructed a Sankey diagram mapping individual genes to their respective pathway assignments (Figure 4b). To visualize the topological organization of these enriched terms and assess their interdependence, we constructed an enrichment map in which nodes represent individual pathways and edge weights reflect pairwise Jaccard similarity of their gene membership (Figure 4c). The resulting network resolves into two principal connectivity regions. The upper region is anchored by pattern recognition receptor signaling pathway, the highest-degree hub bridging inflammatory regulation terms (e.g., positive regulation of NF- κ B, pyroptotic response, IL-1 β production, and inflammatory response). The lower region is organized around antiviral innate immunity, where RIG-I-like receptor signaling pathway, type I interferon-mediated signaling pathway⁶⁰, and cytoplasmic pattern recognition receptor signaling pathway⁶¹ form a tightly interconnected subcluster connected to viral infection pathways including Influenza A, Hepatitis B, Hepatitis C, and Herpes simplex virus 1 infection. The two regions are bridged through shared gene membership in innate immune response and cytosolic DNA-sensing pathway⁶², suggesting that the transcriptional program captured by the model integrates both sterile inflammatory and antiviral defense components into a unified regulatory structure. Apoptotic process and regulation of apoptotic process form a peripheral module linked to the inflammatory core, consistent with the role of caspase-mediated apoptosis⁶³ as a functionally adjacent yet mechanistically distinct outcome of the same upstream danger-sensing machinery.

To further dissect the molecular architecture underlying the enriched inflammatory pathways, we constructed a protein-protein interaction (PPI) network centered on NOD-related inflammasome components (Figure 4d). Nodes are scaled by interaction degree and colored by protein class, encompassing NLR family members (NOD1, NOD2, NLRP1, NLRP3, NLRP6, NLRP7, NLRC4), inflammasome effectors (CASP1, CASP4, CASP5, GSDMD, PYCARD), ALR family sensors (AIM2, IFI16), cytokines (IL1B, IL18, TNF), and signaling adaptors (MAVS, MYD88, STING1, TBK1, IKKBE). CASP1 and PYCARD emerge as the highest-degree hubs, consistent with their central roles in inflammasome assembly and effector activation^{59,64}, while GSDMD bridges this core complex to pyroptotic membrane execution^{65,66}. The network further reveals functional

crosstalk between canonical NOD/NLRP3 signaling and the cGAS-STING innate immune axis⁶¹ through MAVS, TBK1, and STING1, as well as between inflammasome activation and autophagy regulators including ATG5, ATG12, and ATG16L1⁶⁷. These topological features indicate that the pathway enrichment signals identified above reflect a convergent, rather than modular, inflammatory architecture operating in the diabetic context.

2.7 Pathophysiological implications of the identified pathways

Cell Death Signaling in Diabetic β -Cell Dysfunction The diabetic microenvironment introduces metabolic toxicity signals⁶⁸ that are initially detected by intracellular NOD-like receptor pathways⁶⁹ where the heavily weighted genes ERBIN⁷⁰, RNASEL⁴², and NOD2⁷⁰ function as frontline danger sensors (Figure 4e, left panel). ERBIN acts as a critical regulator of NOD2⁷⁰, and the activation of this receptor directly induces cell-autonomous innate immune responses⁷¹ that initiate inflammatory signaling and contribute to insulin resistance⁷⁰. These sensory signals rapidly propagate to hyperactivate the NF- κ B signaling hub through mediators such as the adipokine TNFSF13B (Figure 4e, center panel)⁷², which is abundantly secreted by adipose tissue under obese conditions to engage its receptor TNFRSF13C and stimulate the IKK complex⁷³. The resulting dysregulation of NF- κ B, exacerbated by ERBIN deficiency, drives a systemic chronic inflammatory cascade characterized by the substantial production of pro-inflammatory cytokines including IL-1 β ^{42,74}. This severe inflammatory stress ultimately recruits downstream executioner proteins, specifically the model-identified components CASP4⁶⁵ and BID (Figure 4e, right panel)⁷⁵, to drive irreversible programmed cell death. This terminal phase comprises apoptosis as well as parallel inflammatory modalities including pyroptosis and necroptosis, which directly cause the rupture and death of pancreatic β -cells. The prominence of these mechanisms is robustly supported by the functional enrichment of specific biological processes, particularly pyroptotic cell death, pyroptosome complex assembly, and the positive regulation of the pyroptotic inflammatory response, alongside the necroptotic signaling pathway and the general regulation of the necroptotic process. The subsequent maturation and release of IL-1 β engage IL1R1 to establish a destructive auto-amplification loop^{76,77}, demonstrating that this integrated network of programmed cell death pathways operates as a central pathogenic driver rather than a peripheral inflammatory byproduct in the disease progression⁷⁸.

Lipid Metabolism and Cardiovascular Complications Beyond direct cellular toxicity, the systemic metabolic dysregulation in T2D profoundly alters lipid homeostasis, establishing a critical mechanistic link to cardiovascular complications⁵¹. This pathogenic axis is strongly

supported by the specific enrichment of the lipid and atherosclerosis pathway in our analysis. The functional integration of these enrichment terms reveals a robust intersection between aberrant lipid metabolism and the innate immune networks. Specifically, the accumulation of metabolic byproducts, notably oxidized low-density lipoproteins, acts as a potent stimulus that directly engages pattern recognition receptors⁷⁹. This cellular response to oxidized lipids subsequently triggers the assembly and positive regulation of the NLRP3 inflammasome while simultaneously driving the canonical NF- κ B signaling cascade⁸⁰. Within the vascular microenvironment, this continuous inflammatory signaling exacerbates macrophage activation and foam cell formation, thereby accelerating atherosclerotic progression⁸¹. Consequently, these findings indicate that the lipid-driven inflammatory feedback loop is not merely a secondary manifestation of diabetes but a fundamental driver that mechanistically bridges metabolic dysfunction with severe cardiovascular comorbidities.

