BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

LÊ MẠNH HÙNG

Đề tài:

NÂNG CAO HIỆU QUẢ DỰ ĐOÁN LIÊN KẾT THUỐC-BỆNH DỰA TRÊN SIÊU ĐƯỜNG DẪN VÀ SUY LUẬN BAYES TRÊN MẠNG KHÔNG ĐỒNG NHẤT

LUẬN ÁN TIẾN SĨ NGÀNH CÔNG NGHỆ THÔNG TIN

Hà Nội – 2026

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

LÊ MẠNH HÙNG

Đề tài:

NÂNG CAO HIỆU QUẢ DỰ ĐOÁN LIÊN KẾT THUỐC-BỆNH DỰA TRÊN SIÊU ĐƯỜNG DẪN VÀ SUY LUẬN BAYES TRÊN MẠNG KHÔNG ĐỒNG NHẤT

NGÀNH: CÔNG NGHỆ THÔNG TIN
MÃ SỐ NGÀNH: 9480201

LUẬN ÁN TIẾN SĨ NGÀNH CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. TS. Đào Nam Anh
2. TS. Đặng Xuân Thọ

Hà Nội – 2026

LỜI CAM ĐOAN

Tôi xin cam đoan luận án: "Dự đoán liên kết thuốc – bệnh trong mạng không đồng nhất" là công trình nghiên cứu của chính mình dưới sự hướng dẫn khoa học của tập thể hướng dẫn. Luận án sử dụng thông tin trích dẫn từ nhiều nguồn tham khảo khác nhau và các thông tin trích dẫn được ghi rõ nguồn gốc. Các kết quả nghiên cứu của tôi được công bố chung với các tác giả khác đã được sự nhất trí của đồng tác giả khi đưa vào luận án. Các số liệu, kết quả được trình bày trong luận án là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác ngoài các công trình công bố của tác giả. Luận án được hoàn thành trong thời gian tôi làm nghiên cứu sinh tại Khoa Công nghệ thông tin, Trường Đại học Điện lực.

Hà Nội, ngày...tháng...năm 2026

Tác giả luận án

. **Lê Mạnh Hùng**

LỜI CẢM ƠN

Trước hết, em xin bày tỏ lòng biết ơn sâu sắc tới hai thầy hướng dẫn, TS. Đào Nam Anh và TS. Đặng Xuận Thọ, những người đã tận tình chỉ bảo, định hướng khoa học và luôn đồng hành cùng em trong suốt quá trình nghiên cứu. Sự tâm huyết, trách nhiệm và những tri thức quý báu từ hai thầy là nền tảng quan trọng giúp em hoàn thành luận án này.

Tôi xin trân trọng cảm ơn Ban Giám hiệu Nhà trường và Phòng Quản lý Đào tạo - Trường Đại học Điện lực, đã tạo điều kiện thuận lợi để tôi được học tập, nghiên cứu và phát triển năng lực học thuật. Tôi cũng xin gửi lời cảm ơn tới Ban lãnh đạo cùng các thầy cô Khoa Công nghệ thông tin, Trường Đại học Điện lực, những người đã nhiệt tình giảng dạy, truyền cảm hứng và hỗ trợ tôi trong suốt quá trình học tập.

Xin chân thành cảm ơn các đồng nghiệp, bạn bè và đặc biệt là nhóm AI Study đã luôn đồng hành, chia sẻ và trao đổi kiến thức, góp phần giúp tôi có thêm động lực và kinh nghiệm quý báu.

Tôi vô cùng biết ơn bố mẹ và anh chị em trong gia đình – những người luôn yêu thương, tin tưởng và là chỗ dựa tinh thần to lớn cho tôi. Cuối cùng, tôi xin dành lời tri ân sâu sắc tới vợ và hai con yêu quý, những người đã luôn ở bên cạnh, động viên và tiếp thêm sức mạnh để tôi kiên trì theo đuổi con đường học thuật đầy thử thách này.

Tác giả luận án

. **Lê Mạnh Hùng**

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Bảng 1: Bảng tóm tắt các ký hiệu và từ viết tắt

STT	**Từ viết tắt**	**Nguyên nghĩa**
1	ACC	Độ chính xác tổng thể (Accuracy)
2	AC	Bộ phân loại thích ứng (Adaptive Classifier)
3	ADD	Mối liên hệ thuốc–bệnh (Association Drug–Disease)
4	ADMM	Phương pháp nhân tử hướng luân phiên (Alternating Direction Method of Multipliers)
5	ANN	Mạng nơ-ron nhân tạo (Artificial Neural Network)
6	AUC	Diện tích dưới đường cong ROC (Area Under the Curve)
7	AUPR	Diện tích dưới đường cong Precision–Recall (Area Under the Precision–Recall Curve)
8	BGMSDDA	Khuếch tán đồ thị hai phần với tích hợp đa tương đồng (Bipartite Graph Diffusion with Multiple Similarity Integration)
9	BNNR	Chuẩn hạt nhân Bayes có điều chuẩn (Bayesian Nuclear Norm Regularization)
10	CNN	Mạng nơ-ron tích chập (Convolutional Neural Network)
11	DDI	Tương tác thuốc–bệnh (Drug–Disease Interactions)
12	DRAGNN	Mạng nơ-ron đồ thị tăng cường tái định vị thuốc (Drug Repositioning Augmented Graph Neural Network)
13	DRIMC	Tái định vị thuốc bằng hoàn thiện ma trận quy nạp (Drug Repositioning by Inductive Matrix Completion)
14	DR-IBRW	Tái định vị thuốc bằng bước ngẫu nhiên hai chiều cá nhân (Drug Repositioning based on Individual Bi-random Walk)
15	DRRS	Hệ thống gợi ý tái định vị thuốc (Drug Repositioning Recommender System)
16	DTINet	Mạng tương tác thuốc–đích (Drug–Target Interaction Network)
17	E	cạnh (Edges)

(còn tiếp trang sau)

STT	Từ viết tắt	Nguyên nghĩa
18	FDA	Cục Quản lý Thực phẩm và Dược phẩm Hoa Kỳ (Food and Drug Administration)
19	F1	Chỉ số F1 (F1-score)
20	FN	Âm tính sai (False Negative)
21	FP	Dương tính sai (False Positive)
22	G	đồ thị (Graph)
23	GCN	Mạng nơ-ron tích chập đồ thị (Graph Convolutional Network)
24	G-Mean	Trung bình hình học (Geometric Mean)
25	GNN	Mạng nơ-ron đồ thị (Graph Neural Network)
26	GRGMF	Phân rã ma trận tổng quát có điều chuẩn đồ thị (Graph Regularized Generalized Matrix Factorization)
27	GRLMN	Học biểu diễn đồ thị trên mạng đa phân tử sinh học (Graph Representation Learning Over Multi-Biomolecular Network)
28	GTN	Mạng biến đổi đồ thị (Graph Transformer Network)
29	HIN	Mạng thông tin không đồng nhất (Heterogeneous Information Network)
30	HNet-DNN	Mạng nơ-ron sâu dựa trên mạng không đồng nhất (Heterogeneous Network-Based Deep Neural Network)
31	HNRD	Mạng không đồng nhất cho thuốc–bệnh (Heterogeneous Network for Drug–Disease)
32	KNN	K-láng giềng gần nhất (K-Nearest Neighbors)
33	LAGCN	Mạng nơ-ron tích chập đồ thị với cơ chế chú ý lớp (Layer Attention Graph Convolutional Network)
34	LightGBM	Máy tăng cường độ dốc nhẹ (Light Gradient Boosting Machine)
35	LMF	Phân rã ma trận logistic (Logistic Matrix Factorization)
36	LRSSL	Học không gian thưa có điều chuẩn Laplacian (Laplacian Regularized Sparse Subspace Learning)
37	MBiRW	Đo độ tương đồng và bước ngẫu nhiên hai chiều (Similarity Measures and Bi-random Walk)
38	MCC	Hệ số tương quan Matthews (Matthews Correlation Coefficient)
39	MC	Hoàn thiện ma trận (Matrix Completion)

(còn tiếp trang sau)

STT	Từ viết tắt	Nguyên nghĩa
40	MF	Phân rã ma trận (Matrix Factorization)
41	MSBMF	Phân rã ma trận song tuyến tính đa tương đồng (Multi-similarities Bi-linear Matrix Factorization)
42	NCS	Nghiên cứu sinh
43	NRLMF	Phân rã ma trận logistic có điều chuẩn lân cận (Neighborhood Regularized Logistic Matrix Factorization)
44	OMC	Hoàn thiện ma trận chồng lấn (Overlap Matrix Completion)
45	OSS	Lấy mẫu một phía (One-Sided Selection)
46	PCA	Phân tích thành phần chính (Principal Component Analysis)
47	PRE	Độ chính xác (Precision)
48	RNN	Mạng nơ-ron hồi quy (Recurrent Neural Network)
49	REC	Độ thu hồi (Recall)
50	RF	Rừng ngẫu nhiên (Random Forest)
51	RLMD	Phân rã ma trận logistic có điều chuẩn (Regularized Logistic Matrix Decomposition)
52	RUS	Lấy mẫu ngẫu nhiên dưới mức (Random Undersampling)
53	SCMFDD	Phân rã ma trận ràng buộc tương đồng để dự đoán liên kết thuốc–bệnh (Similarity Constrained Matrix Factorization for Drug–Disease Association Prediction)
54	SE	Độ nhạy (Sensitivity)
55	SSGC	Cắt đồ thị bán giám sát (Semi-Supervised Graph Cut)
56	SMOTE	Kỹ thuật tạo mẫu tổng hợp cho lớp thiểu số (Synthetic Minority Over-sampling Technique)
57	SNF	Hợp nhất mạng tương đồng (Similarity Network Fusion)
58	SPLCMF	Học tự điều tốc với phân rã ma trận có tương quan (Self-Paced Learning with Correlated Matrix Factorization)
59	SP	Độ đặc hiệu (Specificity)
60	SVD	Phân tích giá trị kỳ dị (Singular Value Decomposition)
61	SVM	Máy vector hỗ trợ (Support Vector Machine)
62	t	protein
63	TL-HGBI	Suy luận dựa trên đồ thị không đồng nhất ba lớp (Triple-Layer Heterogeneous Graph-Based Inference)
64	TN	Âm tính đúng (True Negative)

(còn tiếp trang sau)

