یادگیری بازنمایی عناصر زیستی-پزشکی با استفاده از گراف‌های ناهمگن

شادی زینالی مغانجوقی[1]، کارشناسی ارشد، اسماعیل نورانی[1]، دانشیار، عسگر علی بویر[1]، دانشیار

[1]دانشکده فناوری اطلاعات و مهندسی کامپیوتر- دانشگاه شهید مدنی آذربایجان- تبریز- ایران- ac.nourani@azaruniv.ac.ir

چکیده: گراف‌های ناهمگن چارچوبی قدرتمند برای مدل‌سازی و تحلیل مسائل پیچیده دنیای واقعی فراهم می‌کنند. این گراف‌ها با نمایش انواع مختلف گره‌ها و روابط میان آن‌ها، امکان ادغام و تفسیر داده‌های متنوع را مهیا می‌سازند. یادگیری بازنمایی مؤثر از اجزای گراف‌های ناهمگن، یکی از چالش‌های اساسی در توسعه الگوریتم‌های یادگیری ماشین و یادگیری عمیق است، چرا که این بازنمایی‌ها نقش کلیدی در بهبود دقت پیش‌بینی و کشف الگوهای پنهان ایفا می‌کنند. روش‌های سنتی و گراف‌های همگن با توجه به محدودیت در نمایش تنوع داده‌های زیستی در حوزه بیوانفورماتیک عملکرد مطلوبی ندارند. در مقابل، گراف‌های ناهمگن با بهره‌گیری از اطلاعات ساختاری پیچیده، قادر به مدل‌سازی مؤثرتر روابط زیستی هستند. در این مقاله، رویکرد نوین BioGraph2vec برای یادگیری بازنمایی عناصر زیستی-پزشکی با استفاده از گراف‌های ناهمگن ارائه شده است. این روش با ترکیب داده‌های توالی پروتئین‌ها و اطلاعات تعاملات زیستی، گراف‌هایی شامل پروتئین‌های میزبان و پاتوژن و روابط میان آن‌ها ایجاد می‌کند. برای استخراج بازنمایی دقیق از گره‌ها، از مکانیزم‌های توجه و انتقال پیام استفاده می‌شود تا ویژگی‌های مهم هر گره در بستر شبکه شناسایی گردد. در ادامه، بازنمایی‌های به‌دست‌آمده به مدل‌های یادگیری ماشین داده می‌شوند تا تعاملات احتمالی بین پروتئین‌ها پیش‌بینی شوند. ارزیابی روش پیشنهادی بر روی مجموعه داده‌های متنوع نشان می‌دهد که ادغام داده‌های زیستی در قالب گراف‌های ناهمگن، همراه با تکنیک‌های پیشرفته یادگیری بازنمایی، می‌تواند به بهبود تحلیل داده‌های بیوانفورماتیکی و شناسایی الگوهای پیچیده در سیستم‌های زیستی منجر شود.

واژه‌های کلیدی: یادگیری بازنمایی، گراف‌های همگن، عناصر زیستی-پزشکی، یادگیری ماشین

* نویسنده مسئول، ایمیل نویسنده

Representation Learning for Biomedical Entities using Heterogeneous Graphs

Shadi Zeynali Moghanjoughi [1], M.Sc., Esmaeil Nourani [1], Associate Professor, Asgar Alibouyer [1], Associate Professor

[1]Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran, ac.nourani@azaruniv.ac.ir

Abstract: Heterogeneous graphs provide a powerful framework for modeling and analyzing complex real-world problems. These graphs, by representing different types of nodes and their relationships, enable the integration and interpretation of diverse data. Learning effective representations of the components of heterogeneous graphs is one of the fundamental challenges in the development of machine learning and deep learning algorithms, since these representations play a key role in improving prediction accuracy and discovering hidden patterns. In the field of bioinformatics, traditional methods and homogeneous graphs do not perform well due to their limitations in representing the diversity of biological data. In contrast, heterogeneous graphs, by leveraging complex structural information, are capable of more effectively modeling biological relationships. In this paper, we present a novel approach, BioGraph2vec, for learning representations of biomedical entities using heterogeneous graphs. This method combines protein sequence data and biological interaction information to create graphs that include host and pathogen proteins and the relationships between them. To extract precise node representations, attention mechanisms and message passing techniques are utilized to identify the important features of each node within the network. The obtained representations are then provided to machine learning models to predict potential interactions between proteins. Evaluation of this method on diverse datasets shows that integrating biological data in the form of heterogeneous graphs, together with advanced representation learning techniques, can lead to improved bioinformatics data analysis and better identification of complex patterns in biological systems.

Keywords: *Representation Learning; Heterogeneous Graphs; Biomedical Entities; Machine Learning*

*** Corresponding author**

1. مقدمه

گراف‌های ناهمگن[1] یکی از ابزارهای اساسی برای مدل‌سازی و تحلیل سیستم‌های پیچیده محسوب می‌شوند که در آن‌ها موجودیت‌های مختلف با ویژگی‌ها و ماهیت‌های متفاوت، از طریق انواع متنوعی از روابط با یکدیگر تعامل دارند. این گراف‌ها در حوزه‌های گوناگونی از جمله شبکه‌های علمی، شبکه‌های اجتماعی و گراف‌های اقتصادی به‌کار گرفته می‌شوند [1]. تنوع در انواع گره‌ها و روابط، مدل‌سازی و تحلیل چنین گراف‌هایی را به چالشی اساسی در یادگیری ماشین و و علوم داده تبدیل کرده است.

یکی از رویکردهای کلیدی برای بهره‌برداری از ساختارهای پیچیده گراف، یادگیری بازنمایی است که هدف آن استخراج نمایش‌های فشرده و غنی از گره‌ها و یال‌ها با حفظ ساختار و ویژگی‌های آن‌هاست. این مسئله در گراف‌های ناهمگن اهمیت بیشتری دارد، چراکه تنوع زیاد موجودیت‌ها و روابط در آن‌ها باعث می‌شود روش‌های سنتی یادگیری ماشین در استخراج اطلاعات معنادار ناتوان باشند. افزون بر این، اغلب روش‌های فعلی تنها به یک دامنه خاص محدودند و قادر به مدل‌سازی ارتباطات پیچیده میان انواع مختلف موجودیت‌ها نیستند، در حالی که گراف‌های ناهمگن با انعطاف‌پذیری بالا نقش مؤثری در این زمینه ایفا می‌کنند. بنابراین، طراحی روش‌های مؤثر برای یادگیری بازنمایی در این نوع گراف‌ها، گامی ضروری در تحلیل داده‌های پیچیده دنیای واقعی به شمار می‌رود.

در سال‌های اخیر، داده‌های زیستی-پزشکی به‌عنوان منابعی ارزشمند در تحلیل‌های مبتنی بر علوم داده و هوش مصنوعی شناخته شده‌اند و مطالعات متعددی برای بهره‌گیری از آن‌ها در حل مسائل مربوط به سلامت انسان انجام شده است [2]. بیشتر پژوهش‌های گذشته بر گراف‌های همگن متمرکز بوده‌اند، که تنها یک نوع گره و یال را در نظر می‌گیرند؛ اما داده‌های زیست‌پزشکی از تنوع و ساختارهای پیچیده‌تری برخوردارند که با گراف‌های ناهمگن بهتر مدل‌سازی می‌شوند. با توجه به این پیچیدگی، یادگیری بازنمایی مؤثر از این نوع گراف‌ها، هرچند دشوار، امری حیاتی است. چنین بازنمایی‌هایی می‌توانند در مسائل کاربردی مانند پیش‌بینی تعامل میان پروتئین‌های میزبان و پاتوژن یا طراحی سامانه‌های پیشنهاددهنده دارویی نقش مهمی ایفا کنند.

تاکنون روش‌های مختلفی برای یادگیری بازنمایی از گراف‌های همگن ارائه شده‌اند که در آن‌ها تنها یک نوع گره در ساختار گراف وجود دارد. یکی از این روش‌ها، الگوریتم node2vec [3] است که با نگاشت گره‌ها به یک فضای با ابعاد پایین‌تر، اطلاعات معناداری از ساختار همسایگی گره‌ها را حفظ می‌کند. روش DeepWalk [4] نیز با بهره‌گیری از پیاده‌روی‌های تصادفی، بازنمایی برداری پیوسته‌ای از روابط بین گره‌ها ارائه می‌دهد که مبتنی بر اطلاعات محلی شبکه است.

با گسترش یادگیری عمیق، رویکردهای جدیدی تحت عنوان شبکه‌های عصبی گرافی معرفی شده‌اند [5]–[7] که قادرند ضمن تجمیع اطلاعات ساختاری، ویژگی‌های گره‌ها را از طریق انتشار پیام میان گره‌های متصل به‌دست آورند. این روش‌ها به دلیل انعطاف‌پذیری بالا، در بسیاری از کاربردهای گرافی مورد توجه قرار گرفته‌اند.

در همین راستا، یادگیری بازنمایی در گراف‌های ناهمگن نیز به‌طور فزاینده‌ای مورد توجه پژوهشگران قرار گرفته است. بخش زیادی از این الگوریتم‌ها بر پایه مسیر طراحی شده‌اند؛ به‌گونه‌ای

[1] Heterogeneous Graphs

که ارتباطات میان گره‌ها را به صورت توالی‌هایی مدل می‌کنند و از پیاده‌روی‌های تصادفی برای استخراج این توالی‌ها بهره می‌برند. این توالی‌ها سپس به ماتریس‌هایی نظیر ماتریس مجاورت یا شباهت تبدیل شده و به عنوان ورودی برای مدل‌های یادگیری ماشین مورد استفاده قرار می‌گیرند [8]. افزون بر این، برخی رویکردها نیز مبتنی بر شبکه‌های عصبی گرافی توسعه یافته‌اند تا بتوانند ویژگی‌های پیچیده‌تری از ساختار گراف را یاد بگیرند.