Immunological Disorders The pathways related to Inflammatory Bowel Disease (IBD), Rheumatoid Arthritis (RA), and the Intestinal Immune Network for IgA Production, have interconnected aspects that might have implications in the context of Type 2 Diabetes (T2D)⁸². Recent findings suggest a notable link between the metabolites produced by gut microbes and the subsequent inflammation in the intestinal and systemic regions, with the onset and advancement of diabetes⁸³. It's estimated that nearly 90% of type 2 diabetes cases are connected to imbalances in the gut microbiota, known as dysbiosis. This kind of microbial imbalance is also observed in cases of Inflammatory Bowel Disease (IBD)^{84,85}. Aside from that, there's a pathogenic connection between RA and T2D, forming a vicious circle driven by glucose derangement and inflammatory mediators. This suggests that the immunological disturbances in RA could have a cascading effect, leading to or worsening T2D⁸⁶. RA is associated with anomalies in glucose metabolism, primarily insulin resistance, which could progress to T2D. This association suggests a metabolic and immunological crosstalk between the two conditions⁸⁷. And a retrospective cohort study found that there could be shared pathogenesis stemming from the shared genetics of immune-related diseases between T2D and RA, indicating a possible genetic and immunological basis for the association between these conditions⁸⁸. What's more, some research has hinted at novel regulatory pathways within the intestinal immune network for IgA production, which could be relevant in the context of T2D. For instance, an investigation in black rockfish revealed a potential regulatory pathway of the immunity along the intestine-spleen axis, which might be of interest in understanding the immunometabolic interactions in T2D⁸⁹.

Bacterial infections Tuberculosis and Salmonella infection pathways are highly related to type 2 diabetes. Salmonella can cause extra-intestinal focal infections, apart from the common gastrointestinal problems. In some documented cases, individuals with T2D, especially those with poor glycemic control, have been found to be susceptible to severe Salmonella infections. For instance, a case study reported a 30-year-old man with poorly controlled T2D who developed a severe Salmonella skin and soft tissue infection leading to a chest wall abscess. This infection required surgical intervention and the patient was treated with antibiotics alongside insulin therapy to manage his blood sugar levels⁹⁰. In another case, individuals with Type 2 Diabetes are at an increased risk for infections, including the development of abscesses, due to impaired immune responses and altered tissue conditions. High blood sugar levels can impair the function of immune cells, slow down blood circulation, and affect the body's ability to heal, which makes it easier for infections to take hold and develop into abscesses. An abscess, which is a collection of pus that has formed within the tissue, can occur at different sites in the body, and its management can be complicated in individuals with diabetes, requiring meticulous glycemic control alongside infection management⁹¹. Furthermore, diabetes impairs the immune system, making it more challenging for the body to combat TB bacteria⁹². And the evidence shows that diabetes increases the likelihood of developing active TB. Individuals with diabetes have about a two to three-fold higher risk of developing TB compared to those without diabetes⁹³.

3 Discussion

In this study, for the first time, we introduced Graph in Graph (GiG), which is a novel dual-graph AI model and is capable of systematically integrating and interpreting complex medical record phenotypes and high-dimensional omic datasets at the individual patient level. While the availability of clinical records and biological omics has created new avenues for precision healthcare, effective multi-modal integration has remained a significant computational bottleneck. GiG mitigates these challenges by leveraging a dual-graph structure: a person-phenotype graph to capture clinical features alongside patient-specific omics signaling graphs. This design enables seamless representation across the whole-person biological spectrum, bridging the gap between molecular mechanisms and macro-level clinical manifestations.

Applying GiG to the Long Life Family Study (LLFS) cohort demonstrated both strong predictive performance and high model interpretability in evaluating Type 2 Diabetes (T2D), pre-T2D, and

control profiles. The model uncovered key clinical phenotypes and multi-omic biomarkers associated with disease progression. Beyond candidate biomarker identification, GiG successfully pinpointed critical signaling pathways driving metabolic dysfunction and prioritized potential therapeutic drugs, offering valuable mechanistically backed hypotheses for translational research. Overall, the GiG framework provides a scalable, generalizable AI framework for precision medicine. By effectively synthesizing heterogeneous biological and clinical data to drive both diagnosis and biomarker discovery, GiG can be readily adapted to investigate diverse disease contexts with both medical records and omics datasets.

4. Methodology

4.1 Study Population Design

This study is based on multi-omics and clinical phenotype data from the Long Life Family Study (LLFS) cohort. Transcriptomic profiles were available for 1,810 subjects, DNA methylation data (Illumina Methylation450K array, summarized across five gene-regulatory regions—core promoter, proximal promoter, distal promoter, upstream, and downstream) were available for 936 subjects, and clinical phenotype data were available for 4,810 subjects. Subjects were retained in the final analytic sample only if they had complete data across all three modalities—transcriptomics, DNA methylation, and clinical phenotype, including type 2 diabetes (T2D) status—yielding a final cohort of 813 subjects.

Subjects were stratified into three mutually exclusive groups based on established LLFS T2D status classifications: T2D (n=68), pre-T2D (n=104), and No-T2D (n=641). This three-group T2D status served both as the between-group comparison variable and as the prediction target (classification label) for the graph neural network model.

In addition to T2D status, 34 continuous clinical phenotypes were incorporated as phenotype-channel features and used for group-wise statistical comparisons, spanning demographic (age), anthropometric (BMI, waist circumference, weight), cardiovascular (systolic/diastolic blood pressure z-scores, pulse pressure, pulse rate), metabolic (glucose, HbA1c, and lipid profile), pulmonary function (FEV1, FEV6, FEV1/FVC, FVC), cognitive function (MMSE total score, digit span, digit symbol substitution, logical memory, animal naming), and physical performance (gait speed, chair-stand time, grip strength) domains. All continuous phenotypes were z-score standardized, with log- or inverse-normal transformation applied as needed prior to normalization.

4.2 Omics Data

Three data types were used to construct the model. Transcriptomic data consisted of covariate-adjusted expression residuals for 16,419 genes (Ensembl gene IDs) across 1,810 subjects⁹⁴. DNA methylation data, derived from the Illumina Methylation450K array, were organized as gene-level β -values (0–1) across the five regulatory regions described above, each covering 936 subjects. Clinical phenotype and T2D status labels were obtained from the LLFS phenotype database. The dataset was accessible at:

<https://prod.eliteportal.synapse.org/Explore/Projects/DetailsPage?shortName=LLFS>

To prepare these data for machine learning, transcriptomic and methylation probes were first mapped to gene-level identifiers via Ensembl annotation, and the resulting gene sets were intersected across the transcriptomic, methylation, and KEGG pathway network gene lists, yielding a final panel of 1,390 genes (reduced from 16,419). Subject identifiers were then intersected across the phenotype, T2D-label, transcriptomic, and methylation datasets to obtain the final analytic sample of 813 subjects (T2D=68, pre-T2D=104, No-T2D=641). Phenotype data were cleaned and restricted to the visit-1 timepoint, one-hot encoded, and discretized into categorical bins using the 10th/90th percentile thresholds to construct subject–phenotype bipartite subfeature nodes.