STT	Từ viết tắt	Nguyên nghĩa
65	TP	Dương tính đúng (True Positive)
66	TP-NRWRH	Bước ngẫu nhiên khởi động lại hai pha trên mạng không đồng nhất (Two-Pass Network-based Random Walk with Restart on Heterogeneous Network)
67	v	đỉnh (Vertices)
68	ϕ	hàm ánh xạ đối tượng
69	ψ	hàm ánh xạ liên kết
70	$\mathcal{A}$	loại đối tượng
71	$\mathcal{R}$	loại quan hệ
72	$\mathcal{T}_G$	lược đồ mạng (Network schema)
73	$\circ$	toán tử hợp thành/hợp nhất các quan hệ
74	$\rightarrow$	mũi tên/ánh xạ
75	ClusterONE	Phân cụm với sự mở rộng khu vực lân cận chồng chéo(Clustering With Overlapping Neighborhood Expansion)
75	DR-LGBM-MH	Tái định vị thuốc bằng cách sử dụng LightGBM và Meta-Paths trong mạng lưới không đồng nhất (Drug Repositioning using LightGBM and Meta-Paths in Heterogeneous network)
76	HS-TMP	Sự tương đồng đồng nhất và ba nhóm siêu đường dẫn (Homogeneous Similarities and Three Meta-Path)
77	DDA-BNS	Dự đoán mối liên hệ giữa thuốc và bệnh bằng mạng Bayes và phương pháp lấy mẫu (Drug-Disease Association prediction using Bayesian Network and Sampling)
78	AML	Bạch cầu cấp tính (Acute Myeloid Leukemia)
79	EMP-SVD	Tổng hợp siêu đường dẫn và phân tách giá trị kỳ dị (Ensemble Meta-Paths and Singular Value Decomposition)
80	TCA	Thuốc chống trầm cảm ba vòng (Tricyclic Antidepressants)
81	DDAP	Dự đoán liên kết thuốc–bệnh (Drug–Disease Association Prediction)
82	TPR	Tỷ lệ dương tính thực tế được dự đoán đúng
83	FPR	Tỷ lệ âm tính thực tế được dự đoán đúng

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC BẢNG BIỂU

MỞ ĐẦU

Vấn đề nghiên cứu:

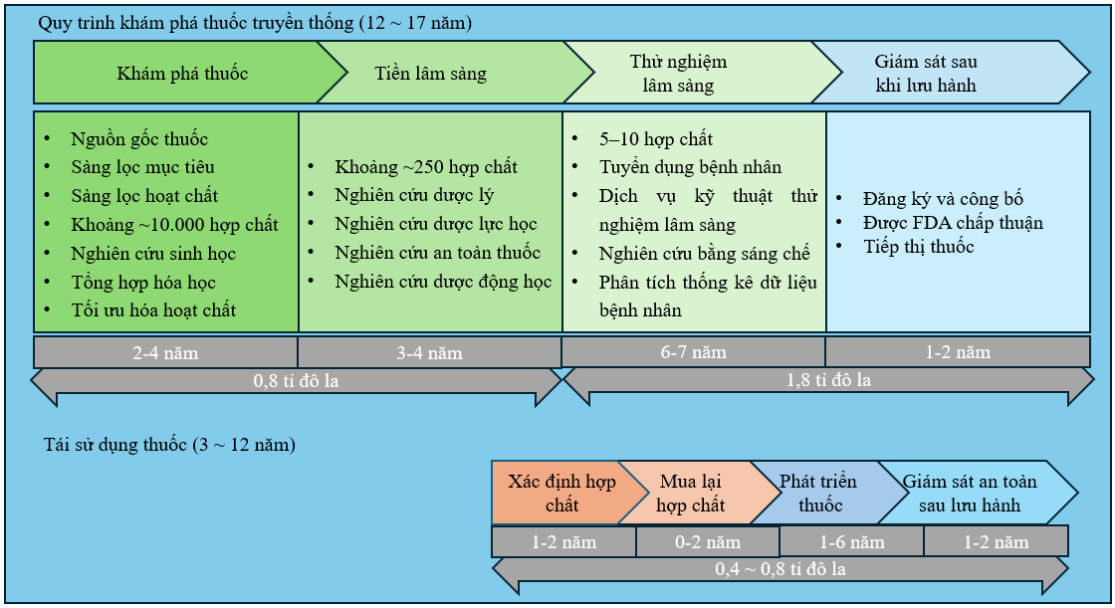
Dự đoán liên kết thuốc–bệnh (Drug–Disease Association Prediction, DDAP) là một bài toán trọng tâm trong lĩnh vực tin sinh học và khoa học dữ liệu y sinh, mang ý nghĩa quan trọng cả về mặt lý thuyết lẫn ứng dụng thực tiễn. Bài toán này nhằm ước lượng khả năng tồn tại các mối quan hệ có ý nghĩa giữa thuốc và bệnh, chẳng hạn như khả năng điều trị, gây tác dụng phụ liên quan, hoặc tiềm năng mở rộng chỉ định điều trị mới cho thuốc.

Trong bối cảnh phát triển dược phẩm hiện đại, DDAP đóng vai trò nền tảng trong việc hỗ trợ tái định vị thuốc (drug repurposing)—một chiến lược được xem là hiệu quả nhằm rút ngắn thời gian phát triển, tiết kiệm chi phí và giảm thiểu rủi ro so với quy trình phát triển thuốc truyền thống. Trên thực tế, phát triển một loại thuốc mới là quá trình phức tạp, kéo dài và tốn kém. Như minh họa trong Hình 1, quy trình này thường kéo dài từ 12 đến 17 năm và bao gồm bốn giai đoạn chính: (1) khám phá thuốc, (2) nghiên cứu tiền lâm sàng, (3) thử nghiệm lâm sàng và (4) theo dõi sau khi lưu hành. Do các yêu cầu nghiêm ngặt về an toàn, hiệu quả và chất lượng trong nghiên cứu trên động vật cũng như thử nghiệm lâm sàng do các cơ quan quản lý đặt ra, thời gian và nguồn lực cần thiết cho quá trình này là rất lớn [1].

Lịch sử phát triển dược phẩm đã cung cấp nhiều bằng chứng thuyết phục cho thấy sự tồn tại của các mối liên hệ thuốc–bệnh vượt ra ngoài chỉ định ban đầu. Những phát hiện này, dù trong nhiều trường hợp mang tính tình cờ, đã góp phần củng cố cơ sở khoa học cho việc nghiên cứu có hệ thống các liên kết thuốc–bệnh mới. Chẳng hạn, Sildenafil ban đầu được phát triển để điều trị bệnh mạch vành nhưng trong quá trình thử nghiệm đã cho thấy tác dụng cải thiện chức năng cương dương, từ đó được tái định vị thành thuốc điều trị rối loạn cương dương [2]. Tương tự, Chlorpromazine, vốn được sử dụng như một thuốc

kháng histamine, đã mở ra kỷ nguyên điều trị các rối loạn loạn thần sau khi tác dụng an thần của nó được Heri Laborit phát hiện vào năm 1952 [3]. Gần đây hơn, trong bối cảnh đại dịch COVID-19, remdesivir—một loại thuốc được phát triển để điều trị Ebola—đã được tái định vị nhằm điều trị COVID-19 [4].

Tuy nhiên, dữ liệu y sinh thực nghiệm thường phân tán, không đầy đủ, phức tạp và đòi hỏi chi phí cao để thu thập, dẫn đến việc nhiều mối liên kết thuốc–bệnh tiềm năng trong thực tế chưa được ghi nhận hoặc xác thực. Điều này đặt ra nhu cầu cấp thiết phải phát triển các phương pháp tính toán hiệu quả nhằm dự đoán các liên kết chưa biết, qua đó hỗ trợ nhà nghiên cứu thu hẹp không gian tìm kiếm và ưu tiên các giả thuyết có tiềm năng cao trước khi tiến hành các thử nghiệm sinh học tốn kém. Do đó, bài toán dự đoán liên kết thuốc–bệnh giữ vai trò then chốt trong việc thúc đẩy quá trình tái định vị thuốc một cách có hệ thống, chính xác và hiệu quả.



Hình 1: So sánh quy trình khám phá thuốc và tái định vị

Dữ liệu y sinh hiện đại có tính đa dạng, không đồng nhất và phân tán, bao gồm nhiều loại thực thể như thuốc, bệnh, gene, protein, pathway hay cấu trúc hóa học, do đó mạng không đồng nhất (HIN) trở thành mô hình đặc biệt phù hợp để tích hợp và biểu diễn các mối quan hệ phức tạp này. Trong bối cảnh đó, nhiều phương pháp học máy và học sâu trên đồ thị đã được phát triển nhằm

cải thiện dự đoán liên kết thuốc–bệnh. Các hướng tiếp cận tiêu biểu gồm: (i) các phương pháp dựa trên độ tương đồng [5, 6, 7], hoặc sử dụng các thước đo khoảng cách như Euclidean, cosine [8, 9]; (ii) các phương pháp phân rã và hoàn thiện ma trận [10, 11, 12, 13] để bù khuyết dữ liệu và suy luận liên kết tiềm năng; (iii) các phương pháp khai thác đồ thị như meta-path, GCN, GNN, GTN [14, 15, 16, 17, 18, 19, 20] nhằm học biểu diễn đặc trưng sâu hơn. Bên cạnh đó, suy luận Bayes là phương pháp đặc biệt phù hợp với dữ liệu y sinh, vốn chứa nhiều yếu tố không chắc chắn, thiếu hụt hoặc nhiễu. Lý thuyết Bayes cho phép mô hình hóa xác suất của một liên kết thuốc–bệnh dựa trên bằng chứng thu được từ các meta-path và các đặc trưng đa nguồn [12]. Mặc dù các phương pháp này đã mang lại nhiều tiến bộ đáng kể trong phát hiện liên kết thuốc–bệnh, các nghiên cứu [21, 22, 23] chỉ ra rằng chúng vẫn đối mặt với các thách thức quan trọng sau đây:

- Thứ nhất, tính đa dạng và phức tạp trong dữ liệu thuốc-bệnh: Việc tích hợp nhiều nguồn thông tin sinh học khác nhau nhằm cải thiện khả năng biểu diễn của mô hình đang được quan tâm nghiên cứu. Tuy nhiên, tính không đồng nhất và sự khác biệt lớn giữa các nguồn dữ liệu này có thể làm suy giảm chất lượng dự đoán. Do đó, việc khai thác và tích hợp hiệu quả dữ liệu đa nguồn hiện vẫn là một thách thức trọng yếu.

- Thứ hai, dữ liệu thưa và mất cân bằng dữ liệu: Dữ liệu y sinh thường có đặc trưng thưa, dẫn đến sự mất cân bằng nghiêm trọng khi các mối quan hệ thuộc lớp thiểu số ít hơn rất nhiều so với lớp đa số. Điều này gây khó khăn trong quá trình huấn luyện, làm tăng nguy cơ mô hình bị thiên lệch về phía dữ liệu đa số.

- Thứ ba, vấn đề âm tính giả: Trong nghiên cứu thuốc – bệnh, nhiều cặp thuốc – bệnh chưa được xác định nhưng có khả năng tồn tại quan hệ điều trị thường bị gán nhãn âm tính. Điều này gây ra sai lệch trong dữ liệu huấn luyện, làm giảm độ chính xác của mô hình và bỏ sót các mối quan hệ tiềm năng.

- Cuối cùng, khả năng giải thích và diễn giải mô hình: Việc xây dựng các mô hình có tính minh bạch và khả năng giải thích cao vẫn là một thách thức quan trọng. Tính giải thích này không chỉ giúp các chuyên gia y tế hiểu rõ kết quả dự đoán mà còn góp phần nâng cao độ tin cậy khi áp dụng trong thực tiễn lâm sàng.

Những hạn chế trên cho thấy nhu cầu cấp thiết phải phát triển các phương pháp tính toán hiệu quả và đáng tin cậy hơn. Một số hướng nghiên cứu tiềm năng bao gồm:

- Thứ nhất, khai thác mối quan hệ tiềm ẩn giữa thuốc và bệnh từ nhiều nguồn dữ liệu sinh học (gen, protein, cấu trúc phân tử, đặc tính hóa học...), qua đó giảm sự phụ thuộc vào dữ liệu quan hệ đã biết.