یکی از رویکردهای مطرح در زمینه یادگیری بازنمایی برای گراف‌های ناهمگن، الگوریتم Methapath2vec است. این روش با بهره‌گیری از پیاده‌روی‌های تصادفی هدایت‌شده [4] و ساختاری موسوم به فرامسیر[۲] [9]، امکان مدل‌سازی مؤثر ارتباطات در گراف‌های ناهمگن را فراهم می‌سازد. فرامسیر، به توالی از گره‌ها گفته می‌شود که از طریق یال‌ها به‌صورت متوالی به هم متصل‌اند و می‌توانند متعلق به انواع مختلفی از موجودیت‌ها باشند.

در این الگوریتم، برای یادگیری بازنمایی برداری گره‌ها، از مدل skip-gram [10] استفاده شده است تا همسایگی ناهمگن هر گره به‌درستی مدل‌سازی شود. مزیت اصلی این روش، حفظ هم‌زمان ساختار شبکه و معنای ارتباطات میان انواع مختلف گره‌هاست. علاوه بر آن، Methapath2vec قادر است بدون نیاز به تعریف دستی قواعد پیچیده، به‌صورت خودکار روابط معنایی درونی میان موجودیت‌ها را در ساختار ناهمگن گراف شناسایی و بازنمایی کند.

روش دیگر برای یادگیری بازنمایی گراف ناهمگن MethaGraph2Vec است [11] که از فراگراف[۳] [12] برای تولید پیاده‌روی تصادفی استفاده می‌کند. یک فراگراف شامل مسیرهای متعددی بین گره‌ها است که هرکدام یک نوع رابطه را توصیف می‌کنند. اگرچه این روش در شناسایی روابط معنایی و استخراج اطلاعات غنی میان گره‌ها عملکرد قابل قبولی دارد، اما در بازیابی دقیق ساختار کلی گراف از کارایی مطلوبی برخوردار نیست.

از روش‌های مبتنی یادگیری عمیق نیز برای استخراج بازنمایی‌های گره‌های یک گراف ناهمگن استفاده می‌شود که برای نمونه می‌توان به روش CARL اشاره کرد [13]. این روش نه تنها اطلاعات توپولوژیکی گراف را دریافت می‌کند، بلکه محتوای معنایی بدون ساختار را نیز در نظر می‌گیرد. در این روش، مدلی بر پایه skip-gram طراحی شده است که هدف آن حفظ نزدیکی ساختاری میان گره‌ها و روابط با انواع مختلف است. پس از آن، دو مکانیزم مؤثر مبتنی بر رمزنگاری معنایی عمیق به کار گرفته می‌شود تا بتوان محتوای بدون ساختار برخی از گره‌های گراف را با هدف استخراج روابط معنایی مناسب ترکیب کرد.

با در نظر گرفتن اینکه فرامسیرهای گوناگون می‌توانند مفاهیم متفاوتی را منتقل کنند، روش HAHE در پاسخ به محدودیت‌های روش‌هایی که وزن یکسانی برای همه گره‌ها در نظر می‌گیرند، معرفی شده است [14]. این روش از ساختاری سلسله‌مراتبی بهره می‌برد که هدف آن استخراج اهمیت گره‌ها و فرامسیرها است. معماری HAHE شامل سه لایه است: لایه نخست اهمیت هر گره را در یک فرامسیر مشخص می‌کند و نقش آن را در مسیر مربوطه نشان می‌دهد. لایه دوم مسئول شناسایی فرامسیرهای تأثیرگذارتر در میان مجموعه فرامسیرهاست. در نهایت، لایه سوم وظیفه محاسبه اطلاعات معنایی مرتبط با گره‌ها را بر عهده دارد.

یکی دیگر از رویکرد های مؤثر در این زمینه استفاده از روش‌های بدون نظارت است که بتوانند ساختار و ویژگی‌های شبکه را بدون نیاز به برچسب‌های دستی حفظ کنند. در این

---

۲ MetaPath

۳ MetaGraph

راســـتا، مدل LHGI با بهره‌گیری از نمونه‌گیری زیرگراف های هدایت شده توسط فرامسیرها توسعه یافته است [15]،این روش با انتخاب زیرگراف‌های مرتبط با مســـیرهای خاص، می تواند شـــبکه را فشـــرده مدل کند، بدون آن‌که اطلاعات ســـاختاری و معنایی گراف از بین برود. استفاده از فرامسیرها به مدل این امکان را می دهد که ارتباطات پیچیده میان انواع مختلف گره ها را به‌صورت خودکار شناسایی و در بازنمایی‌ها بازتاب دهد.

مزیت اصلی LHGI، مقیاس‌پذیری بالا و عدم نیاز به داده‌های برچسب‌خورده است، که آن را برای تحلیل گراف‌های بزرگ و متنوع مناسب می‌سازد. با این حال، تعیین فرامسیرها از پیش نیازمند دانش دامنه و طراحی دقیق است و ممکن است برخی جزئیات محلی شبکه در نمونه‌گیری زیرگراف‌ها از دست برود.

در گراف های ناهمگن، گره ها می‌توانند به‌طور همزمان با همســـایه‌های هم‌نوع و غیرهم‌نوع خود تعامل داشـــته باشـــند، اما بســـیاری از روش‌های موجود تنها به نوع خاصـــی از همســـایگی توجه می‌کنند و قادر به مدل سازی تمام وابستگی‌ها نیستند. در پاسخ به این محدودیت، مدل HeMGNN توسعه یافته است که با استفاده از شبکه عصبی گرافی ترکیبی، اطلاعات هم سایگی هم‌نوع و غیرهم‌نوع را به‌طور همزمان در بازنمایی‌ها لحاظ می‌کند [16]. این روش با تجمیع مؤثر پیام‌ها از گره‌های مختلف، قادر است بازنمایی‌هایی تولید کند که هم ساختار و هم روابط پیچیده میان انواع مختلف موجودیت‌ها را حفظ می‌کند.

از نقاط قوت HeMGNN می‌توان به توانایی مدل‌سازی دقیق وابستگی‌های چندگانه و ارائه embeddingهای باکیفیت برای گراف‌های متنوع اشاره کرد. با این حال، پیچیدگی محاسباتی این مدل نسبت به شبکه‌های عصبی گرافی ساده بیشتر است و نیازمند تنظیمات دقیق در ترکیب اطلاعات همسایگی است.

روش‌های موجود دارای چند مشکل هستند:

- بیش‌تر آن‌ها شامل طراحی فرامسیر برای گراف هستند، و اطلاعات دامنه گراف باید به صورت دستی وارد شود.
- در این دسته از روش‌ها، این فرض وجود دارد که گره‌ها و یال‌ها از نظر ویژگی‌ها و توزیع آماری در یک فضای مشترک قرار دارند.

جدول (1) خلاصه‌ای از روش‌های پیشین و جدید یادگیری بازنمایی در گراف‌های ناهمگن همراه با متدولوژی، مزایا و معایب آن‌ها ارائه می‌دهد. علاوه بر این تحقیقات زیادی در زمینه داده‌های زیستی-پزشکی مانند پیش‌بینی ارتباطات بین پروتئین‌های میزبان و پاتوژن انجام شده است که از اطلاعات مهم شبکه استفاده نکرده‌اند.

با در نظر گرفتن این محدودیت‌ها، هدف اصلی این پژوهش، توسعه یک روش یادگیری بازنمایی برای عناصر زیستی-پزشکی بر پایه گراف‌های ناهمگن است؛ روشی که بتواند بازنمایی‌هایی متناسب با نوع گره و یال ارائه دهد و وابستگی به طراحی دستی فرامسیرها را از میان بردارد بازنمایی‌های حاصل باید قادر باشند به‌طور مؤثر ویژگی‌های عناصر را منعکس کرده و از کارایی و دقت بالایی برخوردار باشند. در این مقاله عناصر زیستی-پزشکی به عنوان یک گراف ناهمگن گردآوری شده‌اند و روشی معرفی شده است که بازنمایی عناصر گراف استخراج شود. در نهایت از بازنمایی‌های بدست آمده، جهت پیش‌بینی ارتباطات بین پروتئین‌های میزبان و پاتوژن استفاده شده است.

برای مقابله با ناهمگنی گراف، مکانیزم اهمیت وابسته به نوع گره‌ها و یال‌ها استفاده شده است. این مدل با تجزیه هر یال به

فراارتباط[٤] (گره مبدأ، یال بین دو گره، گره هدف) تعریف می‌شود. این رویکرد به مدل این امکان را می‌دهد که نمایش‌های خاص و منحصربه‌فردی برای انواع مختلف گره‌ها و یال‌ها ایجاد کند. علاوه بر این، گره‌های متصل از انواع مختلف با وجود فضای توزیعی متفاوت همچنان قادر به تعامل با یکدیگر خواهند بود و بدون توجه به تفاوت‌های توزیعی، می‌توانند پیام‌ها را منتقل کرده و اطلاعات را تجمیع کنند.

جهت ارزیابی مدل پیشنهادی، از مجموعه‌ای از معیارهای مختلف شامل دقت، ضریب همبستگی متیو[٥]، ناحیه زیر منحنی[٦] و F1 استفاده شده است.

