current issues in molecular biology

MDPI

Article

FuDiCo: Gene Fusion-Initiated Path Propagation for Disease Comorbidity Prediction

Ashwag Altayyar and Li Liao *

Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA; ashwag@udel.edu

* Correspondence: liliao@udel.edu

Abstract

Disease comorbidity—the co-occurrence of two or more diseases in the same individual—has gained growing attention due to its association with adverse clinical outcomes and increased treatment complexity. Recent subgraph-based approaches for disease comorbidity prediction model disease modules as subgraphs induced by disease-associated genes in the protein–protein interaction (PPI) network and learn disease representations from subgraph topology. However, these approaches are constrained by incomplete disease–gene annotations, which may obscure important molecular relationships between diseases. Accordingly, disease comorbidity may also be influenced by molecular events beyond annotated disease genes, such as gene fusion events that have emerged as important contributors to disease mechanisms. Motivated by the role of gene fusions in disease development, we introduce Gene Fusion-Initiated Path Propagation for Disease Comorbidity Prediction (FuDiCo), a framework that models comorbidity through influence propagation over the PPI network. FuDiCo represents fusion-associated genes as localized perturbation sources and learns how their influence propagates along interaction paths toward disease subgraphs, thereby capturing propagation patterns that link related diseases and contribute to their comorbidity. Experiments on a benchmark disease comorbidity dataset show that FuDiCo outperforms state-of-the-art methods, achieving statistically significant improvements. These results shed light on the importance of gene fusion events in understanding disease relationships.

Keywords: disease comorbidity prediction; gene fusion; protein–protein interaction network; subgraph representation learning; influence propagation; graph analysis

Check for updates

Academic Editor: Howard H. Yang

Received: 13 April 2026
Revised: 3 June 2026
Accepted: 11 June 2026
Published: 16 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The term comorbidity was introduced in 1970 to describe additional diseases co-occurring with a primary index disease [1]. In patients with chronic diseases, the presence of coexisting diseases before treatment can substantially influence the clinical course, prognosis, and evaluation of treatment outcomes of the index disease [1]. Over the past decades, comorbidity has remained highly prevalent among individuals with chronic diseases. Recent national surveillance data indicate that the prevalence of two or more chronic diseases among U.S. adults increased from 47.3% in 2013 to 51.4% in 2023 [2]. The increasing prevalence of comorbidity underscores the need for computational approaches that can better capture molecular interaction patterns associated with disease comorbidity.

Early studies characterized disease relationships by identifying shared disease-associated genes [3]. However, genetic overlap alone may not fully capture the complexity

of disease relationships, as disease phenotypes typically emerge from the interplay of multiple pathobiological processes within complex molecular interaction networks rather than from abnormalities in a single gene [4]. To effectively model these complex molecular relationships, network medicine conceptualizes diseases as modules within the human protein–protein interaction (PPI) network, where nodes represent gene products, and each disease module is defined by its associated genes [4]. Within the disease module paradigm, disease comorbidity is inferred using a network proximity measure based on the separation between disease modules in the PPI network, with smaller separations indicating stronger comorbidity relationships [5].

Inspired by the concept of network proximity, subsequent approaches embed the PPI network into a lower-dimensional space in which disease-associated gene products are encoded as vectors that preserve network geodesic distances [6,7]. Disease comorbidity is then predicted using disease-level embeddings obtained by aggregating the resulting gene-product vectors. Although these embedding-based approaches effectively capture global connectivity patterns in the PPI network, they represent diseases through global proximity relationships among individual gene products and thus overlook the internal topology of fragmented disease modules, where gene products may form multiple connected components within a module. As a result, these approaches may fail to capture patterns relevant to comorbidity encoded in the component-level topology of fragmented disease modules.

More recent graph deep learning approaches attempt to address these limitations by learning disease module representations from the topology of their fragmented subgraphs. For example, SNN-VGA [8] represents disease modules as fragmented subgraphs and learns topology-aware representations of their connected components through random anchor-patch message passing. FDS-CAP [9] extends this approach by introducing a component-level attention mechanism that adaptively weights connected components according to their contribution to the overall disease-subgraph representation. Advancing this line of work, DisSubFormer [10] replaces random anchor-based message passing with fully learnable subgraph-to-subgraph transformer attention over biologically informed anchor patches.

Despite these advances, subgraph-based modeling approaches learn representations from the topological properties of disease modules within the PPI network that depend on currently annotated disease-associated genes. However, these approaches remain limited by incomplete disease–gene annotation [5] and may therefore fail to fully capture network-level molecular interaction patterns associated with disease comorbidity. Consequently, addressing these limitations requires extending beyond currently annotated disease-associated genes to include additional genes implicated through molecular events such as mutations [11], regulatory disruptions [12], and gene fusions [13]. Among these, gene fusion is defined as the formation of a single transcriptional unit through the juxtaposition of sequences from two previously independent genes, producing a chimeric gene or transcript [14]. Gene fusion events may arise through structural genomic rearrangements, including translocations, inversions, deletions, insertions, or other complex chromosomal alterations that create fusion genes. These fusion genes may subsequently be transcribed into fusion transcripts. Alternatively, fusion transcripts can be produced through RNA-level mechanisms, such as transcriptional read-through between neighboring genes or trans- or cis-splicing of pre-mRNAs [15]. When the open reading frame (ORF) is preserved, these fusion transcripts may be translated into chimeric proteins combining functional domains from distinct genes. Functionally, gene fusions can lead to gain- or loss-of-function effects, including constitutive activation of kinase signaling or deregulation of transcriptional programs [16]. Additionally, gene fusions may disrupt interactions with key cellular regulators,

thereby perturbing PPI networks [17]. Consistent with these functional effects, gene fusions are strongly implicated in human diseases and are increasingly recognized as clinically significant molecular events with diagnostic, prognostic, and therapeutic relevance across diverse pathological conditions [13,16,18,19].

Motivated by the growing recognition of gene fusions as clinically significant molecular events and their impact on protein interactions, we propose FuDiCo, a computational framework that models disease comorbidity as gene fusion-initiated influence propagation over the PPI network. In network medicine, molecular events can induce perturbations that are not confined to directly involved gene products but may propagate through molecular interaction networks to influence otherwise genetically intact interactors [4]. Accordingly, when fusion events affect certain genes, the resulting perturbations may spread through the PPI network, influencing additional interacting gene products beyond those directly implicated in the initial event. FuDiCo models this process by representing each fusion-associated gene as a localized perturbation source and learning how the resulting influence propagates through PPI paths toward disease subgraphs. If fusion-initiated influence propagates toward multiple disease subgraphs, the resulting propagation patterns may capture network-level molecular relationships associated with disease comorbidity.

FuDiCo introduces several methodological innovations aimed at extending disease representation learning beyond the topological properties of currently annotated disease genes. Specifically, FuDiCo captures the propagation of fusion-initiated perturbations through sampled propagation paths within the PPI network. These propagation paths are subsequently encoded using a fusion influence-aware gated recurrent unit (GRU) architecture. In this architecture, the recurrent gating mechanisms are conditioned on diffusion-based reachability between fusion-associated genes and disease-subgraph components. Collectively, these methodological innovations form a unified framework for disease comorbidity prediction through fusion-initiated influence propagation. Figure 1 provides a conceptual illustration of this propagation process within the PPI network.

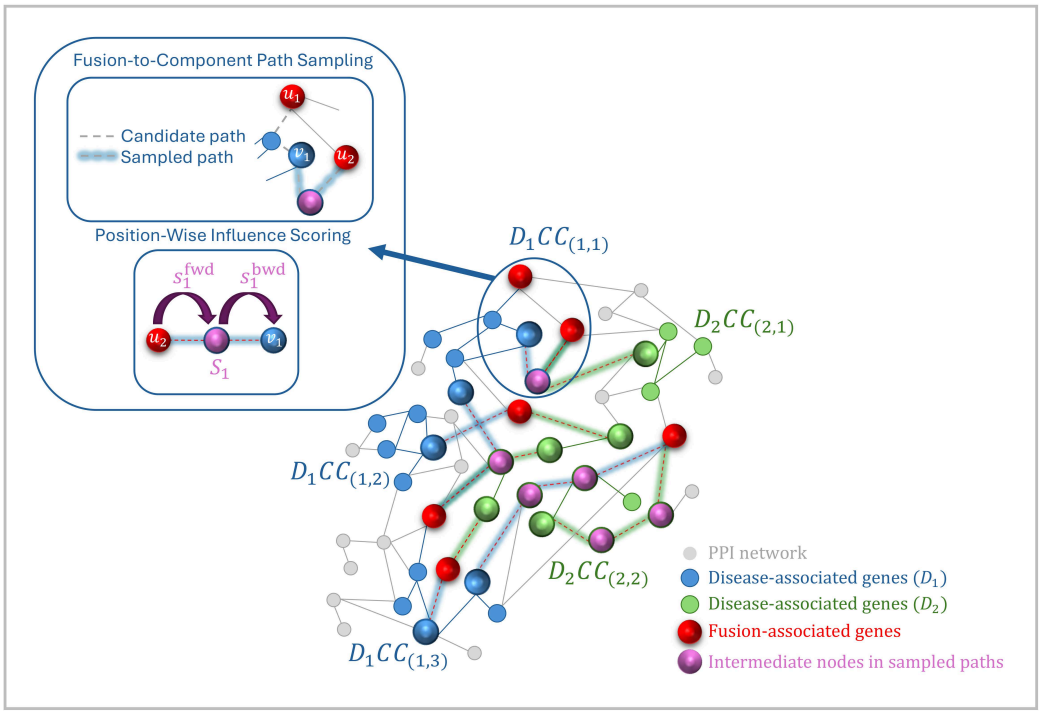


Figure 1. Gene fusion-initiated influence propagation in the protein–protein interaction (PPI) network. The PPI network is modeled as a graph G_{PPI} that serves as the underlying molecular interaction scaffold. Within G_{PPI}, genes associated with disease i form an induced subgraph D_i, which is composed

of connected components $CC_{(i,j)}$, where j is the component index. Fusion-associated genes are represented as localized perturbation source nodes whose influence propagates along interaction paths toward $CC_{(i,j)}$. These propagation paths are subsequently encoded to capture network-level influence patterns that link related diseases and contribute to comorbidity. In the illustrated example of the fusion-to-component path sampling and position-wise influence scoring process, candidate paths connecting fusion-associated genes $u_1, u_2 \in FG_1$ to component node $v_1 \in CC_{(1,1)}$ form a candidate pool for path sampling. Among these candidates, the path connecting u_2 to v_1 is selected and added to $\mathcal{P}^{(2)}_{(1,1)}$, since u_2 achieves the highest diffusion-based reachability score to v_1 (Equation (10)). For the sampled path, the position-wise influence score S_1 is computed using Equation (13) from the forward and backward scores $\left(s_1^{\text{fwd}}, s_1^{\text{bwd}}\right)$ to quantify fusion-initiated influence propagation at position $t = 1$ prior to path encoding.