- Thứ hai, giải quyết dữ liệu thưa và mất cân bằng bằng các kỹ thuật cân bằng dữ liệu đã được kiểm chứng trong nhiều lĩnh vực, từ đó nâng cao độ chính xác và tính ổn định của mô hình.

- Thứ ba, khắc phục ảnh hưởng của các mẫu âm tính giả thông qua cải thiện chất lượng dữ liệu huấn luyện, nhằm tăng hiệu suất và độ tin cậy của các mô hình học máy trong dự đoán quan hệ thuốc–bệnh.

- Thứ tư, nâng cao khả năng diễn giải và tính minh bạch của mô hình, giải thích cơ chế dự đoán liên kết thuốc–bệnh.

Những vấn đề này vừa là thách thức vừa là động lực khoa học, đồng thời là cơ sở để nghiên cứu sinh lựa chọn đề tài luận án tiến sĩ: **"Nâng cao hiệu quả dự đoán liên kết thuốc-bệnh dựa trên siêu đường dẫn và suy luận Bayes trên mạng không đồng nhất"**. Luận án nhằm phát triển một mô hình kết hợp siêu đường dẫn và suy luận Bayes để khắc phục các hạn chế trên, đồng thời nâng cao hiệu quả dự đoán liên kết thuốc–bệnh.

Để hoàn thành luận án, nghiên cứu sinh (NCS) đề ra các mục tiêu cụ thể sau đây:

Mục tiêu nghiên cứu

- **Mục tiêu 1:** Phát triển một mô hình dựa trên việc khai thác siêu đường dẫn và áp dụng suy luận Bayes trên mạng thông tin không đồng nhất. Mục tiêu này nhằm (i) tích hợp hiệu quả các nguồn dữ liệu sinh học đa dạng (thuốc, bệnh, protein...) và (ii) ước lượng xác suất tồn tại liên kết thuốc–bệnh, qua đó nâng cao khả năng diễn giải của mô hình.

- **Mục tiêu 2:** Xây dựng và triển khai các kỹ thuật xử lý dữ liệu chuyên sâu, bao gồm: (i) thuật toán lựa chọn mẫu âm tính đáng tin cậy (thông qua tiền xử lý và gán nhãn thông minh) để hạn chế âm tính giả, và (ii) kỹ thuật cân bằng dữ liệu để giảm thiểu tác động của mất cân bằng lớp.

Nội dung nghiên cứu

Luận án tập trung vào các nội dung chính sau:

- Tổng quan và phân tích các phương pháp học máy, học sâu và học trên mạng không đồng nhất đã và đang được ứng dụng trong bài toán dự đoán mối quan hệ giữa thuốc và bệnh trong tin sinh học.

- Phân tích và đánh giá các thách thức trong bài toán dự đoán liên kết thuốc-bệnh, bao gồm: dữ liệu thưa, mất cân bằng dữ liệu, âm tính giả và hạn chế về khả năng giải thích của mô hình dự đoán.

- Đề xuất khai thác meta-path dựa trên mạng không đồng nhất, cho phép tận dụng các nguồn dữ liệu sinh học khác nhau (thuốc, bệnh, thông tin gen, protein...) để nâng cao hiệu quả dự đoán mối liên kết thuốc – bệnh.

- Đề xuất và phân tích một khung suy luận Bayes để ước lượng xác suất tồn tại liên kết thuốc–bệnh, tích hợp đặc trưng đa nguồn và cấu trúc mạng không đồng nhất.

- Xây dựng hệ thống huấn luyện và đánh giá mô hình, tích hợp các kỹ thuật xử lý dữ liệu và thuật toán học sâu, đảm bảo độ chính xác, độ tin cậy và

khả năng diễn giải.

Phạm vi nghiên cứu

Để hiện thực hóa mục tiêu của luận án, nội dung nghiên cứu được tổ chức theo một chuỗi công việc liên kết chặt chẽ, từ tổng quan phương pháp đến đề xuất mô hình và kiểm chứng thực nghiệm, như sau:

- Nghiên cứu tập trung vào bài toán dự đoán liên kết thuốc-bệnh bằng dựa trên các phương pháp khai thác trên mạng thông tin không đồng nhất
- Phạm vi dữ liệu bao gồm các nguồn dữ liệu công khai và chuẩn hóa trong y sinh như: DrugBank, OMIM, Gottlieb research, v.v. Dữ liệu được tích hợp và chuẩn hóa dưới dạng mạng không đồng nhất drug–protein–disease nhằm khai thác các quan hệ sinh học giữa thuốc, protein và bệnh. Các dạng dữ liệu khác như chuỗi sinh học, cấu trúc phân tử 2D/3D hoặc văn bản sinh học chưa được tích hợp trực tiếp, do phạm vi nghiên cứu của luận án tập trung vào việc khai thác cấu trúc quan hệ trong mạng sinh học không đồng nhất cho bài toán dự đoán drug–disease association.
- Luận án không đi sâu vào thử nghiệm sinh học hay lâm sàng, mà tập trung vào khía cạnh tính toán và thuật toán, nhằm phát triển mô hình có khả năng dự đoán hiệu quả và có tính ứng dụng trong giai đoạn tiền lâm sàng.

Đối tượng nghiên cứu

Để định vị rõ ranh giới và trọng tâm của đề tài, mục sau trình bày đối tượng nghiên cứu của luận án—bao gồm bài toán như sau:

- Các mô hình tính toán và thuật toán học máy, học sâu, và đặc biệt là học trên mạng không đồng nhất được áp dụng trong lĩnh vực tin sinh học.
- Mối quan hệ giữa thuốc và bệnh, bao gồm các cặp thuốc – bệnh đã biết và các mối liên kết tiềm năng được dự đoán dựa trên dữ liệu sinh học.
- Dữ liệu y sinh đa nguồn: thuốc, bệnh, thông tin gen, protein...

Phương pháp nghiên cứu

- Về lý thuyết: Để đạt được mục tiêu nghiên cứu, luận án tập trung tìm hiểu và xây dựng cơ sở lý thuyết về:

  ✓ Mạng dữ liệu không đồng nhất và khái niệm siêu đường dẫn (meta-path), các kỹ thuật khai thác dữ liệu dựa trên meta-path và tích hợp dữ liệu y sinh trên HIN để tính toán mối quan hệ thuốc–bệnh.

  ✓ Nắm vững định lý Bayes và các khái niệm cốt lõi: phân phối tiên nghiệm, phân phối hậu nghiệm, hàm hợp lý, phân phối dự đoán hậu nghiệm; nghiên cứu kỹ thuật suy luận xác suất cho liên kết thuốc–bệnh.

  ✓ Các mối quan hệ tiềm ẩn giữa thuốc, bệnh và các thực thể y sinh khác; khảo cứu phương pháp cân bằng dữ liệu xử lý dữ liệu thưa và mất cân bằng.

  ✓ Kỹ thuật học máy và học sâu: phân loại, phân cụm, lựa chọn đặc trưng để xây dựng mô hình tối ưu và chọn tham số phù hợp.

- Về thực nghiệm:

  Luận án sử dụng ngôn ngữ lập trình Python cùng các công cụ như Spyder, Jupyter Notebook và các thư viện Python chuyên dụng. Dữ liệu được trích xuất từ các nguồn uy tín như DrugBank, OMIM và các bộ dữ liệu y sinh chuẩn khác. Hiệu suất của mô hình đề xuất sẽ được so sánh với các phương pháp hiện có để đánh giá tính chính xác và khả năng tổng quát. Ngoài ra, để nâng cao độ tin cậy, kết quả dự đoán của mô hình sẽ được đối chiếu và xác nhận thông qua các tài liệu khoa học uy tín trong lĩnh vực y sinh.

Đóng góp của luận án

Luận án trình bày hai đóng góp cốt lõi:

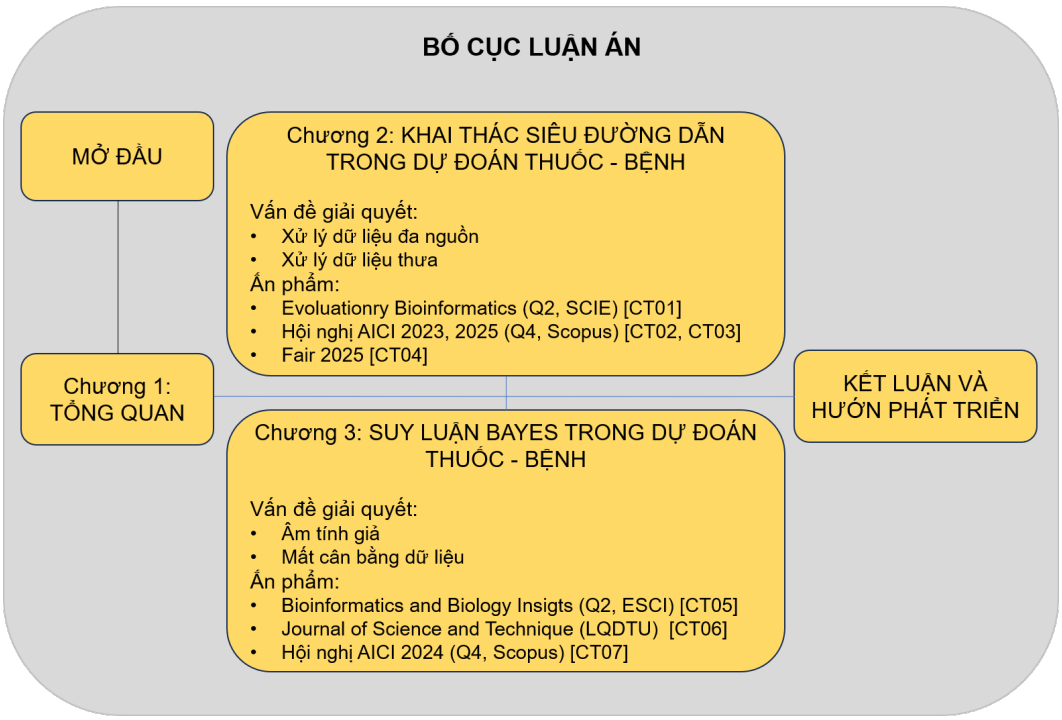
- Đóng góp 1: Khai thác các siêu đường dẫn trong HIN để tính toán mối quan hệ tiềm ẩn giữa thuốc và bệnh từ nhiều nguồn dữ liệu phức tạp, từ đó đề

xuất mô hình dự đoán liên kết thuốc–bệnh tin cậy hơn. Nội dung của đóng góp này được trình bày chi tiết trong Chương 2 của luận án và đã được công bố tại [CT01], [CT02], [CT03], [CT04].

- Đóng góp 2: Đề xuất dùng lý thuyết Bayes phân tích mối liên kết thuốc–bệnh dựa trên HIN thuốc–bệnh và đề xuất phương pháp trích xuất mẫu âm tính chất lượng cao kết hợp kỹ thuật cân bằng dữ liệu cũng như xử lý vấn đề sự mất cân bằng dữ liệu. Phương pháp này góp phần giảm thiểu tỷ lệ sai lệch trong phân loại và nâng cao khả năng tổng quát hóa của mô hình. Nội dung của đóng góp này được trình bày trong Chương 3 và đã được công bố tại: [CT05], [CT06], [CT07].

Cấu trúc luận án

Toàn bộ luận án được trình bày theo cấu trúc logic và chặt chẽ, với mạch nội dung xuyên suốt từ tổng quan lý thuyết, đề xuất phương pháp cho đến thực nghiệm và đánh giá, được minh họa trong Hình 2.



