Journal of Biomedical Informatics 178 (2026) 105028

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

A heterogeneous graph neural network for candidate gene prediction in endometriosis

Zelia Soo [a,*], Hua Lin [b], Mengjia Wu [a], Kairui Guo [a], Mark Grosser [b], Yi Zhang [a], Jie Lu [a]

[a] Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, 61 Broadway, Ultimo, 2007, NSW, Australia

[b] 23Strands, Suite 107 26-32 Pirrama Road, Pyrmont, 2009, NSW, Australia

ARTICLE INFO

Dataset link: https://github.com/zsoo99999/HetBio-CLiP.git

Keywords:
Gene–disease association
Graph neural network
Gene prediction
Endometriosis

ABSTRACT

Objective: Precision medicine applications depend on elucidating the underlying molecular mechanisms of disease. However, many disorders, such as endometriosis, remain poorly characterised genetically due to data scarcity, positive-unlabelled (PU) imbalance, and the heterogeneous structure of biomedical knowledge. This study aims to develop HetBio-CLiP, a novel heterogeneous graph-based contrastive learning (CL) methodology that improves candidate gene prioritisation across heterogeneous biomedical graphs by explicitly addressing positive-unlabelled learning constraints and structural heterogeneity.

Methods: HetBio-CLiP (Heterogeneous Biomedical graph Contrastive Learning with Interactions and Positive-Unlabelled learning) integrates multi-relational genomic, variant, and clinical data from real-world patient cohorts. To address the challenges of data scarcity and class imbalance, the graph neural network (GNN) methodology combines heterogeneous graph CL with PU learning. Furthermore, the model incorporates an interpretable GNNShap explainer to provide transparency at both the feature and edge levels, to assess the biological relevance of the predictions.

Results: HetBio-CLiP achieved superior performance across key evaluation metrics, attaining the highest Area Under the Curve (AUC) of 0.9489 ± 0.04 and Area Under the Precision–Recall Curve (AUPR) of 0.9401 ± 0.05. This outperformed state-of-the-art baselines such as M-SAGEGraph (AUC 0.9149 ± 0.04) and GAT (AUC 0.9138 ± 0.07). In terms of ranking candidate genes, the model achieved the highest Precision@42 (0.6933 ± 0.06) and TP@42 (29.1 ± 2.4) and second-highest Normalised Discounted Cumulative Gain (NDCG@42) (0.827 ± 0.06). Ablation studies confirmed that the integration of PU and contrastive learning was essential for these performance gains. Biologically, the model successfully prioritised known pathophysiology drivers, with top-ranked validated gene predictions including WNT4, GREB1, and ESR1.

Conclusion: HetBio-CLiP presents an effective and interpretable methodology for uncovering candidate gene–disease relationships in data-scarce environments. By combining CL with PU-training, the proposed method addresses critical limitations of current gene prioritisation models and produces biologically meaningful predictions. The model demonstrates a strong balance between ranking accuracy and recall, identifying broader sets of plausible candidates than existing methods. This approach benefits researchers, clinicians, and downstream genomic studies by offering an improved foundation for identifying novel disease-associated genes and guiding future experimental investigation in endometriosis and other complex diseases.

1. Introduction

Endometriosis, a disease that is caused by the presence of endometrial-like tissue outside the uterus, affects approximately 10% of women globally [1]. Despite widespread prevalence, very little is understood about the causes of this disease. The heritability of the disease is estimated to be 50% and genetic variation could account for 26% of endometriosis cases, but only 1.75% of cases have been explained by currently identified loci. This motivates the need for computational models that can recommend novel genes most likely implicated in the disease [2]. By identifying affected genes, the understanding behind the causal mechanisms can provide biologically interpretable hypotheses for validation and therefore help in disease screening or treatment [3,4]. These observations highlight the need for systematic approaches that can prioritise disease-relevant genes in the context of incomplete and heterogeneous evidence.

* Corresponding author.
E-mail address: zeliamei-ze.soo@student.uts.edu.au (Z. Soo).

https://doi.org/10.1016/j.jbi.2026.105028
Received 10 December 2025; Received in revised form 15 February 2026; Accepted 20 March 2026
Available online 23 March 2026
1532-0464/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Candidate gene prioritisation aims to support disease research by stratifying the most promising genes according to their likelihood of involvement in a disease [5]. Early approaches involved studying affected individuals and their families. However, computational methods are now widely used to leverage prior knowledge and existing databases to provide in-silico rankings of potential disease-related genes [5]. Considerable progress has been made in curating gene–disease associations through databases such as ClinVar [6], ClinGen [7], the Human Phenotype Ontology [8], and DisGeNet [9]. While these databases collate evidence-based associated variants and genes associated with various diseases, the findings are usually incomplete or biased towards well-studied conditions, leaving complex and heterogeneous diseases such as endometriosis poorly characterised [3]. Furthermore, computational methods produce predictions that can be difficult to interpret, limiting transparency and clinical trust.

In response to these limitations, recent advances in graph representation learning (GRL) have transformed the way biomedical data can be modelled and analysed [10]. GRL has enabled powerful integration of heterogeneous data that can capture underlying structures and relationships between entities [11]. Graph neural networks (GNNs), a powerful deep learning class of GRL, have demonstrated strong performance across a wide range of biomedical and non-biomedical applications [12]. This ability to learn latent graph features, including topology and attribute, into low-dimensional vectors means that there is a broad applicability in genomics to utilise a full gamut of disease related and gene related information [13]. Gene–disease association prediction could also be conceptualised as a biological recommendation problem, where candidate genes can be recommended to a disease based on their multi-relational evidence. GNNs used in recommender systems, such as LightGCN, learn embeddings that capture high-order relational structures to rank interactions [14]. This same mechanism could be used in this biological setting to uncover disease-specific gene preferences. However, most existing GNN-based methods in gene–disease prediction face three challenges: limited integration of real-world patient genomic data, lack of mechanisms to handle label scarcity, and insufficient interpretability to support clinical or experimental validation.

To address these challenges, we propose a Heterogeneous Biomedical graph Contrastive Learning with Interactions and PU learning (HetBio-CLiP), a heterogeneous graph neural network methodology designed to identify clinically relevant candidate disease genes by integrating contrastive learning (CL), positive-unlabelled (PU) learning, and explainable AI (XAI) over multi-modal biomedical data. CL has emerged as a powerful technique to enhance generalisation under limited supervision by enforcing semantic consistency between biologically similar entities, even when labels are scarce [15]. PU learning addresses the label scarcity problem by learning from known positive associations while treating the remaining unlabelled data as uncertain rather than negative, which better reflects the reality of incomplete biomedical knowledge in genomic data [16]. Finally, XAI provides transparency and interpretability by identifying which genomic features, interactions, or biological pathways most strongly contribute to the model's predictions, thus enhances the clinical utility of HetBio-CLiP's findings [17]. By highlighting key features and the decisions made by the model that can be evaluated by human professionals [18]. The framework is demonstrated in the context of endometriosis but can be broadly applicable to other diseases characterised by incomplete genomic knowledge. HetBio-CLiP incorporates real-world genomic and transcriptomic data into a unified heterogeneous graph, enabling robust, interpretable, and generalisable gene–disease prediction.

The main methodological contributions of this study are as follows:

- We produced a novel heterogeneous graph construction pipeline that integrates multi-source biomedical information—including knowledge-based associations, single-cell co-expression networks, and patient-level genomic variants—into a unified representation suitable for learning and inference.
- We developed a semi-supervised learning framework that integrates contrastive learning and positive unlabelled learning, to enhance prediction accuracy under limited and noisy labelling conditions, a common challenge in translational bioinformatics.
- We developed an explainable graph learning module, integrating SHAP and GNNShap, to provide interpretable feature- and edge-level attributions that align with biological knowledge and support expert validation.

The structure of this paper is as follows: Section 2 presents the related work. Section 3 summarises the methodology. Section 4 outlines the experiment undertaken to test the methodology. Section 5 presents and discusses the results from the experiments. The last section offers conclusions and directions for further research.

Statement of Significance	
Problem or issue:	Discovering genes associated with complex diseases such as endometriosis is limited by sparse labelled data, strong positive-unlabelled class imbalance, and the heterogeneous nature of biomedical knowledge distributed across sources. These limitations hinder the effectiveness of network-based gene prioritisation methods.
What is already known:	Existing network-based and graph neural network models have been applied to disease gene prioritisation but often rely on fixed supervision and struggle to capture the complexity of multi-type biomedical relationships, resulting in reduced robustness and limiting their ability to utilise unlabelled genes well.
What this paper adds:	We introduce HetBio-CLiP, a contrastive learning framework that integrates heterogeneous biomedical networks with positive-unlabelled learning. By jointly modelling genes, variants, diseases, and transcriptomic relationships, with real-world patient cohort data from UKBB and AoU and single cell RNA-seq, HetBio-CLiP improves gene prioritisation performance and re-prioritises biologically relevant endometriosis-associated genes absent from curated training labels.
Who would benefit:	Researchers studying the molecular mechanisms of endometriosis, bioinformatician and biomedical AI developers developing disease gene prioritisation methods, and clinicians seeking mechanistic insight into endometriosis will benefit from this framework and its application.

2. Related work

To establish our methodology, the following areas need to be studied: GNNs and their application in gene–disease associations, XAI in gene–disease association GNNs, and GNNs in recommender systems.

2.1. Graph neural networks and their applications in gene–disease association

Many GNNs have been applied towards the gene–disease association problem [19]. Homogeneous graphs have only one type of node and

edge, while heterogeneous graphs have multiple types of nodes and edges [12]. Graph convolutional network (GCN) learns node embeddings by implementing convolutional operations on a graph based on the properties of neighbouring nodes [20]. Schulte-Sasse et al. used a 3 layer GCN in their gene–disease association model EMOGI for cancer genes, by adding multiomics data as feature vectors in a protein–protein interaction graph [21]. Graph Sample and Aggregate (GraphSAGE) is a framework for generating node embeddings in large graphs [22]. Its ability to generate embeddings from unseen nodes allow it to capture global structures and feature information. Mastropietro et al. used a 7-layer GraphSAGE model for cancer genes in a PPI [13]. Graph Attention Networks (GAT) is an extension of GCN whereby it assigns weights to different neighbours with masked self-attention layers [23]. This allows the model to adapt weights towards important neighbours for higher accuracy. Many gene–disease association graphs utilise this, though mainly for lncRNA or miRNA purposes [15,24].