جدول (1): مقایسه روش‌های بازنمایی گراف

روش/مدل	متدولوژی	مزایا	معایب
Node2vec [3]	نگاشت گره‌ها به فضای با ابعاد پایین‌تر با حفظ ساختار همسایگی؛ استفاده از پیاده‌روی تصادفی	ساده و سریع؛ حفظ اطلاعات محلی شبکه	مناسب برای گراف‌های همگن
Deepwalk [4]	پیاده‌روی تصادفی همراه با مدل skip-gram برای استخراج بازنمایی	حفظ روابط محلی گره‌ها؛ پیاده‌سازی آسان	محدود به اطلاعات محلی؛ کارایی پایین در گراف‌های ناهمگن بزرگ
Methapath2vec [4]	پیاده‌روی تصادفی هدایت‌شده توسط فرامسیرها همراه با skip-gram	حفظ ساختار و معنای ارتباطات میان انواع گره‌ها؛ شناسایی خودکار روابط معنایی	نیازمند تعریف فرامسیر؛ احتمال از بین رفتن اطلاعات شبکه‌های بزرگ به دلیل پیچیدگی بالا
MethaGraph2Vec [11]	استفاده از فراگراف برای تولید پیاده‌روی تصادفی	شناسایی روابط معنایی و استخراج اطلاعات غنی میان گره‌ها	ضعیف در بازیابی دقیق ساختار کلی گراف
CARL [13]	ترکیب اطلاعات توپولوژیکی و محتوای معنایی بدون ساختار؛ با مدل skip-gram	حفظ نزدیکی ساختاری و معنایی؛ مناسب گراف‌های با محتوای بدون ساختار	پیچیدگی بالای محاسباتی؛ نیاز به تنظیمات دقیق در رمزنگاری معنایی
HAHE [14]	ساختار سلسله‌مراتبی برای اهمیت‌دهی به گره‌ها و فرامسیرها؛ سه لایه تحلیل اهمیت و محاسبه اطلاعات معنایی	استخراج اهمیت گره‌ها و مسیرها؛ بهبود کیفیت بازنمایی تولید شده	پیچیدگی مدل بالا؛ نیازمند داده‌های با کیفیت برای آموزش
LHGI (2023) [15]	نمونه‌گیری زیرگراف هدایت‌شده توسط فرامسیر با یادگیری بازنمایی بدون نظارت	بدون نیاز به برچسب؛ حفظ معنا و ساختار شبکه؛ مقیاس‌پذیری بالا	نیاز به تعیین فرامسیرها از پیش؛ احتمال از بین رفتن جزئیات شبکه محلی
HeMGNN (2023)[[16]	ترکیب پیام همسایه‌های هم‌نوع و غیرهم‌نوع در بازنمایی	مدل‌سازی دقیق وابستگی‌های چندگانه؛ حفظ روابط پیچیده	پیچیدگی محاسباتی بالا؛ نیاز به تنظیمات دقیق ترکیب اطلاعات

٤ MetaRelation

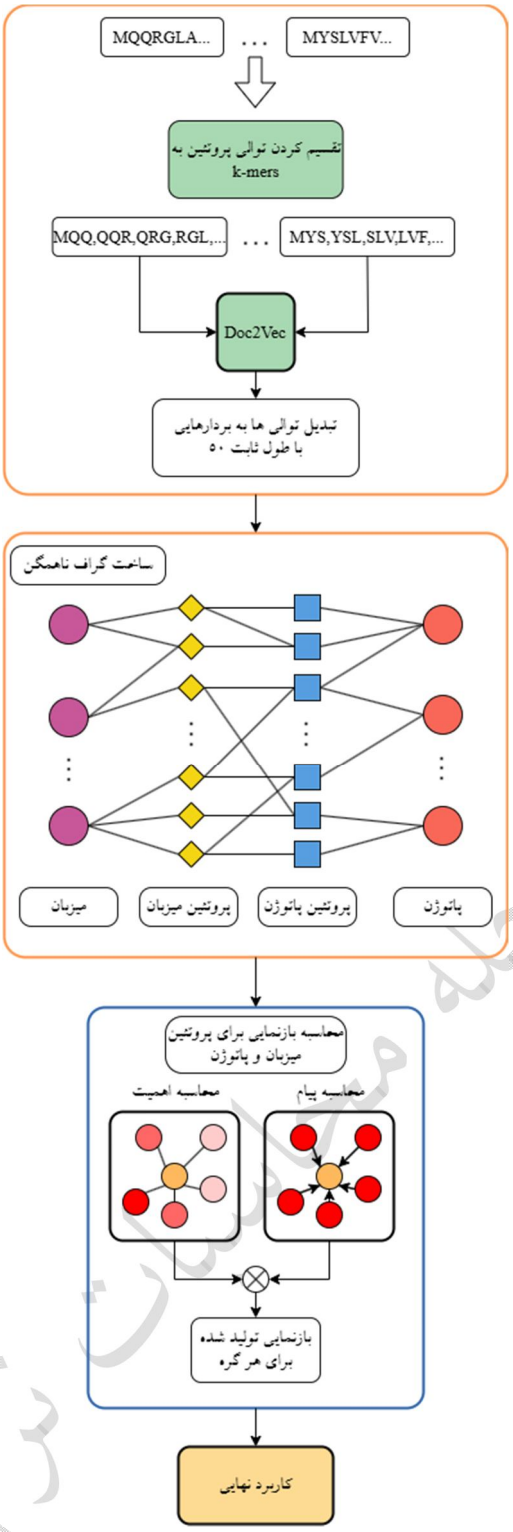
٥ Matthews Correlation Coefficient

٦ AUC (Area Under Curve)

پس از تولید بازنمایی‌ها برای گره‌ها، این بازنمایی‌ها به ماشین‌های یادگیری مختلفی مانند شبکه عصبی[۷]، جنگل تصادفی[۸] و ماشین بردار پشتیبان[۹] وارد شده‌اند. به منظور بررسی عملکرد مدل در مقیاس‌های مختلف، مجموعه داده‌ها به بخش‌های متنوعی تقسیم شده و مدل با میزبان ها و پاتوژن‌های مختلف آزمایش گردیده می‌شود. نتایج تجربی نشان می‌دهند که مدل پیشنهادی در پیش‌بینی تعاملات پروتئینی، عملکردی بهتر از روش‌های موجود داشته و قادر است اطلاعات ساختاری گراف ناهمگن را به‌طور مؤثری نمایش دهد. این مدل نه تنها در مقایسه با روش‌های ارائه شده قبلی نتایج بهتری ارائه کرده بلکه می‌تواند به‌طور مؤثری در تحلیل و پیش‌بینی تعاملات پروتئینی در داده‌های زیستی کاربرد داشته باشد.

2. روش پیشنهادی

در این مقاله همانطور که در شکل (1) نشان داده شده است، ارتباطات بین عناصر زیستی-پزشکی که می‌تواند شامل پروتئین‌های میزبان و پروتئین‌های پاتوژن، بیماری، دارو، میزبان و پاتوژن باشد، به‌صورت یک گراف ناهمگن مدل‌سازی شده‌اند. هدف از این مدل‌سازی، بهره‌گیری از اطلاعات غنی ساختار توالی پروتئین در کنار سایر ویژگی‌های مهم عناصر یاد شده برای بهبود پیش‌بینی ارتباطات بین پروتئین‌های میزبان و پاتوژن است. ازاین‌رو، روشی موردنیاز است که بتواند اطلاعات موجود در گراف را به‌طور مؤثر استخراج کند.



شکل (1): بلوک دیاگرام روش پیشنهادی

---

۷ Neural Network

۸ Random Forest

۹ Support vector machine

1,2. مفاهیم پایه

گراف ناهمگن به‌صورت یک گراف جهت‌دار تعریف می‌شود که به فرم $G = (V, E, A, R)$ نمایش داده می شود، به‌گونه‌ای که $v \in V$ نمایانگر گره‌ها و $e \in E$ یال‌ها هستند. در این ساختار، گره‌ها و یال‌ها می‌توانند از انواع گوناگون باشند و با نگاشت‌هایی به فرم $\tau(v) : V \to A$ و $\Phi(e) : E \to R$ مشخص می‌شوند، به‌طوری‌که $|A| + |R| > 2$.

برای هر یال $e = (s, t)$ که از گره مبدأ s به گره مبدأ t متصل است، یک فراارتباط به‌صورت سه‌تایی $\langle \tau(s), \Phi(e), \tau(t) \rangle$ تعریف می‌گردد. همچنین، یال معکوس از t به s با نماد $\Phi^{-1}(e)$ نمایش داده می‌شود. در مدل‌سازی گراف‌های ناهمگن در دنیای واقعی، فرض می‌شود که انواع مختلفی از ارتباطات میان گره‌های گوناگون ممکن است وجود داشته باشد.

در سال‌های اخیر شبکه‌های عصبی گراف در تحلیل داده‌های رابطه‌محور عملکرد موفقی از خود نشان داده اند [18]، [19]. به‌طور کلی می‌توان این شبکه‌ها را به عنوان ابزاری برای استفاده از ساختار گراف به منظور ساخت نمودار محاسباتی در فرآیند ارسال پیام در نظر گرفت [20]، که در آن اطلاعات محلی هر گره گردآوری شده و به بازنمایی فشرده‌ای تبدیل می‌شود. ساختار کلی شبکه‌های عصبی گراف را می‌توان به‌صورت زیر بیان کرد:

با فرض اینکه $H^l(t)$ نمایش گره هدف t در لایه lام شبکه است، رابطه (1) نحوه به‌دست آوردن نمایش از لایه $l-1$ام به لایه lام به را بیان می‌کند.

$$H^l(t) \leftarrow aggregate\big(extract(H^{l-1}(s); H^{l-1}(t), e)\big) \quad (1)$$

به ازای هر $s \in N(t), e \in E(s,t)$ به‌طوری‌که $N(t)$ تمام گره‌های در ارتباط با گره هدف و $E(s,t)$ تمام یال‌های از گره مبدأ s به گره هدف t را شامل می‌شود.

در این شبکه‌ها، دو عملگر کلیدی وجود دارند: $aggregate$ و $extract$. وظیفه‌ی $extract$ ، جمع‌آوری اطلاعات از گره‌های همسایه است، در حالی‌که $aggregate$ داده‌های دریافتی از این گره‌ها را با استفاده از عملگرهایی مانند میانگین یا جمع، ترکیب و خلاصه‌سازی می‌کند. البته در این مرحله می‌توان از توابع ادغام پیشرفته‌تر و روش‌های نرمال‌سازی دقیق‌تری نیز بهره گرفت.