2. Materials and Methods

2.1. Materials

2.1.1. Protein–Protein Interaction Network

The protein–protein interaction network used in this study consists of 13,460 protein-coding genes connected by 141,296 unique interactions between their encoded protein products [5]. Gene identifiers are mapped to their corresponding protein products using UniProtKB release 2025_03 [20] to represent interactions at the protein level.

2.1.2. Disease–Gene Associations

The disease–gene association dataset used in this study was originally reported in [5]. In that study, disease–gene annotations from Online Mendelian Inheritance in Man [21] and UniProtKB/Swiss-Prot [22] were combined with genome-wide association study data from the Phenotype–Genotype Integrator [23], where only associations meeting a genome-wide significance threshold of $p \leq 5 \times 10^{-8}$ were retained. The dataset was further restricted to diseases with at least 20 associated genes and to genes with available interaction data, resulting in a final set of 299 diseases and 3173 unique associated genes.

2.1.3. Relative Risk-Based Comorbid Disease Pairs

Disease-pair relationships are obtained from the analysis of the Medicare dataset reported in [5], which comprises medical records of disease history for 30 million individuals aged 65 years and older. The analysis identified 10,743 disease pairs, each quantified by a relative risk (RR) value that measures co-occurrence beyond what is expected under independence. For a disease pair $\left(D_i, D_j\right)$, the relative risk is defined as

$$\text{RR}_{D_iD_j} = \frac{C_{D_iD_j} \cdot N}{P_{D_i} \cdot P_{D_j}}, \tag{1}$$

where $C_{D_iD_j}$ is the number of patients diagnosed with both diseases, N represents the total population size, and P_{D_i} and P_{D_j} denote the number of patients diagnosed with the respective diseases. In this study, comorbidity is operationally determined using a predefined threshold θ applied to the RR values, such that a disease pair $\left(D_i, D_j\right)$ is classified as comorbid if $\text{RR}_{D_iD_j} > \theta$.

2.1.4. Fusion Gene Dataset

The fusion gene dataset is obtained from FusionGDB 2.0 [24]. This resource integrates fusion gene information from two major databases: ChiTaRS 5.0 [25], a database of chimeric transcripts matched with druggable fusions and 3D chromatin maps, and ChimerDB 4.0 [26], an updated and expanded database of curated fusion genes. ChiTaRS contains

50,360 fusion genes, whereas ChimerDB reports 52,737. The union of these two datasets yields 102,647 unique fusion genes in FusionGDB 2.0. ORF annotations are included in FusionGDB 2.0 for fusion events, with categories assigned according to breakpoint location within gene regions of both gene partners, namely coding sequence (CDS), untranslated region, or intron. For fusion events in which both breakpoints occur within coding sequences (CDS–CDS), ORF status is determined from the full-length fusion transcript sequence. If the length of the resulting sequence is a multiple of three, the fusion is classified as in-frame, indicating preservation of the reading frame. Otherwise, it is classified as frame-shift, reflecting disruption of the reading frame. This classification is biologically significant, as preservation or disruption of the reading frame determines whether a continuous coding sequence is maintained in the fusion transcript. When such a continuous coding sequence is preserved, the fusion transcript may be translated into a chimeric protein. Accordingly, we retain ORF annotations for CDS–CDS fusion events, including 16,273 in-frame and 17,803 frame-shift fusions.

2.2. Methods

2.2.1. Problem Definition

The PPI network is modeled as an undirected graph $G_{\mathrm{PPI}} = (V, E)$, where $V = \{1, \cdots, n\}$ denotes the set of protein-coding genes (nodes), and $E \subseteq V \times V$ represents the undirected edge set, where each edge $(u, v) \in E$ indicates a protein–protein interaction between the proteins encoded by genes u and v. The graph is equivalently described by an adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, where a matrix element $a_{uv} = 1$ if an edge exists between nodes u and v, and $a_{uv} = 0$ otherwise.

A node embedding matrix $\mathbf{F} \in \mathbb{R}^{n \times d}$ contains the initial embedding vectors for all nodes in G_{PPI}, where the v-th row corresponds to the embedding vector of node v, denoted by $\mathbf{f}_v \in \mathbb{R}^d$. These embedding vectors are derived from a pretrained Evolutionary Scale Modeling (ESM-2) protein language model [27], which generates representations based on the amino acid sequence of the protein encoded by each gene. Each disease is represented by an induced subgraph $D_i = (V_i, E_i)$, where $V_i \subset V$ is the set of nodes corresponding to genes associated with the disease, and $E_i = \{(u, v) \in E |\ u, v \in V_i\}$ defines the subset of PPI edges connecting those nodes. Because disease-associated genes may not form a single connected region in G_{PPI}, D_i may consist of multiple connected components defined as $D_i^{(\mathrm{CC})} = \left\{CC_{(i,1)}, \cdots, CC_{(i,n_i)}\right\}$, where n_i is the number of connected components. For each component $CC_{(i,j)}$, we compute an initial component representation $\mathbf{h}^{\mathrm{init}}_{CC_{(i,j)}} \in \mathbb{R}^{d_{CC}}$ by aggregating the embedding vectors of its nodes. Given a disease-subgraph set $\mathcal{D} = \{D_1, \cdots, D_m\}$, FuDiCo learns disease-subgraph representations $\{\mathbf{h}_{D_i}\}_{i=1}^{m}$ for comorbidity prediction, where $\mathbf{h}_{D_i} \in \mathbb{R}^{d_D}$ and m is the number of disease subgraphs.

2.2.2. Model Design Overview

We consider the PPI graph G_{PPI} as the underlying molecular interaction scaffold that supports gene fusion-initiated influence propagation, where each disease is represented as an induced subgraph D_i within G_{PPI}. On this scaffold, FuDiCo defines genes involved in gene fusion events as sources of perturbation and models how their influence propagates across paths of varying lengths toward disease subgraphs. The resulting propagation patterns are then encoded as disease-subgraph representations (Figure 2). An algorithmic overview of FuDiCo is provided in Algorithm A1 (Appendix A).

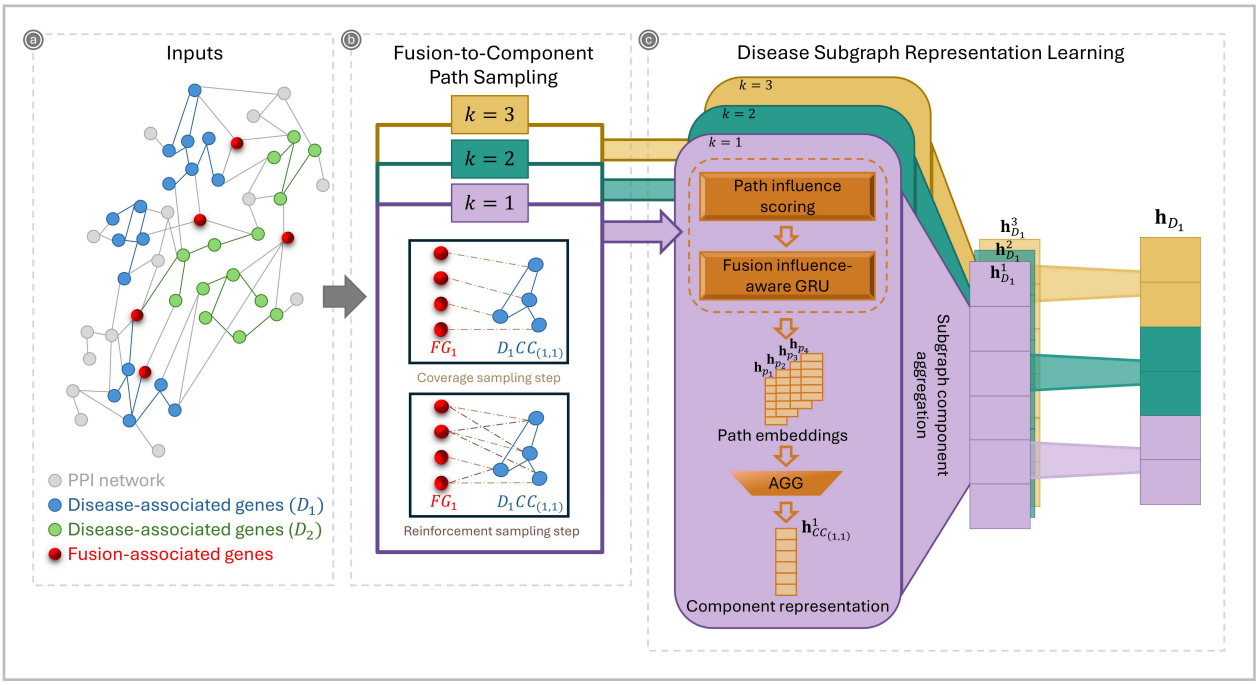


Figure 2. FuDiCo architecture. (**a**) Inputs: the protein–protein interaction (PPI) network, disease-associated genes forming disease subgraphs D_i with their connected components $CC_{(i,j)}$, and the corresponding fusion-associated gene sets FG_i; (**b**) fusion-to-component path sampling: for each path length $k \in \{1, 2, 3\}$, paths from fusion-associated genes in FG_i to nodes within each connected component $CC_{(i,j)}$ of D_i are sampled using a two-step sampling procedure; (**c**) disease-subgraph representation learning: sampled paths are scored and encoded using a fusion influence-aware GRU to generate path embeddings $\mathbf{h}_p$. These path embeddings are aggregated to obtain component representations $\mathbf{h}^k_{CC_{(i,j)}}$, which are subsequently aggregated into path-length-specific disease-subgraph representations $\mathbf{h}^k_{D_i}$. The resulting representations are then combined to produce the final disease-subgraph representation $\mathbf{h}_{D_i}$ for comorbidity prediction.