Hình 2: Bố cục luận án

Trong các phần tiếp theo của luận án, nội dung từng phần được bố cục

một cách logic và có cấu trúc chặt chẽ như sau:

Chương 1: Tổng quan.

Chương 1 giới thiệu bài toán dự đoán liên kết thuốc–bệnh, phân tích các thách thức chính: dữ liệu đa tạp, thưa thớt, mất cân bằng, âm tính giả và khả năng giải thích mô hình. Đồng thời, chương trình bày các khái niệm cơ bản về HIN, các phương pháp khai thác như meta-path và suy luận Bayes, cùng các thuật toán học máy, học sâu ứng dụng trong bài toán. Phần tổng quan nghiên cứu hệ thống hóa các hướng tiếp cận hiện có, chỉ ra ưu điểm, hạn chế và khoảng trống, từ đó xác định mục tiêu và hướng tiếp cận cho luận án.

Chương 2: Khai thác siêu đường dẫn trong dự đoán thuốc–bệnh.

Chương này, đề xuất sử dụng phương pháp khai thác và kết hợp các siêu đường dẫn trong mạng thuốc–protein–bệnh. Đầu tiên, luận án xây dựng các mô hình dự đoán tổng hợp dựa trên kỹ thuật bỏ phiếu từ các meta-path cơ sở. Sau đó, mở rộng mạng bằng cách tích hợp thêm các quan hệ đồng nhất (thuốc–thuốc, protein–protein, bệnh–bệnh) và đề xuất ba siêu đường dẫn mới. Các đặc trưng được trích xuất qua giảm chiều dữ liệu, sau đó áp dụng mô hình phân loại để dự đoán liên kết. Nội dung chương dựa trên các công trình đã công bố [**CT01**], [**CT02**], [**CT03**], [**CT04**].

Chương 3: Suy luận bayes trong dự đoán liên kết thuốc - bệnh.

Chương này đánh dấu một bước chuyển phương pháp luận quan trọng: từ việc khai thác đặc trưng đồ thị sang mô hình hóa xác suất nguyên lý dựa trên Suy luận Bayes. Cách tiếp cận mới này cho phép trực tiếp mô hình hóa các cơ chế lan truyền quan hệ thuốc-protein-bệnh và định lượng được độ không chắc chắn. Để giải quyết triệt để các vấn đề nền tảng về chất lượng dữ liệu, luận án đề xuất: (1) một thuật toán lọc mẫu Âm tính chất lượng cao (HNS) dựa trên xác suất Bayes để loại bỏ âm tính giả, và (2) việc áp dụng kỹ thuật Gaussian-SMOTE tối ưu để cân bằng dữ liệu sau khi đã được làm sạch. Việc kết hợp tuần tự hai kỹ thuật này tạo ra hiệu ứng cộng hưởng, dẫn đến một tập dữ liệu huấn luyện tối ưu, giúp mô hình đạt được độ chính xác và khả năng tổng

quát hóa vượt trội, được kiểm chứng qua các thí nghiệm toàn diện và công bố tại [**CT05**], [**CT06**], [**CT07**].

Kết luận và hướng phát triển.

Trong phần này, luận án tóm tắt các đóng góp chính của nghiên cứu, đồng thời phân tích ưu điểm và hạn chế của các phương pháp đã được đề xuất. Trên cơ sở đó, luận án đưa ra các hướng cải thiện và phát triển trong tương lai để tiếp tục nâng cao hiệu quả và khả năng ứng dụng của mô hình trong các bài toán dự đoán mối quan hệ thuốc–bệnh.

CHƯƠNG 1. TỔNG QUAN

Mạng thông tin không đồng nhất (HIN) là mô hình dữ liệu có khả năng mô tả phong phú các loại thực thể và quan hệ phức tạp trong nhiều lĩnh vực, đặc biệt trong y – sinh học. Trong bối cảnh đó, dự đoán liên kết thuốc – bệnh nổi lên như một nhiệm vụ cốt lõi hỗ trợ tái định vị thuốc. Các loại thực thể như thuốc, protein và bệnh, cùng các quan hệ đặc trưng (thuốc điều trị bệnh, thuốc liên kết với protein, protein gây ra bệnh) tạo nên cấu trúc HIN giàu thông tin nhưng cũng đặt ra thách thức về mô hình hóa và khai thác hiệu quả.

Chương 1 của luận án giới thiệu bài toán dự đoán liên kết thuốc – bệnh dựa trên HIN, trình bày các khái niệm cơ bản làm nền tảng cho các chương tiếp theo, và tổng quan các hướng nghiên cứu hiện nay. Đồng thời, chương này mô tả phương pháp tiếp cận tổng quát của luận án — kết hợp khai thác meta-path, tích hợp biểu diễn đặc trưng và áp dụng các kỹ thuật học máy tiên tiến để cải thiện chất lượng dự đoán. Cuối cùng, chương đưa ra các nhận định tổng quan nhằm định vị rõ mục tiêu của luận án trong bối cảnh nghiên cứu hiện tại.

1.1. Bài toán dự đoán liên kết thuốc-bệnh

Bài toán dự đoán liên kết thuốc – bệnh trong HIN đóng một vai trò quan trọng trong việc phát hiện các chỉ định mới tiềm năng cho các dược phẩm hiện có.

Mục tiêu cốt lõi của bài toán là đánh giá khả năng tồn tại của một liên kết giữa một cặp thuốc–bệnh dựa trên các thông tin quan hệ gián tiếp có sẵn trong mạng (ví dụ: thông qua protein, gene), ngay cả khi dữ liệu trực tiếp về tương tác đó còn thiếu hoặc chưa được xác minh. Dự đoán liên kết trong mạng thông tin không đồng nhất là bài toán nhằm ước lượng khả năng tồn tại của một liên kết E_k thuộc loại quan hệ mục tiêu giữa hai đối tượng $v_i \in V_i$ và $v_j \in V_j$

trong mạng dữ liệu không đồng nhất G [24].

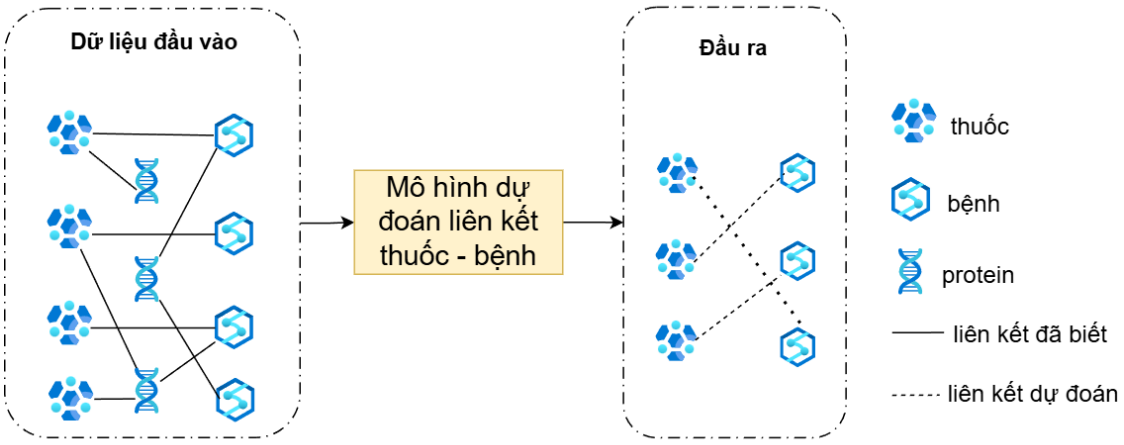
- **Đầu vào:** Một mạng dữ liệu không đồng nhất $G = (V, E)$, trong đó V là tập các đối tượng (ví dụ: thuốc, bệnh, gen, protein, ...) và E là tập các loại quan hệ sinh học đã biết giữa chúng (ví dụ: thuốc - bệnh, bệnh - protein, thuốc - protein Associations, ...).

- **Xử lý:** Áp dụng các phương pháp khai thác đồ thị, kỹ thuật phân rã hoặc hoàn thiện ma trận, hoặc các mô hình học sâu để phân lớp nhằm dự đoán khả năng tồn tại của liên kết mục tiêu.

- **Đầu ra:** Với mỗi cặp thuốc–bệnh tiềm năng (v_d, v_b), mô hình đưa ra dự đoán về sự tồn tại của cạnh:

$$p(E_k) = \begin{cases} 1, & \text{nếu liên kết } E_k \text{ có khả năng tồn tại;} \\ 0, & \text{ngược lại.} \end{cases}$$

Quy trình tổng quan của bài toán được minh họa trong Hình 1.1 và có thể mô tả như sau:



Hình 1.1: Mô tả bài toán dự đoán liên kết thuốc–bệnh.

Trong bối cảnh HIN, các nút đại diện cho các thực thể như thuốc, bệnh, protein, trong khi các cạnh mô tả quan hệ giữa chúng (ví dụ: thuốc - bệnh, thuốc – protein, bệnh – protein). Những quan hệ này có thể phản ánh cơ chế tác động dược lý hoặc sự tương tác sinh học nền tảng. Tuy nhiên, một thách thức lớn là

thông tin về kiểu nút và kiểu quan hệ thường không đầy đủ hoặc chưa được chú giải rõ ràng, khiến cho bài toán dự đoán trở nên phức tạp.

1.2. Các khái niệm liên quan

1.2.1. Mạng thông tin

Một mạng thông tin (Information Network) là một mô hình trừu tượng hóa thế giới thực, trong đó các đối tượng (Entities) được biểu diễn dưới dạng các đỉnh (Vertices) và các tương tác hoặc mối quan hệ giữa chúng được biểu diễn dưới dạng các cạnh (Edges). Cách mô hình hóa dựa trên đồ thị này cho phép biểu diễn có hệ thống các cấu trúc dữ liệu phức tạp, từ đó hỗ trợ hiệu quả cho các nhiệm vụ phân tích mối quan hệ, suy luận tri thức và dự đoán các liên kết tiềm năng trong nhiều lĩnh vực nghiên cứu khác nhau.

Để phân tích sâu hơn cấu trúc và ngữ nghĩa của mạng thông tin, các mạng này thường được phân loại dựa trên mức độ đa dạng của các thành phần cấu thành. Việc phân biệt giữa mạng thông tin đồng nhất và mạng thông tin không đồng nhất là bước quan trọng ban đầu, giúp lựa chọn các phương pháp biểu diễn và khai thác phù hợp với bản chất của dữ liệu.

1.2.2. Mạng thông tin đồng nhất và không đồng nhất

Mạng thông tin đồng nhất

Một mạng thông tin [25] được biểu diễn là đồ thị $G = (V, E)$ với một hàm ánh xạ loại đối tượng $\phi : V \to A$ và một hàm ánh xạ loại liên kết $\psi : E \to R$. Mỗi đối tượng $v \in V$ thuộc về một loại đối tượng cụ thể trong tập A, tức là $\phi(v) \in A$, và mỗi liên kết $e \in E$ thuộc về một loại quan hệ cụ thể trong tập các quan hệ R, tức là $\psi(e) \in R$. Nếu hai liên kết thuộc cùng một loại quan hệ, thì chúng sẽ chia sẻ cùng loại đối tượng bắt đầu và loại đối tượng kết thúc.