در شبکه‌های عصبی گراف مبتنی بر مکانیزم اهمیت نیز برای به‌دست آوردن بازنمایی از دو مفهوم اهمیت و پیام استفاده می‌شود. به‌طور کلی روش‌های مبتنی بر مکانیزم اهمیت توسط رابطه (۲) تعریف می‌شوند.

$$H^l(t) \leftarrow Aggregate\big(Attention(s,t).Message(s)\big) \quad (2)$$

این روش بر پایه سه عملگر اصلی عمل می‌کند که وظایف هر یک به‌صورت زیر تعریف می‌شود:

- $Attention$ میزان اهمیت و نقش گره‌های مبدأ را در ارتباطات مشخص می‌سازد.
- $Message$ وظیفه استخراج پیام‌ها یا همان اطلاعات کلیدی گره‌های مبدأ را بر عهده دارد.
- $aggregate$ نیز با در نظر گرفتن وزن‌های تعیین‌شده توسط $Attention$ ، پیام‌های دریافتی از همسایگان را ترکیب کرده و برای گره هدف تجمیع می‌کند

برای مثال در ساده‌ترین مکانیزم اهمیت در شبکه عصبی گراف [23] هر کدام از عملگرها توسط روابط (3)، (4) و (5) تعریف شده‌اند.

$$Attention_{GAT}(s,t) = \text{softmax}\left(\vec{a}\left(WH^{l-1}[t] \,||\, WH^{l-1}[s]\right)\right) \quad (3)$$

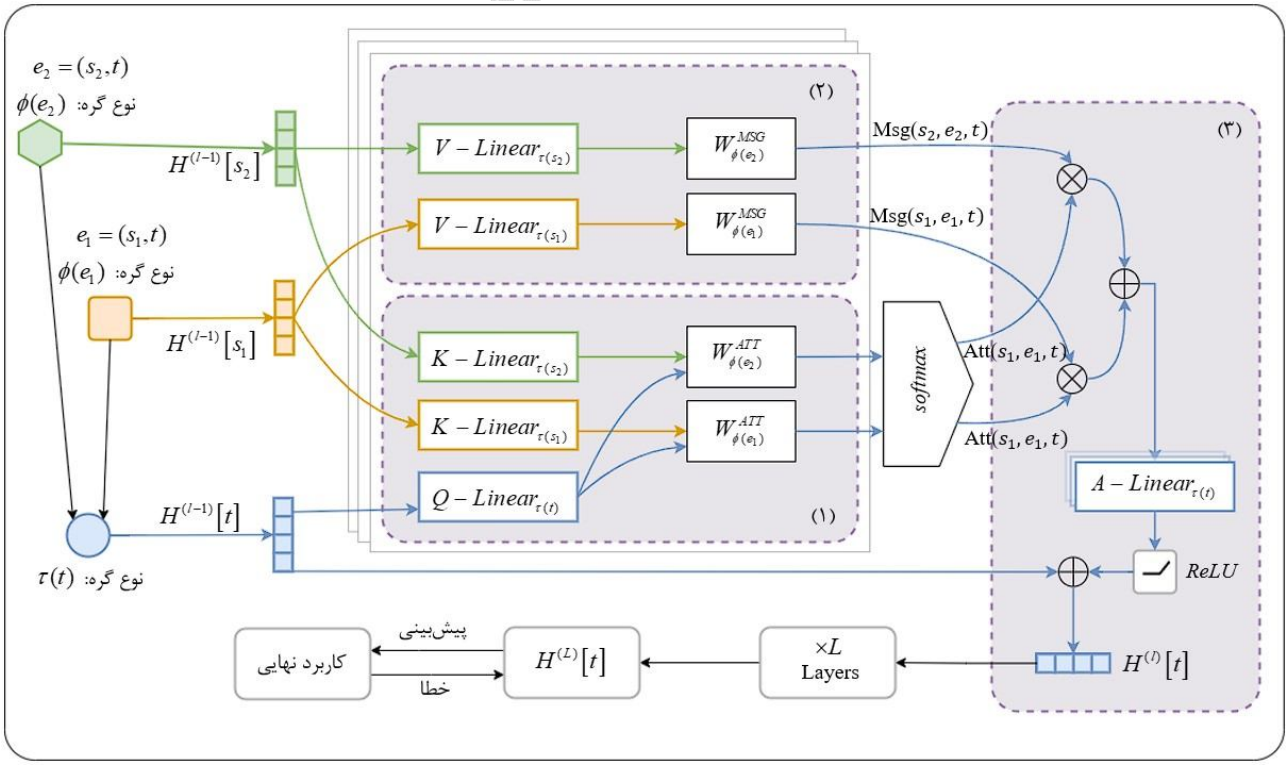
$$Message_{GAT}(s) = WH^{l-1}[s] \quad (4)$$

$$Aggregate_{GAT}(.) = \sigma(\text{Mean}(.)) \quad (5)$$

در شبکه عصبی گراف [23] از یک روش جمع ساده برای محاسبه اهمیت استفاده می‌شود. همچنین از یک وزن یکسان برای محاسبه پیام بهره برده می‌شود. در این حالت فرض می‌شود که گره‌های مبدأ و مقصد توزیع ویژگی یکسانی دارند. اما این فرض برای گراف‌های ناهمگن ممکن است نادرست باشد. با در نظر گرفتن این محدودیت، اهمیت برای هر فرارتباط $\langle \tau(s), \Phi(e), \tau(t) \rangle$ محاسبه می‌شود.

2,2. ساختار مدل پیشنهادی

در این بخش، هر یک از مراحل روش پیشنهادی که در شکل (1) آورده شده، به‌صورت مستقل و با جزئیات مورد بررسی قرار گرفته است. این مراحل شامل استخراج بردار عددی برای پروتئین‌ها، محاسبه میزان اهمیت، محاسبه پیام و ترکیب پیام‌ها می‌باشد. جزئیات مربوط به مراحل محاسبه میزان اهمیت، محاسبه پیام و ترکیب پیام‌ها به‌صورت دقیق‌تر در شکل (2) ارائه شده‌اند. در نهایت، بازنمایی حاصل برای گره‌های هدف در یک مسئله یادگیری به کار گرفته می‌شود. در این مطالعه، از بازنمایی‌های استخراج‌شده با روش پیشنهادی BioGraph2vec برای پیش‌بینی ارتباطات بین پروتئین‌های میزبان و پاتوژن استفاده شده است، درحالی که پروتئین‌ها به عنوان گره‌های هدف در نظر گرفته می‌شوند.



شکل (2): ساختار محاسبه اهمیت، پیام و بازنمایی در روش پیشنهادی

1,2,2. رمزگذاری توالی پروتئین

در اغلب پژوهش‌ها، توالی پروتئینی به‌عنوان منبع داده‌ی اصلی برای آموزش مدل‌ها مورد استفاده قرار می‌گیرد و تمام فرآیندهای پیش‌پردازش بر روی همین داده‌ها صورت می‌پذیرد. با توجه به اینکه طول توالی پروتئین‌ها در نمونه‌های مختلف یکسان نیست، لازم است از روشی بهره گرفت که این توالی‌ها را به بردارهایی با اندازه یکسان تبدیل کند.

در این پژوهش، سه روش مختلف برای تولید بازنمایی اولیه‌ی توالی پروتئین‌ها مورد استفاده قرار گرفته‌اند: Doc2Vec، ProtBert و ESM-2.

در روش Doc2Vec [20]، که در شکل (1) نیز نشان داده شده است، هر توالی پروتئینی به‌عنوان یک سند متنی در نظر گرفته می‌شود. برخلاف اسناد متنی معمول، توالی‌های پروتئینی از رشته‌هایی طولانی و بدون تقسیم‌بندی تشکیل شده‌اند. برای رفع این مسئله، توالی‌ها به بخش‌هایی با طول یکسان (3-mer) تقسیم می‌شوند و مدل Doc2Vec با استفاده از این قطعات تکه‌تکه‌شده از توالی‌های پروتئینی میزبان و پاتوژن آموزش می‌بیند. به‌منظور تعیین طول بهینه‌ی بردار خروجی، آزمایش‌هایی با طول‌های مختلف انجام شد و نتایج نشان داد که طول 50 بهترین کارایی را دارد. در نهایت، هر توالی پروتئین به برداری با اندازه‌ی یکنواخت تبدیل شد که به‌عنوان ورودی مراحل بعدی مدل مورد استفاده قرار گرفت.این مدل با دریافت مجموعه‌ای از اسناد، برای هر سند یک بردار ویژگی تولید می‌کند. در این پژوهش، هر توالی پروتئینی به‌صورت یک سند یکپارچه در نظر گرفته شده است. برخلاف اسناد معمول که از چندین کلمه تشکیل شده‌اند، توالی‌های پروتئینی به‌صورت رشته‌هایی طولانی و بدون تقسیم‌بندی هستند.

برای بهره‌گیری از مدل های ز بانی عمیق آموزش‌د یده بر داده‌های زیستی، دو روش پیشرفته‌تر نیز به‌کار گرفته شدند. روش ProtBert [21] بر پایه‌ی معماری BERT طراحی شده و با استفاده از تعداد زیادی توالی پروتئینی موجود در پایگاه UniRef100 آموزش دیده است. این مدل قادر است الگوهای نحوی و معنایی موجود در توالی‌های آمینواسیدی را همانند درک زبانی مدل‌های زبانی طبیعی استخراج کند. در این پژوهش، از خروجی لایه‌ی نهایی مدل ProtBert برای هر توالی پروتئینی به‌عنوان بازنمایی اولیه استفاده شده است.