The utility of heterogeneous graphs in GNN-based methodologies have attracted significant attention due to their prevalence in real-world scenarios and ability to model complex relationships. The ability to use multi-dimensional and complementary information representation has a huge advantage in disease gene prediction over homogeneous data [12]. Yang et al. developed HerGePred, a low-dimensional vector representation of nodes in a heterogeneous disease-gene network [2]. Jia et al. developed ModulePred, which utilises a heterogeneous module network of genes, diseases, protein complexes and protein interactions. By using L3 link prediction algorithms to augment the PPI network, ModulePred was able to integrate biological motivations into its prediction of candidate genes for different diseases [25]. Sebastian et al. aimed to determine the mechanisms behind Alzheimer's disease (AD) genes, and the control they might have over the entire gene network [26]. By using multi-omics data alongside PPI and GO networks in a graph diffusion network, they found that a majority of AD genes are highly connected with other genes related to AD.

A limitation for these GNN-based methods is that it is reliant on labelled data since it is supervised learning. This is further exacerbated since gene discovery can be labelled as a PU learning problem, where all unknown genes cannot be assigned a negative label [16]. There have been some PU algorithms developed for the gene–disease problem, but not in GNN; PUDI [27] and ProDiGe [28]. To overcome this, methods such as graph contrastive learning (GCL) have been developed. This is a type of self-supervised clustering technique that reduces the distance between positive pairs and increases the distance between negative pairs in the embedding space. Li et al. based their GCL model on PU learning to reduce biased and improve performance of the model in node classification, node clustering, and link prediction [29].

2.2. Explainable AI in gene–disease association graph neural networks

One of the main challenges of applying deep learning in healthcare is the lack of interpretability and explainability [30]. Improving our understanding of how models generate predictions can enhance their real-world applicability and trustworthiness. This has driven significant efforts to develop explainable AI (XAI) methods for interpreting neural network outputs, including those of graph neural networks (GNNs) [31]. Existing XAI methods for GNNs can be broadly grouped into model-level and instance-level explainers.

Model level methods explain general behaviour and provide input independent explanations for the whole graph, such as XGNN [32]. Instance level methods explain input-dependent explanations that identify important features for the prediction. Most of the XAI algorithms developed are categorised as instance level. GNNExplainer was the first method designed for XAI on GNNs. It provides explanation for important nodes and edges by optimising soft masks to maximise mutual information between the original predictions and new predictions, thereby highlighting the subgraph structures most influential to the model's output [31]. Layer-wise relevance propagation (LRP) is an explainability technique that traces a model's prediction backward through its layers to identify the key factors that contributed to the decision [33]. The EMOGI algorithm used LRP to identify the particular omics-level data or PPI modules used to classify a gene as cancerous. By identifying the contributing features of each gene as well as the most important direct interaction in PPI, they were able to provide more mechanistic insights into oncogenesis [21]. In EMGNN, the IG module in Captum was used. It identifies the most important node and edge features through decomposition. It also used a gene set enrichment analysis (GSEA) to understand the biological mechanisms, ranking each gene from the aggregate maximum feature importance of each node. From this, it was able to find which feature vectors were most important [34].

Most XAI methods still emphasise isolated nodes or edges, and may overlook the joint contribution of connected subgraphs. Dai et al. addresses this by performing key subgraph retrieval based on Euclidean distances between node embeddings [35]. These results show that subgraph level reasoning provides more stable and meaningful explanations rather than node-level perturbation approaches. There are also quite a few XAI algorithms that are based on Shapley values, including GraphSVX, SubgraphX, and GNNShap. GraphSVX explains the node and node features for a prediction, constructing a surrogate model on a perturbed dataset to achieve this [36]. SubgraphX explores subgraphs and generates the subgraph with the most important features [37]. GNNShap is a model-agnostic method for interpreting machine learning methods that provides an importance score for all relevant edges when performing prediction of a target node [38]. In XGDAG, three XAI algorithms: GNNExplainer, GraphSVX, and SubgraphX, were tested to find the most influential nodes for prediction and found that the GNNExplainer and GraphSVX performed best [13].

2.3. Graph neural networks in recommender systems

Graph neural networks have also become foundational in modern recommender systems, where the task is to model interactions between users and items to predict future preferences or rank unseen items. By representing users, items, and their interactions as nodes and edges, GNN-based recommenders can effectively capture higher-order collaborative signals beyond direct connections [39]. This has helped overcome the problems of sparsity and scalability that standard recommender systems typically experience, enabling richer representations of user-item relationships [40].

Early work such as Neural Graph Collaborative Filtering (NGCF) used a GCN to leverage message passing on bipartite graphs to capture higher-order connectivity. The Light Graph Convolutional Network (LightGCN) improved upon NGCF by only using the most important components of the GCN in the model since it was found that adding additional operations degraded recommendation performance [14]. By focusing on neighbourhood aggregation LightGCN was able to improve top-K recommendation accuracy. This recommender system framework naturally aligns with the gene–disease association problem, where diseases correspond to users and genes to items being recommended. By adopting this recommender perspective, biomedical graph learning models may better prioritise disease-relevant genes through ranking mechanisms rather than rigid binary classification, thereby aligning with the discovery-oriented nature of genomics research.

3. Methodology

This section outlines the methodology of HetBio-CLiP, outlining the issues that have been identified by the related works in Section 2. Section 3.1 leverages diverse genomic and transcriptomic data to construct a biologically informed graph, where nodes represent entities such as genes, variants, and diseases, and edges capture their functional and clinical relationships. Section 3.2 outlines how we employ a graph neural network (GNN) backbone to learn low-dimensional embeddings

Table 1
A summary of graph network node and associated attributes.

Node type	Node attributes	Attribute descriptions
Gene	Gene full name	Complete, formal name of the gene.
	Gene NCBI type	Classification of the gene (e.g., protein-coding) from NCBI.
	Gene pLI and DPI	**pLI:** Probability of Loss-of-Function Intolerance measures sensitivity to mutations. **DPI:** Disease Pleiotropy Index measures diversity of associated diseases based on MeSH terms.
	Protein class	Functional group the gene product belongs to (e.g., enzymes).
	Functional annotations	Tags describing biological pathways or molecular functions (e.g., GO/KEGG).
	Encoded category	Embedded version of the gene category for use in machine learning.
	Gene summary as vector	Numerical representation of the gene summary for machine learning.
	Percentage upregulated in scRNA	The proportion of cells in a single-cell RNA sequencing dataset where this gene shows significantly increased expression
Variant	Chromosome and position dbSNP class Variant summary vector	Specific coordinate where variation occurs. Type of variation such as SNP or indel. Numerical embedding of variant features for machine learning.
	Allele frequency	Frequency calculated across whole-genome or whole-exome datasets.
	Polyphen and SIFT scores	In-silico prediction algorithms showing of the impact of an amino acid substitution on protein function
	Real-world cohort	Frequency observed in clinical or patient populations.
Disease	Genes associated Variants associated	Total count of distinct genes linked to the disease. Total count of specific genetic variations linked to the disease.
	Disease class	Categorisation based on etiology, pathology, or affected body system.

that capture the multi-relational structure of the data. The model is trained using a contrastive objective to enhance representation quality and a PU-based loss to handle label uncertainty. Finally, an explainability module is applied to interpret the model's predictions and identify biologically meaningful associations.

3.1. Heterogeneous graph construction

To model gene–disease associations in endometriosis, we constructed a heterogeneous graph incorporating multiple biomedical entities from multiple databases. This includes diseases, genes, DNA variants, RNA, and proteins. These entities and their interactions provide a diverse biological context, which is critical for identifying novel gene–disease associations. We integrated multiple public and curated datasets to capture these relationships comprehensively. summarise the associated attributes of all nodes and edges.

Gene–disease, variant-disease, and disease-disease associations were obtained from DISGENET, using the query "endometriosis" [9]. The associated features from this data were also used as node and edge attributes. Further detail about the disease nodes are found in Appendix A. To enrich gene-level information specific to endometriosis, we applied the literature-based natural language processing tool BiblioEngine [41]. Using a naïve PubMed query "endometriosis AND (gene* OR genom*)", BiblioEngine extracted candidate gene names for endometriosis from relevant papers. These genes were then used to retrieve protein–protein interactions (PPI) from STRING-DB. These were used to provide additional gene node attributes [42]. Variants related to endometriosis were also sourced from the polygenic risk score catalogue, (PGS ID: PGS001866) [43].

To incorporate real-world genetic variation, we included patient-level genomic data from two large biobanks. The UK BioBank (UKBB) is a large-scale comprehensive biomedical database, which started in 2006, containing whole genome sequences and other genetic data, along with other labelled health information from 500,000 UK participants. The All of Us (AoU) Research Program is a large-scale longitudinal study based in the USA that aims to generate whole genome sequence data for one million participants from diverse ancestries [44]. In this study, 7035 endometriosis cases from the UKBB and 4036 from AoU were included. These individuals were identified using the ICD9 (617) or ICD10 (N80) endometriosis labels. The single nucleotide polymorphisms identified from PGS001866 were extracted from the gVCF files in the cohort databases. BCFtools, a widely used C-based bioinformatics toolkit was used to separate out the chosen variants [45]. Candidate genes from the BiblioEngine search were also used to identify structural variants, such as small insertions and deletions (indels) and copy number variants (CNVs), which were incorporated as additional variant nodes. The genes from BiblioEngine were mapped to their chromosomal coordinates using SQL queries to identify their "chr:pos" ranges. The gVCF-derived variant files were then filtered to retain only variants located within these gene regions and with a genotype quality (GQ) score ≥ 20, a common quality threshold in genomic studies. Each variant node, representing an observed genomic variant present in individuals diagnosed with endometriosis, is connected to its corresponding gene based on genomic location. Variants and their population-based statistics (e.g. allele frequency, allele presence) were also linked to disease nodes to enrich the case-specific genetic evidence without assuming direct causality. By this, variant nodes will serve as intermediaries that propagate disease-related information to gene nodes during message passing. Genes with multiple variants observed across independent cohorts may accumulate stronger disease-relevant representations, enabling the model to leverage real-world genetic variants alongside curated knowledge sources. Additionally, to capture transcriptional context, we integrated single-cell RNA sequencing data from the E-MTAB-14039 dataset from BioStudies. This contained 103 samples from endometriosis and healthy cell tissues [46]. Standard filtering and clustering techniques were done using the R package Seurat (v. 5.3.0) [47], and DoubletFinder (v. 2.0.6) [48], which were

Table 2
A summary of graph network edge and associated attributes.