2.2.3. Fusion-to-Component Diffusion Reachability

For each disease subgraph D_i, we first determine the fusion-associated genes that function as sources of fusion-initiated influence propagation. These genes are derived from fusion events that occur either within D_i, where both partners belong to V_i, or across its boundary, involving at least one gene in V_i. Because both cases can introduce disease-relevant influence, we unify the participating fusion-associated genes into a single source set FG_i, which defines the initiating genes for influence propagation toward the disease subgraph D_i. From this unified source set, influence propagates node-to-node along interaction paths in G_{PPI}, reaching nodes in each connected component $CC_{(i,j)}$ of D_i. Formally, interaction paths are defined by

$$\{(v_0, \cdots, v_k) : v_0 \in FG_i,\ v_k \in CC_{(i,j)},\ (v_{t-1}, v_t) \in E,\ \text{for}\, t = 1, \cdots, k\}, \tag{2}$$

To quantify fusion-initiated influence propagation from fusion-associated genes to nodes in each connected component, we define a symmetric degree-normalized diffusion operator $\mathbf{P}$ on G_{PPI} as follows:

$$\mathbf{P} = \mathbf{Deg}^{-\frac{1}{2}}\mathbf{A}\,\mathbf{Deg}^{-\frac{1}{2}}, \tag{3}$$

where $\mathbf{A}$ is the adjacency matrix of G_{PPI}, and $\mathbf{Deg} \in \mathbb{R}^{n \times n}$ is the diagonal degree matrix with entries:

$$Deg_{uv} = \begin{cases} \deg(u), & u = v \\ 0, & u \neq v \end{cases}, \text{ where } \deg(u) = \sum_{w=1}^{n} a_{uw}, \tag{4}$$

Node-to-node connectivity in G_{PPI} is measured by the number of walks between them, which follows the recursive relation of matrix powers [28]:

$$\left(\mathbf{A}^{k+1}\right)_{uv} = \left(\mathbf{A}^{k}\mathbf{A}\right)_{uv} = \sum_{w=1}^{n} \left(\mathbf{A}^{k}\right)_{uw} a_{wv}, \tag{5}$$

where $\left(\mathbf{A}^{k+1}\right)_{uv}$ gives the number of walks of length $k+1$ from node u to node v. Since the diffusion operator $\mathbf{P}$ is a normalized form of $\mathbf{A}$, $\left(\mathbf{P}^{k}\right)_{uv}$ captures degree-normalized walks of length k between nodes. However, considering only a single walk length k may fail to capture the full propagation extent, since influence may propagate across walks of varying lengths. Accordingly, we define a truncated diffusion operator to aggregate degree-normalized walks across different lengths as follows:

$$\mathbf{T}_K = \sum_{k=1}^{K} w_k \mathbf{P}^k, \tag{6}$$

where $K \in \mathbb{N}$ denotes the maximum walk length and $\{w_k\}_{k=1}^{K}$ are normalized length weights given by $w_k = \frac{\beta^k}{\sum_{i=1}^{K} \beta^i}$ for $\beta \in (0,1)$, such that $\sum_{k=1}^{K} w_k = 1$. Smaller values of β place proportionally greater weight on shorter walks, thereby emphasizing local influence propagation.

Using the truncated diffusion operator, we characterize the diffusion-based reachability for any node pair (u, v) as

$$R_K(u \to v) \triangleq \left(\mathbf{T}_K\right)_{uv}, \tag{7}$$

For a fusion-associated gene $u \in FG_i$ and a node $v \in CC_{(i,j)}$, $R_K(u \to v)$ quantifies the strength of fusion-initiated influence that can propagate from u to v through interaction paths of length at most K in G_{PPI}. For length-specific analysis, we also use $R_k(u \to v)$ with $k \leq K$, which is computed by limiting the diffusion operator to walks of length up to k.

2.2.4. Fusion-to-Component Path Sampling

For each fusion-associated gene $u \in FG_i$ and each node $v \in V_i$ in the disease subgraph D_i, we define a pool of candidate simple paths of length $k \in \{1, \cdots, K\}$ from u to v in G_{PPI}, through which fusion-initiated influence may propagate:

$$\mathcal{C}_i^{(k)}(u, v) = \left\{ p = (v_0, \cdots, v_k) \,|v_0 = u,\ v_k = v,\ (v_{t-1}, v_t) \in E,\ \text{ for } t = 1, \cdots, k \right\}, \tag{8}$$

For each component $CC_{(i,j)}$ within D_i and path length $k \in \{1, \cdots, K\}$, FuDiCo samples a set of paths from the candidate pool that (i) ensures coverage across component nodes by selecting at least one path per node, and (ii) allocates the remaining budget to additional paths connecting high-reachability fusion-associated genes to component nodes. Accordingly, a sampled path set for component $CC_{(i,j)}$ is defined from the candidate paths as follows:

$$\mathcal{P}_{(i,j)}^{(k)} \subseteq \left\{ p \mid \exists\, u \in FG_i,\ v \in CC_{(i,j)} : p \in \mathcal{C}_i^{(k)}(u, v) \right\}, |\mathcal{P}_{(i,j)}^{(k)}| \leq B, \tag{9}$$

where $\mathcal{P}_{(i,j)}^{(k)}$ denotes the sampled set of paths of length k connecting fusion-associated genes in FG_i to nodes in the component $CC_{(i,j)}$, and B is a component-level budget that caps the

number of sampled paths for each component and path length, ensuring computational tractability through controlled path sampling. Specifically, the sampled path set $\mathcal{P}^{(k)}_{(i,j)}$ is constructed using a two-step sampling procedure that balances coverage and reinforcement:

- Coverage sampling step: For each component node $v \in CC_{(i,j)}$, we select the fusion-associated gene $u \in FG_i$ that maximizes diffusion-based reachability (Equation (7)) to v among fusion-associated genes with at least one candidate path of length k ending at v:

$$u^\star(v) = \arg \max_{\substack{u \in FG_i \\ \mathcal{C}^{(k)}_i(u,v) \neq \varnothing}} R_k(u \to v), \tag{10}$$

  We then add at most one path from the selected fusion-associated gene $u^\star(v)$ ending at v to the sampled path set $\mathcal{P}^{(k)}_{(i,j)}$.
- Reinforcement sampling step: If $|\mathcal{P}^{(k)}_{(i,j)}| < B$ after the coverage step, we expand the sampled path set $\mathcal{P}^{(k)}_{(i,j)}$ by allocating the remaining budget to additional paths connecting high-reachability fusion-associated genes to component nodes, while preserving diversity across component nodes.

2.2.5. Fusion-to-Component Path Influence Scoring

For each path $p = (v_0, \cdots, v_k)$ in the sampled path set $\mathcal{P}^{(k)}_{(i,j)}$, where k represents the path length, we quantify fusion-initiated influence propagation at each path position t, with v_t corresponding to the associated node. In particular, we assess whether fusion-initiated influence (i) reaches v_t within t hops and (ii) can further propagate from v_t to the component endpoint node v_k within the remaining $k - t$ hops. Therefore, path-level propagation is characterized through position-level influence reception and subsequent influence propagation along the path. This position-level influence is quantified using diffusion-based reachability over G_{PPI} (Equation (7)), through forward and backward scores defined at position t as follows:

$$s_t^{\text{fwd}} = R_t(v_0 \to v_t), \tag{11}$$

$$s_t^{\text{bwd}} = R_{k-t}(v_t \to v_k), \tag{12}$$

The forward score s_t^{fwd} quantifies the strength of influence propagation from the fusion-associated gene source v_0 to v_t, while the backward score s_t^{bwd} measures the ability of the remaining $k - t$ hops to propagate that influence from v_t toward the component endpoint node v_k. Because the forward and backward scores jointly characterize the extent to which position t receives and propagates fusion-initiated influence along the path, we combine them to define the position-wise influence score S_t as

$$S_t = \frac{2s_t^{\text{fwd}} s_t^{\text{bwd}} + \lambda\left(s_t^{\text{fwd}} + s_t^{\text{bwd}}\right)}{s_t^{\text{fwd}} + s_t^{\text{bwd}} + 2\lambda}, \tag{13}$$

The central term $\frac{2s_t^{\text{fwd}} s_t^{\text{bwd}}}{s_t^{\text{fwd}} + s_t^{\text{bwd}}}$ corresponds to the harmonic mean between the forward and backward scores, ensuring that strong influence is assigned only when both directions contribute to propagation along the path. However, because the values of s_t^{fwd} and s_t^{bwd} may vary substantially across positions along the path, this harmonic term can become

unstable when one of the scores approaches zero. To stabilize this behavior, we incorporate a scale-adaptive pseudocount λ into S_t, defined as follows:

$$\lambda = \rho\left(s_t^{\text{fwd}} + s_t^{\text{bwd}}\right), \rho = \frac{\tau}{1-2\tau}, \tau \in (0,\ 0.5), \tag{14}$$

The position-wise influence score S_t satisfies several desirable properties including: (i) symmetry, $S_t\left(s_t^{\text{fwd}}, s_t^{\text{bwd}}\right) = S_t\left(s_t^{\text{bwd}}, s_t^{\text{fwd}}\right)$; (ii) boundedness, $\min\left\{s_t^{\text{fwd}}, s_t^{\text{bwd}}\right\} \leq S_t \leq \max\left\{s_t^{\text{fwd}}, s_t^{\text{bwd}}\right\}$; (iii) scale equivariance, such that for any $c > 0$, $S_t\left(cs_t^{\text{fwd}}, cs_t^{\text{bwd}}\right) = cS_t$; (iv) a fixed-point property, $S_t(s,s) = s$; and (v) controlled one-sided behavior, $S_t\left(s_t^{\text{fwd}}, 0\right) = \tau s_t^{\text{fwd}}$, $S_t\left(0, s_t^{\text{bwd}}\right) = \tau s_t^{\text{bwd}}$.