Gene expression and methylation feature matrices were each normalized and concatenated into a six-channel gene-node feature representation (one transcriptomic channel plus five methylation-region channels); normalized phenotype features were concatenated along the subject dimension to form the final node feature matrix. Gene–gene edges were derived from the KEGG pathway database (1,390 gene nodes; 9,199 edges), and subject–phenotype bipartite edges were derived from the discretized phenotype categories (813 subject nodes, 122 phenotype subfeature nodes; 32,600 edges). For exploratory group comparisons, each gene- and CpG-region-level feature was tested across the three pairwise T2D-status comparisons (T2D vs. No-T2D, T2D vs. pre-T2D, pre-T2D vs. No-T2D) using the Mann–Whitney U test, with multiple-testing correction by the Benjamini–Hochberg procedure.

4.3 Model of Graph in Graph (GiG)

Problem formulation To construct this graph neural network framework, we considered embedding the graph neural network as a node in another graph, which will be used as the giant graph in this framework. In this study, we embedded the transcriptomic graphs as nodes in the subject-feature graph. Hence, for the patient-feature graph, $G = (V, E)$, where $V = \{P, F\}$, $P = \{P_1, P_2, \dots, P_m, \dots, P_M\}$, $F =$

$\{(F_{S_1}^{(1)}, F_{S_1}^{(2)}, \dots, F_{S_1}^{(n_{S_1})}), (F_{S_2}^{(1)}, F_{S_2}^{(2)}, \dots, F_{S_2}^{(n_{S_2})}), \dots, (F_{S_q}^{(1)}, F_{S_q}^{(2)}, \dots, F_{S_q}^{(n_{S_q})}), \dots, (F_{S_Q}^{(1)}, F_{S_Q}^{(2)}, \dots, F_{S_Q}^{(n_{S_Q})})\}$ and $|V| = M + \sum_{q=1}^Q n_{S_q}$. To be more specific, there are two types of nodes in the graph. Nodes set P is the set containing patient nodes, and there are M patients in the graph G . And nodes set F is the set containing feature nodes, and there are Q groups of subfeatures. In certain subfeature groups S_q , there are n_{S_q} feature nodes. And we have patient phenotype feature matrix $H_{Pheno} \in \mathbb{R}^{M \times Q}$ and subfeature matrix $H_{Sub} \in \mathbb{R}^{(\sum_{q=1}^Q n_{S_q}) \times Q}$. To implement the embedded graph in specific patient nodes P_m , the embedded graph $G_{P_m} = (V_{P_m}, E_{P_m})$ will be built, where $|V_{P_m}| = N$ and all of E_{P_m} share the same graph structure with E_g . In any of the embedded graph P_m , the nodes set (i.e., the genes set) contains the transcriptomic data with node/gene feature matrix $X_{P_m}^{(0)} \in \mathbb{R}^{N \times d^{(0)}}$, where $d^{(0)}$ is the number of dimensions for the node initial feature. For the gene feature in the patient cohort, the gene feature matrix will be $X_g = [X_{P_1} | X_{P_2} | \dots | X_{P_m} | \dots | X_{P_M}]$, where $\mathbb{R}^{M \times N \times d^{(0)}}$.

Hence, the Graph in Graph model (GiG) will predict the type of patients for the cohort with the phenotype and genotype input H_{Pheno}, H_{Sub}, X_g ,

$$O = \text{GiG}(E, E_g, H_{Pheno}, H_{Sub}, X_g)$$

, where $O \in \mathbb{R}^{C \times L^{(pheno)}}$ and C is the number of classes for the downstream task.

GiG Framework Architecture The overall architecture of the GiG framework is illustrated in Figure 1a. To embed the graph as a feature vector for the node P_m in the patient-feature graph, which will be the graphical model in the gene graph with $G_{P_m} = (V_{P_m}, E_{P_m})$, the attention-based model by graph transformer⁹⁵ will be applied here for interpretation possibility through message passing over layers $1, 2, \dots, l^{(g)}, \dots, L^{(g)}$:

$$(X_{P_m}^{(l^{(g)})})_i = \beta^{(g)} \cdot W_1^{(l^{(g)})} (X_{P_m}^{(l^{(g)}-1)})_i + (1 - \beta^{(g)}) \cdot \sum_{j \in \mathcal{N}(i)} (\alpha_{P_m}^{(l^{(g)})})_{ij} W_2^{(l^{(g)})} (X_{P_m}^{(l^{(g)}-1)})_j$$

$$(\alpha_{P_m}^{(l^{(g)})})_{ij} = \text{softmax}\left(\frac{\text{SUM}[(W_3^{(l^{(g)})}(X_{P_m}^{(l^{(g)}-1)})_i)^T(W_4^{(l^{(g)})}(X_{P_m}^{(l^{(g)}-1)})_j)]}{\sqrt{d^{(l^{(g)})}}}\right)$$

In the $l^{(g)}$ layer, the input node features is the set of N nodes/genes with $X_{P_m}^{(l^{(g)}-1)} = \{(X_{P_m}^{(l^{(g)}-1)})_1, (X_{P_m}^{(l^{(g)}-1)})_2, \dots, (X_{P_m}^{(l^{(g)}-1)})_n, \dots, (X_{P_m}^{(l^{(g)}-1)})_N\}$, where $(X_{P_m}^{(l^{(g)}-1)})_n \in \mathbb{R}^{d^{(l^{(g)}-1)} \times 1}$, $W_1^{(l^{(g)})} \in \mathbb{R}^{d^{(l^{(g)})} \times d^{(l^{(g)}-1)}}$, $W_2^{(l^{(g)})} \in \mathbb{R}^{d^{(l^{(g)})} \times d^{(l^{(g)}-1)}}$, $W_3^{(l^{(g)})} \in \mathbb{R}^{(h^{(l^{(g)})} \times c^{(l^{(g)})}) \times d^{(l^{(g)}-1)}}$, $W_4^{(l^{(g)})} \in \mathbb{R}^{(h^{(l^{(g)})} \times c^{(l^{(g)})}) \times d^{(l^{(g)}-1)}}$ and there are $L^{(g)}$ layers in total. $h^{(l^{(g)})}$ and $c^{(l^{(g)})}$ are the number of heads and channels in the $l^{(g)}$ layer. Here, the SUM function will aggregate all embedded features. Thus, the final attention coefficient between node i and node j will be $(\alpha_{P_m}^{(l^{(g)})})_{ij} \in \mathbb{R}$ in the $l^{(g)}$ layer. To mitigate the dilution of attention weights over many features, we subsequently apply a threshold to prune low-weight edges and retain only the top-weighted nodes for interpretation.