Edge type	Edge attributes	Attribute descriptions
Gene–Gene	Co-expression	Score that shows gene pairs with similar expression patterns, suggesting functional linkage
	Experimentally determined interaction	Physical protein–protein interactions (PPI) that have been validated through laboratory assays
	Combined score	Computed by combining the probabilities from the different evidence channels and corrected for the probability of randomly observing an interaction
	Gene fusion	Identifies open reading frames that result from gene fusion events between genes
	Phylogenetic co-occurrence	Score that shows the amount of conservation of the linked homologous protein across species
	Homology	Sequence similarity between genes that suggests they share a common evolutionary ancestor
	Neighbourhood on chromosome	Association score to gene pairs that are located close to each other on the chromosome in the same orientation
	Co-expression in cell type	Specific mRNA correlation restricted to a particular cell population
Gene-Variant	Variant is in this gene	The genomic coordinate of the variant falls within the start and end positions of the gene
Gene–Disease	Evidence level	A metric used by ClinGen and GenCC that measures the strength of evidence of a gene–disease relationship that correlates to a qualitative classification
	Evidence index	Indicates the existence of contradictory results in publications supporting the variant-disease associations
	Pathway names and ID	Specific biological pathways (e.g., KEGG or Reactome) that both the gene and the disease are significantly involved in
	Number of PMIDs in evidence	The total count of research papers in PubMed that link the gene and the disease
	GDA score	Gene–Disease association is a value from DISGENET that quantifies the strength of this association
Variant-Disease	VDA score	Variant-Disease association is a value from DISGENET that quantifies the strength of this association
	Evidence index VDA	Indicates the existence of contradictory results in publications supporting the variant-disease associations
Disease-Disease	Number of shared genes	The count of distinct genes that are significantly associated with both diseases
	Number of shared variants	The count of specific genetic variations that are common to both diseases
	Jaccard index genes	A statistical measure of similarity between the sets of genes associated with each disease
	Jaccard index variants	A statistical measure of similarity between the sets of variants associated with each disease

used to create cell-type-specific co-expression networks. Gene pairs exceeding a correlation threshold (r greater > 0.7) were incorporated as ‘‘co-expressed-in-celltyp’’ relations, enabling the model to capture transcriptional context within the heterogeneous graph.

To harmonise these disparate data types for use into this heterogeneous graph, node and edge attributes were normalised within their respective modalities prior to graph construction. The choice of normalisation method was guided by the statistical properties of the feature type. For example, continuous features such as PPI confidence scores which are bounded confidence measures used min–max normalisation, while gene–gene co-expression correlations used z-score normalisation since this is an unbounded feature. Binary and categorical attributes were retained without transformation. Features such as genomic variant data were also retained since they had been normalised during pre-processing. Non-numerical node and edge attributes were encoded using one hot encoding or BioWordVec word embeddings [49]. No manual cross-modal weighting was applied between different data sources. Each edge type retained its own semantic meaning, and heterogeneous relationships were learned jointly by the GNN through type-specific message passing. This ensured that heuristic weighting across disparate modalities was avoided and allowed the final GNN methodology to learn the relative importance of each different evidence source in an end-to-end manner.

The final graph network was constructed using PyTorch for ease of using it as input into the neural network in the later steps. Altogether, the final graph consisted of 2776 genes nodes, 33 disease nodes, and 610 variant nodes, with genomic and functional attributes assigned to gene and variant nodes. Out of the 2776 gene nodes, 42 of these had been externally validated as relating to endometriosis from a genome-wide association meta-analysis study [50]. The construction of this network is found in Fig. 1A.

3.2. HetBio-CLiP

3.2.1. Graph construction

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ be a heterogeneous graph, where:

- $\mathcal{V}$ is the set of nodes, partitioned into different types, $\mathcal{V} = U_t$, where U_t denotes the types of nodes and t denotes a specific node type. Genes nodes are $\mathcal{V} = U_{gene}$, variant nodes are $\mathcal{V} = U_{variant}$, and disease nodes are $\mathcal{V} = U_{disease}$.
- $\mathcal{E}$ is the set of edges, where each edge $(i, j, r) \in E$ connects node i to node j via relation $r \in \mathcal{R}$, the set of edge types.

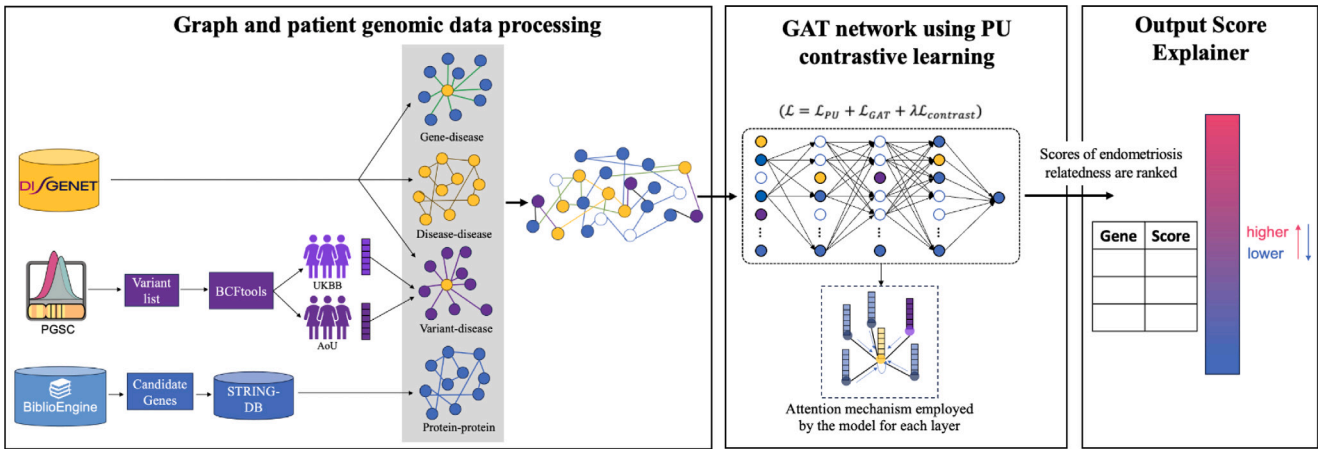


Fig. 1. Overview of graph and data processing, GAT network, and the output score explainer. (A) depicts the data processing process using databases and genomic data, with graph construction incorporating all collated data. (B) depicts the GAT network using PU and CL, highlighting the novel loss function of HetBio-CLiP. (C) depicts the final output score and the associated SHAP explainer module to identify important features.

- Each node i has a feature vector $\mathbf{h}_i \in \mathbb{R}^d$ where d is the feature dimension.

The attention mechanism in GAT is extended to heterogeneous graphs by computing relation-specific attention scores. For a given relation r, the attention coefficient between nodes i and j is defined as:

$$a_{ij}^{(r)} = \frac{\exp\left(\text{LeakyReLu}\left(\mathbf{a}^{(r)T}\left[\mathbf{W}^{(r)}\mathbf{h}_i \parallel \mathbf{W}^{(r)}\mathbf{h}_j\right)\right)}{\sum_{k\in\mathcal{N}_i^r}\exp\left(\text{LeakyReLu}\left(\mathbf{a}^{(r)T}\left[\mathbf{W}^{(r)}\mathbf{h}_i \parallel \mathbf{W}^{(r)}\mathbf{h}_k\right)\right)\right)} \tag{1}$$

where:

- $\mathbf{W}^{(r)}$ is a learnable transformation matrix for relation r.
- $\mathbf{a}^{(r)}$ is a relation-specific attention vector.
- $\mathcal{N}_i^r$ is the set of neighbours of node i connected by relation r.
- $\parallel$ denotes concatenation.

To aggregate information across different relation types, a weighted sum of transformed features is computed as:

$$h_i' = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r} a_{ij}^{(r)}\mathbf{W}^{(r)}\mathbf{h}_j\right) \tag{2}$$

where σ is a non-linear activation LeakyReLu.

To train the heterogeneous GAT model, we optimise a cross-entropy loss function over the labelled gene nodes. Given the predicted class probabilities for a gene node, and its ground truth label, the loss function is defined as:

$$\mathcal{L}_{GAT} = -\sum_{i\in\mathcal{V}_L} y_i \log\hat{y}_i + (1-y_i)\log(1-\hat{y}_i) \tag{3}$$

where:

- $\mathcal{V}_L$ denotes the set of labelled gene nodes.
- y_i is the binary label indicating whether the gene is associated with endometriosis.
- $\hat{y}_i$ is the predicted probability of the gene being associated with endometriosis.

During training, the loss is backpropagated through the network to update the learnable parameters of the model, including the relation-specific transformation matrices $\mathbf{W}^{(r)}$ and attention vectors $\mathbf{a}^{(r)}$. The model parameters are optimised using the Adam optimiser, with early stopping applied based on validation loss to prevent overfitting.

3.2.2. Contrastive attention learning

CL is incorporated to ensure that gene nodes with similar properties have similar attention distributions, while dissimilar gene nodes have different distributions. Given a set of positive pairs P (similar nodes) and negative pairs N (dissimilar nodes), we define contrastive attention loss as:

$$\mathcal{L}_{contrast} = \sum_{(i,j)\in P}\log\sigma\left(\alpha_i^T\alpha_j\right) + \sum_{(i,k)\in N}\log\sigma\left(-\alpha_i^T\alpha_k\right) \tag{4}$$

where:

- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.
- α is the attention distribution of node i, i.e. $\alpha_i = \left[\alpha_{ij}^{(r)}\right]$ for all $j \in \mathcal{N}_i^r$ and $r \in \mathcal{R}$.
- P is constructed based on domain knowledge (e.g. genes with known associations to endometriosis).
- N consists of structurally or functionally different nodes.

If no labels are available, CL was used to encourage robust gene representations by defining positive and negative node pairs based on graph structure and relational context. Positive pairs were generated as augmented views of the same gene node obtained by perturbing feature representations or neighbourhood information. For example, genes sharing at least one common neighbour within a 1-hop neighbourhood, such as a common disease association, shared variant, or direct gene–gene interaction. Negative pairs were sampled from genes nodes separated beyond a fixed k-hop neighbourhood (k = 2), as this distance typically delineates local functional clusters in biological networks [51], and no shared disease or variant associations. For example, a hormone gene linked to the reproductive pathway and an immune related gene linked to an inflammatory response. This ensures that negative pairs represent functionally and topologically distinct entities within the heterogeneous graph. This negative sampling strategy reduces the likelihood of sampling similar genes as negative, which could weaken the contrastive supervision.