2.2.6. Path Encoding with Fusion Influence-Aware GRU

Given the fusion-to-component path set $\mathcal{P}_{(i,j)}^{(k)}$ with quantified position-wise influence scores along each path, FuDiCo encodes these paths into fixed-length embeddings conditioned on the propagated fusion-initiated influence. For each position $t \in \{0, \cdots, k\}$ within the path, we represent the node v_t by its embedding $\mathbf{f}_{v_t}$ and associate it with the position-wise influence score S_t, which measures the strength of influence propagation at that position. However, propagation strength should capture both the influence at the current position t and the persistence of propagation across preceding positions along the path. Therefore, we introduce an accumulated influence score $\overline{S}_t$ that aggregates the influence propagation strength across preceding positions up to t as follows:

$$\overline{S}_t = (1-\gamma)S_t + \gamma\overline{S}_{t-1}, \tag{15}$$

where $\gamma \in (0,1)$ is a propagation parameter controlling the contribution of previously accumulated influence $\overline{S}_{t-1}$ to $\overline{S}_t$, such that larger γ increases the contribution of past influence, whereas smaller γ places greater emphasis on the current influence score S_t. The accumulated influence score $\overline{S}_k$ at the final position k thereby defines a path-wise influence score, capturing the overall strength and consistency of fusion-initiated influence propagated along the path from v_0 to v_k.

We encode each path using a fusion influence-aware GRU, built upon the standard GRU architecture [29,30], with gates explicitly conditioned on fusion-initiated influence scores. The fusion influence-aware GRU processes path positions sequentially, maintaining a hidden state that encodes the fusion-initiated influence propagated up to each position. The hidden state at position t, denoted by $\mathbf{h}_t \in \mathbb{R}^H$, is computed as a gated combination between the previous state $\mathbf{h}_{t-1} \in \mathbb{R}^H$ and a candidate state $\tilde{\mathbf{h}}_t \in \mathbb{R}^H$:

$$\mathbf{h}_t = (1-\mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t, \tag{16}$$

where $\odot$ denotes element-wise multiplication, and $\mathbf{z}_t \in (0,1)^H$ is the update gate computed as

$$\mathbf{z}_t = \sigma(\mathbf{W}_z\mathbf{f}_{v_t} + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{V}_z S_t), \tag{17}$$

where $\mathbf{f}_{v_t}$ represents the embedding of node v_t, $\mathbf{W}_z$, $\mathbf{U}_z$, and $\mathbf{V}_z$ are update-gate learnable parameters, and $\sigma(\cdot)$ is the sigmoid activation function. The update gate is conditioned on the position-wise influence score S_t, allowing positions with strong propagated influence to exert greater control over the hidden state update. Biologically, path positions corresponding to molecular interactions receiving stronger propagated fusion-initiated influence contribute more strongly to the encoded path representation. As a result, the model prioritizes interaction regions that may be more relevant to disease-associated molecular perturbations underlying disease comorbidity.

Similarly, the reset gate $\mathbf{r}_t \in (0,1)^H$ regulates how much previously propagated fusion-initiated influence is incorporated when forming the candidate state, and is computed as

$$\mathbf{r}_t = \sigma\big(\mathbf{W}_r \mathbf{f}_{v_t} + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{V}_r \overline{S}_t\big), \tag{18}$$

where $\mathbf{W}_r$, $\mathbf{U}_r$, and $\mathbf{V}_r$ are reset-gate learnable parameters. The reset gate is conditioned on the accumulated influence score $\overline{S}_t$, modulating the extent to which previously propagated fusion-initiated influence encoded in $\mathbf{h}_{t-1}$ contributes to the computation of the candidate state $\tilde{\mathbf{h}}_t$. This design enables the model to preserve persistent fusion-initiated influence propagation patterns across consecutive biologically connected interaction paths.

The candidate state combines the current node embedding, the previous hidden state modulated by the reset gate, and the position-wise influence score S_t:

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_n \mathbf{f}_{v_t} + \mathbf{U}_n(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{V}_n S_t), \tag{19}$$

where $\mathbf{W}_n, \mathbf{U}_n$, and $\mathbf{V}_n$ are learnable parameters and $\tanh(\cdot)$ denotes the hyperbolic tangent activation function. Incorporating S_t into the candidate state computation enables positions along the path with strong influence propagation to exert a stronger impact on the candidate state representation.

After processing all positions along the path p, the final hidden state $\mathbf{h}_k$ at position k is taken as the path embedding $\mathbf{h}_p \in \mathbb{R}^H$. In addition, the accumulated influence score $\overline{S}_k$ at the final position k defines the path-wise influence score $\overline{S}_p$. Together, the pair $\big(\mathbf{h}_p, \overline{S}_p\big)$ is subsequently used to construct disease-subgraph representations for comorbidity prediction. Accordingly, all fusion influence-aware GRU parameters, including those of the update gate, reset gate, and candidate state transformations (i.e., $\mathbf{W}_z, \mathbf{U}_z, \mathbf{V}_z$, $\mathbf{W}_r, \mathbf{U}_r, \mathbf{V}_r, \mathbf{W}_n, \mathbf{U}_n$, and $\mathbf{V}_n$), are learned jointly in an end-to-end manner via backpropagation from the final comorbidity prediction loss (Equation (25)). This training procedure thereby learns path representations that capture fusion-initiated influence propagation patterns in the PPI network relevant to disease comorbidity.

2.2.7. Disease-Subgraph Representation and Comorbidity Prediction

Disease-Subgraph Representation. Given the sampled fusion-to-component paths, the corresponding path embeddings, and the associated path-wise influence scores, we construct disease-subgraph representations that encode fusion-initiated influence propagation across their connected components. For each component $CC_{(i,j)}$ of the disease subgraph D_i, we use sampled path sets $\mathcal{P}^{(k)}_{(i,j)}$ corresponding to different path lengths $k \in \{1, \cdots, K\}$. Each path $p \in \mathcal{P}^{(k)}_{(i,j)}$ is associated with a path embedding $\mathbf{h}_p$ and its corresponding influence score $\overline{S}_p$. The score $\overline{S}_p$ is then normalized via a softmax over the influence scores of all paths in $\mathcal{P}^{(k)}_{(i,j)}$ to yield a path-level attention weight:

$$\alpha_p = \frac{\exp\big(\overline{S}_p\big)}{\sum_{\acute{p} \in \mathcal{P}^{(k)}_{(i,j)}} \exp\big(\overline{S}_{\acute{p}}\big)}, \tag{20}$$

A weighted representation for component $CC_{(i,j)}$ at path length k is computed as the attention-weighted sum of path embeddings in $\mathcal{P}^{(k)}_{(i,j)}$, allowing paths with stronger and more persistent fusion-initiated influence propagation to contribute more strongly to the component representation:

$$\mathbf{h}^k_{CC_{(i,j)}} = \sum_{p \in \mathcal{P}^{(k)}_{(i,j)}} \alpha_p \mathbf{h}_p, \tag{21}$$

The representations of the connected components within D_i are further aggregated to obtain a disease-subgraph representation corresponding to path length k:

$$\mathbf{h}_{D_i}^{k} = \mathrm{AGG}_{\mathrm{CC}}\left(\mathbf{h}_{CC_{(i,1)}}^{k}, \cdots, \mathbf{h}_{CC_{(i,n_i)}}^{k}\right), \tag{22}$$

Finally, the overall disease-subgraph representation is constructed by aggregating the path-length-specific disease-subgraph representations across path lengths k:

$$\mathbf{h}_{D_i} = \mathrm{READOUT}\left(\mathbf{h}_{D_i}^{1}, \cdots, \mathbf{h}_{D_i}^{K}\right), \tag{23}$$

Comorbidity Prediction. For a pair of diseases D_i and D_j, FuDiCo estimates their comorbidity by passing the concatenated disease-subgraph representations through a two-layer multilayer perceptron (MLP) followed by a sigmoid activation:

$$\hat{y} = \sigma\left(\mathbf{W}_2\mathrm{ReLU}\left(\mathbf{W}_1\mathrm{concat}\left(\mathbf{h}_{D_i}, \mathbf{h}_{D_j}\right) + \mathbf{b}_1\right) + \mathbf{b}_2\right), \tag{24}$$

where $\hat{y} \in (0,1)$ denotes the predicted probability that diseases D_i and D_j are comorbid, $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices, and $\mathbf{b}_1$ and $\mathbf{b}_2$ are learnable bias vectors of the MLP. The rectified linear unit (ReLU) is used as the activation function.

Training Objective. Let $\left\{\left(D_{i_n}, D_{j_n}, y_n\right)\right\}_{n=1}^{N}$ be the set of disease pairs with comorbidity labels, where N is the total number of pairs in the dataset. For the n-th pair, $y_n \in \{0,1\}$ represents the ground-truth comorbidity label, with $y_n = 1$ indicating that diseases D_{i_n} and D_{j_n} are comorbid and $y_n = 0$ otherwise. Given this set and the predicted comorbidity probabilities, the model is trained by minimizing the binary cross-entropy (BCE) loss:

$$L_{\mathrm{BCE}} = -\frac{1}{N}\sum_{n=1}^{N}(y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)), \tag{25}$$

2.3. Experimental Setup

2.3.1. Dataset

We trained and evaluated FuDiCo using the data sources described in Section 2.1. These sources include: (i) the PPI network, which serves as the base interaction graph; (ii) disease-associated genes represented as induced disease subgraphs of the PPI network; (iii) clinically reported disease pairs with RR values for deriving ground-truth comorbidity labels; and (iv) fusion gene data for modeling fusion-initiated influence.

Following prior work, we set the RR threshold to $\theta = 0$ and classify a disease pair (D_i, D_j) as comorbid if $\mathrm{RR}_{D_iD_j} > \theta$, yielding 8874 comorbid disease pairs out of 10,743 total pairs. The 8874 comorbid disease pairs were partitioned at the disease-pair level into 80% training, 10% validation, and 10% testing splits. Under this setting, disease pairs were partitioned disjointly across splits, while individual diseases may appear in multiple splits through different disease-pair combinations. Nevertheless, FuDiCo does not learn disease representations directly from disease-pair comorbidity relations, but rather through fusion-initiated influence propagation within the PPI network, which reduces the risk of information leakage arising from disease overlap across splits. For training and evaluation, negative disease pairs corresponding to 25% of the number of comorbid disease pairs within each split were sampled with replacement from disease pairs not included in the observed comorbid set. This sampling ratio was selected to preserve class imbalance characteristics commonly observed in real-world clinical and biological association datasets while increasing the proportion of negative samples to make the prediction task more challenging during training and evaluation. In addition, using substantially fewer negative samples relative to positive comorbid pairs could bias model optimization

toward the majority positive class and amplify the effect of the highly positive-skewed class distribution on evaluation metrics and reported model performance. Therefore, to ensure fair comparative evaluation under the selected sampling configuration, the same train/validation/test partitions and negative sampling protocol were consistently applied across all the compared methods.