همچنین، از مدل ESM-2 [22] که یکی از نسل‌های جدید مدل‌های زبانی ترنسفورمری برای داده‌های پروتئینی است، استفاده شد. نسخه‌ی مورد استفاده در این پژوهش ESM2-t12-35M-UR50D می‌باشد که یک مدل 12 لایه با حدود 35 میلیون پارامتر است. این مدل قادر است وابستگی‌های طولانی‌مدت میان اسیدهای آمینه را در سطح توالی درک کرده و بازنمایی‌های زیستی دقیق و فشرده‌ای ارائه دهد. در این پژوهش، خروجی لایه‌ی نهایی مدل ESM-2 برای هر توالی پروتئینی به‌عنوان بازنمایی اولیه استخراج و در مرحله‌ی بعدی مدل پیشنهادی مورد استفاده قرار گرفت.

به این ترتیب، هر سه روش Doc2Vec، ProtBert و ESM-2 به‌صورت مستقل برای تولید بردارهای اولیه‌ی پروتئین‌ها به‌کار گرفته شدند تا تأثیر نوع بازنمایی بر عملکرد کلی مدل پیشنهادی مورد بررسی قرار گیرد.

2,2,2. مکانیزم محاسبه اهمیت فراارتباط

در مدل پیشنهادی، هدف اصلی از مکانیزم اهمیت شناسایی میزان اهمیت هر گره مبدأ نسبت به گره هدف است تا تأثیر هر ارتباط

به‌صورت پویا در فرآیند یادگیری بازنمایی لحاظ شود. نخستین گام در فرآیند یادگیری بازنمایی برای گره هدف، محاسبه میزان اهمیت میان گره مبدأ s و گره هدف t است. اغلب روش‌های متداول برای تعیین وزن اهمیت، به گره‌هایی که نقش مهم‌تری دارند وزن بالاتری اختصاص می‌دهند، اما این روش‌ها معمولا فرض می‌کنند که گره‌های مبدأ و هدف در یک فضای توزیعی یکسان قرار دادند؛ در حالی‌که این فرض در بیشتر مسائل دنیای واقعی صادق نیستند و منجر به کاهش دقت مدل می‌شوند.

با توجه به این محدودیت و برای به حداکثر رساندن اشتراک‌گذاری اطلاعات بین گره‌های مختلف و حفظ ویژگی‌های ارتباطات خاص، روش محاسبه اهمیت با ایده گرفتن از مقاله [17] ارائه شده است. در این روش گره هدف t و تمامی گرهای همسایگی آن $s \in N(t)$ در فضای توزیعی متفاوتی هستند و اهمیت متقابل آن‌ها با در نظر گرفتن فراارتباط $\langle \tau(s), \Phi(e), \tau(t) \rangle$ محاسبه می‌شود. در این تحقیق با ایده گرفتن از معماری [25] گره هدف t به بردار پرسش و گره مبدأ s به بردار کلید نگاشت می‌شود. سپس ضرب نقطه‌ای آن‌ها به عنوان وزن اهمیت محاسبه می‌شود.

تفاوت اصلی میان روش پیشنهادی و تبدیل‌کننده ساده ارائه شده در [25] در این است که در نسخه ساده از یک طرح نگاشت یکنواخت برای کلمات استفاده می‌کند، اما در روش فعلی لازم است برای هر فراارتباط، وزن اهمیت منحصربه‌فردی تعریف شود. در این چارچوب، ماتریس وزن مربوط به یک فراارتباط به‌صورت جداگانه برای گره هدف، گره مبدأ و نوع یال بین آن‌ها پارامترسازی می‌گردد. به این ترتیب، وزن اهمیت مربوط به هر یال $\mathrm{e} = (s,t)$ مطابق رابطه (6) محاسبه خواهد شد.

$$\mathbf{Attention}(s,e,t) = \underset{\forall s \in N(t)}{\mathrm{softmax}} \left(\underset{i \in [1,h]}{\|} ATT-head^i(s,e,t) \right) \tag{6}$$

برای افزایش توانایی مدل در یادگیری روابط پیچیده، از مکانیزم چندسری استفاده شده است تا تعاملات از دیدگاه‌های مختلف ساختاری و معنایی بررسی شوند که محاسبات آن طبق رابطه (7) انجام می‌گیرد.

در نخستین گام، برای به‌دست آوردن اهمیت در مرحله i، گره مبدأ s از نوع $\tau(s)$ با استفاده از تبدیل خطی اختصاصی $\mathrm{K-linear}^i_{\tau(s)}$ به بردار کلید $K^i(s)$ تبدیل می‌شود که در فرمول (8) بیان شده است. لازم به ذکر است که هر نوع گره مبدأ داری نگاشتی مجزا و مختص به خود است. به‌طور مشابه گره هدف t از نوع $\tau(s)$ نیز با تبدیل خطی $\mathrm{Q-linear}^i_{\tau(t)}$ که در رابطه (9) آمده است به بردار پرسش iام تبدیل خواهد شد.

$$ATT-head^i(s,e,t) = \left(K^i(s)\, W^{ATT}_{\phi(e)}\, Q^i(t)\right) \cdot \mu_{\langle \tau(s), \Phi(e), \tau(t) \rangle} \tag{7}$$

$$K^i(s) = \mathrm{K-linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right) \tag{8}$$

$$Q^i(t) = \mathrm{Q-linear}^i_{\tau(t)}\left(H^{(l-1)}[t]\right) \tag{9}$$

در گام بعدی میزان شباهت بین بردارهای گره هدف و گره مبدأ محاسبه می‌شود. یکی از جنبه‌های مهم در گراف‌های ناهمگن، تنوع ارتباطات میان جفت‌گره‌های مختلف است؛ از این رو، برای هر نوع یال $\Phi(e)$، یک ماتریس وزن به‌صورت $W^{ATT}_{\phi(e)}$ در نظر گرفته می‌شود. این کار موجب می‌شود مدل بتواند حتی برای گرهای یکسان، تفاوت‌های معنایی ناشی از انواع ارتباط را تشخیص دهد. همچنین، از آن‌جا که تمام روابط به یک اندازه برای گره هدف مفید نیستند، از یک تنسور μ قابل یادگیری جهت نمایش میزان اهمیت کلی ارتباط سه‌تایی استفاده شده است.

پارامترهای $W^{ATT}_{\phi(e)}$ و μ به‌صورت مستقل برای هر نوع یال و گره یاد گرفته می‌شوند. این تمایز باعث می‌شود مدل بتواند تأثیر

انواع مختلف ارتباطات (مانند ارتباطات درون‌گونه‌ای یا بین‌گونه‌ای) را به‌صورت مجزا در نظر بگیرد. در پایان خروجی تمام سرهای اهمیت با یکدیگر ترکیب می‌شود تا اهمیت برای هر جفت‌گره به‌دست آید. سپس وزن‌های اهمیت همسایگی‌های $N(t)$ برای گره هدف t با اعمال تابع softmax نهایی‌سازی می‌شوند.

وزن‌های اهمیت در مدل پیشنهادی به‌صورت خودکار و از طریق فرآیند یادگیری به‌روزرسانی می‌شوند. این وزن‌ها با استفاده از تابع softmax نرمال‌سازی شده و در هر مرحله از آموزش، مدل با توجه به خطاهایی که در پیش‌بینی دارد، مقدار این وزن‌ها را طوری تغییر می‌دهد که عملکرد کلی بهتر شود. در نتیجه، گره‌هایی که تأثیر بیشتری در پیش‌بینی ارتباطات دارند، وزن بالاتری می‌گیرند.

3,2,2. انتقال پیام

به‌صورت همزمان و موازی با محاسبه وزن اهمیت بین گره مبدأ و هدف، اطلاعات ساختاری و ویژگی‌های معنایی گره‌های مبدأ نیز به گره‌های هدف منتقل می‌گردند. مشابه فرایند تعیین وزن اهمیت، در مرحله انتقال پیام نیز مفهوم فراارتباط به عنوان یک مؤلفه کلیدی لحاظ شده است. مطابق روابط (10) و (11) پیام قابل انتقال برای جفت گره $e = (s,t)$ محاسبه می‌گردد.

$$\mathbf{Message}(s,e,t) = \| \, MSG - head^i(s,e,t) \tag{10}$$

$$MSG - head^i(s,e,t) = M - \text{linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right) W^{MSG}_{\phi(e)} \tag{11}$$

برای گرفتن پیام مرحله lام، در ابتدا گره مبدأ از نوع $\tau(s)$ را با تبدیل‌کننده خطی $\text{M} - \text{linear}^i_{\tau(s)}$ به بردار پیام iام نگاشت می‌کنیم. سپس از ماتریس وزن $W^{MSG}_{\phi(e)}$ برای دخیل کردن تاثیر نوع یال استفاده می‌کنیم. در نهایت تمامی پیام‌ها به هم متصل می‌شوند و پیام نهایی برای یک جفت‌گره به‌دست می‌آید.