CL was applied exclusively to gene nodes, as gene prioritisation constitutes the primary prediction objective in this study. Unlike gene nodes, variant nodes are very sparse and cohort-specific, and disease nodes represents high-level diagnostic categories that can aggregate diverse molecular mechanisms. Thus, by choosing to use gene nodes only, additional noise or bias is avoided.

3.2.3. Node and edge masking

Node masking is applied to control which nodes contribute to model training and evaluation. A mask vector $m_i \in [0, 1]$ is assigned to each

node.

$$\mathbf{h}_i^{masked} = m_i \mathbf{h}_i \tag{5}$$

If $m_i = 0$, the node's features are masked, preventing it from influencing the training process. Edge masking selectively removes edges from the training process by defining an edge mask function m_{ij} such that,

$$\alpha_{ij}^{(r),masked} = m_{ij} \alpha_{ij}^{(r)} \tag{6}$$

where $m_{ij} \in [0, 1]$ indicates whether the edge is considered during attention computation.

3.2.4. Positive-unlabelled learning framework

In this model we also employ a PU learning framework where only a subset of nodes is labelled as positive, and the remaining nodes are treated as unlabelled rather than explicitly negative. The PU learning objective is defined as:

$$\mathcal{L}_{PU} = \pi_P \mathbb{E}_{i \sim P} \left[l \left(f(\mathbf{h}_i), 1 \right) \right] + \mathbb{E}_{i \sim U} \left[l \left(f(\mathbf{h}_i), 0 \right) \right] \tag{7}$$

where:

- π_P is the prior probability of a node being positive.
- P denotes the set of known positive nodes.
- U is the unlabelled set which may contain both positive and negative nodes.
- l is a classification loss function, cross entropy.

π_P is a hyperparameter that is tuned via grid search since the true proportion of genes is unknown and unlabelled nodes may contain latent positives. Specifically, the values tested were in the range $[0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2]$ corresponding to the observed proportion of validated positive genes which is about 4%–5% and plausible higher values to account for potential positive nodes among the unlabelled set. While the lower values represent conservative estimates, the inclusion of 0.1–0.2 serves as a robust upper bound to account for latent genetic associations. The final value was chosen based on AUROC performance on a held-out validation set. These results are found in Appendix B. PU learning ensures that the model does not assume unlabelled nodes are not necessarily negative, mitigating label bias and improving generalisation to truly positive nodes. The overall loss function integrates PU learning, standard GAT loss and CL:

$$\mathcal{L} = \mathcal{L}_{PU} + \mathcal{L}_{GAT} + \lambda \mathcal{L}_{contrast} \tag{8}$$

where λ is a hyperparameter controlling the strength of the contrastive loss. The final output is a probability vector of all genes with their scores in relation to endometriosis.

3.3. Experimentation

In the experiments, 5-fold cross validation was adopted to evaluate the performance of HetBio-CLiP. Mean and standard deviation are recorded for each metric. All 42 previously validated endometriosis-associated genes [50] were labelled as positive nodes, while the remaining genes were treated as unlabelled. Node and edge masking strategies were used during training to prevent information leakage. Hyperparameter optimisation was conducted using a grid search, tuning the number of layers, learning rate, weight decay, dropout rate, the number of hidden units, as well as the contrastive loss weight, and prior estimation for PU learning $\pi_P \in [0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2]$. The model was trained using the cross-entropy loss function and the Adam optimiser.

Although the model was formulated as a semi-supervised node classification task, the predicted probabilities inherently form a ranked list of gene recommendations for endometriosis. Each gene node outputs a prediction score reflecting its likelihood of disease association, which can be interpreted as a recommendation score in a disease-gene ranking context. Therefore, the primary evaluation metrics used in this study to assess performance were the area under the curve (AUC), area under the precision–recall curve (AUPR), and mean reciprocal rank (MRR). These metrics quantify the model's ability to prioritise validated disease-associated genes among unlabelled candidates. These provide a fair and interpretable measure of recommendation quality, while also being appropriate for PU learning where negative labels are uncertain. Additionally, Precision@K (Prec@K) and Normalised Discounted Cumulative Gain (NDCG@K) were adopted to evaluate the ranking quality of the top recommendations, where K is the validated set of positive genes, 42. The result of how these validated genes are ranked among all genes in the prediction emphasises the model's ability to position true positives among the highest-ranked candidates and reflect real-world use where only a subset of top predictions is experimentally verifiable. True Positive @ 42 (TP@42) was used to show the number of validated endometriosis genes recovered among the top 42 ranked predictions. Finally, precision, recall, and F1 were also measured as secondary metrics for completeness. These metrics have limited interpretability under PU settings, since unlabelled genes would be classified as "true negative", which is incorrect since there may be unrecognised positives. The primary metrics are defined as follows:

$$AUC = \int_0^1 TPR(FPR) d(FPR) \tag{9}$$

$$AUPR = \int_0^1 Precision(Recall) d(Recall) \tag{10}$$

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{predicted rank of } i} \tag{11}$$

The baselines used were Random Walk with restart heterogeneous (RWRH) [12], GCN (from EMOGI) [21], GAT, GraphSAGE (from XGDAG) [13], HerGePred [2], and ModulePred [25]. These methods were selected due to their standard usage in bioinformatics and their relevance to our research objectives.

As the baseline methods are originally developed under different graph assumptions, each was adapted to operate on the same underlying dataset while preserving its original modelling constraints. All baselines were trained and evaluated using the same gene–disease subgraph extracted from our heterogeneous network where possible and 5-fold cross-validation splits. No baseline was provided access to additional node or edges beyond what was available to HetBio-CLiP. Baselines that do not support PU learning were trained under the standard assumption that unlabelled genes represent negative examples, consistent with their original formulations.

Homogeneous GNN baselines (GCN and GraphSAGE) were restricted to a homogeneous gene-only induced subgraph, where only gene nodes and gene–gene interaction edges were used. Other nodes and cross-type edges were excluded as these models do not natively support heterogeneous message passing. This restriction to a homogeneous subgraph limits the amount of information these baselines receive compared to the rest of the baselines, and this is further discussed in Section 4.1.

HerGePred and ModulePred are designed for heterogeneous graph structures so were evaluated using their original formulations, subject to the availability of input data. Both HerGePred and ModulePred were applied to the same heterogeneous graph as HetBio-CLiP. However, the original ModulePred framework used protein complexes, which was not available in our dataset and therefore omitted. All other components of the models were used as described in the original studies.

To further evaluate the ranking performance of HetBio-CLiP from the perspective of recommender systems, we compared our model with LightGCN [14] and the improved SAGEgraph methodology from [39]. In these adaptations, disease nodes were treated as users and gene nodes as items, with existing gene–disease associations used as implicit positive feedback. Standard hyperparameter settings from their respective papers were implemented.

Table 3
Comparison of performance metrics for validated endometriosis genes.

Methods	AUC	AUPR	MRR
RWRH	0.8563 ± 0.06	0.7450 ± 0.08	0.7381 ± 0.08
HerGePred	0.8661 ± 0.05	0.8355 ± 0.08	0.8095 ± 0.07
EMOGI	0.8993 ± 0.05	0.8973 ± 0.07	0.9047 ± 0.07
GAT	0.9138 ± 0.07	0.9061 ± 0.11	0.9284 ± 0.10
XGDAG	0.9012 ± 0.04	0.7999 ± 0.07	0.8809 ± 0.06
ModPred	0.8928 ± 0.06	0.9186 ± 0.11	0.8809 ± 0.11
LightGCN	0.9009 ± 0.04	0.8511 ± 0.06	0.8717 ± 0.07
M-SAGEGraph	0.9149 ± 0.04	0.8562 ± 0.06	0.9001 ± 0.06
HetBio-CLiP (Ours)	**0.9489 ± 0.04**	**0.9401 ± 0.05**	**0.9284 ± 0.06**
Methods	Prec@42	NDCG@42	TP@42
RWRH	0.3694 ± 0.09	0.6785 ± 0.08	15.5 ± 3.9
HerGePred	0.3820 ± 0.08	0.6910 ± 0.08	16.0 ± 3.4
EMOGI	0.4192 ± 0.08	0.7021 ± 0.08	17.6 ± 3.4
GAT	0.5474 ± 0.12	0.7992 ± 0.11	23 ± 5.2
XGDAG	0.4976 ± 0.06	0.7803 ± 0.06	20.9 ± 2.7
ModPred	0.5388 ± 0.12	0.8088 ± 0.11	22.4 ± 5.1
LightGCN	0.6094 ± 0.08	0.8166 ± 0.08	25.6 ± 3.5
M-SAGEGraph	0.6175 ± 0.07	**0.8292 ± 0.06**	25.9 ± 3.0
HetBio-CLiP (Ours)	**0.6933 ± 0.06**	0.8274 ± 0.06	**29.1 ± 2.4**

3.4. Model interpretability

To understand and interpret the decision-making process of HetBio-CLiP, SHAP (SHapley Additive exPlanations) values were used [52]. Given a trained GNN with a prediction function $f : \mathcal{G} \to \mathbb{R}$ where $\mathcal{G}$ represents the input graph, GNN assigns importance scores to each node by computing the Shapley values. To understand the edges, GNNShap was adapted for use in heterogeneous graphs [38].

4. Results and discussion

4.1. Validated gene results

To evaluate the predictive performance of the proposed HetBio-CLiP, we first examined the known validated endometriosis genes from [50] against the entire base of the gene nodes of the graph and compared the results with the baselines. To compare, the aggregate performance metrics are shown in Table 3. Precision, recall and F1 score metrics are found in Appendix B, Table C.8.

Our proposed model, HetBio-CLiP, achieved the highest performance across most key evaluation metrics, demonstrating its effectiveness in prioritising disease-associated genes under a PU setting. All results are reported as mean ± standard deviation across five cross-validation folds. Specifically, HetBio-CLiP attained the highest AUC (0.9459 ± 0.04), AUPR (0.9401 ± 0.05), MRR (0.9284 ± 0.06), and Precision@42 (0.6933 ± 0.06). Among the baseline methods, the simple GAT achieved an similar MRR (0.9284 ± 0.10) but exhibited larger variability across folds. M-SAGEGraph and LightGCN demonstrated competitive but consistently lower performance.

When focused on ranking-based metrics, which are particularly relevant in recommendation settings, HetBio-CLiP achieved the highest TP@42 (29.1 ± 2.4) and the second-highest NDCG@42 (0.8274 ± 0.06) whereas M-SAGEgraph attained the highest NDCG@42 (0.8292) and second-highest Precision@42 (0.6175) and TP@42 (25.9). This complementary performance suggests that HetBio-CLiP excels at retrieving the highest proportion of true positives within the top-ranked candidates, while M-SAGEgraph slightly outperforms in terms of rank ordering among positives.