The fusion gene dataset used in FuDiCo is obtained from FusionGDB 2.0, which reports 16,273 in-frame and 17,803 frame-shift fusion events. After combining these two categories and removing duplicates, we obtained 26,956 fusion gene pairs involving 12,751 unique genes. These fusion gene pairs were further filtered to retain only pairs with both partner genes present in the PPI network. The final set comprised 18,360 fusion gene pairs and 9247 genes, covering 68.70% of the genes in the PPI network.

2.3.2. Pretraining Gene Embeddings

Each node in G_{PPI}, corresponding to a protein-coding gene, was initialized using protein-sequence embeddings derived from ESM-2. Genes were mapped to UniProtKB protein accessions using the UniProt ID Mapping service, and the associated amino acid sequences for these accessions were retrieved. The sequences were encoded using the 33-layer ESM-2 transformer model (≈650 M parameters) via the Hugging Face Transformers library [31], generating residue-level embeddings that were subsequently mean-pooled to obtain fixed-length sequence representations. When multiple protein sequences corresponded to a single gene, their embeddings were averaged to form a unified gene-level embedding $\mathbf{f}_v \in \mathbb{R}^{1280}$. Each gene embedding $\mathbf{f}_v$, representing the initial node embedding, was then transformed through path-length-specific learnable linear projections from $\mathbb{R}^{1280}$ to $\mathbb{R}^{128}$. These projections were optimized jointly in an end-to-end manner via backpropagation from the final comorbidity prediction loss (Equation (25)).

2.3.3. FuDiCo Training for Disease Comorbidity Prediction

FuDiCo limits path lengths to $k \in \{1, 2, 3\}$, where $K = 3$ denotes the maximum path length. For each path length k, a component-level path budget of $B = 100$ was applied to control the number of sampled paths. Fusion-initiated influence propagation along these sampled paths was quantified through diffusion-based reachability scores $R_K(u \to v)$ derived from the truncated diffusion operator with normalized length weights w_k, where $\beta = 0.5$ (Equation (6)). These reachability scores were subsequently used to compute the position-wise influence scores S_t, with a smoothing parameter $\tau = 0.2$. The accumulated influence score $\overline{S}_t$ was then obtained using a globally shared propagation parameter γ across all the disease subgraphs, propagation paths, and path lengths. The parameter γ was defined as $\gamma = \sigma(\gamma_{\text{raw}})$, where γ_{raw} is a learnable raw parameter initialized to 0.5. The raw parameter γ_{raw} was optimized jointly with all the model parameters via end-to-end backpropagation from the comorbidity prediction objective (Equation (25)).

Initial component representations $\mathbf{h}^{\text{init}}_{CC_{(i,j)}} \in \mathbb{R}^{128}$ were first computed for connected components within the disease subgraph D_i via mean aggregation of node embedding vectors within each component. For each path length k, the sampled fusion-to-component propagation paths were encoded using a fusion influence-aware GRU with a hidden-state dimension of 128, yielding path embeddings $\mathbf{h}_p \in \mathbb{R}^{128}$. Within each component, an attention-weighted sum of path embeddings was computed to construct a component-level representation. The resulting component representations were then averaged to obtain a path-length-specific disease-subgraph representation. Disease-subgraph representations corresponding to different path lengths k were subsequently concatenated to form the final disease-subgraph representation $\mathbf{h}_{D_i} \in \mathbb{R}^{32K}$.

FuDiCo was implemented using PyTorch Geometric v2.6.0 [32] within the PyTorch Lightning v2.4.0 training framework [33]. The model was optimized using the AdamW optimizer [34] with weight decay 2.58×10^{-4}. The learning rate was scheduled according to the OneCycleLR policy [35], starting from 8.03×10^{-4} and increasing to a peak value of 1.41×10^{-3} during training. Gradient clipping with a threshold of 2.56×10^{-1} was additionally applied to stabilize optimization. Training was performed for up to 20 epochs with a batch size of 32, and the model with the highest validation area under the receiver operating characteristic curve (AUROC) was selected. All the experiments were conducted on a machine equipped with an NVIDIA GeForce RTX 4070 GPU with 12 GB VRAM, a 24-core Intel Core i9 CPU at 2.00 GHz, and 32 GB RAM.

2.3.4. Comparison on Disease Comorbidity Prediction

We benchmarked FuDiCo against three recent state-of-the-art methods for disease comorbidity prediction: Biologically Supervised Embedding (BSE) [7], FDS-CAP, and DisSubFormer.

BSE derives gene embeddings for all genes in the PPI network by mapping them into a biologically supervised lower-dimensional space while preserving global network structure. Disease-level representations are then obtained by aggregating disease-associated gene embeddings, which are used as input to a supervised support vector machine classifier for comorbidity prediction.

FDS-CAP is a two-stage graph-based deep learning framework that models disease modules as fragmented subgraphs within the PPI network. In the first stage, a subgraph neural network performs topology-aware message passing between connected components and randomly sampled anchor patches, producing component embeddings that are aggregated into disease embeddings using an attention mechanism. In the second stage, these embeddings serve as node features for a variational graph autoencoder applied to an RR-derived human disease network to predict missing comorbidity links.

DisSubFormer further advances subgraph-based disease modeling by introducing a subgraph transformer that learns disease representations from fragmented subgraphs. The model first generates unified gene embeddings by integrating PPI-derived molecular interactions with Gene Ontology-based functional information. It then applies subgraph-to-subgraph attention between biologically informed anchor patches and connected components to learn disease-subgraph representations that capture their topological properties for comorbidity prediction.

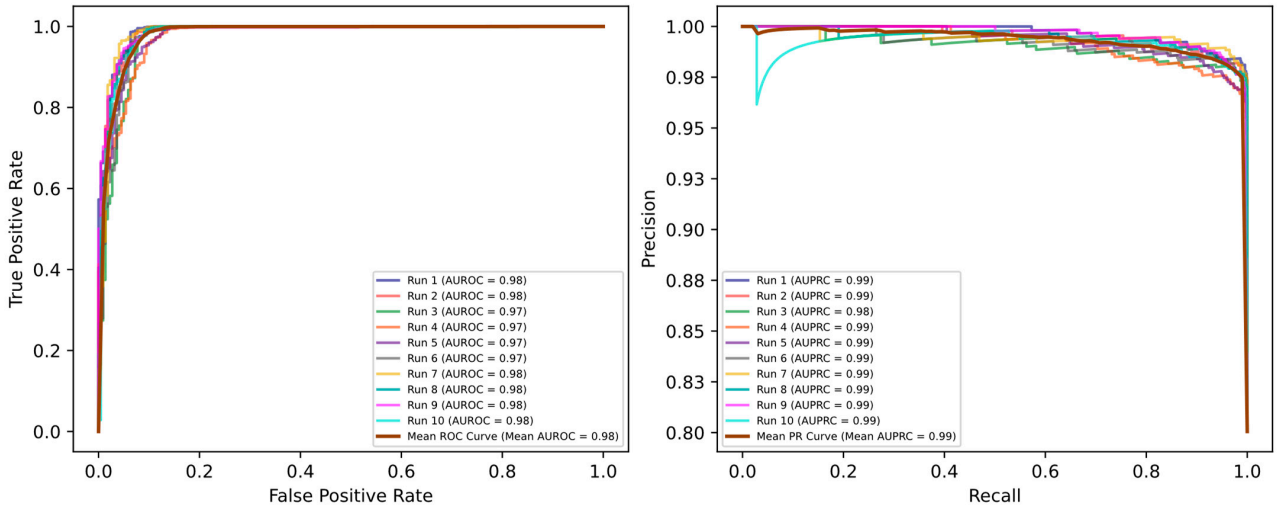
3. Results

3.1. Performance on Disease Comorbidity Prediction

We evaluated FuDiCo for disease comorbidity prediction across ten independent runs using four standard evaluation metrics, including AUROC, accuracy, F1 score, and average precision (AP). As summarized in Table 1, FuDiCo demonstrates consistently strong performance across all the evaluation metrics, achieving mean values of 0.9815 ± 0.0052 for AUROC, 0.9728 ± 0.0046 for accuracy, 0.9828 ± 0.0029 for F1 score, and 0.9940 ± 0.0024 for AP. The relatively small standard deviations observed across these runs indicate that the proposed model is stable and produces reliable predictions. To further examine performance consistency, Figure 3 presents the ROC and precision–recall (PR) curves for individual runs of FuDiCo together with their corresponding mean curves. The curves exhibit minimal variation and substantial overlap, indicating stable performance across independent runs.

Table 1. Performance comparison of FuDiCo and state-of-the-art models for disease comorbidity prediction. Values in bold indicate the best performance for each metric. Reported *p*-values correspond to statistical comparisons between FuDiCo and each model.