With the embedded node features, $X_{P_m}^{(L^{(g)})} \in \mathbb{R}^{N \times d^{L^{(g)}}}$, in the last layer, the pooling method based assignment matrix⁹⁶ will be used for transforming the N nodes/genes into lower dimensions with K critical features by

$$H_{P_m} = \text{FLATTEN}[SX_{P_m}^{(L^{(g)})}]$$

, where $S \in \mathbb{R}^{K \times N}$, $H_{P_m} \in \mathbb{R}^{(K \times d^{L^{(g)}})}$.

After collecting the embedded graph from gene features of the individual patient, we can get the node/ patient features with $H_P = [H_{P_1}|H_{P_2}|\dots|H_{P_m}|\dots|H_{P_M}]$, where $H_P \in \mathbb{R}^{M \times (K \times d^{L^{(g)}})}$ is the patient embedded genomic feature. From the initial patient-feature graph G , we have patient phenotype feature matrix $H_{Pheno} \in \mathbb{R}^{M \times Q}$ and subfeature matrix $H_{Sub} \in \mathbb{R}^{(\sum_{q=1}^Q n_{s_q}) \times Q}$. By concatenating the embedded genomic features and phenotype features for all patients, we have $Z_P = [H_P|H_{Pheno}]$, where $Z_P \in \mathbb{R}^{M \times (K \times d^{L^{(g)}} + Q)}$. Leveraging the zero padding subfeature matrix, $H_{Pad} \in \mathbb{R}^{(\sum_{q=1}^Q n_{s_q}) \times (K \times d^{L^{(g)}})}$, we will have subfeature concatenated matrix $Z_{Sub} = [H_{Pad}|H_{Sub}]$, where $Z_{Sub} \in \mathbb{R}^{(\sum_{q=1}^Q n_{s_q}) \times (K \times d^{L^{(g)}} + Q)}$. In this way, the initial feature of the patient-feature graph with concatenated feature matrix will be $Z^{(0)} = [Z_P|Z_{Sub}]$, where $Z^{(0)} \in \mathbb{R}^{|V| \times (K \times d^{L^{(g)}} + Q)}$. For convenience, we will use $d^{(0)} = K \times d^{L^{(g)}} + Q$ and $|V| = R$.

In the end, the graph transformer will also be applied on the patient-feature graph message passing to embed the node features over layers $1, 2, \dots, l^{(pheno)}, \dots, L^{(pheno)}$. For certain layer $l^{(pheno)}$, the message passing by the transformer will have the same procedure for the gene graph. **Finally**, we will have the embedded node features in the $L^{(pheno)}$ layer with $Z^{(L^{(pheno)})} \in \mathbb{R}^{R \times d^{L^{(pheno)}}}$. As we will make the node classification task, the linear transformation will be made by

$$O = \text{OUTPUT}(Z^{(L^{(pheno)})})$$

, where $O \in \mathbb{R}^{C \times L^{(pheno)}}$ and C is the number of classes for the downstream task.

4.4 Experimental Settings and Classification Results

For the gene graph, we set the number of messages passing layer $L^{(g)} = 3$. For the patient-feature graph, we set the number of messages passing layer $L^{(pheno)} = 3$. And for the rest of model parameters, you can check the **Table 8** for reference.

Table 8. Model parameter and hyper-parameter settings

Patient-Feature Graph Parameters	Settings	Gene Graph Parameters	Settings
$[d^{(0)}, d^{(1)}, d^{(2)}, d^{(3)}]$	$[1, 3, 3, 3]$	$d^{(0)} = K \times d^{(3)} + Q$	96
K	18	$[d^{(0)}, d^{(1)}, d^{(2)}, d^{(3)}]$	$[96, 18, 18, 18]$
$[h^{(1)}, h^{(2)}, h^{(3)}]$	$[1, 1, 1]$	$[h^{(1)}, h^{(2)}, h^{(3)}]$	$[1, 1, 1]$
$[c^{(1)}, c^{(2)}, c^{(3)}]$	$[3, 3, 3]$	$[c^{(1)}, c^{(2)}, c^{(3)}]$	$[18, 18, 18]$
M	1405	N	3409
Q	42	C	3
R	1527		

Here, 5-fold cross-validation was used to run the experiments. Each used the same hyperparameters with the Adam optimizer, set with $\text{beta} = [0.8, 0.999]$, $\text{eps} = 1\text{e-}7$, $\text{weight_decay} = 1\text{e-}10$. The multi-step learning rate schedule function in Pytorch was used by setting the stage with [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1250] and $\text{gamma} = 0.9$ with learning rate starting from 0.01. To evaluate the contribution of each data modality and benchmark the dual-graph architecture against alternative GNN designs, we compared the full GiG model with two baseline architectures trained on the same combined input (GAT, GIN; Table 9). Across 5-fold cross-validation, the full GiG model achieved the highest mean test accuracy ($80.57\% \pm 0.79\%$), outperforming the baseline architectures (GAT: $75.15\% \pm 2.84\%$; GIN: $77.25\% \pm 0.54\%$). These results indicate that jointly modeling phenotype and genotype within the GiG dual-graph architecture provides complementary predictive information beyond GNN architecture alone.

Table 9. Test prediction accuracy of GiG compared with baseline GNN architectures (GAT, GIN) for three-class T2D/pre-T2D/No-T2D classification on the LLFS dataset using 5-fold cross-validation.