Mạng thông tin đồng nhất là trường hợp đặc biệt khi chỉ tồn tại một loại đối tượng và một loại quan hệ ($|A| = 1, |R| = 1$). Ví dụ Hình 1.2(a) biểu diễn

mạng tương tác thuốc-thuốc là một mạng đồng nhất, trong đó mỗi đỉnh biểu diễn một thuốc và mỗi cạnh biểu diễn một tương tác "tương tự" giữa hai thuốc. Loại mạng này cho phép áp dụng trực tiếp nhiều thuật toán đồ thị truyền thống nhưng gặp hạn chế khi cần mô hình hóa các hệ thống đa thực thể.

Mạng thông tin không đồng nhất

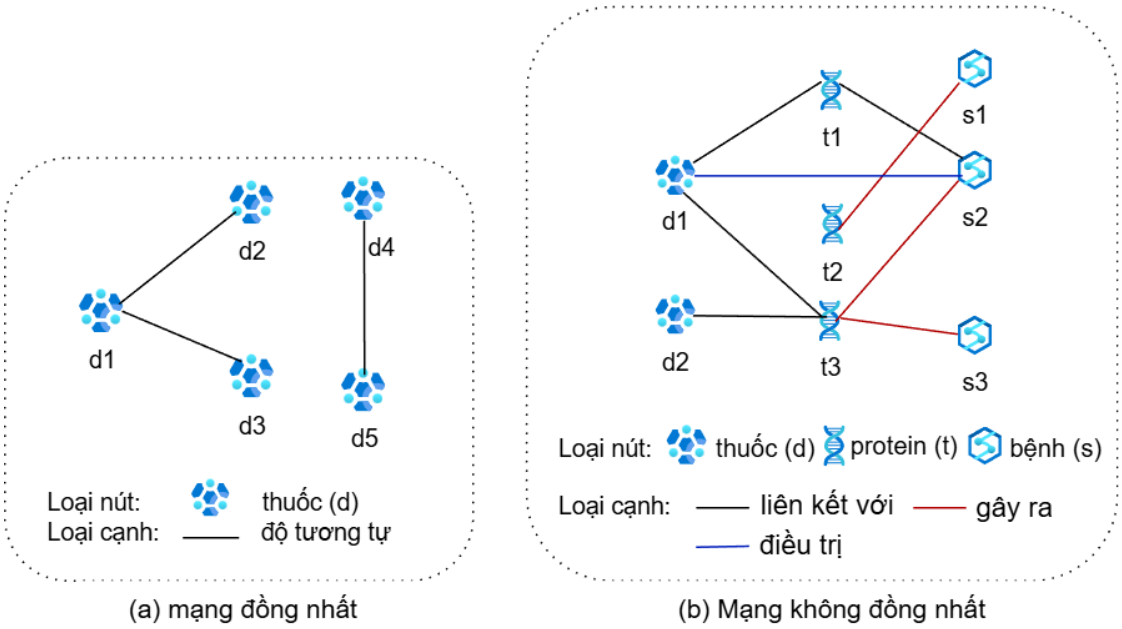
Một mạng thông tin được gọi là mạng thông tin không đồng nhất [25] khi tồn tại nhiều hơn một loại đối tượng ($|A| > 1$) hoặc nhiều hơn một loại quan hệ ($|R| > 1$). Ngược lại, mạng thông tin được gọi là mạng thông tin đồng nhất.

Mạng thông tin không đồng nhất (Heterogeneous Information Network – HIN) cho phép tích hợp nhiều loại thực thể và nhiều kiểu quan hệ khác nhau trong cùng một khuôn khổ thống nhất. Trong bối cảnh thuốc–protein–bệnh, Hình 1.2(a) biểu diễn mạng không đồng nhất thuốc-protein-bệnh. Trong đó các mối liên kết thuốc-bệnh là "điều trị" (thuốc điều trị bệnh), liên kết thuốc-protein là "liên kết với" (thuốc liên kết với protein) và liên kết protein-bệnh là "gây ra" (protein gây ra bệnh hoặc bệnh gây ra bởi protein). Nhờ đó, HIN phản ánh sát hơn bản chất phức tạp của dữ liệu y sinh và tạo điều kiện khai thác các quan hệ gián tiếp giữa thuốc và bệnh. Trong phần tiếp theo, mạng thông tin không đồng nhất sẽ được gọi tắt là mạng không đồng nhất.

1.2.3. Lược đồ mạng

Lược đồ mạng [25], được ký hiệu là $TG = (A, R)$, là một dạng chuẩn tổng quát của một mạng thông tin $G = (V, E)$ với phép ánh xạ loại đối tượng $\phi : V \to A$ và phép ánh xạ loại liên kết $\psi : E \to R$. Lược đồ mạng là một đồ thị có hướng được định nghĩa trên tập các loại đối tượng A, trong đó các cạnh biểu diễn các loại quan hệ R.

Lược đồ mạng đóng vai trò như bản thiết kế tổng quát cho HIN, xác định các kiểu thực thể và các quan hệ hợp lệ giữa chúng. Lược đồ trong Hình 1.3 minh họa cấu trúc của một HIN, thể hiện các loại đối tượng và mối quan hệ



(a) mạng đồng nhất (b) Mạng không đồng nhất

Hình 1.2: Ví dụ về mạng đồng nhất và mạng không đồng nhất

giữa chúng, bao gồm ba loại đối tượng {thuốc, protein, bệnh} và ba loại quan hệ tương ứng { điều trị, gây ra, liên kết với}. Lược đồ này định hướng quá trình xây dựng và khai thác các mẫu quan hệ phức tạp trong mạng.

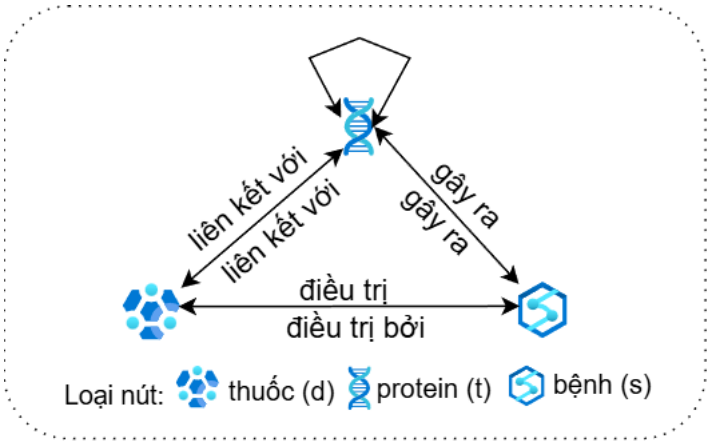
1.2.4. Siêu đường dẫn

Một siêu đường dẫn P [25] là một con đường được định nghĩa trên lược đồ mạng $TG = (A, R)$, và được ký hiệu dưới dạng $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, trong đó xác định một quan hệ tổng hợp $R = R_1 \circ R_2 \circ \cdots \circ R_l$ giữa các loại đối tượng $A_1, A_2, \ldots, A_{l+1}$, với ký hiệu $\circ$ biểu thị phép hợp nhất các quan hệ.

Trong mạng thuốc–protein–bệnh, các siêu đường dẫn cho phép mô hình hóa các mối liên hệ gián tiếp giữa thuốc và bệnh. Hình 1.4 minh họa một ví dụ về siêu đường dẫn trong một mạng không đồng nhất, với các loại nút như thuốc, bệnh, và protein. Cụ thể, có hai siêu đường dẫn:

Siêu đường dẫn đầu tiên (trên cùng):

- thuốc → protein → bệnh
- Đây là một mối quan hệ nối kết giữa thuốc và bệnh thông qua các protein. Siêu đường dẫn này phản ánh mối liên hệ gián tiếp giữa thuốc và bệnh



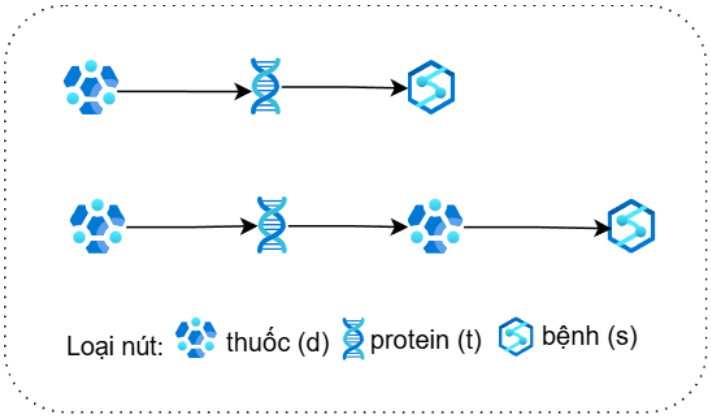
Hình 1.3: Ví dụ về lược đồ mạng.

thông qua protein trung gian, cho thấy khả năng thuốc có tác động đến bệnh thông qua cùng cơ chế sinh học.

Siêu đường dẫn thứ hai (dưới cùng):

- thuốc → protein → thuốc → bệnh
- Đây là một con đường dài hơn, mô tả một mối liên hệ gián tiếp giữa hai loại thuốc và một bệnh, dựa trên cơ chế tác động sinh học chung thông qua các protein và các thuốc trung gian.

Việc khai thác các siêu đường dẫn như vậy giúp tăng cường khả năng suy luận và dự đoán các liên kết thuốc–bệnh tiềm năng.



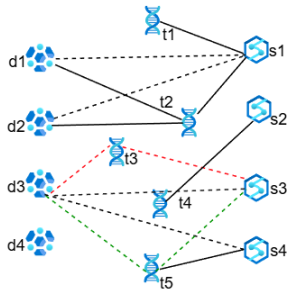
Hình 1.4: Ví dụ về lược đồ mạng.

1.2.5. Ma trận kết hợp của siêu đường dẫn

Cho một mạng thông tin không đồng nhất $G = (V, E)$ với lược đồ mạng TG [25]. Với một siêu đường dẫn $P = T_1 \to T_2 \to \cdots \to T_k$, ma trận kết hợp (commuting matrix) của P được xác định bởi tích các ma trận kề tương ứng:

$$X = A_{T_1T_2} A_{T_2T_3} \cdots A_{T_{k-1}T_k}, \tag{1.1}$$

trong đó $A_{T_iT_j}$ là ma trận kề biểu diễn mối quan hệ giữa các đối tượng thuộc loại T_i và T_j. Phần tử $X(i, j)$ thể hiện số lượng các thể hiện đường đi từ đối tượng $u_i \in T_1$ đến đối tượng $v_j \in T_k$ tuân theo siêu đường dẫn P.



a. Mạng không đồng nhất d-t-s

A_{ds}

	s1	s2	s3	s4
d1	1	0	0	0
d2	1	0	0	0
d3	0	0	**2**	1
d4	0	0	0	0

=

A_{dt}

	t1	t2	t3	t4	t5
d1	0	1	0	0	0
d2	0	1	0	0	0
d3	0	0	1	0	1
d4	0	0	0	0	0

X

A_{st}^{T}

	s1	s2	s3	S4
t1	1	0	0	0
t2	1	0	0	0
t3	0	0	1	0
t4	0	1	0	0
t5	0	0	1	1

b. Ma trận kết hợp của đường đi từ d-t-s

Hình 1.5: Ví dụ về ma trận kết hợp của siêu đường dẫn

Hình 1.5 minh họa ví dụ về ma trận kết hợp trong mạng thông tin không đồng nhất thuốc–protein–bệnh. Cụ thể, Hình 1.5(a) biểu diễn mạng tích hợp gồm ba loại đối tượng: thuốc (d), protein (t) và bệnh (s), với các ma trận kề tương ứng A_{dt}, A_{ts} và A_{ds}. Từ mạng này, Hình 1.5(b) thể hiện ma trận kết hợp A_{ds} tương ứng với siêu đường dẫn $d \to t \to s$, trong đó mỗi phần tử phản ánh mức độ liên kết gián tiếp giữa thuốc và bệnh thông qua protein trung gian.