Models	Evaluation Metrics							
	AUROC	***p*-Value**	**Accuracy**	***p*-Value**	**F1**	***p*-Value**	**AP**	***p*-Value**
FuDiCo (Ours)	**0.9815** ± 0.0052	-	**0.9728** ± 0.0046	-	**0.9828** ± 0.0029	-	**0.9940** ± 0.0024	-
DisSubFormer	0.9703 ± 0.0054	1.57×10^{-4}	0.9606 ± 0.0056	2.76×10^{-5}	0.9756 ± 0.0035	5.53×10^{-5}	0.9886 ± 0.0030	3.10×10^{-4}
FDS-CAP	0.9288 ± 0.0122	1.59×10^{-7}	0.9229 ± 0.0095	8.55×10^{-8}	0.9524 ± 0.0060	1.26×10^{-7}	0.9704 ± 0.0066	9.84×10^{-7}
BSE	0.9194 ± 0.0170	2.67×10^{-7}	0.9052 ± 0.0050	3.69×10^{-12}	0.9440 ± 0.0028	4.31×10^{-12}	0.9665 ± 0.0076	8.40×10^{-7}

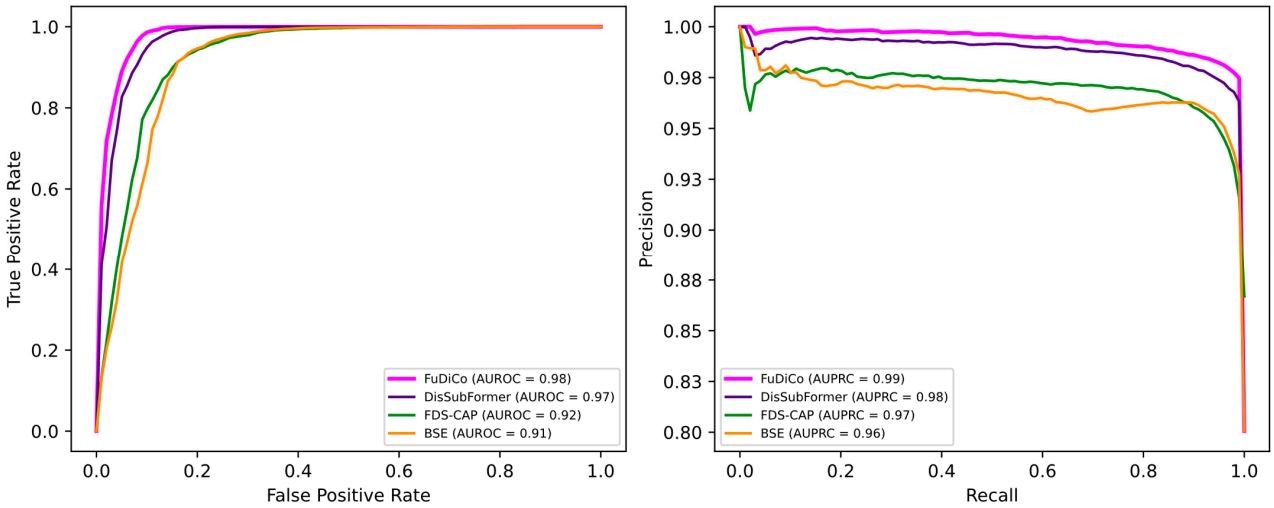


(**a**) (**b**)

Figure 3. Receiver operating characteristic (ROC) and precision–recall (PR) curves across ten independent runs of FuDiCo: (**a**) ROC curves for individual runs together with the mean ROC curve and the corresponding area under the receiver operating characteristic curve (AUROC) values; (**b**) PR curves for individual runs together with the mean PR curve and the corresponding area under the precision–recall curve (AUPRC) values.

We compared FuDiCo with three state-of-the-art methods, namely DisSubFormer, FDS-CAP, and BSE, as shown in Table 1. For all the methods, experiments were repeated over ten independent runs using identical test sets across all the methods to ensure a fair comparison, and the mean performance across runs was reported. Compared to DisSubFormer, FuDiCo improves AUROC by 1.16%, accuracy by 1.26%, F1 score by 0.73%, and AP by 0.54%. Although these performance gains are modest, they remain statistically significant across all the evaluation metrics, with *p*-values of 1.57×10^{-4} for AUROC, 2.76×10^{-5} for accuracy, 5.53×10^{-5} for F1 score, and 3.10×10^{-4} for AP. These results indicate that the observed improvements are unlikely to arise from random variation across runs. The performance gains become more substantial when comparing FuDiCo with FDS-CAP. FuDiCo achieves statistically significant improvements of 5.68% in AUROC $(p = 1.59 \times 10^{-7})$, 5.41% in accuracy $(p = 8.55 \times 10^{-8})$, 3.19% in F1 score $(p = 1.26 \times 10^{-7})$, and 2.43% in AP $(p = 9.84 \times 10^{-7})$. Relative to BSE, the performance gap further widens, with statistically significant increases of 6.76% in AUROC $(p = 2.67 \times 10^{-7})$, 7.47% in accuracy $(p = 3.69 \times 10^{-12})$, 4.11% in F1 score

$(p = 4.31 \times 10^{-12})$, and 2.85% in AP $(p = 8.40 \times 10^{-7})$. Figure 4 further illustrates these comparisons by showing the mean ROC and PR curves across runs for FuDiCo and the state-of-the-art methods. From these curves, FuDiCo maintains higher true positive rates across most false positive rate ranges and higher precision across a broad range of recall values, indicating more reliable identification of comorbid disease pairs. Moreover, the ROC and PR curves of FuDiCo remain consistently above those of the compared methods without noticeable curve crossings, demonstrating stable superiority across different decision thresholds rather than improvements observed only at specific operating points.



(**a**) (**b**)

Figure 4. Mean receiver operating characteristic (ROC) and precision–recall (PR) curves across ten independent runs comparing FuDiCo with state-of-the-art models, including DisSubFormer, FDS-CAP, and BSE. (**a**) Mean ROC curves for all models together with the corresponding area under the receiver operating characteristic curve (AUROC) values; (**b**) Mean PR curves for all models together with the corresponding area under the precision–recall curve (AUPRC) values.

The experimental results show that FuDiCo consistently outperforms state-of-the-art methods across all evaluation metrics for disease comorbidity prediction. These performance improvements may be attributed to the explicit incorporation of fusion-initiated influence propagation within the PPI network. By modeling such propagation, FuDiCo learns disease-subgraph representations that better capture disease relationships than approaches relying solely on the topology of known disease–gene associations.

3.2. Ablation Study

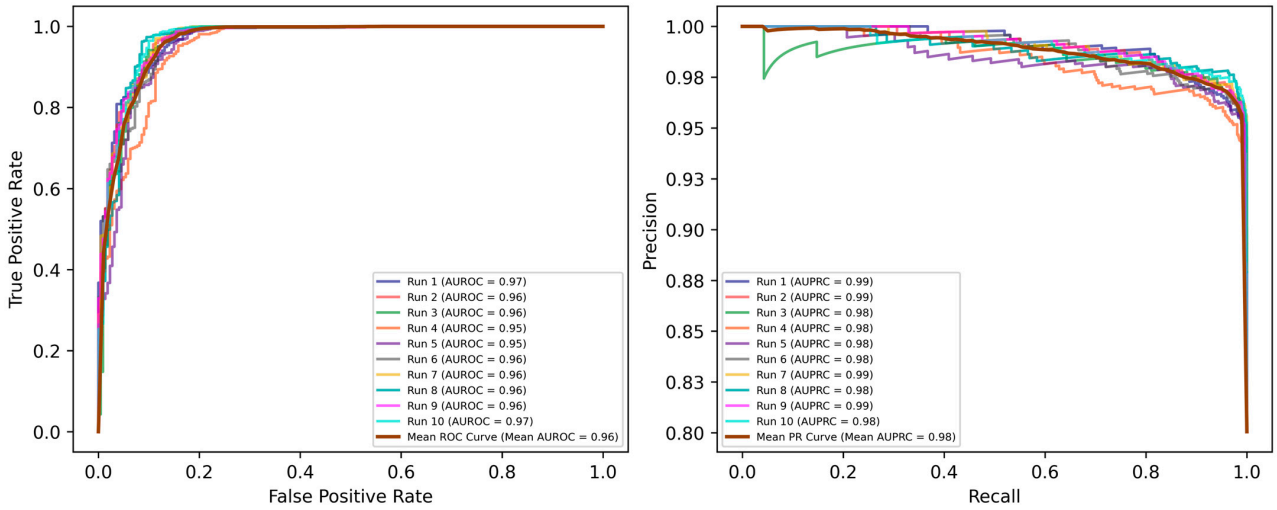
We conducted an ablation study to evaluate the contribution of the proposed fusion influence-aware GRU to modeling fusion-initiated influence propagation and improving disease comorbidity prediction performance. Specifically, the fusion influence-aware GRU was replaced with a standard GRU architecture that does not incorporate propagation influence scores during sequential propagation path encoding. Under this ablation setting, the fusion influence-aware gating modifications were removed. The position-wise influence score S_t was excluded from the update gate and candidate state computations (Equations (17) and (19)), while the accumulated influence score $\overline{S}_t$ (Equation (18)) was excluded from the reset gate computation. Without these influence-aware modifications, the sampled fusion-to-component propagation paths were encoded using only node embedding sequences derived from the node embedding matrix $\mathbf{F}$. Accordingly, we trained and evaluated FuDiCo with the standard GRU variant while preserving the same propagation path sampling strategy, connected-component aggregation procedure, and overall

disease-subgraph representation and comorbidity prediction pipeline used in FuDiCo with the fusion influence-aware GRU.

Table 2 summarizes the experimental results comparing FuDiCo with the fusion influence-aware GRU and the standard GRU variant for disease comorbidity prediction across ten independent runs using the standard evaluation metrics. The results demonstrate that removing the fusion influence-aware gating modifications leads to a statistically significant reduction in disease comorbidity prediction performance across all evaluation metrics. Specifically, AUROC, accuracy, F1 score, and AUPRC decreased by 1.67% ($p = 2.60 \times 10^{-6}$), 1.88% ($p = 3.11 \times 10^{-7}$), 1.12% ($p = 2.56 \times 10^{-7}$), and 0.54% ($p = 1.10 \times 10^{-4}$), respectively. To further examine the performance of the standard GRU variant, Figure 5 presents the ROC and PR curves of FuDiCo with the standard GRU variant across ten independent runs together with the corresponding AUROC and AUPRC values.

Table 2. Performance comparison of FuDiCo with the fusion influence-aware GRU and the standard GRU variant for disease comorbidity prediction. Values in bold indicate the best performance for each metric. Reported *p*-values correspond to statistical comparisons between the two FuDiCo variants.