These results collectively indicate that HetBio-CLiP achieves a strong balance between ranking accuracy and recall, which are particularly critical in PU learning settings where the unlabelled data are not guaranteed to be true negatives. The high AUC and AUPR scores reflect the model's capacity to effectively distinguish between likely positive and unlabelled genes, even when the number of confirmed positive samples is small. The improvement in MRR, Precision@42, TP@42, and NDCG@42 demonstrates that HetBio-CLiP ranks true endometriosis-associated genes closer to the top of the list, which is an essential feature in biological recommendation tasks where only the top few candidates are experimentally validated.

Of all the homogeneous baselines, GraphSAGE performed the best, but had lower AUPR and AUC values and higher variability, which indicate that it may perform less consistently across different thresholds, reflecting limited robustness in probabilistic ranking. The restriction to a gene-only subgraph, limiting information propagation across diverse entities probably contributed to the lower performance. Furthermore, they treated unlabelled genes as negative, unlike HetBio-CLiP's PU methodology. These limitations underscore the value of heterogeneous graph representations and the incorporation of complementary information for more accurate prediction of disease-associated genes.

Although ModulePred achieved slightly higher precision, this metric alone does not fully reflect performance under the PU assumption. High precision can indicate conservative predictions that exclude uncertain but potentially relevant genes. Conversely, HetBio-CLiP's superior recall and overall ranking metrics suggest that it can identify a broader set of plausible candidates while maintaining strong discrimination. The absence of the protein complex data which is included in the original ModulePred may have limited its predictive performance in our dataset. However, the comparison using the same network shows that HetBio-CLiP still provides a meaningful improvement in the identification of candidate endometriosis-associated genes.

To assess the biological plausibility of the learned representations, we examined the ranking of validated endometriosis-associated genes produced by HetBio-CLiP. The top three validated gene predictions identified were WNT4, GREB1, and ESR1. These genes have all previously implicated in endometriosis pathophysiology. These genes play key roles in hormonal regulation, oestrogen signalling, and cell proliferation, aligning closely with known mechanisms underlying endometrial tissue growth and inflammatory response [53]. Their presence among the highest-ranked predictions demonstrates that HetBio-CLiP not only achieves high statistical accuracy but also captures biologically meaningful interactions.

Overall, these findings indicate that HetBio-CLiP achieves strong and stable gene prioritisation performance and exhibits robustness and interpretability advantages in the PU context. Its performance consistency across metrics highlights its capacity to integrate multi-relational genomic and clinical evidence effectively, thereby offering a reliable framework for prioritising underexplored candidate genes for endometriosis and potentially other complex diseases.

4.2. Ablation study

To systematically analyse the effectiveness of the important elements of HetBio-CLiP, namely CL and positive unlabelled learning (PU), an ablation study was conducted. We compared the proposed HetBio-CLiP method with three ablation variants that remove different variations of components from the original, namely -CL (without the CL component), -PU (without the PU component), and -CLPU (without either CL or PU). Each variation was trained and evaluated under the same 5-fold cross-validation setup and hyperparameter configuration as the full model. These variants were compared in Table 4. More metrics can be found in Table C.10

The results show a consistent decline in performance with the removal of these components, confirming their complementary importance to the overall framework. The full HetBio-CLiP model achieved the best results across all primary metrics, followed by the -PU variant, then the -CL variant, with the -CLPU variant performing the worst overall.

This ranking indicates that the PU module contributes slightly more to overall predictive performance, particularly in handling the label uncertainty inherent in the PU setting. By contrast, the CL module

Table 4
Comparison of performance metrics for ablation study.

Methods	AUC	AUPR	MRR
Original	**0.9489** ± 0.04	**0.9401** ± 0.05	**0.9284** ± 0.06
-PU	0.9361 ± 0.06	0.9355 ± 0.09	0.9195 ± 0.08
-CL	0.9330 ± 0.06	0.9073 ± 0.08	0.9047 ± 0.09
-CLPU	0.9138 ± 0.07	0.9061 ± 0.11	0.9284 ± 0.10
Methods	Prec@42	NDCG@42	TP@42
Original	**0.6933** ± **0.06**	**0.8274** ± 0.06	**29.1** ± **2.4**
-PU	0.5711 ± 0.09	0.7999 ± 0.09	24.0 ± 3.8
-CL	0.5592 ± 0.08	0.8147 ± 0.08	23.5 ± 3.5
-CLPU	0.5474 ± 0.12	0.7992 ± 0.11	22.9 ± 5.2

primarily enhances representation quality, improving the discrimination between biologically similar but functionally distinct gene nodes. When both modules were removed, the model's performance dropped sharply, suggesting that their synergy is critical for capturing latent biological relationships and for improving robustness against noisy or sparse supervision. Thus, the combined effect of CL and PU learning leads to superior generalisation, more stable rankings, and higher interpretability, confirming that HetBio-CLiP's modular design effectively integrates representation learning and uncertainty-aware optimisation.

4.3. Model robustness evaluation

To assess whether model performance is driven by proximity to well-studied endometriosis-associated genes, we conducted a multi-tiered topological robustness analysis. We constructed balanced evaluation sets stratified by graph distance between validated and unlabelled genes. The first set used unlabelled genes that are closely connected to validated genes (1-hop), the second is remotely connected (2–3 hop), and the third is unrelated genes (≥4-hop). The experimental results are provided in Appendix C in Table C.9.

For the 1-hop neighbour experiment, the reasonable recall (0.5714 ± 0.09) and AUC (0.6168 ± 0.09) demonstrated that even when HetBio-CLiP is restricted to a challenging subset, where negative samples share immediate adjacency with the positive set, it can distinguish a validated endometriosis gene from its immediate neighbours in the graph. The 2–3 hop neighbour experiment had a recall (0.5714 ± 0.10) and AUC (0.5918 ± 0.09), similar to the 1-hop experiment. This highlights that HetBio-CLiP is still able to utilise the set of genomic features around it rather than the distance to known endometriosis-associated genes. Lastly, the ≥4-hop experiment had the lowest recall (0.3571 ± 0.10) and AUC (0.3951 ± 0.10). The decreased in AUC shows that the model does not rely solely on connectivity bias or hub effects, but rather on shared relational and functional context. However, performance degradation for highly distant genes indicates that, like most graph-based approaches, the model depends on informative topological structure and may be less effective for sparsely connected candidates.

4.4. Unvalidated gene results

In this section, we examine genes prioritised by HetBio-CLiP that were not included in the validated endometriosis positive gene labels. The remaining gene nodes have not been annotated as confirmed related to endometriosis. These 2734 unvalidated genes nodes were compiled by rank according to all model's scores. The top 10 genes and their scores, along with the rest of the baselines are shown in Table 5. Based on the high recovery of validated genes, HetBio-CLiP could provide useful rankings of candidate genes pertaining to endometriosis.

From the HetBio-CLiP model, the top 5 genes were IL10, VEGFA, MIR145, PGR, and KRAS. Many of these genes also had a high score in the other baselines, suggesting that these genes have a high likelihood of relation to endometriosis. Overall, the most similar common gene

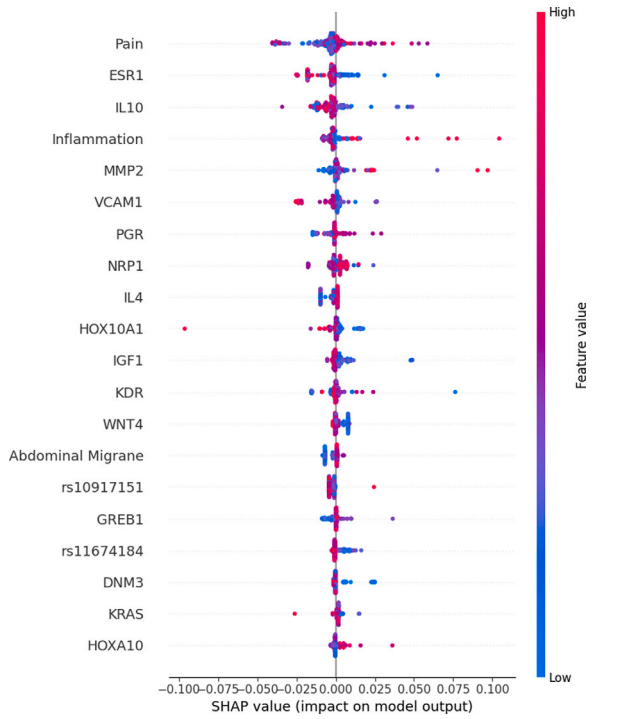


Fig. 2. The summary of the top 20 features overall. Red shows the nodes that positively contribute. Blue shows the nodes that negatively contribute. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

identified across all baselines was VEGFA (7), MIR145 (6), PGR (6), KRAS (6), and MAPK1 (5). Some of the genes identified (PGR, IL10, KRAS) have already been linked to endometriosis with high levels of evidence, as seen in the DISGENET database. Thus, MIR145, VEGFA, and MAPK1 could all be genes of high interest in further research of endometriosis that have not been identified previously.

MIR145 is a microRNA involved in regulating cell differentiation, proliferation, and apoptosis. This can affect expression of various genes that control cell adhesion,invasion and inflammatory responses. In the context of endometriosis, altered expression of MIR145 has been reported, suggesting that it may play a role in the abnormal behaviour of endometrial cells [54]. Li et al. also found that there was dysregulation of endometrial MIR145 in subfertile women [55].

VEGFA (vascular endothelial growth factor A) is an established key driver of angiogenesis, the process by which new blood vessels form [56]. Elevated levels of VEGFA have been observed in endometriotic tissues during menstruation, where they potentially contribute to the development of abnormal vasculature [57]. This enhanced angiogenesis not only supports lesion survival but also increases vascular permeability, which can exacerbate inflammation and pain. As a result, VEGFA is considered a potential therapeutic target; inhibiting its signalling pathway may help limit the growth and persistence of endometriotic lesions [58].

MAPK1, also known as extracellular signal-regulated kinase 2 (ERK2), is a critical component of the MAPK/ERK signalling pathway, which regulates cell growth, differentiation, and survival [59]. In endometriosis, dysregulation of MAPK1 activity has been linked to the enhanced proliferation and survival of ectopic endometrial cells. Aberrant MAPK1 signalling may promote chronic inflammation and resistance to apoptosis within endometriotic lesions, thereby contributing to their persistence and progression [53].