Ví dụ, phần tử $A_{ds}[3, 3] = 2$ cho biết tồn tại hai đường đi khác nhau kết nối thuốc d_3 với bệnh s_3 theo siêu đường dẫn đã xét, cụ thể là các chuỗi $d_3 \to t_3 \to s_3$ và $d_3 \to t_5 \to s_3$, như minh họa trong Hình 1.5(a). Các giá trị trong ma trận kết hợp do đó có thể được xem như các đặc trưng định lượng, phản ánh cường độ liên kết ngữ nghĩa giữa thuốc và bệnh dựa trên cấu trúc mạng.

1.2.6. Mạng thông tin không đồng nhất thuốc–protein–bệnh

Trong luận án, không gian nghiên cứu được xây dựng dựa trên mạng thông tin không đồng nhất thuốc–protein–bệnh, hình thành từ việc tích hợp ba mạng con: mạng thuốc–bệnh, mạng thuốc–protein và mạng bệnh–protein. Mạng tích hợp này bao gồm các nút đại diện cho thuốc, protein và bệnh, cùng với các cạnh biểu diễn mối quan hệ điều trị, tương tác sinh hóa và liên quan sinh học. *Mạng thông tin thuốc–protein–bệnh được biểu diễn dưới dạng một đồ thị vô hướng:* $G_{dst} = (V_{dst}, E_{dst})$, *trong đó* $V_{dst} = D \cup S \cup T$ *là tập hợp các nút, và* $E_{dst} = E_{ds} \cup E_{dt} \cup E_{st}$ *là tập hợp các cạnh.*

Mạng thuốc–protein–bệnh cung cấp nền tảng để khai thác các quan hệ gián tiếp giữa thuốc và bệnh thông qua protein trung gian. Trên mạng này, các siêu đường dẫn được sử dụng như các mẫu ngữ nghĩa cốt lõi nhằm xây dựng đặc trưng và phát triển các mô hình dự đoán liên kết thuốc–bệnh trong các chương tiếp theo.

1.2.7. Dữ liệu thưa

Dữ liệu thưa (Sparse Data) là một thách thức đáng kể trong lĩnh vực dự đoán liên kết thuốc-bệnh (Drug-Disease Association - DDA), đặc biệt là trong tái định vị thuốc (drug repositioning) [26, 12, 27, 28]. Trong ngữ cảnh dự đoán mối liên hệ giữa thuốc và bệnh, dữ liệu thưa thường xuất hiện do:

- Thông tin không đầy đủ: Nhiều phương pháp dự đoán dựa trên độ tương đồng (similarity-based methods) đòi hỏi dữ liệu đầu vào đầy đủ để tính toán chính xác. Tuy nhiên, trong thực tế, dữ liệu về thuốc và bệnh thường thiếu hụt (incomplete data), khiến cho việc tính toán các độ đo tương đồng trở nên khó khăn hoặc không đáng tin cậy [26].

- Thiếu mẫu âm xác thực: Trong các tập dữ liệu cặp thuốc-bệnh, chỉ có một số tương tác đã được xác minh (mẫu dương) và rất nhiều cặp chưa được dán nhãn. Để huấn luyện mô hình, nhiều phương pháp đơn giản coi các

mẫu không được dán nhãn là mẫu âm, điều này có thể dẫn đến việc đưa vào nhiễu nhân tạo (Artificial Noises) [26], [29].

1.2.8. Mẫu âm tính giả

Trong bài toán dự đoán liên kết thuốc–bệnh phục vụ tái định vị thuốc, các mô hình học máy thường được huấn luyện trên một tập dữ liệu bao gồm hai loại cặp: cặp thuốc–bệnh đã được xác nhận là có liên kết (mẫu dương) và cặp được giả định là không có liên kết (mẫu âm). Vấn đề đặt ra là: trong hầu hết các cơ sở dữ liệu hiện nay, số lượng mẫu dương được xác minh thực nghiệm chiếm tỷ lệ rất nhỏ so với không gian tổ hợp thuốc–bệnh có thể có. Do đó, phần lớn dữ liệu huấn luyện là các cặp chưa xác minh, và mặc định được gán nhãn là âm tính.

Tuy nhiên, việc gán nhãn âm tính cho tất cả các cặp chưa được xác minh là một giả định không hoàn toàn chính xác. Trong số các cặp này có thể tồn tại nhiều liên kết thực sự, nhưng chưa được công bố hoặc phát hiện do giới hạn của nghiên cứu thực nghiệm hiện tại. Việc đánh đồng toàn bộ nhóm chưa xác minh là "không có liên kết" đã tạo ra một vấn đề nghiêm trọng: sự xuất hiện của mẫu âm tính giả.

Việc tồn tại một lượng lớn mẫu âm tính giả không chỉ làm giảm chất lượng dữ liệu mà còn làm trầm trọng thêm vấn đề mất cân bằng dữ liệu vốn đã tồn tại giữa lớp dương (ít) và lớp âm (rất nhiều).

1.2.9. Mất cân bằng dữ liệu

Một đặc điểm nổi bật của bài toán dự đoán liên kết thuốc – bệnh là sự mất cân bằng dữ liệu nghiêm trọng giữa hai lớp: số lượng cặp thuốc – bệnh đã được xác minh thực nghiệm (lớp dương) rất nhỏ so với số lượng cặp chưa được xác minh (lớp âm). Mất cân bằng dữ liệu xảy ra khi một hoặc nhiều lớp – thường là lớp quan trọng cần phát hiện – có số lượng mẫu quá ít so với các lớp còn lại.

Trong dự đoán liên kết thuốc – bệnh, lớp dương đại diện cho các liên kết thực sự, cũng chính là mục tiêu trọng tâm của tái định vị thuốc, lại rơi vào tình trạng thiểu số nghiêm trọng. Ngược lại, lớp âm chiếm đa số dữ liệu nhưng chứa nhiều nhiễu, do trong đó có thể tồn tại các trường hợp âm tính giả.

Khi được huấn luyện trên tập dữ liệu mất cân bằng như vậy, các mô hình học máy có xu hướng thiên lệch về lớp chiếm đa số (lớp âm). Kết quả là mô hình có thể đạt độ chính xác tổng thể cao, nhờ dự đoán đúng phần lớn các mẫu âm, nhưng lại thất bại trong việc phát hiện lớp dương. Điều này làm suy giảm nghiêm trọng giá trị ứng dụng của mô hình trong tái định vị thuốc và dự đoán mối liên kết tiềm năng giữa thuốc – bệnh.

1.2.10. Phân tách giá trị kỳ dị

Phân tách giá trị kỳ dị (Singular Value Decomposition – SVD) là một phép phân tách ma trận quan trọng trong đại số tuyến tính. Giả sử $X \in \mathbb{R}^{n \times m}$, với m là số hàng (số mẫu) và n là số cột (số đặc trưng). SVD phân tách ma trận X thành ba ma trận:

$$X = U \cdot \Sigma \cdot V^T \tag{1.2}$$

- $U \in \mathbb{R}^{n \times n}$ với các cột là các vector kỳ dị bên trái, tạo thành một cơ sở trực chuẩn cho các hồ sơ biểu hiện thử nghiệm.
- $\Sigma \in \mathbb{R}^{n \times m}$, là ma trận đường chéo chứa các giá trị kỳ dị không âm $\sigma_1, \sigma_2, \ldots, \sigma_k$ trên đường chéo chính, các phần tử khác đều bằng 0, thể hiện mức độ quan trọng của từng thành phần trong ma trận dữ liệu.
- $V^T \in \mathbb{R}^{m \times m}$ với các dòng là các vector kỳ dị bên phải (right singular vectors), tạo thành một cơ sở trực chuẩn cho các phản ứng phiên mã của gen.

Mỗi meta-path tạo ra một ma trận có kích thước $n \times m$, trong đó n là số lượng thuốc và m là số lượng bệnh. Đối với mỗi cặp thuốc – bệnh, vector đặc trưng được xây dựng có kích thước $n + m$, bao gồm tất cả thông tin từ cả thuốc và bệnh. Tuy nhiên, kích thước này có thể trở nên quá lớn, làm giảm độ tin cậy

của mô hình do lượng thông tin cần xử lý quá nhiều, dẫn đến hiện tượng quá khớp và giảm hiệu suất dự đoán. Để giải quyết vấn đề này, người ta thường sử dụng phép phân tích giá trị kỳ dị giảm chiều (reduced SVD – trong luận án này gọi chung là SVD) để tập trung vào các thành phần chính có ý nghĩa thống kê và sinh học.

1.3. Cơ sở lý thuyết suy luận Bayes trong tái định vị thuốc

Trong phần này, ký hiệu d, t và s lần lượt đại diện cho thuốc, protein và bệnh, là ba thực thể trung tâm của bài toán tái định vị thuốc dựa trên suy luận Bayes toàn diện [30]. Mục tiêu của tái định vị thuốc là ước lượng xác suất một thuốc tiềm năng d có thể điều trị một bệnh s, được ký hiệu bởi:

$$p(d \mid s). \tag{1.3}$$

Quá trình dự đoán tương tác thuốc–bệnh có thể được mô hình hóa như bài toán khai thác đồ thị trên một mạng không đồng nhất, trong đó các nút biểu diễn thuốc, protein và bệnh. Protein đóng vai trò cầu nối quan trọng vì thuốc thường tác động thông qua tương tác với protein, trong khi protein lại liên quan trực tiếp đến sự xuất hiện hoặc tiến triển của nhiều bệnh. Thực tế cho thấy nhiều bệnh có nguồn gốc từ các đột biến tại vùng giao diện liên kết hoặc từ các biến đổi dị lập thể trong cấu trúc protein [31].