Models	Evaluation Metrics							
	AUROC	***p*-Value**	**Accuracy**	***p*-Value**	**F1**	***p*-Value**	**AP**	***p*-Value**
FuDiCo (fusion influence-aware GRU)	**0.9815** ± 0.0052	-	**0.9728** ± 0.0046	-	**0.9828** ± 0.0029	-	**0.9940** ± 0.0024	-
FuDiCo (standard GRU)	0.9654 ± 0.0070	2.60×10^{-6}	0.9548 ± 0.0049	3.11×10^{-7}	0.9719 ± 0.0030	2.56×10^{-7}	0.9887 ± 0.0028	1.10×10^{-4}



(**a**) (**b**)

Figure 5. Receiver operating characteristic (ROC) and precision–recall (PR) curves across ten independent runs of FuDiCo with the standard GRU variant: (**a**) ROC curves for individual runs together with the mean ROC curve and the corresponding area under the receiver operating characteristic curve (AUROC) values; (**b**) PR curves for individual runs together with the mean PR curve and the corresponding area under the precision–recall curve (AUPRC) values.

Beyond the predictive performance, these results further indicate that conditioning the GRU gating mechanism on propagation influence scores enables FuDiCo to prioritize biologically relevant propagation patterns along sampled fusion-to-component paths. Consequently, these prioritized propagation patterns are incorporated into the learned disease-subgraph representations, allowing FuDiCo to better capture comorbid disease relationships associated with fusion-initiated influence propagation.

3.3. Scalability and Computational Complexity

The computational complexity of FuDiCo primarily scales with the number of disease subgraphs, connected components, fusion-associated genes, and sampled propagation paths. Accordingly, the overall computational workflow can be divided into two main stages: (1) fusion-to-component path sampling and influence propagation estimation, and (2) sequential propagation path encoding using the fusion influence-aware GRU.

During fusion-to-component path sampling, the number of candidate simple propagation paths may grow substantially with increasing network connectivity and propagation depth, potentially leading to considerable overhead in dense biological interaction networks. To maintain computational tractability, FuDiCo adopts bounded propagation path sampling using fixed path-length and path-budget constraints defined in Section 2.3.3. These constraints limit both the length and number of sampled propagation paths processed during training and inference, thereby controlling memory usage and runtime complexity. In addition, several preprocessing steps, including fusion-to-component path sampling, diffusion-based reachability score computation, and position-wise and accumulated influence score calculation, are performed prior to model training. As a result, this preprocessing accelerates model optimization and hyperparameter tuning.

At the architectural level, FuDiCo models biologically relevant propagation patterns between fusion-associated genes and disease-subgraph connected components rather than modeling exhaustive propagation paths between all nodes in the PPI network and connected components. This design reduces redundant computations by restricting propagation modeling to biologically relevant paths, providing a favorable tradeoff between biological expressiveness and computational efficiency.

In the sequential propagation path encoding stage, the proposed fusion influence-aware GRU scales linearly with both propagation path length and the number of sampled propagation paths, which remains computationally tractable under the bounded path-sampling strategy. To empirically assess runtime scalability, we measured the mean propagation path encoding runtime per training batch across all training epochs, together with the mean encoding runtime associated with each propagation path length k. The mean propagation path encoding runtime was 0.83 s per batch, of which 4%, 23%, and 73% corresponded to propagation path lengths $k = 1$, $k = 2$, and $k = 3$, respectively. These results indicate that computational cost increases with propagation path length due to the additional sequential operations required for longer fusion-to-component propagation paths, while remaining computationally manageable under the bounded path-sampling strategy used in FuDiCo.

4. Discussion

As described in Section 2.3.3, FuDiCo limits propagation path lengths to $k \in \{1, 2, 3\}$ to control the rapid growth in candidate simple paths and reduce computational complexity. To evaluate whether the imposed path length restriction preserves biologically meaningful fusion-initiated influence propagation, we analyzed the distribution of shortest propagation paths between fusion-associated genes and disease-subgraph connected component nodes across all 299 disease subgraphs. Among the 23,302 analyzed nodes in disease-subgraph connected components, all the nodes were reachable from at least one fusion-associated gene within the corresponding disease subgraph through paths in the global PPI network when no path length cutoff is imposed. Under the imposed path length restriction, 99.86% of reachable nodes remained connected to fusion-associated genes within the selected propagation path length cutoff of $k \leq 3$, whereas only 0.14% required longer paths.

We further analyzed connected-component-level reachability between disease-subgraph connected components and fusion-associated genes under the imposed propagation path

length cutoff. Among the 13,232 analyzed connected components, 13,199 retained at least one node reachable from a fusion-associated gene within $k \leq 3$, yielding 99.75% disease-subgraph connected component coverage. The remaining 33 excluded connected components each contain only a single disease-component node requiring propagation paths longer than $k = 3$, indicating that no larger multi-node disease component is entirely excluded by the imposed path length restriction. Accordingly, these findings suggest that biologically meaningful fusion-initiated influence propagation is primarily concentrated within short-range molecular interaction regions of the PPI network. The imposed path length restriction therefore preserves nearly all biologically reachable disease-subgraph connected components while reducing the potential noise associated with long-range propagation patterns in large-scale PPI networks.

Although the reachability analysis shows that nearly all disease-subgraph connected components are reachable within the selected propagation path length cutoff, a small subset of components still lacks valid sampled fusion-to-component paths under the imposed short-range propagation restriction. Such cases may arise from the incomplete coverage of currently available PPI networks and fusion gene datasets, which can limit short-range propagation patterns captured by FuDiCo. For these connected components, FuDiCo falls back to a gated connected-component representation of the form $\mathbf{g}^{(k)} \odot \mathbf{h}_{CC_{(i,j)}}^{\text{init},k}$, where $\mathbf{g}^{(k)} = \sigma(\boldsymbol{\alpha}_k)$ denotes a path-length-specific gate and $\boldsymbol{\alpha}_k \in \mathbb{R}^{d_{CC}}$ is a learnable parameter vector shared across all connected components at path length k. This gating mechanism controls the extent to which the initial component representation is retained when no valid fusion-to-component paths are available. The gating parameters are optimized jointly with all model parameters via end-to-end backpropagation from the comorbidity prediction objective (Equation (25)). While this gating design preserves model stability and ensures that all components within disease subgraphs are retained, it may slightly reduce the amount of path-specific propagation information available for a small subset of disease subgraphs.

Another limitation of the current study is that FuDiCo does not explicitly distinguish between different biological categories of fusion events, such as germline and somatic events, because such annotations are not available in the utilized fusion gene dataset. However, our method does not assume that fusion events directly drive or causally determine disease comorbidity relationships across all disease contexts. Rather, the proposed framework models how fusion-initiated influence propagation patterns within the PPI network may capture molecular relationships associated with disease comorbidity.

In addition, the dataset is partitioned at the disease-pair level to train and evaluate FuDiCo on the disease comorbidity prediction task, allowing individual diseases to appear in multiple dataset splits through different disease-pair combinations. Although this disease-pair partitioning strategy is commonly adopted in disease relationship prediction studies, disease-level partitioning may provide a more stringent assessment of model generalization to previously unseen diseases. Furthermore, the selected negative sampling strategy may influence the distribution of positive and negative disease pairs used to assess model performance. Future work may investigate the effect of alternative negative sampling configurations on disease comorbidity prediction performance, particularly under different ratios of positive and negative disease pairs and comorbidity label definitions. For example, varying RR thresholds for defining comorbidity labels could alter the balance between positive and negative disease pairs and provide additional perspectives on model robustness.

5. Conclusions

In this study, we introduced FuDiCo, a novel framework for disease comorbidity prediction based on gene fusion-initiated influence propagation over the PPI network.

Unlike prior methods that rely on the topology or global proximity of annotated disease genes to learn disease representations, FuDiCo integrates gene fusion events, network diffusion, and disease subgraphs to model fusion-initiated influence propagation along sampled paths in the PPI network toward disease subgraphs. The resulting propagation paths are then encoded using a fusion influence-aware GRU encoder to learn disease-subgraph representations that capture the underlying propagation patterns associated with disease comorbidity. Experimental results on a benchmark disease comorbidity dataset demonstrate that FuDiCo achieves robust and consistent state-of-the-art performance for disease comorbidity prediction, outperforming recent methods including DisSubFormer, FDS-CAP, and BSE across multiple evaluation metrics. Beyond predictive performance, these results suggest that gene fusions open new avenues for investigating disease relationships through molecular interaction networks. Future work may further extend this direction by examining how fusion-initiated influence varies across different biological categories of fusion events, including germline and somatic fusion events. Such analyses may reveal category-specific patterns of influence propagation and support future studies of network-based disease comorbidity modeling.

Author Contributions: Conceptualization, A.A. and L.L.; Methodology, A.A.; Software, A.A.; Validation, A.A.; Formal analysis, L.L.; Investigation, A.A. and L.L.; Resources, L.L.; Data curation, A.A.; Writing—original draft, A.A.; Writing—review & editing, A.A. and L.L.; Visualization, A.A.; Supervision, L.L.; Project administration, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw datasets used in this study are publicly available from previously published sources, including the PPI network, disease–gene associations, and disease-pair datasets reported in [5], as well as the fusion gene dataset from FusionGDB 2.0 [24]. The raw datasets, processed datasets generated in this study, and the FuDiCo source code are publicly available in the following repository: https://github.com/Ashwag-ta/FuDiCo.git, accessed on 3 June 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average precision
AUROC	Area under the receiver operating characteristic curve
BCE	Binary cross-entropy
CDS	Coding sequence
ESM-2	Evolutionary Scale Modeling 2
GRU	Gated recurrent unit
MLP	Multilayer perceptron
ORF	Open reading frame
PPI	Protein–protein interaction
PR	Precision–recall
ReLU	Rectified linear unit
ROC	Receiver operating characteristic
RR	Relative risk

Appendix A

This appendix presents the detailed algorithm of FuDiCo.

Algorithm A1. FuDiCo: Gene Fusion-Initiated Path Propagation for Disease Comorbidity Prediction. Forward (fwd) and backward (bwd) scores quantify the strength of fusion-initiated influence propagation along fusion-to-component paths. GRU denotes a fusion influence-aware gated recurrent unit for propagation path encoding. A multilayer perceptron (MLP) is used to predict the probability of comorbidity for a pair of diseases.