The recovery of VEGFA, MIR145, and MAPK1 is significant not as a novel gene discovery, but as an indication of the model's ability to detect associations that exist in biological reality and literature but

Table 5
Top 10 unvalidated genes with endometriosis-related scores for each method.

RWRH	HerGePred	GCN	GAT	GraphSAGE
MIR21 (1.000)	PGR (0.9794)	PGR (0.9989)	MIR21 (0.9879)	MAPK3 (1.000)
MAOA (1.000)	VEGFA (0.9759)	VEGFA (0.9784)	MIR145 (0.9866)	MAPK1 (1.000)
SMARCD3 (0.9989)	MIR145 (0.9692)	SERPINE1 (0.9578)	VDR (0.9849)	VEGFA (1.000)
VEGFA (0.9821)	TNF (0.9444)	ESR1 (0.9522)	IL1B (0.9776)	PGR (0.9776)
VDR (0.9692)	ESR2 (0.9441)	CST12P (0.9432)	CYP19A1 (0.9648)	PRL (0.9692)
MALAT1 (0.8596)	IL2 (0.9387)	KRAS (0.9119)	MAPK1 (0.9648)	IL1B (0.9549)
CST12P (0.8596)	CCND1 (0.9001)	MIR145 (0.9116)	MAPK3 (0.9648)	CYP19A1 (0.9521)
RNU1-1 (0.8555)	APOE (0.8299)	IL2 (0.8838)	CHD6 (0.9211)	TNF (0.8779)
MIR126 (0.8245)	CYP19A1 (0.8299)	CST12P (0.8675)	AMH (0.9017)	KRAS (0.8534)
MIR145 (0.7466)	KRAS (0.8275)	LAMC1 (0.8245)	PTPN22 (0.8748)	AMH (0.8534)

ModulePred	LightGCN	M-SAGEGraph	HetBio-CLiP (Ours)
MAPK1 (0.9989)	PGR (0.9942)	MAPK1 (0.9999)	IL10 (0.9849)
PGR (0.9986)	FGF9 (0.9846)	IL10 (0.9849)	VEGFA (0.9810)
LAMC1 (0.9849)	IL10 (0.9801)	PRL (0.9834)	MIR145 (0.9799)
MMP2 (0.9843)	MIR145 (0.9794)	CYP19A1 (0.9813)	PGR (0.9799)
VEGFA (0.9821)	CCND1 (0.9772)	VEGFA (0.9702)	MIR126 (0.9698)
MIR145 (0.9721)	TNF (0.9745)	IL1B (0.9666)	CST12P (0.9697)
FGF9 (0.9721)	MIR21 (0.9696)	KRAS (0.9649)	TNF (0.9695)
CST12P (0.9699)	VDR (0.9630)	AMH (0.9505)	MAPK1 (0.9555)
SOXZ2 (0.9520)	MAPK3 (0.9612)	IL2 (0.9499)	KRAS (0.9512)
KDM3A (0.9333)	KRAS (0.9544)	IGFBP1 (0.9182)	POSTN (0.9300)

have not yet been captured in the structured database labels used for training. These results confirms that HetBio-CLiP can successfully prioritise biologically meaningful candidates that are currently missing from standard curation efforts, supporting its utility for hypothesis generation in endometriosis research.

4.5. Results from the SHAP explainer module

With these genes of interest identified, SHAP and the adapted GNNShap module was used. Using the SHAP summary plot (Fig. 2), the top twenty nodes of importance across the heterogeneous graph are graphically represented, showing the feature's impact on the model in descending order and the magnitude and direction of each feature's effect on the prediction.

Pain, Inflammation are established connections with endometriosis and have many overlapping gene and variant connections, as seen in [50] and also in the disease-disease associations from DISGENET. ESR1 and IL10 are genes that have many connections within the PPI network, and have known connections with endometriosis as well.

We used VEGFA as a test to investigate the explanations SHAP and GNNShap were giving to individual node decisions. SHAP was used to identify the main features used for the particular node of interest, VEGFA, as seen in Fig. 3A. The adapted GNNShap was used to identify the main edge relationships as a subgraph for a node of interest, as seen in Fig. 3B. The most important nodes identified from Fig. 3A were KDR and pain in the positive direction, and COL18A1 in the negative direction, and this is similarly depicted in the subgraph (Fig. 3B). VEGFA has been positively linked to pain, one of the symptoms/diseases linked to endometriosis from DISGENET and a very common associated comorbidity [50]. VEGFA binds to KDR (also called VEGFR2) in the angiogenesis pathway, with Steinthorsdottir et al. identifying two loci nearby KDR that is positively associated with endometriosis [60], while VEGFA is inhibited by COL18A1 [61,62]. All the associations provided by SHAP analysis show that VEGFA could be a strong candidate gene for further endometriosis research.

5. Conclusion

In this paper, we introduce HetBio-CLiP, a general heterogeneous graph neural network methodology that integrates CL and PU learning to predict candidate disease-associated genes. Unlike conventional classification frameworks, HetBio-CLiP reformulates the gene–disease association discovery problem as a biomedical recommendation problem, where the model prioritises candidate genes based on multi-relational genomic, variant, and clinical evidence. The integration of genetic data from the UKBB and the AoU databases, alongside single-cell RNA-sequencing data from BioStudies provides real patient samples across multiple genetic ancestries that enrich the heterogeneous graph embeddings. This recommender-inspired formulation enables the model to prioritise candidate genes through a ranking-based learning process.

HetBio-CLiP demonstrates superior predictive performance compared with other state-of-the-art baselines, achieving the highest AUC, AUPR, and MRR values across validated endometriosis genes. The combination of CL and PU learning enables the model to capture latent biological relationships even in sparsely labelled datasets, while its XAI module provides transparent and interpretable reasoning by highlighting the most influential nodes and edges. This methodology allows for identifying key genes and interactions that drive the model's prediction of disease relevance, allowing further research to validate model outputs against known pathways and literature, bridging computational discovery with biological interpretation.

Despite these advances, there are still some limitations to our model. A key limitation of this approach is the reliance on a small set of 42 validated endometriosis-associated genes for training and evaluation. Performance may decrease for sparsely connected genes or those lacking sufficient contextual evidence in current biomedical databases. While mitigation strategies such as 5-fold cross-validation, and node and edge masking were used to avoid bias, this small subset could still increase overfitting and introduce bias predictions towards well-studied nodes. Future work should focus on expanding and integrating additional multi-omics data, including transcriptomic profiles and epigenetic modifications to capture a more comprehensive view of disease mechanisms.

As the gene–disease associations identified by HetBio-CLiP represent computational hypotheses, subsequent experimental validation is essential to substantiate their biological relevance and reinforce the trustworthiness of the prioritised candidates. Future research incorporating laboratory-based confirmation would significantly refine the interpretability of these findings. Another area of future work includes exploring the use of CL for the variant and disease nodes using node-type specific augmentation strategies and objectives tailored to their distinct semantic roles. Beyond endometriosis, HetBio-CLiP is readily generalisable to other complex diseases with limited known gene associations, such as polycystic ovary syndrome or rare genetic disorders, and could benefit from transfer learning approaches to adapt to novel

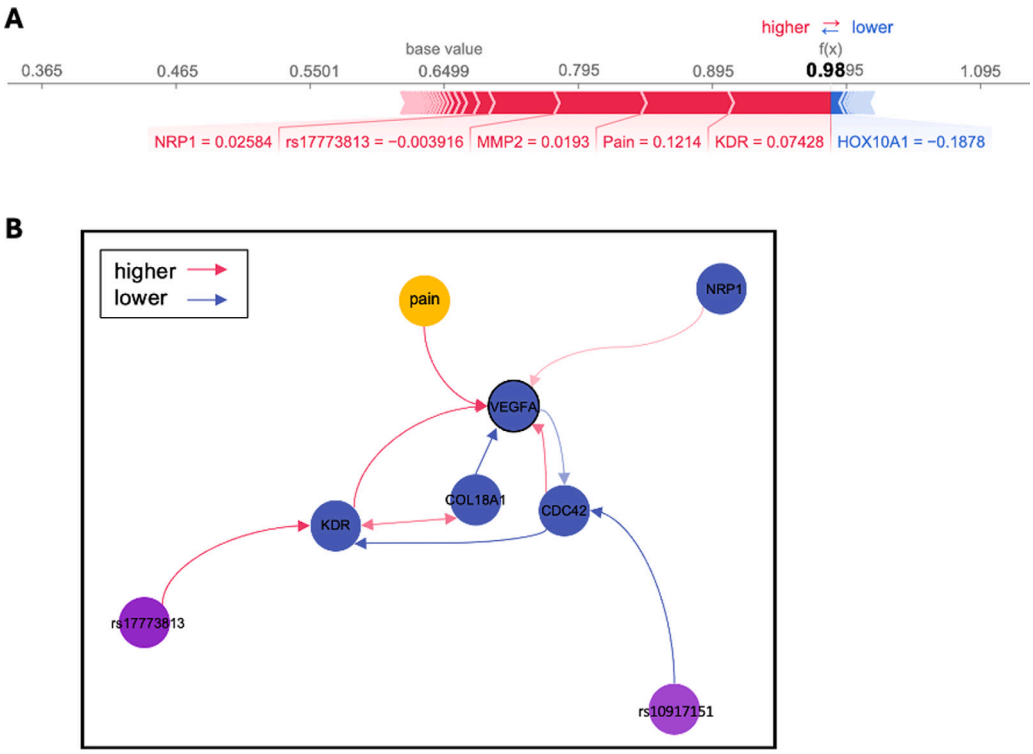


Fig. 3. The SHAP results for VEGFA. (A) is the individual SHAP plot given for VEGFA. Red shows the nodes that positively contributes to the score. Blue shows the nodes that are negatively contributing to the score. (B) is the subgraph given for VEGFA. Yellow nodes are disease nodes, blue nodes are gene nodes, purple nodes are variant nodes. Red lines show positive contributions to the score. Blue lines show the negative contributions to the score. Increasing opacity of the lines indicates the strength of that contribution. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

biomedical applications with scarce labelled data. Overall, HetBio-CLiP provides a robust, interpretable, and generalisable framework for candidate gene prioritisation, combining heterogeneous graph learning with PU-aware recommendation and explainable AI, and represents a step forward in computational strategies for translational genomics.