Data Split	GiG	GAT	GIN
Fold 1	81.48%	76.54%	76.54%
Fold 2	80.86%	70.99%	77.78%
Fold 3	80.25%	73.46%	77.16%
Fold 4	80.86%	77.78%	77.78%
Fold 5	79.39%	76.97%	76.97%
Average	$80.57\% \pm 0.79\%$	$75.15\% \pm 2.84\%$	$77.25\% \pm 0.54\%$

4.5 Drug-Evidence Augmentation of Prioritized Genes

To improve the translational interpretability of the GiG-prioritized signaling network, we implement a downstream drug-evidence augmentation module that maps prioritized genes to target-level pharmacologic evidence. This module is a post-GiG layer: it does not modify model

architecture, optimization, or classification performance, but adds a structured evidence layer to network-level mechanistic outputs.

The module takes as input the top 142 GiG-prioritized genes derived from the core T2D/pre-T2D network. In this input set, all target symbols are unique; therefore, no additional deduplication step is required in the manuscript workflow. Each target is propagated directly to evidence mining.

Evidence retrieval follows a rule-constrained, citation-auditable pipeline (Figure 1b). For each target, we query ChEMBL human-target records, mechanism records, and molecule annotations. Candidate target-drug links are retained only when mechanism-level records with traceable literature support are available. Linked PubMed records are then parsed to extract citation metadata (PMID/DOI). Development status is harmonized by a controlled vocabulary (approved, investigational, preclinical, unknown), with phase-based normalization when source annotations differ.

To minimize unsupported claims, evidence generation is separated from evidence validation. In the generation stage, candidate rows are assembled into a structured ledger with five fields: Target, Candidate Drug, Drug Class, Evidence Citation (must include PMID/DOI/Author-Year), and Development Status. In the validation stage, each non-placeholder row undergoes lightweight citation-consistency checks against PubMed metadata and mechanism context. Rows that fail confidence criteria are removed. Targets without remaining validated non-placeholder evidence are automatically reverted to a standardized placeholder row ("No validated drug"; "No direct evidence found"; status = unknown). This fallback policy prevents over-interpretation when direct evidence is insufficient.

The workflow is executed in batched mode to maintain deterministic parsing and auditable intermediate outputs. Reproducibility assets include batch-level draft ledgers, batch-level validated ledgers, row-level validation reports, and a merged final validated ledger with summary statistics. Raw API payloads (target/mechanism/molecule and PubMed XML) are cached for inspection. After ledger construction, validated target-drug rows are projected onto the GiG core network to create drug-augmented overlays while preserving original gene-gene topology and node-significance encoding. This projection supports direct interpretation of

translationally supported targets within the inflammatory and innate-immune network architecture identified by GiG.

Acknowledgment

We extend our gratitude to the Long Life Family Study consortium, including the participants, investigators, and administrative and clinical staff. This study was supported by NIA 4R33AG078799-02, NIA 1R21AG078799-01A1, NIA AG063893.

Data Availability

The data supporting this study are available from the NIA Long Life Family Study via the Synapse platform. Access details and application procedures can be found at <https://prod.eliteportal.synapse.org/Explore/Projects/DetailsPage?shortName=LLFS>.

References

1. Molla G, Bitew M. Revolutionizing Personalized Medicine: Synergy with Multi-Omics Data Generation, Main Hurdles, and Future Perspectives. *Biomedicines*. 2024;12(12):2750. doi:10.3390/biomedicines12122750
2. Mataraso SJ, Espinosa CA, Seong D, et al. A machine learning approach to leveraging electronic health records for enhanced omics analysis. *Nat Mach Intell*. 2025;7(2):293-306. doi:10.1038/s42256-024-00974-9
3. Wojczynski MK, Jiuan Lin S, Sebastiani P, et al. NIA Long Life Family Study: Objectives, Design, and Heritability of Cross-Sectional and Longitudinal Phenotypes. Le Couteur D, ed. *The Journals of Gerontology: Series A*. 2022;77(4):717-727. doi:10.1093/gerona/glab333
4. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Med*. 2018;1(1):18. doi:10.1038/s41746-018-0029-1
5. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform*. 2018;22(5):1589-1604. doi:10.1109/JBHI.2017.2767063
6. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep*. 2016;6(1):26094. doi:10.1038/srep26094
7. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z
8. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2
9. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3(3):199-217. doi:10.1038/s42256-021-00307-0
10. Li F, Zhang H, Huang D, et al. AI for Scientific Discovery in Omics Data-Driven Precision Medicine. *Mo Med*. 2026;123(1):55-66.