4,2,2. ترکیب پیام‌ها با وزن اهمیت

پس از محاسبه وزن اهمیت و پیام برای هر جفت‌گره، لازم است پیام‌های گره‌های مبدأ در گره‌های مقصد ادغام شوند. از آن‌جا که تابع softmax در مرحله دوم، مجموع وزن مربوط به هر فراارتباط را تعیین کرده است، در این مرحله می‌توان به‌سادگی از بردار اهمیت به‌عنوان وزن برای پیام‌ها بهره گرفت. در نهایت، بردار $\widetilde{H}^{(l)}[t]$ مطابق رابطه (12) محاسبه می‌شود.

$$\widetilde{H}^{(l)} = \oplus \left(\boldsymbol{Attention}(s,e,t) \,.\, \boldsymbol{Message}(s,e,t)\right) \tag{12}$$

در این مرحله، تمامی اطلاعات مربوط به همسایگان گره هدف t این گره ادغام می‌شود. گام نهایی شامل تبدیل بردار حاصل به فضای توزیع اولیه مرتبط با گره از نوع $\tau(t)$ است. برای این منظور، با الگوبرداری از روش ارائه‌شده در [22]، یک نگاشت خطی به‌صورت $\text{A} - \text{linear}^i_{\tau(t)}$ مطابق رابطه (13) بر بردار $\widetilde{H}^{(l)}[t]$ اعمال می‌گردد.

$$H^{(l)}[t] = \text{A} - \text{linear}^i_{\tau(t)}\left(\sigma\left(\widetilde{H}^{(l)}[t]\right)\right) + H^{(l-1)}[t] \tag{13}$$

به‌این صورت، خروجی لایه l برای گره به‌صورت $H^{(l)}[t]$ محاسبه می‌شود. این خروجی می‌تواند به عنوان ورودی برای وظایف پایین‌دستی مانند دسته‌بندی یا پیش‌بینی به کار گرفته شود.

3. پیاده سازی روش پیشنهادی

در این بخش، روش پیاده‌سازی BioGraph2vec، مجموعه داده برای انجام پژوهش و مراحل یادگیری شرح داده خواهند شد. در این پژوهش پیش‌بینی ارتباطات بین پروتئین‌های میزبان و پاتوژن

به‌عنوان روشی برای سنجش بازنمایی‌های ایجاد شده استفاده می‌شود.

1,3. مجموعه داده

در این پژوهش از مجموعه داده مرجع [27] استفاده شده که این مجموعه داده دارای 12157 ارتباط بین 29 میزبان و 332 ویروس است که جزئیات آن در جدول (2) ارائه شده است. به دلیل تعدد تعاملات مربوط به پروتئین‌های های میزبان انسانی در مقایسه با سایر میزبان‌ها، تعاملات میان میزبان‌های مختلف و پاتوژن‌ها به‌صورت جداگانه گروه‌بندی شده‌اند.

جدول (2): تعداد ارتباطات مجموعه داده اول

میزبان	میزبان اصلی (Taxonomy ID)	تعداد ارتباطات میزبان و پاتوژن	تعداد ویروس‌های در ارتباط
انسان	Homo sapiens (9606)	11491	246
	Mus musculus (10090)	191	89
	Bos taurus (9913)	125	32
	Rattus norvegicus (10116)	86	19
حیوان	Sus scrofa (9823)	57	10
	Gallus gallus (9031)	15	9
	Equus caballus (9796)	7	6
	Drosophila melanogaster (7227)	4	3
	Canis lupus familiaris (9615)	3	1
گیاه	Arabidopsis thaliana (3702)	17	11
	Escherichia coli K-12 (83333)	78	9
باکتری	Streptococcus pneumonia (170187)	49	2
	Pseudomonas aeruginosa (208963)	13	4
	Escherichia coli (562)	3	1
غیره	15 میزبان	18	15
مجموع	29	12157	332

علاوه بر مجموعه داده مذکور، در یک آزمایش دیگر از داده‌های پایگاه PHISTO [28] نیز بهره گرفته شده است. این مجموعه داده شامل پروتئین‌های مربوط به میزبان انسانی است. در این پژوهش، صرفا از پروتئین‌های پاتوژن‌های از نوع ویروس استفاده شده‌اند. همچنین برای بررسی ویروس Sars-cov-2 از مجموعه داده Gordon [29] استفاده شده است. این مجموعه داده دارای 332 ارتباط بین پروتئین‌های میزبان انسان و 27 عدد پروتئین ویروس Sars-cov-2 را شامل می‌شود.

پیش‌بینی ارتباطات بین پروتئین‌های میزبان و پاتوژن نیازمند داده‌های مثبت و منفی است. همچنین وجود ارتباط بین پروتئین‌ها به‌عنوان داده‌های مثبت و عدم وجود ارتباط به عنوان داده منفی تلقی می‌گردد. در این مطالعه، برای تولید نمونه‌های منفی، از روش نمونه‌گیری تصادفی استفاده شده است. نمونه‌های منفی از میان جفت پروتئین‌هایی انتخاب می‌شوند که در مجموعه داده مثبت حضور ندارند. به‌منظور حفظ تعادل داده‌ها، تعداد نمونه‌های منفی با تعداد نمونه‌های مثبت هم‌تراز شده است. این انتخاب تصادفی ساده و کارآمد است و هیچ سوگیری خاصی ایجاد نمی‌کند، ضمن اینکه امکان آموزش مدل برای تمایز بین نمونه‌های مثبت و منفی را به‌صورت مؤثر فراهم می‌کند.

2,3. نتایج

همانطور که قبلا اشاره شد، برای یکسان کردن طول توالی پروتئین‌ها و بدست آوردن بازنمایی اولیه از روش Doc2Vec، ESM-2 و ProtBert استفاده گردید تا تأثیر نوع بازنمایی بر دقت پیش‌بینی ارتباطات میزبان-ویروس مورد بررسی قرار گیرد.

پس از ساخت گراف ناهمگن، این گراف به مدل یادگیری پیشنهادی داده می‌شود. در این مرحله ابتدا بازنمایی‌ها برای گره‌ها تولید شده و سپس بازنمایی‌های تولید شده به ماشینی یادگیری داده می‌شوند. به عنوان ماشین یادگیری نهایی از شبکه عصبی،

جنگل تصادفی و ماشین بردار پشتیبان استفاده شده است. معیارهای مختلفی برای ارزیابی مدل پیشنهادی وجود دارد؛ مانند معیار دقت ، ضریب همبستگی متیو و F1 که با روابط(14)، (15)و (16) تعریف می‌شوند:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$MCC = \frac{(TP \times TN) + (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (15)$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (16)$$

مجموعه داده‌های معرفی شده در قسمت قبل برای آموزش مدل، به بخش‌های مختلفی تقسیم‌بندی می‌شوند تا بتوان روش پیشنهادی را با پاتوژن‌های مختلف و همچنین جدید نیز آزمایش کرد. داده‌های تقسیم شده برای آزمایش ویروس جدید به شکل زیر هستند:

- TR1: ارتباطات بین پروتئین‌های فقط میزبان انسان و پروتئین‌های تمامی ویروس‌ها به جز H1N1
- TR2: ارتباطات بین پروتئین‌های انسان و کلیه‌ی ویروس‌ها به جز Ebola
- TR3: ارتباطات بین پروتئین‌های کلیه‌ی میزبان‌ها و کلیه‌ی ویروس‌ها به جز H1N1
- TR4: ارتباطات بین پروتئین‌های کلیه‌ی میزبان‌ها و کلیه‌ی ویروس‌ها به جز Ebola
- TS1: ارتباطات بین پروتئین‌های میزبان انسان و ویروس H1N1
- TS2: ارتباطات بین پروتئین‌های میزبان انسان و ویروس Ebola

در جدول (3) نتایج آزمایش مدل پیشنهادی با سناریوهای پیشنهادی آورده شده است. نتایج به‌دست‌آمده نشان می‌دهند که استفاده از ProtBert در تمام معیارهای ارزیابی شامل ACC، MCC، AUC و F1 منجر به بهبود قابل‌توجهی شده است که در شکل (3) نیز قابل مشاهده است. این موضوع بیانگر آن است که ProtBert توانسته بازنمایی غنی‌تر و معنی‌دارتری از توالی‌های پروتئینی ارائه دهد و در نتیجه تفکیک بهتری میان تعاملات مثبت و منفی ایجاد کند. در مقابل، روش Doc2Vec که مبنای کلاسیک‌تری دارد عملکرد ضعیف‌تری نشان داد و ESM معمولاً بین دو روش دیگر قرار گرفت.

به‌منظور مقایسه عملکرد کلی روش پیشنهادی با پژوهش‌های پیشین، بهترین ترکیب مدل یعنی BioGraph2Vec با بازنمایی اولیه ProtBert با روش ارائه‌شده در مقاله [27] مقایسه گردید و نتایج در جدول (4) قابل مشاهده هستند. نتایج مقایسه در چهار سناریوی آموزشی-آزمایشی نشان می‌دهد که مدل پیشنهادی در تمامی معیارها از روش مرجع عملکرد بهتری داشته است. بهبود چشمگیر در معیارهای AUC و MCC نشان می‌دهد که BioGraph2Vec توانایی بیشتری در شناسایی الگوهای واقعی تعاملات دارد و پایداری بیشتری در برابر داده‌های نامتوازن از خود نشان می‌دهد. با این حال، عملکرد مدل در سناریوهایی که داده‌های مربوط به ویروس Ebola در مجموعه آزمون قرار داشتند اندکی کاهش یافت، که می‌تواند ناشی از پیچیدگی بیشتر الگوهای تعامل این ویروس با میزبان‌های خود باشد. در مقابل، نتایج به‌دست‌آمده برای ویروس H1N1 از تعمیم‌پذیری بالای مدل نسبت به پاتوژن‌های با الگوهای رفتاری مشابه حکایت دارد.