Dựa trên quy tắc tích của xác suất (product rule) và xét cấu trúc phụ thuộc phổ biến giữa ba thực thể thuốc–protein–bệnh, ta có thể phân rã xác suất chung của bộ ba (d, t, s) thành các xác suất có điều kiện. Giả sử rằng thuốc d phụ thuộc vào protein t và protein t phụ thuộc vào bệnh s, khi đó:

$$p(d \wedge t \wedge s) = p(d, t, s) = p(d \mid t)\, p(t \mid s)\, p(s). \tag{1.4}$$

Sử dụng định nghĩa của xác suất có điều kiện và quy tắc cộng (marginalization), ta thu được:

$$p(d \mid s) = \frac{p(d,s)}{p(s)} = \sum_t \frac{p(d,t,s)}{p(s)}. \tag{1.5}$$

Thay thế biểu thức $p(d,t,s)$ từ (1.4) vào (1.5), ta thu được dạng biểu diễn theo mối quan hệ thuốc–protein và protein–bệnh:

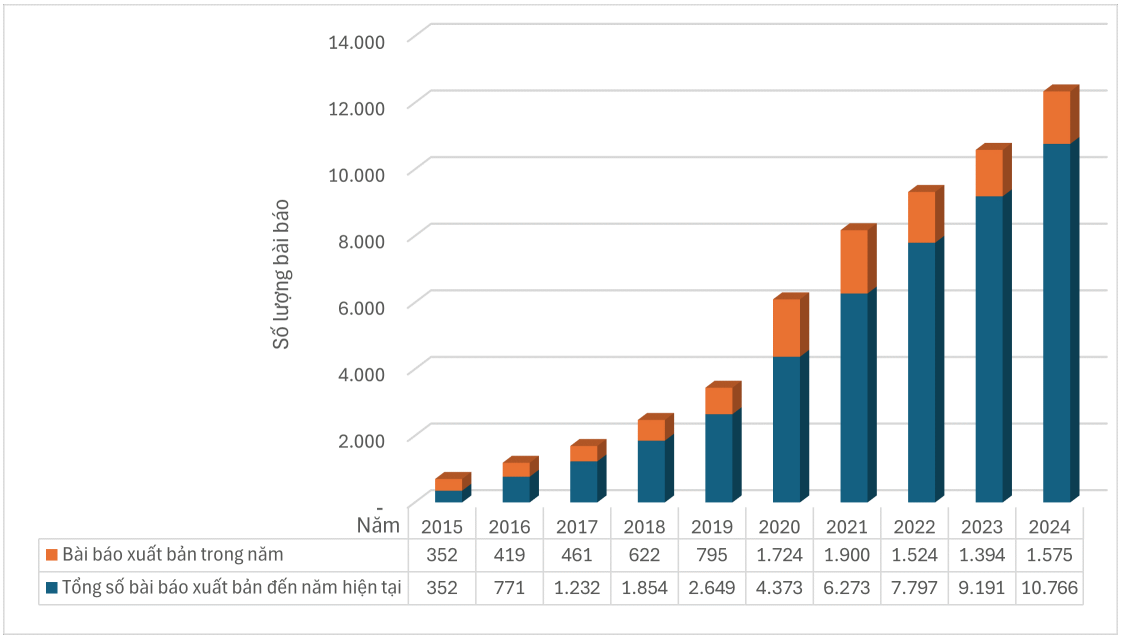
$$p(d \mid s) = \sum_t p(d \mid t)\, p(t \mid s). \tag{1.6}$$

Biểu thức (1.6) cho thấy vai trò trung tâm của protein trong quá trình suy luận Bayes: xác suất thuốc có thể điều trị bệnh được xác định thông qua tích của (i) xác suất thuốc tương tác với protein và (ii) xác suất protein liên quan đến bệnh. Do đó, suy luận Bayes cung cấp một cơ chế giải thích minh bạch và nhất quán để đánh giá các ứng viên thuốc mới dựa trên thông tin đã được xác nhận về thuốc, protein và bệnh.

1.4. Tổng quan về tình hình nghiên cứu

Hình 1.6 trình bày phân bố theo năm của số lượng bài báo khoa học liên quan đến "drug repositioning" hoặc "drug repurposing" được tìm thấy trên PubMed trong giai đoạn 2015–2024. Dữ liệu được thu thập thông qua truy vấn sử dụng hai từ khóa chính nêu trên. Có thể thấy số lượng công bố tăng trưởng mạnh trong thập kỷ qua, đặc biệt từ năm 2020 trở đi. Cụ thể, giai đoạn 2015–2019, số bài báo mỗi năm đều dưới 800; từ năm 2020 số lượng công bố tăng vọt, đạt 1.724 bài (2020) và 1.900 bài (2021). Đến năm 2024, số bài báo vẫn duy trì ở mức cao (1.575 bài), đưa tổng số công bố trong giai đoạn 10 năm lên 10.766 bài.

Xu hướng tăng trưởng này phản ánh hai yếu tố quan trọng. Thứ nhất, nhu cầu cấp bách về các chiến lược điều trị mới trong bối cảnh những thách thức y tế toàn cầu như COVID-19 đã thúc đẩy mạnh mẽ các nghiên cứu tái định vị thuốc. Thứ hai, sự phát triển bùng nổ của các cơ sở dữ liệu sinh – dược học quy mô lớn và các phương pháp tính toán hiện đại dựa trên trí tuệ nhân



Năm	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Bài báo xuất bản trong năm	352	419	461	622	795	1.724	1.900	1.524	1.394	1.575
Tổng số bài báo xuất bản đến năm hiện tại	352	771	1.232	1.854	2.649	4.373	6.273	7.797	9.191	10.766

Hình 1.6: Phân bố bài báo từ năm 2015 đến năm 2024

tạo, học máy và phân tích dữ liệu lớn đã tạo điều kiện thuận lợi cho việc sàng lọc và dự đoán các tương tác thuốc–bệnh tiềm năng ở quy mô toàn cục.

HIN là một dạng mạng đồ thị bao gồm nhiều loại đối tượng (nút) và nhiều loại mối quan hệ (cạnh) khác nhau (Sun và cộng sự, 2011). Trong bối cảnh y sinh, các thực thể thường gặp gồm: thuốc, bệnh, gen, protein, tác dụng phụ, con đường chuyển hóa..., cùng các mối quan hệ giữa chúng (thuốc–tương tác–protein, gen–liên quan–bệnh, thuốc–điều trị–bệnh,...). Ưu điểm của HIN là cho phép biểu diễn đồng thời cấu trúc liên kết và ngữ nghĩa phong phú của dữ liệu; mỗi loại nút và cạnh mang một ý nghĩa sinh học riêng, giúp các thuật toán có thể khai thác không chỉ topo mạng mà còn bản chất sinh học của tương tác. Khi mô hình hóa bài toán dự đoán liên kết thuốc–bệnh trong HIN, ta có thể tận dụng đồng thời nhiều nguồn kiến thức dị thể, vượt xa việc chỉ dùng một ma trận tương tác nhị phân đơn lẻ.

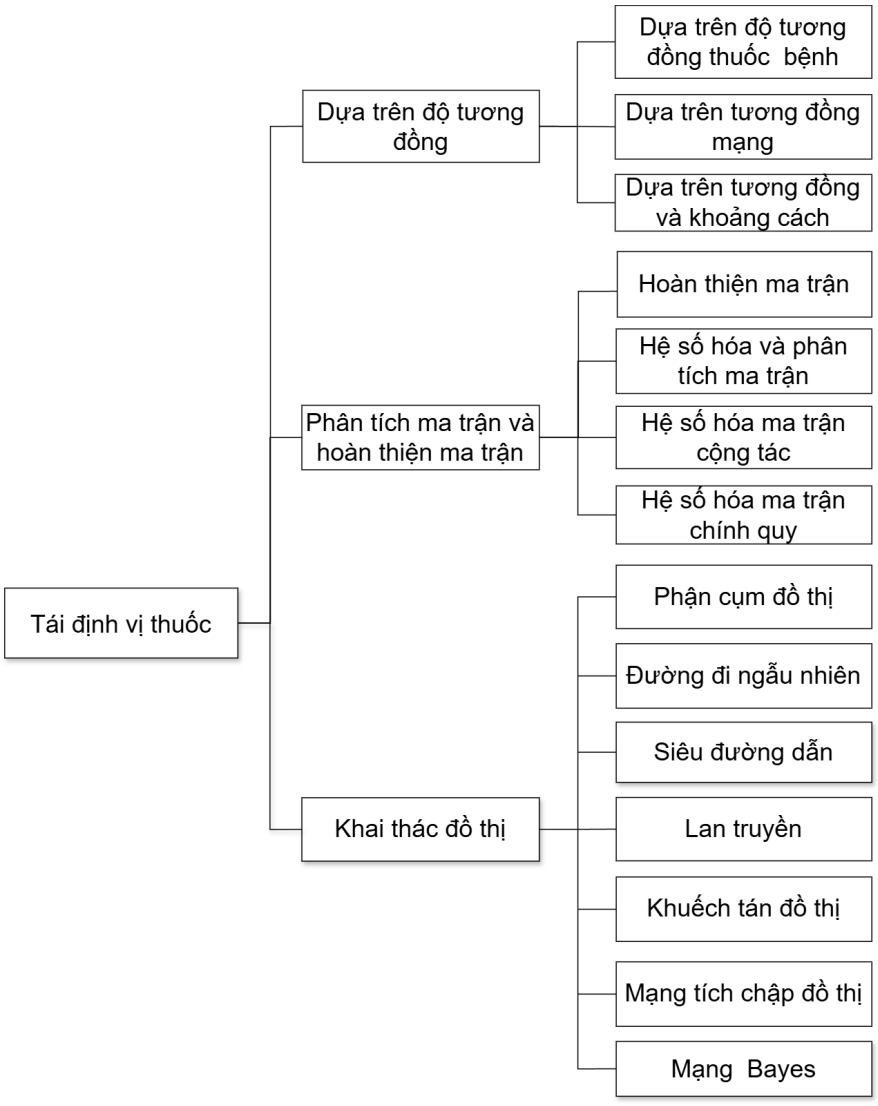
Theo các nghiên cứu tổng quan của Maryam Bagherian [32], Yoonbee Kim [33] và Lijun Cai [34], các phương pháp dự đoán liên kết trong lĩnh vực tái định vị thuốc có thể chia thành bốn nhóm chính:

1. Các phương pháp dựa trên độ tương đồng

2. Các phương pháp dựa trên hoàn thiện và phân rã ma trận

3. Các phương pháp dựa trên khai thác đồ thị

Mỗi nhóm phương pháp có những ưu điểm và hạn chế riêng. Hình 1.7 mô tả tổng quan các nhóm phương pháp dự đoán liên kết thuốc–bệnh; nội dung chi tiết được trình bày ở các mục tiếp theo.



Hình 1.7: Tổng quan về tình hình nghiên cứu

1.4.1. Các mô hình dựa trên tương đồng và khoảng cách

Các phương pháp dựa trên mạng và tương đồng chủ yếu xuất phát từ giả thuyết *guilt-by-association* [26, 35] : các thực thể có đặc trưng hoặc ngữ cảnh

tương tự (thuốc–thuốc, bệnh–bệnh, thuốc–mục tiêu. . .) có xu hướng chia sẻ các liên kết sinh học gần nhau. Trong bối cảnh dự đoán liên kết thuốc–bệnh, điều này có thể hiểu là: (i) hai thuốc có cấu trúc hóa học hoặc phổ tác dụng sinh học tương tự thường điều trị các bệnh tương tự; (ii) hai bệnh có cơ chế sinh bệnh hoặc kiểu hình lâm sàng gần nhau thường được điều trị bởi các thuốc tương tự; và (iii) thuốc có xu hướng tác động lên các protein nằm trong cùng mạng lưới phân tử với các protein gây bệnh. Phần này trình bày các nhóm phương pháp chính dựa trên tương đồng và khoảng cách, cùng với đánh giá hạn chế của chúng.

Phương pháp dựa trên độ tương đồng

Các phương pháp dựa trên độ tương đồng khai thác các đặc trưng đã biết của thuốc và bệnh, bao gồm cấu trúc hóa học, hồ sơ tác động sinh học, mục tiêu phân tử và các chú giải ngữ nghĩa, nhằm dự đoán các liên kết thuốc–bệnh tiềm năng.