11. Mitchener L, Yiu A, Chang B, et al. Kosmos: An AI Scientist for Autonomous Discovery. *arXiv*. Preprint posted online 2025. doi:10.48550/ARXIV.2511.02824
12. Huang D, Li H, Li W, et al. OmniCellAgent: An AI Scientist for Omic-Driven Scientific Discovery.
13. Zhang H, Xu T, Cao D, et al. OmniCellTOSG: The First Cell Text-Omic Signaling Graphs Dataset for Graph Language Foundation Modeling. *Res Sq*. Published online February 17, 2026;rs.3.rs-8774770. doi:10.21203/rs.3.rs-8774770/v1
14. Zhang H, Liang S, Xu T, et al. BioMedGraphica: An All-in-One Platform for Joint Textual Biomedical Prior Knowledge and Numeric Graph Generation. *bioRxiv: The Preprint Server for Biology*. Preprint posted online September 17, 2025:2024.12.05.627020. doi:10.1101/2024.12.05.627020
15. Huang K, Zhang S, Wang H, et al. Biomni: A General-Purpose Biomedical AI Agent. *Bioinformatics*. Preprint posted online June 2, 2025. doi:10.1101/2025.05.30.656746
16. Gottweis J, Weng WH, Daryin A, et al. Accelerating scientific discovery with Co-Scientist. *Nature*. 2026;655(8122):487-496. doi:10.1038/s41586-026-10644-y
17. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Advances in neural information processing systems*. 2017;30.
18. Li CY, Liang X, Hu Z, Xing EP. Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. *AAAI*. 2019;33(01):6666-6673. doi:10.1609/aaai.v33i01.33016666
19. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. *arXiv*. Preprint posted online February 4, 2018:arXiv:1710.10903. doi:10.48550/arXiv.1710.10903
20. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:160902907*. Published online 2016.
21. Zhang H, Goedegebuure SP, Ding L, et al. M3NetFlow: A multi-scale multi-hop graph AI model for integrative multi-omic data analysis. *iScience*. 2025;28(3):111920. doi:10.1016/j.isci.2025.111920
22. Feng J, Song H, Province M, et al. PathFinder: a novel graph transformer model to infer multi-cell intra- and inter-cellular signaling pathways and communications. *Front Cell Neurosci*. 2024;18:1369242. doi:10.3389/fncel.2024.1369242
23. Dong Z, Zhao Q, Payne PRO, et al. Highly accurate disease diagnosis and highly reproducible biomarker identification with PathFormer. *Research Square*. Preprint posted online November 16, 2023;rs. 3.rs-3576068. doi:10.21203/rs.3.rs-3576068/v1
24. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans Neural Netw Learning Syst*. 2021;32(1):4-24. doi:10.1109/TNNLS.2020.2978386
25. Hinton GE, Osindero S, Teh YW. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*. 2006;18(7):1527-1554. doi:10.1162/neco.2006.18.7.1527
26. Doersch C. Tutorial on Variational Autoencoders. *arXiv*. Preprint posted online January 3, 2021:arXiv:1606.05908. doi:10.48550/arXiv.1606.05908
27. Kingma DP, Welling M. Auto-Encoding Variational Bayes. *arXiv*. Preprint posted online December 10, 2022:arXiv:1312.6114. doi:10.48550/arXiv.1312.6114
28. Chiu YC, Zheng S, Wang LJ, et al. Predicting and characterizing a cancer dependency map of tumors with deep learning. *Sci Adv*. 2021;7(34):eabh1275. doi:10.1126/sciadv.abh1275
29. Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*. Vol 2. IEEE; 2005:729-734. doi:10.1109/IJCNN.2005.1555942
30. Lazaros K, Koumadorakis DE, Vlamos P, Vrahatis AG. Graph Neural Network approaches for single-cell data: A recent overview. *arXiv*. Preprint posted online October 14, 2023:arXiv:2310.09561. doi:10.48550/arXiv.2310.09561

31. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. *Hum Genomics*. 2022;16(1):26. doi:10.1186/s40246-022-00396-x
32. Zhang XM, Liang L, Liu L, Tang MJ. Graph Neural Networks and Their Current Applications in Bioinformatics. *Front Genet*. 2021;12:690049. doi:10.3389/fgene.2021.690049
33. Zhang H, Chen Y, Payne P, Li F. Using DeepSignalingFlow to mine signaling flows interpreting mechanism of synergy of cocktails. *NPJ Syst Biol Appl*. 2024;10(1):92. doi:10.1038/s41540-024-00421-w
34. Zhang H, Cao D, Chen Z, et al. mosGraphGen: a novel tool to generate multi-omics signaling graphs to facilitate integrative and interpretable graph AI model development. *bioRxiv: The Preprint Server for Biology*. Preprint posted online August 27, 2024:2024.05.15.594360. doi:10.1101/2024.05.15.594360
35. Zhang H, Huang D, Chen E, et al. mosGraphGPT: a foundation model for multi-omic signaling graphs using generative AI. *bioRxiv: The Preprint Server for Biology*. Preprint posted online August 6, 2024:2024.08.01.606222. doi:10.1101/2024.08.01.606222
36. Li X, Xiao GY, Guo T, Song YJ, Li QM. Potential therapeutic role of pyroptosis mediated by the NLRP3 inflammasome in type 2 diabetes and its complications. *Front Endocrinol (Lausanne)*. 2022;13:986565. doi:10.3389/fendo.2022.986565
37. Li X, Xiao GY, Guo T, Song YJ, Li QM. Potential therapeutic role of pyroptosis mediated by the NLRP3 inflammasome in type 2 diabetes and its complications. *Front Endocrinol (Lausanne)*. 2022;13:986565. doi:10.3389/fendo.2022.986565
38. Tsetsos F, Roumeliotis A, Tsekmekidou X, et al. Genetic variation in CARD8, a gene coding for an NLRP3 inflammasome-associated protein, alters the genetic risk for diabetic nephropathy in the context of type 2 diabetes mellitus. *Diab Vasc Dis Res*. 2020;17(6):1479164120970892. doi:10.1177/1479164120970892
39. Kim JS, Kim HK, Lee J, et al. Inhibition of CD82 improves colitis by increasing NLRP3 deubiquitination by BRCC3. *Cell Mol Immunol*. 2023;20(2):189-200. doi:10.1038/s41423-022-00971-1
40. Pillon NJ, Chan KL, Zhang S, et al. Saturated fatty acids activate caspase-4/5 in human monocytes, triggering IL-1 β and IL-18 release. *Am J Physiol Endocrinol Metab*. 2016;311(5):E825-E835. doi:10.1152/ajpendo.00296.2016
41. Ding Y, Chen L, Xu J, Feng Y, Liu Q. APAF1 Silencing Ameliorates Diabetic Retinopathy by Suppressing Inflammation, Oxidative Stress, and Caspase-3/GSDME-Dependent Pyroptosis. *Diabetes Metab Syndr Obes*. 2024;17:1635-1649. doi:10.2147/DMSO.S449049
42. McDonald C, Chen FF, Ollendorff V, et al. A Role for Erbin in the Regulation of Nod2-dependent NF- κ B Signaling. *Journal of Biological Chemistry*. 2005;280(48):40301-40309. doi:10.1074/jbc.M508538200
43. Maedler K, Sergeev P, Ris F, et al. Glucose-induced beta cell production of IL-1 β contributes to glucotoxicity in human pancreatic islets. *J Clin Invest*. 2002;110(6):851-860. doi:10.1172/JCI15318
44. Xiang Y, Wang Z, Hui Q, Gwinn M, Vaccarino V, Sun YV. DNA Methylation of TXNIP Independently Associated with Inflammation and Diabetes Mellitus in Twins. *Twin Res Hum Genet*. 2021;24(5):273-280. doi:10.1017/thg.2021.42
45. Carlos D, Pérez MM, Leite JA, et al. NOD2 Deficiency Promotes Intestinal CD4⁺ T Lymphocyte Imbalance, Metainflammation, and Aggravates Type 2 Diabetes in Murine Model. *Front Immunol*. 2020;11:1265. doi:10.3389/fimmu.2020.01265
46. Shao L, Zhang P, Zhang Y, Ma A. Inflammatory unbalance of TLR3 and TLR4 in PCI patients with or without type 2 diabetes mellitus. *Immunol Lett*. 2014;161(1):81-88. doi:10.1016/j.imlet.2014.05.003
47. He W, Mu X, Wu X, et al. The cGAS-STING pathway: a therapeutic target in diabetes and its complications. *Burns & Trauma*. 2024;12:tkad050. doi:10.1093/burnst/tkad050