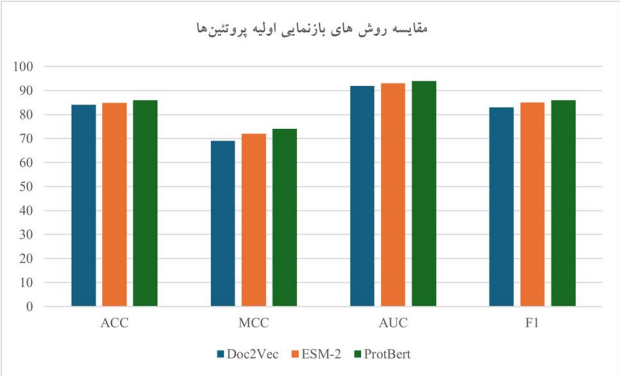
جدول (3): نتایج نهایی مدل آموزش دیده با ویروس‌های جدید

مدل تولید بازنمایی اولیه													
ProtBert				ESM`-2				Doc2Vec					
F1 (%)	AUC (%)	MCC (%)	ACC (%)	F1 (%)	AUC (%)	MCC (%)	ACC (%)	F1 (%)	AUC (%)	MCC (%)	ACC (%)	طبقه بند نهایی	داده
80,5	85	58,4	73,6	78,9	83,9	54,1	74,2	78,8	92,4	76,2	87,8	شبکه عصبی	TR1 TS1
81,3	88,7	59,7	78,6	79,4	88,5	55,2	76,2	81,4	88,2	69,7	83,7	جنگل تصادفی	
87	93,1	73,1	86,5	84,8	90,6	68,7	84,3	93,2	98,5	72,1	85,4	ماشین بردار پشتیبان	
83,9	82,7	66,1	81,1	83,9	83	66,1	81,1	83,4	93,9	74,6	85,5	شبکه عصبی	TR2 TS2
87,3	96,3	73,8	85,8	84,8	93,8	68,3	82,5	80,4	91,4	61,6	80,8	جنگل تصادفی	
86,4	94,1	71,9	85,8	85,3	90,5	69,3	0,844	80,2	91	70,9	83,4	ماشین بردار پشتیبان	
78,9	84	54,3	74	78,7	82,8	53,4	74,2	88,3	92,1	75,9	87,7	شبکه عصبی	TR3 TS1
80,2	89,1	57	74	80,6	88,2	58,1	77,6	80,7	90	66,3	82,5	جنگل تصادفی	
86,3	93	72,2	1.86	82,6	90,7	64	81,9	77,8	87,8	66,7	81,3	ماشین بردار پشتیبان	
83,1	82,2	64,5	80,1	83,1	81,8	64,5	80,1	80,6	93,5	71,4	83,8	شبکه عصبی	TR4 TS2
83,1	95,2	64,5	80,1	83,9	93,6	66,1	81,1	76,8	90,1	58,5	78,8	جنگل تصادفی	
83,7	89,2	66,4	83,1	83,6	87,4	65,4	82,1	74,2	90,1	64,6	79,5	ماشین بردار پشتیبان	

برای بررسی اهمیت مکانیزم توجه و انتقال پیام در مدل پیشنهادی برای گراف ناهمگن، بازنمایی‌های پیش‌آموزش‌دیده مانند Doc2Vec با بردارهای تصادفی جایگزین شدند، در حالی که ساختار اصلی مدل پیشنهادی که شامل مکانیزم توجه و انتقال پیام بود بدون تغییر باقی ماند. همان‌طور که انتظار می‌رفت، عملکرد مدل نسبت به حالت استفاده از بازنمایی‌های واقعی کاهش یافت؛ با این حال، مدل پیشنهادی همچنان توانست اطلاعات ساختاری گراف را به خوبی استخراج کند و به دقت و معیارهای عملکرد قابل توجهی برسد (ACC=72، F1=77,5، MCC=50,4، AUC=83,1). این آزمایش با مجموعه داده TR1 و TS1 و با طبقه بندی نهایی شبکه عصبی انجام گرفت. این نتایج نشان می‌دهند که توانایی مدل در بهره‌گیری از ساختار گراف مستقل از کیفیت بازنمایی‌های ورودی است، ولی بازنمایی های از پیش‌آموزش‌دیده با توالی پروتئین ها باعث افزایش دقت پیش‌بینی می‌شوند.

همانطور که در شکل (5) قابل مشاهده است، تأثیر نوع طبقه‌بند نهایی مورد ارزیابی قرار گرفت است. نتایج نشان داد که هر سه طبقه‌بند عملکرد قابل قبولی دارند اما در اغلب موارد، SVM بهترین مقادیر ACC، AUC و F1 را ارائه داده است. این امر نشان‌دهنده آن است که فضای بازنمایی استخراج‌شده توسط مدل پیشنهادی دارای ساختاری تفکیک‌پذیر است و طبقه‌بندهای مبتنی بر مرز تصمیم خطی مانند SVM می‌توانند از این ویژگی بیشترین بهره را ببرند. شبکه عصبی نیز در برخی سناریوها به‌ویژه در داده‌های پیچیده‌تر عملکرد مشابهی با SVM داشته است، در حالی‌که جنگل تصادفی معمولاً عملکردی میان‌رده و پایدار از خود نشان داده است.

علاوه بر آزمایش‌های ارائه شده، عملکرد BioGraph2vec با داده‌های مربوط به ویروس Sars-Cov-2 نیز مورد ارزیابی قرار گرفته است. برای تحلیل نتایج این آزمایش، از معیار AUC-PR استفاده شده است. مدل پیشنهادی، با استفاده از مجموعه داده

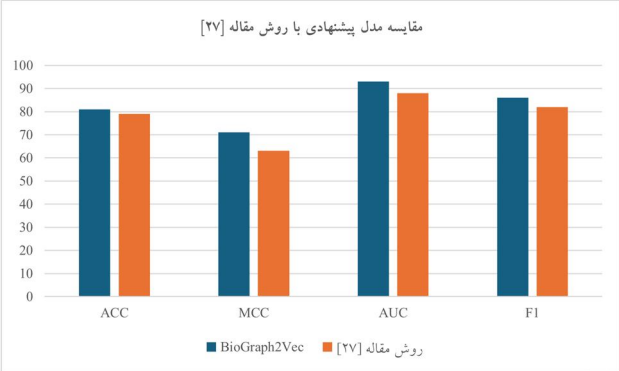


شکل (3): مقایسه عملکرد سه روش بازنمایی اولیه توالی پروتئین‌ها (Doc2Vec، ESM-2، ProtBert) با میانگین نتایج چهار سناریو

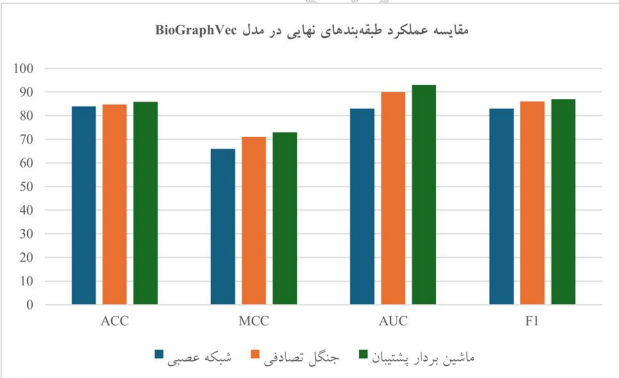
جدول (4): مقایسه نتایج مدل آموزش دیده با ترکیب مدل ProtVec با مقاله [27]

داده	طبقه بند نهایی	ACC (%)	MCC (%)	AUC (%)	F1 (%)
TR1 TS1	شبکه عصبی	73,6	58,4	85	80,5
	جنگل تصادفی	78,6	59,7	88,7	81,3
	ماشین بردار پشتیبان	86,5	73,1	93,1	87
	روش مقاله [27]	77,95	0,721	0,895	83,2
TR2 TS2	شبکه عصبی	81,1	66,1	82,7	83,9
	جنگل تصادفی	85,8	73,8	96,3	87,3
	ماشین بردار پشتیبان	85,8	71,9	94,1	86,4
	روش مقاله [27]	78	0,579	0,867	80,47
TR3 TS1	شبکه عصبی	74	54,3	84	78,9
	جنگل تصادفی	74	57	89,1	80,2
	ماشین بردار پشتیبان	1.86	72,2	93	86,3
	روش مقاله [27]	77,43	0,564	0,884	79,76
TR4 TS2	شبکه عصبی	80,1	64,5	82,2	83,1
	جنگل تصادفی	80,1	64,5	95,2	83,1
	ماشین بردار پشتیبان	83,1	66,4	89,2	83,7
	روش مقاله [27]	81,67	0,656	0,89	**83,77**

آموزشی اولیه و مجموعه داده آزمون Gordon، به میزان دقت 44% دست یافته است. این درحالی است که در روش پیشنهادی تعداد نمونه‌های منفی 10 برابر نمونه‌های مثبت بوده و این داده‌های منفی به‌صورت کاملاً تصادفی تولید شده‌اند. در مقابل، در پژوهش [30] در صد قابل‌توجهی از نمونه‌های منفی از میان پروتئین‌هایی انتخاب شده‌اند که دارای حداقل میزان شباهت هستند، که در این شرایط مدل مذکور به دقت 67% دست یافته است. با وجود این تفاوت‌ها در نحوه تولید داده‌های منفی، مدل پیشنهادی توانسته است عملکرد رقابتی قابل‌توجهی را ارائه دهد. همچنین، در پژوهش [5]، روشی مبتنی بر شبکه‌های عصبی گراف کانوولوشنی ارائه شده است که با انتخاب تصادفی داده‌های منفی، دقتی برابر با 43% به دست آورده است.



شکل (4): مقایسه عملکرد مدل پیشنهادی BioGraph2Vec (با بازنمایی ProtBert) و روش مقاله [27] بر روی میانگین نتایج چهار مجموعه داده‌ی آزمایشی



شکل (5): مقایسه عملکرد سه طبقه‌بند نهایی مورد استفاده در مدل BioGraph2Vec شامل شبکه عصبی، جنگل تصادفی و ماشین بردار پشتیبان بر روی میانگین نتایج چهار مجموعه داده‌ی آزمایشی

4. بحث و نتیجه‌گیری

در این پژوهش، مدل پیشنهادی BioGraph2Vec به‌عنوان چارچوبی مبتنی بر یادگیری بازنمایی در گراف‌های ناهمگن برای پیش‌بینی تعاملات میان پروتئین‌های میزبان و پاتوژن معرفی شد. هدف از این مدل، ترکیب ویژگی‌های توالی پروتئین‌ها با ساختار گراف زیستی به‌منظور بهبود درک روابط میان‌گونه‌ای بود. داده‌های مورد استفاده شامل مجموعه‌ای از تعاملات واقعی بین ویروس‌ها و میزبان‌های مختلف بودند که به‌گونه‌ای تقسیم شدند تا ویروس‌های موجود در داده‌ی آزمون در داده‌ی آموزش حضور نداشته باشند؛ به این ترتیب، قابلیت تعمیم مدل نسبت به پاتوژن‌های جدید ارزیابی گردید.