Input: PPI graph $G_{\text{PPI}} = (V, E)$; Node embeddings $\{\mathbf{f}_v | v \in V\}$; Disease subgraph set $\mathcal{D} = \{D_1, \cdots, D_m\}$, where each disease subgraph D_i consists of connected components $D_i^{(\text{CC})} = \left\{CC_{(i,1)}, \cdots, CC_{(i,n_i)}\right\}$ and fusion-associated gene set FG_i; Maximum path length K; Path budget B; Disease pair set $\left\{\left(D_{i_n}, D_{j_n}\right)\right\}_{n=1}^{N}$.
Output: Disease subgraph representations $\left\{\mathbf{h}_{D_i}\right\}_{i=1}^{m}$ and predicted comorbidity probabilities $\{\hat{y}_n\}_{n=1}^{N}$ for disease pairs.
Model Parameters: Learnable propagation parameter γ; Parameters of the fusion influence-aware GRU θ_{GRU}; Parameters of the path-length-specific gate $\mathbf{g}^{(k)}$; MLP parameters $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$.

```
for disease subgraph i = 1, ··· , m do
  for path length k = 1, ··· , K do
    for connected component j = 1, ··· , n_i do
      h^{init,k}_{CC_(i,j)} = AGG_node({ f_v | v ∈ CC_(i,j) })
      P^(k)_(i,j) ← SamplePaths(FG_i, CC_(i,j), k, B),  |P^(k)_(i,j)| ≤ B        // See Section 2.2.4
      for each path p = (v_0, ··· , v_k) ∈ P^(k)_(i,j) do
        for position t = 0, ··· , k do
          S_t ⟵ PositionInfluence(s^fwd_t, s^bwd_t)                              // See Section 2.2.5
          S̄_t = (1 − γ) S_t + γ S̄_{t−1}
        end
        h_p ⟵ GRU({f_{v_t}}^k_{t=0}, {S_t}^k_{t=0}, {S̄_t}^k_{t=0})               // See Section 2.2.6
        S̄_p ⟵ PathInfluence(S̄_k)
      end
      {α_p}_{p∈P^(k)_(i,j)} = softmax({S̄_p}_{p∈P^(k)_(i,j)})
      h^k_{CC_(i,j)} ← { Σ_{p∈P^(k)_(i,j)} α_p h_p,   P^(k)_(i,j) ≠ ∅
                       { g^(k) ⊙ h^{init,k}_{CC_(i,j)},   otherwise
    end
    h^k_{D_i} = AGG_CC(h^k_{CC_(i,1)}, ··· , h^k_{CC_(i,n_i)})                   // Aggregate subgraph components
  end
  h_{D_i} = READOUT(h^1_{D_i}, ··· , h^K_{D_i})                                  // Aggregate across path lengths
end
for disease pair n = 1, ··· , N do
  ŷ_n = σ(W_2 ReLU(W_1 concat(h_{D_{i_n}}, h_{D_{j_n}}) + b_1) + b_2)            // Comorbidity probability
end
```

References

1. Feinstein, A.R. The Pre-Therapeutic Classification of Co-Morbidity in Chronic Disease. *J. Chronic Dis.* **1970**, *23*, 455–468. [CrossRef]
2. Watson, K.B.; Wiltz, J.L.; Nhim, K.; Kaufmann, R.B.; Thomas, C.W.; Greenlund, K.J. Trends in Multiple Chronic Conditions Among US Adults, by Life Stage, Behavioral Risk Factor Surveillance System, 2013–2023. *Prev. Chronic Dis.* **2025**, *22*, 240539. [CrossRef]
3. Goh, K.-I.; Cusick, M.E.; Valle, D.; Childs, B.; Vidal, M.; Barabási, A.-L. The Human Disease Network. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 8685–8690. [CrossRef] [PubMed]

4. Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network Medicine: A Network-Based Approach to Human Disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef]
5. Menche, J.; Sharma, A.; Kitsak, M.; Ghiassian, S.D.; Vidal, M.; Loscalzo, J.; Barabási, A.-L. Uncovering Disease-Disease Relationships through the Incomplete Interactome. *Science* **2015**, *347*, 1257601. [CrossRef] [PubMed]
6. Akram, P.; Liao, L. Prediction of Comorbid Diseases Using Weighted Geometric Embedding of Human Interactome. *BMC Med. Genom.* **2019**, *12*, 161. [CrossRef]
7. Qin, X.; Liao, L. Improving Disease Comorbidity Prediction with Biologically Supervised Graph Embedding. In *Computational Advances in Bio and Medical Sciences*; Bansal, M.S., Chen, W., Khudyakov, Y., Măndoiu, I.I., Moussa, M.R., Patterson, M., Rajasekaran, S., Skums, P., Thankachan, S.V., Zelikovsky, A., Eds.; Springer: Cham, Switzerland, 2025; Volume 14548, pp. 178–190. [CrossRef]
8. Altayyar, A.; Liao, L. Link Prediction in Disease–Disease Interactions Network Using a Hybrid Deep Learning Model. In *Computational Advances in Bio and Medical Sciences*; Alser, M., Bansal, M.S., Khudyakov, Y., Mangul, S., Mandoiu, I.I., Moussa, M.R., Patterson, M., Rajasekaran, S., Skums, P., Yooseph, S., et al., Eds.; Springer: Cham, Switzerland, 2026; Volume 15599, pp. 90–102. [CrossRef]
9. Altayyar, A.; Liao, L. FDS-CAP: Modeling Fragmented Disease Subgraphs with Component-Level Attention for Comorbidity Prediction. *J. Comput. Biol.* **2026**, *33*, 611–623. [CrossRef]
10. Altayyar, A.; Liao, L. DisSubFormer: A Subgraph Transformer Model for Disease Subgraph Representation and Comorbidity Prediction. *Comput. Biol. Chem.* **2026**, *122*, 108935. [CrossRef]
11. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A.; Kinzler, K.W. Cancer Genome Landscapes. *Science* **2013**, *339*, 1546–1558. [CrossRef]
12. Albert, F.W.; Kruglyak, L. The Role of Regulatory Variation in Complex Traits and Disease. *Nat. Rev. Genet.* **2015**, *16*, 197–212. [CrossRef] [PubMed]
13. Mitelman, F.; Johansson, B.; Mertens, F. The Impact of Translocations and Gene Fusions on Cancer Causation. *Nat. Rev. Cancer* **2007**, *7*, 233–245. [CrossRef]
14. Kaessmann, H. Origins, Evolution, and Phenotypic Impact of New Genes. *Genome Res.* **2010**, *20*, 1313–1326. [CrossRef]
15. Latysheva, N.S.; Babu, M.M. Discovering and Understanding Oncogenic Gene Fusions through Data Intensive Computational Approaches. *Nucleic Acids Res.* **2016**, *44*, 4487–4503. [CrossRef] [PubMed]
16. Mertens, F.; Johansson, B.; Fioretos, T.; Mitelman, F. The Emerging Complexity of Gene Fusions in Cancer. *Nat. Rev. Cancer* **2015**, *15*, 371–381. [CrossRef]
17. Kumar, H.; Tang, L.-Y.; Yang, C.; Kim, P. FusionPDB: A Knowledgebase of Human Fusion Proteins. *Nucleic Acids Res.* **2024**, *52*, D1289–D1304. [CrossRef] [PubMed]
18. Lupski, J.R.; Stankiewicz, P. Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes. *PLoS Genet.* **2005**, *1*, e49. [CrossRef] [PubMed]
19. Oliver, G.R.; Tang, X.; Schultz-Rogers, L.E.; Vidal-Folch, N.; Jenkinson, W.G.; Schwab, T.L.; Gaonkar, K.; Cousin, M.A.; Nair, A.; Basu, S.; et al. A Tailored Approach to Fusion Transcript Identification Increases Diagnosis of Rare Inherited Disease. *PLoS ONE* **2019**, *14*, e0223337. [CrossRef]
20. The UniProt Consortium. UniProt Release 2025_03. Available online: https://www.uniprot.org/release-notes/2025-06-18-release (accessed on 25 June 2025).
21. Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders. *Nucleic Acids Res.* **2005**, *33*, D514–D517. [CrossRef]
22. Mottaz, A.; Yip, Y.L.; Ruch, P.; Veuthey, A.-L. Mapping Proteins to Disease Terminologies: From UniProt to MeSH. *BMC Bioinform.* **2008**, *9*, S3. [CrossRef]
23. Ramos, E.M.; Hoffman, D.; Junkins, H.A.; Maglott, D.; Phan, L.; Sherry, S.T.; Feolo, M.; Hindorff, L.A. Phenotype–Genotype Integrator (PheGenI): Synthesizing Genome-Wide Association Study (GWAS) Data with Existing Genomic Resources. *Eur. J. Hum. Genet.* **2014**, *22*, 144–147. [CrossRef]
24. Kim, P.; Tan, H.; Liu, J.; Lee, H.; Jung, H.; Kumar, H.; Zhou, X. FusionGDB 2.0: Fusion Gene Annotation Updates Aided by Deep Learning. *Nucleic Acids Res.* **2022**, *50*, D1221–D1230. [CrossRef] [PubMed]
25. Balamurali, D.; Gorohovski, A.; Detroja, R.; Palande, V.; Raviv-Shay, D.; Frenkel-Morgenstern, M. ChiTaRS 5.0: The Comprehensive Database of Chimeric Transcripts Matched with Druggable Fusions and 3D Chromatin Maps. *Nucleic Acids Res.* **2019**, *48*, gkz1025. [CrossRef]
26. Jang, Y.E.; Jang, I.; Kim, S.; Cho, S.; Kim, D.; Kim, K.; Kim, J.; Hwang, J.; Kim, S.; Kim, J.; et al. ChimerDB 4.0: An Updated and Expanded Database of Fusion Genes. *Nucleic Acids Res.* **2019**, *48*, gkz1013. [CrossRef] [PubMed]
27. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, 1123–1130. [CrossRef]
28. Godsil, C.; Royle, G.F. *Algebraic Graph Theory*; Springer: New York, NY, USA, 2001.

29. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1724–1734. [CrossRef]
30. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**. [CrossRef]
31. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 38–45. [CrossRef]
32. Fey, M.; Lenssen, J.E. Fast Graph Representation Learning with PyTorch Geometric. *arXiv* **2019**. [CrossRef]
33. Falcon, W.; PyTorch Lightning Team. *PyTorch Lightning*, Version 2.4.0; Zenodo: Geneve, Switzerland, 2024.
34. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
35. Smith, L.N. A Disciplined Approach to Neural Network Hyper-Parameters: Part 1—Learning Rate, Batch Size, Momentum, and Weight Decay. *arXiv* **2018**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.