48. Dhafer AFAK. The IFIH1 gene polymorphism (Ala946Thr) association with the type 2 diabetes mellitus and hypertension among Iraqi patients. *Health Biotechnology and Biopharma journal*. Published online October 3, 2021. doi:10.22034/HBB.2021.15
49. Gu M, Ge J, Pan Q, Hu N, Hua F. Salivary microbiome variations in type 2 diabetes mellitus patients with different stages of periodontitis. *BMC Oral Health*. 2024;24(1):1424. doi:10.1186/s12903-024-05135-3
50. Yu T, Xu B, Bao M, et al. Identification of potential biomarkers and pathways associated with carotid atherosclerotic plaques in type 2 diabetes mellitus: A transcriptomics study. *Front Endocrinol*. 2022;13:981100. doi:10.3389/fendo.2022.981100
51. Meier DT, De Paula Souza J, Donath MY. Targeting the NLRP3 inflammasome–IL-1 β pathway in type 2 diabetes and obesity. *Diabetologia*. 2025;68(1):3-16. doi:10.1007/s00125-024-06306-1
52. Horitani S, Ueda Y, Kamioka Y, et al. The critical role of Rap1-GAPs Rasa3 and Sip1 in T cells for pulmonary transit and egress from the lymph nodes. *Front Immunol*. 2023;14:1234747. doi:10.3389/fimmu.2023.1234747
53. Qing F, Liu Z. Interferon regulatory factor 7 in inflammation, cancer and infection. *Front Immunol*. 2023;14:1190841. doi:10.3389/fimmu.2023.1190841
54. van Os BW, Vos WG, Bosmans LA, et al. CD40L modulates CD4+ T-cell activation through receptor for activated C kinase 1. *Eur J Immunol*. 2023;53(12):e2350520. doi:10.1002/eji.202350520
55. Lork M, Verhelst K, Beyaert R. CYLD, A20 and OTULIN deubiquitinases in NF- κ B signaling and cell death: so similar, yet so different. *Cell Death Differ*. 2017;24(7):1172-1183. doi:10.1038/cdd.2017.46
56. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*. 2000;28(1):27-30.
57. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556
58. The Gene Ontology Consortium, Aleksander SA, Balhoff JP, et al. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*. 2026;54(D1):D1779-D1792. doi:10.1093/nar/gkaf1292
59. Swanson KV, Deng M, Ting JPY. The NLRP3 inflammasome: molecular activation and regulation to therapeutics. *Nat Rev Immunol*. 2019;19(8):477-489. doi:10.1038/s41577-019-0165-0
60. Dai J, Zhou P, Li S, Qiu HJ. New Insights into the Crosstalk among the Interferon and Inflammatory Signaling Pathways in Response to Viral Infections: Defense or Homeostasis. *Viruses*. 2022;14(12):2798. doi:10.3390/v14122798
61. Liu J, Zhou J, Luan Y, et al. cGAS-STING, inflammasomes and pyroptosis: an overview of crosstalk mechanism of activation and regulation. *Cell Commun Signal*. 2024;22(1):22. doi:10.1186/s12964-023-01466-w
62. Rehwinkel J, Gack MU. RIG-I-like receptors: their regulation and roles in RNA sensing. *Nat Rev Immunol*. 2020;20(9):537-551. doi:10.1038/s41577-020-0288-3
63. Yuan M, Liu T, Cai A, et al. Emerging connectivity of programmed cell death pathways and pulmonary vascular remodelling during pulmonary hypertension. *J Cellular Molecular Medi*. 2024;28(16):e70003. doi:10.1111/jcmm.70003
64. Hoss F, Rodriguez-Alcazar JF, Latz E. Assembly and regulation of ASC specks. *Cell Mol Life Sci*. 2017;74(7):1211-1229. doi:10.1007/s00018-016-2396-6
65. Shi J, Zhao Y, Wang K, et al. Cleavage of GSDMD by inflammatory caspases determines pyroptotic cell death. *Nature*. 2015;526(7575):660-665. doi:10.1038/nature15514
66. Dai Z, Liu WC, Chen XY, Wang X, Li JL, Zhang X. Gasdermin D-mediated pyroptosis: mechanisms, diseases, and inhibitors. *Front Immunol*. 2023;14:1178662. doi:10.3389/fimmu.2023.1178662