نتایج نشان داد که بازنمایی‌های استخراج شده توسط مدل BioGraph2Vec توانسته‌اند ساختارهای پیچیده‌ی گراف ناهمگن را به‌طور مؤثر مدل‌سازی کنند. در میان روش‌های بازنمایی اولیه، استفاده از ProtBert منجر به بهبود قابل‌توجه در تمامی معیارهای ارزیابی شامل Accuracy، MCC، AUC و F1 شد که بیانگر توان بالاتر این مدل در استخراج ویژگی‌های زیستی معنادار است. مقایسه‌ی روش پیشنهادی با پژوهش مرجع [27] نیز نشان داد که BioGraph2Vec (Model2) در تمامی معیارها عملکرد بهتری دارد و در معیارهای MCC و AUC پیشرفت چشمگیری داشته است. همچنین نتایج حاکی از آن است که تعاملات مربوط به ویروس Ebola پیچیدگی بیشتری دارند و چالش‌برانگیزتر از ویروس H1N1 هستند.

در بخش تحلیل طبقه‌بندها، سه الگوریتم شبکه عصبی، جنگل تصادفی و ماشین بردار پشتیبان بررسی شدند. نتایج نشان دادند که هر سه طبقه‌بند عملکرد مناسبی دارند اما SVM به‌طور میانگین بهترین مقادیر ACC، AUC و F1 را ارائه کرده است. این امر نشان می‌دهد که بازنمایی‌های تولیدشده توسط BioGraph2Vec از ساختار تفکیک‌پذیر بالایی برخوردارند.

برای ارزیابی توان تعمیم مدل نسبت به پاتوژن‌های جدید، داده‌های ویروس SARS-CoV-2 که یک بیماری واگیرداراست و زندگی میلیاردها نفر در سراسر جهان را تحت تاثیر قرار داده است نیز مورد آزمایش قرار گرفتند[31]. مدل پیشنهادی در این بخش به دقت 44٪ برای AUC–PR دست یافت، در حالی‌که پژوهش‌های [30] و [5] به ترتیب مقادیر 37%و 43% را گزارش کرده‌اند. با توجه به تولید تصادفی داده‌های منفی در این پژوهش، این نتایج نشان‌دهنده‌ی عملکرد رقابتی مدل پیشنهادی است.

به‌طور کلی، یافته‌ها نشان می‌دهند که ترکیب بازنمایی‌های توالی با گراف ناهمگن در چارچوب BioGraph2Vec منجر به بهبود چشمگیر در پیش‌بینی تعاملات میزبان–ویروس شده است. توسعه‌ی نسخه‌های کاراتر از این مدل و ترکیب آن با روش‌های گراف عمیق یا مدل‌های ترنسفورمری می‌تواند مسیرهای آینده‌ی ارزشمندی برای افزایش دقت و تعمیم‌پذیری در تحلیل داده‌های زیستی فراهم کند.

با وجود نتایج امیدوارکننده، برخی چالش‌ها و محدودیت‌ها نیز وجود دارند. از جمله اینکه هنگام استفاده از این مدل برای گراف‌های زیستی دیگر، ممکن است به‌دلیل تفاوت در نوع گره‌ها و ویژگی‌های ورودی، نیاز به بازطراحی یا تنظیم مجدد بخشی از مدل وجود داشته باشد. علاوه بر این، گسترش مدل به داده‌های زیستی متنوع‌تر مانند تعاملات ژنی، شبکه‌های پروتئینی میان‌گونه‌ای و یا مسائل دیگر می‌تواند به درک عمیق‌تر از قابلیت تعمیم مدل کمک کند.

به‌عنوان مسیرهای آینده، می‌توان توسعه نسخه‌های کاراتر و سبک‌تر از BioGraph2Vec برای مجموعه‌داده‌های بزرگ‌تر، و همچنین بررسی ترکیب آن با روش‌های دیگر یادگیری عمیق را به‌منظور بهبود دقت و تعمیم‌پذیری پیشنهاد کرد.

مراجع

[1] S.M. Vahidipour and A. Mohammadi, “Link prediction in scientific networks using machine learning and weighted graphs,” Soft Comput. J., vol. 13, no. 2, pp. 128-153, 2024, doi: 10.22052/scj.2024.253476.1181 [In Persian].

[2] Z. Gao *et al.*, "edge2vec: Representation learning using edge semantics for biomedical knowledge discovery," *BMC Bioinformatics*, vol. 20, no. 1, p. 306, 2019/06/10 2019, doi: 10.1186/s12859-019-2914-2.

[3] A. Grover and J. Leskovec, "node2vec: Scalable Feature Learning for Networks," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016. [Online]. Available: https://doi.org/10.1145/2939672.2939754.

[4] Dong, Y.; Chawla, N. V.; Swami, A. "metapath2vec: Scalable Representation Learning for Heterogeneous Networks". *In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, 2017

[5] M. B. Koca, E. Nourani, F. Abbasoğlu, İ. Karadeniz, and F. E. Sevilgen" ,Graph convolutional network based virus-human protein-protein interaction prediction for novel viruses," *Computational Biology and Chemistry*, vol. 101, p. 107755, 2022.

[6] S.M. Vahidipour, D. Daneshmand, and M.A. Zarif, “A review of knowledge graph embedding methods,” Soft Comput. J., vol. 14, no. 2, pp. 2-19, 2024, doi: 10.22052/scj.2024.254464.1224 [In Persian].

[7] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu" ,A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4-24, 2021, doi: 10.1109/TNNLS.2020.2978386.

[8] Y. Xie, B. Yu, S. Lv, C. Zhang, G. Wang, and M. Gong, "A survey on heterogeneous network representation learning," *Pattern Recognition*, vol. 116, p. 107936, 2021/08/01/ 2021, doi: https://doi.org/10.1016/j.patcog.2021.107936.

[9] C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang, "Discovering Meta-Paths in Large Heterogeneous Information Networks," presented at the Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 2015. [Online]. Available: https://doi.org/10.1145/2736277.2741123.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[11] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "MetaGraph2Vec: Complex Semantic Path Augmented Heterogeneous Network Embedding," in *Advances in Knowledge Discovery and Data Mining*, Cham, D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, Eds., 2018// 2018: Springer International Publishing, pp. 196-208.

[12] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li, "Meta Structure: Computing Relevance in Large Heterogeneous Information Networks," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016. [Online]. Available: https://doi.org/10.114۲۹۳۹۶۷۲,۲۹۳۹۸۱۵/۵.

[13] C. Zhang, A. Swami, and N. V. Chawla, "CARL: Content-aware representation learning for heterogeneous networks," *arXiv preprint arXiv:1805.04983*, 2018.

[14] S. Zhou, J. Bu, X. Wang, J. Chen, and C. Wang, "HAHE: Hierarchical attentive heterogeneous information network embedding," *arXiv preprint arXiv:1902.01475*, 2019.

[15] Zhong, H.; Wang, M.; Zhang, X. "Unsupervised Embedding Learning for Large-Scale Heterogeneous Networks Based on Metapath Graph Sampling". *Entropy* 2023, 25, 297.

[16] Zhong, Hongwei, Mingyang Wang, and Xinyue Zhang. 2023. "HeMGNN: Heterogeneous Network

vol. 19, no. 6, p. 568, 2018/08/13 2018, doi: 10.1186/s12864-018-4924-2.

[28] S. Durmuş Tekir *et al.*" ,PHISTO: pathogen–host interaction search tool," *Bioinformatics*, vol. 29, no. 10, pp. 1357-1358, 2013, doi: 10.1093/bioinformatics/btt137.

[29] D. E. Gordon *et al.*, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," *Nature*, vol. ,٥٨٣ no. 7816, pp. 459-468, 2020/07/01 2020, doi: 10.1038/s41586-020-2286-9.

[30] M. Kshirsagar *et al.*, "Protein sequence models for prediction and comparative analysis of the SARS-CoV-2 —human interactome," in *Biocomputing 2021*, pp. 154-165.

[31] M. Mousavi, S. Hosseini, and M. Omidi, “An improved deep neural network algorithm for COVID-19 detection in the Internet of Things,” Soft Comput. J., vol. 12, no. 2, pp. 54-71, 2024, doi: 10.22052/scj.2023.248686.1117 [In Persian]

Embedding Based on a Mixed Graph Neural Network" *Electronics* 12, no. 9: 2124.

[17] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous Graph Transformer," presented at the Proceedings of The Web Conference 2020, Taipei, Taiwan, 2020. [Online]. Available: https://doi.org/10.1145/3366423.3380027.

[18] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[19] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[20] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*, 2017: PMLR, pp. 1263-1272.

[21] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., et al. “ProtTrans: Towards cracking the language of life’s code through self-supervised deep learning and high performance computing”. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 2021.

[22] Lin, Z., Akin, H., Rao, R., et al. “Evolutionary-scale prediction of atomic-level protein structure with a language model”. *bioRxiv preprint* 2023.

[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[24] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," presented at the Proceedings of the 31st International Conference on Machine Learning, Proceedings of Machine Learning Research, 2014. [Online]. Available: https://proceedings.mlr.press/v32/le14.html.

[25] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-٧٧٨

[27] X. Zhou, B. Park, D. Choi, and K. Han, "A generalized approach to predicting protein-protein interactions between virus and host," *BMC Genomics*,