CRediT authorship contribution statement

Zelia Soo: Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Hua Lin:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Mengjia Wu:** Writing – review & editing, Visualization, Methodology, Investigation. **Kairui Guo:** Writing – review & editing, Conceptualization. **Mark Grosser:** Writing – review & editing. **Yi Zhang:** Writing – review & editing, Supervision. **Jie Lu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the Australian Research Council (ARC) under LP210100414. This research has been conducted using the UK Biobank Resource under Project Number 105141. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.

Appendix A. Disease node composition and categorisation

The disease nodes used in the final graph were based solely on their presence in the DISGENET search using the "endometriosis" query and were not filtered by organ system or presumed clinical relevance, but were compiled based on their genetic or variant overlaps and includes endometriosis as well as a range of genetically associated conditions spanning reproductive disorders, pain and migraine-related phenotypes, inflammatory and autoimmune diseases, cardiovascular conditions, and mood disorders. These disease entities were not manually selected or grouped by the authors. The list of diseases with their category and concept ID is provided in Table A.6.

Appendix B. Hyperparameter testing

We performed a sensitivity analysis on the positive class prior π_P used in the PU loss. AUROC performance peaked in the range π_P = 0.04–0.05, which aligns with the expected prevalence of disease-associated genes when accounting for latent positives among the unlabelled nodes. Performance remained stable across this range, with the lowest variance across the 5-folds, and gradually declined for larger priors ($\pi_P \geq 0.1$), indicating robustness to moderate misspecification of the class prior. The full range of values are found in Table B.7 and visualised in Fig. B.4

Appendix C. Additional performance metrics for validated endometriosis genes results

See Tables C.8–C.10.

Table A.6
Disease nodes included in the heterogeneous graph.

Disease	Categories	Concept ID
Prostatic Neoplasms	Neoplasms, Urogenital diseases	C0033578
Mammary Neoplasms	Neoplasms, skin and connective tissue diseases	C1458155
Mental depression	Behaviour and behaviour mechanisms	C0011581
Depressive disorder	Mental disorders	
Obesity	Pathological conditions, signs and symptoms, nutritional and metabolic diseases	C0028754
Unipolar depression	Mental disorders	C0041696
Myocardial infarction	Cardiovascular diseases	C0027051
Major depressive disorder	Mental disorders	C1269683
Asthma	Immune system diseases, Respiratory tract diseases	C0004096
Rheumatoid arthritis	Skin and connective tissue diseases, Immune system diseases, Musculoskeletal diseases	C0003873
Mood disorders	Mental disorders	C0525045
Adenocarcinoma	Neoplasms	C0001418
Endometrial Neoplasms	Neoplasms, Urogenital diseases	C0014170
Systemic Scleroderma	Skin and connective tissue diseases	C0036421
Arteriosclerosis	Cardiovascular diseases	C0003850
Pain	Psychological phenomena, pathological conditions, signs and symptoms	C0030193
Cardiovascular diseases	Cardiovascular diseases	C0007222
Migraine disorders	Nervous system diseases	C0149931
THOC6-related syndrome	Mental disorders, Behaviour and behaviour mechanisms, Nervous system diseases, Congenital, Hereditary, and Neonatal diseases and abnormalities, Musculoskeletal diseases, Pathological conditions, Signs and symptoms	C3150939
Abdominal migraine	Nervous system diseases	C0270858
Female infertility	Urogenital diseases, Pathological conditions	C0021361
Sick headaches	Nervous system diseases	C0700438
Involutional paraphrenia	Mental Disorders	C1571983
Inflammation	Pathological condition, Signs and symptoms	C0021368
Hemicrania migraine	Nervous system diseases	C0018984
Uterine fibroids	Neoplasms	C0042133
Endometrioma	Urogenital diseases	C0269102
Acute confusional migraine	Nervous system diseases	C0521664
Status migrainosus	Nervous system diseases	C0338489
Psychosis, Involutional	Mental disorders	C1571984
Coronary artery disease	Cardiovascular diseases	C1956346
Cervical migraine syndrome	Nervous system diseases	C0007852
Involutional depression	Mental disorders	C0011574

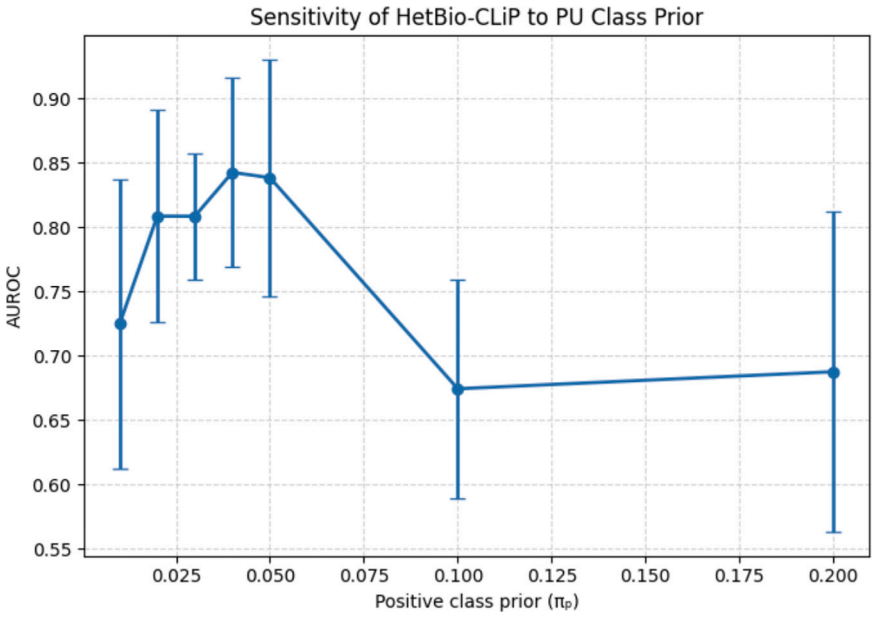


Fig. B.4. PU prior grid search results.

Table B.7
PU class prior grid search result.

π_P value	AUROC ± std.
0.01	0.7248 ± 0.1128
0.02	0.8088 ± 0.0823
0.03	0.8087 ± 0.0489
0.04	0.8428 ± 0.0730
0.05	0.8387 ± 0.0920
0.10	0.6744 ± 0.0849
0.20	0.6876 ± 0.1247

Table C.8
Comparison of performance metrics for validated endometriosis genes.

Methods	Precision	Recall	F1
RWRH	0.6993 ± 0.08	0.8057 ± 0.09	0.7409 ± 0.08
HerGePred	0.7705 ± 0.07	0.8151 ± 0.07	0.7961 ± 0.06
EMOGI	0.7627 ± 0.07	0.9004 ± 0.08	0.7577 ± 0.08
GAT	0.8848 ± 0.10	0.9151 ± 0.10	0.7060 ± 0.11
XGDAG	0.8970 ± 0.05	0.8836 ± 0.06	**0.9147 ± 0.05**
ModPred	**0.9459 ± 0.06**	0.8905 ± 0.07	0.8952 ± 0.06
LightGCN	0.8566 ± 0.04	0.7592 ± 0.04	0.8312 ± 0.04
M-SAGEGraph	0.8747 ± 0.05	0.7411 ± 0.04	0.8559 ± 0.05
HetBio-CLiP (Ours)	0.8916 ± 0.05	**0.9250 ± 0.05**	0.9079 ± 0.06

Table C.9
Performance of HetBio-CLiP under connectivity-stratified evaluation sets.

Dataset	Recall	AUC
1-hop	0.5714 ± 0.09	0.6168 ± 0.09
2-3 hop	0.5714 ± 0.10	0.5918 ± 0.09
≥4-hop	0.3571 ± 0.10	0.3951 ± 0.10

Table C.10
Comparison of performance metrics for ablation study.

Methods	Precision	Recall	F1
Original	0.8916 ± 0.05	**0.9250 ± 0.05**	**0.9079 ± 0.06**
-PU	0.8355 ± 0.07	0.8095 ± 0.09	0.7747 ± 0.8
-CL	**0.8973 ± 0.06**	0.9047 ± 0.09	0.7692 ± 0.08
-PUCL	0.8848 ± 0.10	0.9151 ± 0.10	0.7060 ± 0.11

Data availability

The source code used to implement HetBio-CLiP and reproduce the experiments is publicly available at: https://github.com/zsoo99999/HetBio-CLiP.git

Access to UK Biobank and All of Us data is subject to their respective application processes.