67. Biasizzo M, Kopitar-Jerala N. Interplay Between NLRP3 Inflammasome and Autophagy. *Front Immunol*. 2020;11:591803. doi:10.3389/fimmu.2020.591803
68. Cedillo YE, Sammy MJ, Taylor MG, Hanick CJ, Peterson CM, Gower BA. Mitochondrial DNA DAMPs, Inflammation, and Insulin Sensitivity After Dietary Interventions in Adults with Type 2 Diabetes. *Nutrients*. 2025;17(20):3248. doi:10.3390/nu17203248
69. Oh S, Choi YK, Lee S. Pattern recognition receptors and inflammasome: Now and beyond. *Molecules and Cells*. 2025;48(8):100239. doi:10.1016/j.mocell.2025.100239
70. Du P, Fan B, Han H, et al. NOD2 promotes renal injury by exacerbating inflammation and podocyte insulin resistance in diabetic nephropathy. *Kidney International*. 2013;84(2):265-276. doi:10.1038/ki.2013.113
71. Tamrakar AK, Schertzer JD, Chiu TT, et al. NOD2 Activation Induces Muscle Cell-Autonomous Innate Immune Responses and Insulin Resistance. *Endocrinology*. 2010;151(12):5624-5637. doi:10.1210/en.2010-0437
72. Kim DH, Do MS. BAFF knockout improves systemic inflammation via regulating adipose tissue distribution in high-fat diet-induced obesity. *Exp Mol Med*. 2015;47(1):e129-e129. doi:10.1038/emmm.2014.98
73. Hamada M, Abe M, Miyake T, et al. B Cell-Activating Factor Controls the Production of Adipokines and Induces Insulin Resistance. *Obesity*. 2011;19(10):1915-1922. doi:10.1038/oby.2011.165
74. Jing G, Zuo J, Fang Q, et al. Erbin protects against sepsis-associated encephalopathy by attenuating microglia pyroptosis via IRE1 α /Xbp1s-Ca²⁺ axis. *J Neuroinflammation*. 2022;19(1):237. doi:10.1186/s12974-022-02598-5
75. Makinwa Y, Luo Y, Musich PR, Zou Y. Canonical and Noncanonical Functions of the BH3 Domain Protein Bid in Apoptosis, Oncogenesis, Cancer Therapeutics, and Aging. *Cancers*. 2024;16(12):2199. doi:10.3390/cancers16122199
76. Romero A, Dongil P, Valencia I, et al. Pharmacological Blockade of NLRP3 Inflammasome/IL-1 β -Positive Loop Mitigates Endothelial Cell Senescence and Dysfunction. *Aging and disease*. 2022;13(1):284. doi:10.14336/AD.2021.0617
77. Paik S, Kim JK, Silwal P, Sasakawa C, Jo EK. An update on the regulatory mechanisms of NLRP3 inflammasome activation. *Cell Mol Immunol*. 2021;18(5):1141-1160. doi:10.1038/s41423-021-00670-3
78. Grebe A, Hoss F, Latz E. NLRP3 Inflammasome and the IL-1 Pathway in Atherosclerosis. *Circulation Research*. 2018;122(12):1722-1740. doi:10.1161/CIRCRESAHA.118.311362
79. Cervantes J, Kanter JE. Monocyte and macrophage foam cells in diabetes-accelerated atherosclerosis. *Front Cardiovasc Med*. 2023;10:1213177. doi:10.3389/fcvm.2023.1213177
80. Qian X, Sun J, Li F, et al. Inflammasomes as the molecular hub of cardiovascular-metabolic-immune comorbidity networks. *Front Immunol*. 2025;16:1690443. doi:10.3389/fimmu.2025.1690443
81. Blagov AV, Markin AM, Bogatyreva AI, Tolstik TV, Sukhorukov VN, Orekhov AN. The Role of Macrophages in the Pathogenesis of Atherosclerosis. *Cells*. 2023;12(4):522. doi:10.3390/cells12040522
82. Doroszkiewicz J, Mroczko J, Winkel I, Mroczko B. Metabolic and Immune System Dysregulation: Unraveling the Connections between Alzheimer's Disease, Diabetes, Inflammatory Bowel Diseases, and Rheumatoid Arthritis. *J Clin Med*. 2024;13(17):5057. doi:10.3390/jcm13175057
83. Walin Y, Danqing Z, Wen C, Kun W. Research Progress on the Molecular Mechanisms by which Gut Microbiota Metabolites Affect the Development of Type 2 Diabetes Mellitus. *Acta Laser Biology Sinica*. 2025;34(2):113-120.
84. Arora A, Behl T, Sehgal A, et al. Unravelling the involvement of gut microbiota in type 2 diabetes mellitus. *Life Sciences*. 2021;273:119311. doi:10.1016/j.lfs.2021.119311

85. Joossens M, Huys G, Cnockaert M, et al. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut*. 2011;60(5):631-637. doi:10.1136/gut.2010.223263
86. Di Muzio C, Cipriani P, Ruscitti P. Rheumatoid Arthritis Treatment Options and Type 2 Diabetes: Unravelling the Association. *BioDrugs*. 2022;36(6):673-685. doi:10.1007/s40259-022-00561-7
87. Verma AK, Bhatt D, Goyal Y, et al. Association of Rheumatoid Arthritis with Diabetic Comorbidity: Correlating Accelerated Insulin Resistance to Inflammatory Responses in Patients. *JMDH*. 2021;Volume 14:809-820. doi:10.2147/JMDH.S285469
88. Zemedikun DT, Gokhale K, Chandan JS, et al. Type 2 diabetes mellitus, glycaemic control, associated therapies and risk of rheumatoid arthritis: a retrospective cohort study. *Rheumatology*. 2021;60(12):5567-5575. doi:10.1093/rheumatology/keab148
89. Luck H, Khan S, Kim JH, et al. Gut-associated IgA+ immune cells regulate obesity-related insulin resistance. *Nat Commun*. 2019;10(1):3650. doi:10.1038/s41467-019-11370-y
90. Chiao HY, Wang CY, Wang CH. Salmonella Abscess of the Anterior Chest Wall in a Patient With Type 2 Diabetes and Poor Glycemic Control: A Case Report. *Ostomy Wound Manage*. 2016;62(3):46-49.
91. Pastagia M, Jenkins SG. Salmonella neck abscess as an opportunistic infection in diabetes mellitus. *Case Rep Infect Dis*. 2013;2013:708419. doi:10.1155/2013/708419
92. Yorke E, Atiase Y, Akpalu J, Sarfo-Kantanka O, Boima V, Dey ID. The Bidirectional Relationship between Tuberculosis and Diabetes. *Tuberc Res Treat*. 2017;2017:1702578. doi:10.1155/2017/1702578
93. van Crevel R, Critchley JA. The Interaction of Diabetes and Tuberculosis: Translating Research to Policy and Practice. *Trop Med Infect Dis*. 2021;6(1):8. doi:10.3390/tropicalmed6010008
94. Acharya S, Moghaddam VA, Jung WJ, et al. Finding Significant Hits in Networks: a network-based tool for analyzing gene-level *P*-values to identify significant genes missed by standard methods. *Briefings in Bioinformatics*. 2026;27(2):bbag061. doi:10.1093/bib/bbag061
95. Shi Y, Huang Z, Feng S, Zhong H, Wang W, Sun Y. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization; 2021:1548-1554. doi:10.24963/ijcai.2021/214
96. Ying Z, You J, Morris C, Ren X, Hamilton W, Leskovec J. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*. 2018;31.