References

[1] D.M. Chiorean, et al., New insights into genetics of endometriosis-a comprehensive literature review, Diagn. (Basel) 13 (13) (2023).
[2] K. Yang, et al., HerGePred: Heterogeneous network embedding representation for disease gene prediction, IEEE J. Biomed. Health Informatics 23 (4) (2019) 1805–1815.
[3] K.T. Zondervan, C.M. Becker, S.A. Missmer, Endometriosis, N. Engl. J. Med. 382 (13) (2020) 1244–1256.
[4] K.A. Hansen, K.M. Eyster, Genetics and genomics of endometriosis, Clin. Obstet. Gynecol. 53 (2) (2010) 403–412.
[5] Y. Moreau, L.-C. Tranchevent, Computational tools for prioritizing candidate genes: boosting disease gene discovery, Nature Rev. Genet. 13 (8) (2012) 523–536, [Online]. Available: https://doi.org/10.1038/nrg3253.
[6] M.J. Landrum, et al., ClinVar: Public archive of interpretations of clinically relevant variants, Nucleic Acids Res. 44 (D1) (2015) D862–D868.
[7] H.L. Rehm, et al., ClinGen — The clinical genome resource, New Engl. J. Med. 372 (23) (2015) 2235–2242.
[8] S. Köhler, et al., The human phenotype ontology in 2017, Nucleic Acids Res. 45 (D1) (2017) D865–d876.
[9] J. Piñero, et al., The DisGeNET knowledge platform for disease genomics: 2019 update, Nucleic Acids Res. 48 (D1) (2019) D845–D855, [Online]. Available: https://doi.org/10.1093/nar/gkz1021.
[10] K. Guo, et al., Artificial intelligence-driven biomedical genomics, Knowl.-Based Syst. 279 (2023) 110937, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705123006871.
[11] Z. Zhang, P. Cui, W. Zhu, Deep learning on graphs: A survey, IEEE Trans. Knowl. Data Eng. 34 (1) (2022) 249–270.
[12] S.K. Ata, M. Wu, Y. Fang, L. Ou-Yang, C.K. Kwoh, X.-L. Li, Recent advances in network-based methods for disease gene prediction, Brief. Bioinform. 22 (4) (2020).
[13] A. Mastropietro, G. De Carlo, A. Anagnostopoulos, XGDAG: Explainable gene–disease associations via graph neural networks, Bioinformatics 39 (8) (2023) [Online]. Available: https://doi.org/10.1093/bioinformatics/btad482.
[14] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, LightGCN: Simplifying and powering graph convolution network for recommendation, 2020, CoRR abs/2002.02126, [Online]. Available: https://arxiv.org/abs/2002.02126.
[15] X. Zhao, J. Wu, X. Zhao, M. Yin, Multi-view contrastive heterogeneous graph attention network for lncRNA–disease association prediction, Brief. Bioinform. 24 (1) (2022) [Online]. Available: https://doi.org/10.1093/bib/bbac548.
[16] F. Li, et al., Positive-unlabeled learning in bioinformatics and computational biology: A brief review, Brief. Bioinform. 23 (1) (2021) [Online]. Available: https://doi.org/10.1093/bib/bbab461.
[17] A. van Hilten, S. Katz, E. Saccenti, W.J. Niessen, G.V. Roshchupkin, Designing interpretable deep learning applications for functional genomics: A quantitative analysis, Brief. Bioinform. 25 (5) (2024) [Online]. Available: https://doi.org/10.1093/bib/bbae449.
[18] J.B. Awotunde, E.A. Adeniyi, G.J. Ajamu, G.B. Balogun, F.A. Taofeek-Ibrahim, Explainable artificial intelligence in genomic sequence for healthcare systems prediction, in: S. Mishra, A. González-Briones, A.K. Bhoi, P.K. Mallick, J.M. Corchado (Eds.), Connected E-Health: Integrated IoT and Cloud Computing, Springer International Publishing, Cham, 2022, pp. 417–437, [Online]. Available: https://doi.org/10.1007/978-3-030-97929-4_19.
[19] B. Pfeifer, A. Saranti, A. Holzinger, GNN-SubNet: Disease subnetwork detection with explainable graph neural networks, Bioinformatics 38 (Supplement 2) (2022) ii120–ii126, [Online]. Available: https://doi.org/10.1093/bioinformatics/btac478.
[20] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.
[21] R. Schulte-Sasse, S. Budach, D. Hnisz, A. Marsico, Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms, Nat. Mach. Intell. 3 (6) (2021) 513–526, [Online]. Available: https://doi.org/10.1038/s42256-021-00325-y.
[22] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Adv. Neural Inf. Process. Syst. 30 (2017).
[23] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, Stat 1050 (20) (2017) 10–48550.
[24] B.W. Zhao, et al., Motif-aware miRNA-disease association prediction via hierarchical attention network, IEEE J. Biomed. Health Informatics 28 (7) (2024) 4281–4294.
[25] X. Jia, et al., A deep learning framework for predicting disease-gene associations with functional modules and graph augmentation, BMC Bioinformatics 25 (1) (2024) 214, [Online]. Available: https://doi.org/10.1186/s12859-024-05841-3.
[26] S. Sebastian, S. Roy, J. Kalita, Network-based analysis of Alzheimer's Disease genes using multi-omics network integration with graph diffusion, J. Biomed. Informatics 164 (2025) 104797.
[27] P. Yang, X.L. Li, J.P. Mei, C.K. Kwoh, S.K. Ng, Positive-unlabeled learning for disease gene identification, Bioinformatics 28 (20) (2012) 2640–2647, [Online]. Available: https://doi.org/10.1093/bioinformatics/bts504.
[28] F. Mordelet, J.-P. Vert, ProDiGe: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples, BMC Bioinformatics 12 (1) (2011) 389, [Online]. Available: https://doi.org/10.1186/1471-2105-12-389.
[29] Z. Li, J. Wang, J. Liang, Debiased graph contrastive learning based on positive and unlabeled learning, Int. J. Mach. Learn. Cybern. 15 (6) (2024) 2527–2538, [Online]. Available: https://doi.org/10.1007/s13042-023-02046-7.
[30] E. Steerling, E. Siira, P. Nilsen, P. Svedberg, J. Nygren, Implementing AI in healthcare-the relevance of trust: A scoping review, Front. Health Serv. 3 (2023) 1211150.
[31] Z. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, Gnnexplainer: Generating explanations for graph neural networks, Adv. Neural Inf. Process. Syst. 32 (2019).
[32] H. Yuan, J. Tang, X. Hu, S. Ji, XGNN: Towards model-level explanations of graph neural networks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 430–438, [Online]. Available: https://doi.org/10.1145/3394486.3403085.
[33] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: An overview, in: W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, 2019, pp. 193–209, [Online]. Available: https://doi.org/10.1007/978-3-030-28954-6_10.
[34] M. Chatzianastasis, M. Vazirgiannis, Z. Zhang, Explainable multilayer graph neural network for cancer gene prediction, Bioinformatics 39 (11) (2023) [Online]. Available: https://doi.org/10.1093/bioinformatics/btad643.

[35] L. Dai, D. Xu, Y. Gao, A graph neural network explainability strategy driven by key subgraph connectivity, J. Biomed. Informatics 165 (2025) 104813, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046425000425.

[36] A. Duval, F.D. Malliaros, GraphSVX: Shapley value explanations for graph neural networks, 2021, pp. 302–318, http://dx.doi.org/10.1007/978-3-030-86520-7_19, [Online]. Available: https://doi.org/10.1007/978-3-030-86520-7_19.

[37] H. Yuan, H. Yu, J. Wang, K. Li, S. Ji, On explainability of graph neural networks via subgraph explorations, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 12241–12252.

[38] S. Akkas, A. Azad, GNNShap: Scalable and accurate GNN explanation using Shapley values, in: Proceedings of the ACM Web Conference 2024, 2024, [Online]. Available: http://dx.doi.org/10.1145/3589334.3645599.

[39] Y. Afoudi, M. Lazaar, S. Hmaidi, An enhanced recommender system based on heterogeneous graph link prediction, Eng. Appl. Artif. Intell. 124 (2023) 106553, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623007376.

[40] T.J. Lakshmi, S.D. Bhavani, Link prediction approach to recommender systems, Computing 106 (7) (2024) 2157–2183, [Online]. Available: https://doi.org/10.1007/s00607-023-01227-0.

[41] M. Wu, Y. Zhang, H. Lin, M. Grosser, G. Zhang, J. Lu, BiblioEngine: An AI-empowered platform for disease genetic knowledge mining, in: Health Information Science, Springer Nature Singapore, 2023, pp. 187–198.

[42] D. Szklarczyk, et al., The STRING database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic Acids Res. 51 (D1) (2023) D638–d646.

[43] S.A. Lambert, et al., The polygenic score catalog: new functionality and tools to enable FAIR research, 2024, http://dx.doi.org/10.1101/2024.05.29.24307783, MedRxiv.

[44] A.G. Bick, et al., Genomic data in the all of us research program, Nature 627 (8003) (2024) 340–346, [Online]. Available: https://doi.org/10.1038/s41586-023-06957-x.

[45] P. Danecek, et al., Twelve years of SAMtools and BCFtools, GigaScience 10 (2) (2021) arXiv:https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf, [Online]. Available: https://doi.org/10.1093/gigascience/giab008, giab008.

[46] L. Garcia-Alonso, R. Vento-Tormo, sc/snRNAseq of human endometrium from controls and endometriosis patients, 2024, https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-14039.

[47] Y. Hao, et al., Integrated analysis of multimodal single-cell data, Cell (2021) [Online]. Available: https://doi.org/10.1016/j.cell.2021.04.048.

[48] C.S. McGinnis, L.M. Murrow, Z.J. Gartner, DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors, Cell Syst. 8 (4) (2019) 329–337.e4, [Online]. Available: https://doi.org/10.1016/j.cels.2019.03.003.

[49] Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, BioWordVec, improving biomedical word embeddings with subword information and mesh, Sci. Data 6 (1) (2019) 52, [Online]. Available: https://doi.org/10.1038/s41597-019-0055-0.

[50] N. Rahmioglu, et al., The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions, Nature Genet. 55 (3) (2023) 423–436, [Online]. Available: https://doi.org/10.1038/s41588-023-01323-z.

[51] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2017, arXiv.org.

[52] S.M. Lundberg, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (10) (2018) 749–760, [Online]. Available: https://doi.org/10.1038/s41551-018-0304-0.

[53] M. Zhang, et al., Research advances in endometriosis-related signaling pathways: A review, Biomed. Pharmacother. 164 (2023) 114909, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0753332223006996.

[54] P. Kumari, I. Sharma, S.C. Saha, R. Srinivasan, P. Bhardwaj, Role of serum microRNAs as biomarkers for endometriosis, endometrioid carcinoma of ovary & endometrioid endometrial cancer, Indian J. Med. Res. 156 (3) (2022) 516–523.

[55] X. Li, et al., Extracellular vesicles-derived MicroRNA-145-5p is upregulated in the uterine fluid of women with endometriosis and impedes mouse and human blastocyst development, J. Ovarian Res. 17 (1) (2024) 253, [Online]. Available: https://doi.org/10.1186/s13048-024-01579-x.

[56] J. Donnez, P. Smoes, S. Gillerot, F. Casanas-Roux, M. Nisolle, Vascular endothelial growth factor (VEGF) in endometriosis, Hum. Reprod. 13 (6) (1998) 1686–1690.

[57] K. Danastas, E.J. Miller, A.J. Hey-Cunningham, C.R. Murphy, L.A. Lindsay, Expression of vascular endothelial growth factor A isoforms is dysregulated in women with endometriosis, Reprod. Fertil. Dev. 30 (4) (2018) 651–657.

[58] M.S. Chung, S.J. Han, Endometriosis-associated angiogenesis and anti-angiogenic therapy for endometriosis, Front. Glob. Women's Health 3 (2022) 856316.

[59] G. Bora, A. Yaba, The role of mitogen-activated protein kinase signaling pathway in endometriosis, J. Obs. Gynaecol. Res. 47 (5) (2021) 1610–1623.

[60] V. Steinthorsdottir, et al., Common variants upstream of KDR encoding VEGFR2 and in TTC39B associate with endometriosis, Nat. Commun. 7 (1) (2016) 12350, [Online]. Available: https://doi.org/10.1038/ncomms12350.

[61] H.N. Allen, R. Khanna, Unveiling the dichotomic nature of VEGFA in neuropathic pain: A path to targeted therapeutics, Neurotherapeutics 22 (2) (2025) e00534, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1878747925000121.

[62] L. Seppinen, T. Pihlajaniemi, The multiple functions of collagen XVIII in development and disease, Matrix Biol. 30 (2) (2011) 83–92, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0945053X10001770.