Cheng và cộng sự (2018) [5] sử dụng thước đo độ lân cận trên mạng (network proximity metric) để đánh giá mối quan hệ giữa các mục tiêu thuốc trong mạng tương tác protein–protein, qua đó phát hiện các cơ chế tiềm năng cho chỉ định thuốc mới, tác dụng phụ hoặc hiệu quả bổ trợ của các thuốc đã được phê duyệt. Sibilio và cộng sự (2021) [7] đề xuất một thước đo độ tương đồng mạng mới để định lượng mức độ tương tác giữa mục tiêu thuốc và các protein liên quan đến bệnh, tập trung vào các cặp thuốc–bệnh nằm trong cùng vùng lân cận của mạng.

Nhìn chung, nhóm phương pháp này cho thấy hiệu quả trong việc phân biệt các cặp thuốc–bệnh đã được xác nhận và tiềm năng, đồng thời có khả năng diễn giải tương đối cao do dựa trên các quan hệ tương đồng rõ ràng. Tuy nhiên, hiệu quả của các phương pháp này phụ thuộc mạnh vào chất lượng và mức độ đầy đủ của dữ liệu tương đồng, dễ gặp hiện tượng quá khớp khi dữ liệu dương và âm không cân bằng, và bị hạn chế khi dự đoán cho các thuốc mới chưa có thông tin về mục tiêu phân tử.

Phương pháp dựa trên độ tương đồng mạng

Các phương pháp dựa trên độ tương đồng mạng tập trung khai thác cấu trúc topo của mạng nhằm suy luận các mối quan hệ tiềm ẩn giữa các thực thể thuộc những lớp dữ liệu khác nhau.

Chen và cộng sự (2015) [36] phát triển hai phương pháp suy luận dựa trên cấu trúc topo mạng để dự đoán các liên kết thuốc–bệnh tiềm năng, được mở rộng từ các kỹ thuật gợi ý trong hệ thống khuyến nghị do Zhou và cộng sự (2010) [37] đề xuất. Qi và Liang (2021) [38] đề xuất một khung nhúng mạng trong mạng đa lớp, kết hợp với kỹ thuật lấy mẫu âm nhằm xử lý tình trạng mất cân bằng giữa dữ liệu dương và âm, trên cơ sở phương pháp lọc cộng tác. Wang và cộng sự (2022) [39] đề xuất mô hình kết hợp tuyến tính các độ tương đồng giữa thuốc–mục tiêu và mục tiêu–thuốc, tận dụng lọc cộng tác cùng với thông tin cấu trúc mạng đã biết để tính điểm tương đồng.

Các phương pháp này cho phép khai thác hiệu quả các đặc trưng tiềm ẩn từ cấu trúc mạng và hỗ trợ suy luận các liên kết mới vượt ra ngoài các tương tác trực tiếp đã được quan sát. Tuy nhiên, chúng vẫn phụ thuộc vào các ma trận tương đồng ban đầu, dễ bị ảnh hưởng bởi nhiễu dữ liệu và gặp khó khăn khi xử lý các thực thể mới chưa có thông tin tương đồng, còn gọi là bài toán khởi đầu lạnh (cold-start).

Các phương pháp dựa trên tương đồng và khoảng cách

Một nhóm phổ biến khác là các phương pháp dựa trên độ tương đồng và khoảng cách, trong đó các hàm đo Euclidean, cosine hoặc các độ đo thiết kế riêng cho cấu trúc hóa học, tính chất dược lý hoặc đặc trưng topo trong mạng đa phần tử (*Multipartite Network*) được sử dụng để đánh giá mức độ gần gũi giữa các thực thể [8, 9, 7]. Các phương pháp này thường tìm các cặp gần nhất trong không gian đặc trưng hoặc trên mạng để suy luận các liên kết tiềm năng.

Ưu điểm của nhóm này là mô hình đơn giản, dễ triển khai và dễ diễn giải. Tuy nhiên, do lượng tương tác đã biết và các đặc trưng chú giải còn hạn

chế, phần lớn dữ liệu vẫn ở trạng thái chưa gán nhãn, dẫn đến hiện tượng thưa và hạn chế khả năng suy luận cho các thuốc hoặc bệnh chưa được nghiên cứu nhiều.

Đánh giá chung

Nhóm phương pháp dựa trên tương đồng và khoảng cách cung cấp một nền tảng logic quan trọng cho việc suy luận các liên kết thuốc–bệnh, đồng thời hỗ trợ khám phá các cơ chế dược lý tiềm năng. Tuy nhiên, chúng tồn tại các hạn chế chung: (i) phụ thuộc vào các thông tin tương đồng đã biết của thuốc và mục tiêu; (ii) dễ quá khớp khi dữ liệu dương/âm không cân bằng; và (iii) khó mở rộng cho các thực thể mới hoặc ít dữ liệu. Vì vậy, xu hướng gần đây là kết hợp các phương pháp này với phân rã ma trận, khai thác đồ thị hoặc học sâu để khắc phục các hạn chế trên.

1.4.2. Các mô hình dựa trên hoàn thiện và phân rã ma trận

Trong bối cảnh tái định vị thuốc, việc dự đoán các tương tác thuốc–bệnh thường được biểu diễn dưới dạng một ma trận nhị phân hoặc bán liên tục, trong đó mỗi phần tử biểu thị sự tồn tại (hoặc mức độ tin cậy) của một liên kết. Do phần lớn các phần tử chưa được quan sát, một hướng tiếp cận tự nhiên là xem bài toán này như một bài toán hoàn thiện ma trận (Matrix Completion – MC) hoặc phân rã ma trận (Matrix Factorization – MF). Các phương pháp này tận dụng thông tin từ ma trận tương tác đã biết, đồng thời khai thác cấu trúc tiềm ẩn để suy luận các phần tử chưa biết, phù hợp với hoàn cảnh dữ liệu thiếu nhãn phổ biến trong y dược học.

Hoàn thiện ma trận

Hoàn thiện ma trận (MC) là kỹ thuật điền các phần tử chưa biết trong ma trận dựa trên các quan sát hiện có, thường dưới giả định ma trận có hạng thấp. Trong tái định vị thuốc, MC được dùng để suy luận các chỉ định thuốc

tiềm năng bằng cách bổ sung các giá trị còn thiếu trong ma trận tương tác thuốc–bệnh.

Yang và cộng sự (2019) [40] đề xuất phương pháp OMC (Overlap Matrix Completion), trong đó xử lý hai mạng riêng biệt cho thuốc và bệnh thay vì một mạng duy nhất. OMC2 mở rộng OMC bằng cách kết hợp ma trận tương tác với ma trận tương đồng thuốc/bệnh và sử dụng BNNR (Bayesian Nuclear Norm Regularization) để đảm bảo giá trị dự đoán nằm trong khoảng [0,1]. OMC3 tiếp tục mở rộng bằng việc tích hợp thông tin thuốc–protein và bệnh–protein, nâng cao độ chính xác dự đoán. Yan và cộng sự (2022) [41] kết hợp MC với học đa quan sát (multi-view learning) và điều chuẩn Laplacian (Laplacian graph regularization) để xây dựng các ma trận tương đồng tích hợp, từ đó suy luận các chỉ định tiềm năng.

Nhìn chung, MC hiệu quả trong việc tận dụng cấu trúc hạng thấp và bổ sung các phần tử chưa biết, nhưng dễ rơi vào cực tiểu cục bộ khi dữ liệu nhiễu hoặc thiếu nhãn nặng, đồng thời gặp khó khăn trong việc xây dựng các độ đo tương đồng tối ưu cho cả thuốc và bệnh.

Phân rã ma trận

Phân rã ma trận (MF) phân tách ma trận tương tác thuốc–bệnh thành tích của hai (hoặc nhiều) ma trận hạng thấp hơn, cho phép học các nhân tố tiềm ẩn đại diện cho thuốc và bệnh. Phân rã ma trận Logistic (LMF) là một biến thể quan trọng, trong đó hàm logistic được sử dụng để mô hình hóa xác suất tương tác.

Zhang và cộng sự (2020) [42] đề xuất mô hình DRIMC (Drug Repositioning by Inductive Matrix Completion), sử dụng thông tin đặc trưng phụ trợ (side information) của thuốc và bệnh. Nhóm tác giả kết hợp các độ tương đồng hóa học, miền mục tiêu, chú giải và GIP kernel để xây dựng mạng tương đồng, sau đó cập nhật các ma trận nhân tố tiềm ẩn bằng LMF và dùng tích của chúng để dự đoán các cặp thuốc–bệnh tiềm năng. Zhang và cộng sự (2018) [43] đề xuất SCMFDD (Similarity Constrained Matrix Factorization for Drug–Disease

Association Prediction), tích hợp thông tin cấu trúc thuốc và ngữ nghĩa bệnh vào ràng buộc tương đồng, nhằm nâng cao độ chính xác dự đoán.

Ban và cộng sự (2019) [44] đề xuất NRLMF (Neighborhood Regularized Logistic Matrix Factorization), sử dụng cơ chế *rescoring* dựa trên phân phối Bernoulli để cải thiện khả năng dự đoán cho các cặp thuốc–bệnh có ít dữ liệu tương tác. Wang và Yan (2020) [45] áp dụng RLMD (Regularized Logistic Matrix Decomposition), trong đó kết hợp nhiều loại độ tương đồng giữa thuốc và bệnh để nâng cao hiệu quả dự đoán.

LMF và các biến thể tối ưu trực tiếp bài toán dự đoán nhị phân, đặc biệt hiệu quả khi dữ liệu thưa. Tuy nhiên, chúng đòi hỏi lựa chọn tham số điều chuẩn hợp lý để tránh quá khớp hoặc thiếu khớp, và vẫn gặp hạn chế trong việc tích hợp đầy đủ cấu trúc mạng HIN phức tạp.

Hệ số hóa ma trận cộng tác

Các phương pháp hệ số hóa ma trận cộng tác (CMF) và các biến thể như MSBMF khai thác mối quan hệ cộng tác giữa thuốc và bệnh bằng cách chiếu chúng vào một không gian đặc trưng hạng thấp chung, đồng thời tích hợp nhiều ma trận tương đồng bổ sung.

Yang và cộng sự (2020) [46] đề xuất mô hình Multi-similarities Bi-linear Matrix Factorization (MSBMF), sử dụng nhiều ma trận tương đồng thuốc–thuốc và bệnh–bệnh để tối ưu ma trận đặc trưng của thuốc và bệnh. Phương pháp áp dụng ADMM (Alternating Direction Method of Multipliers) để tối ưu từng biến khi cố định các biến còn lại, và dự đoán các cặp thuốc–bệnh dựa trên tích các ma trận đặc trưng. Xia và cộng sự (2019) [47] đề xuất SPLCMF (Self-Paced Learning with Correlated Matrix Factorization), kết hợp cơ chế học tự điều chỉnh theo tiến độ (Self-paced Learning) với xấp xỉ hạng thấp có trọng số và tích hợp nhiều mạng liên quan vào Regularized Least Squares (RLS) để cải thiện độ chính xác dự đoán.

Các phương pháp CMF và MSBMF cho phép tích hợp linh hoạt nhiều

nguồn dữ liệu và khai thác hiệu quả mối quan hệ cộng tác giữa thuốc và bệnh. Tuy nhiên, độ phức tạp tính toán cao và yêu cầu lựa chọn tham số cẩn trọng là những hạn chế chính.

Hệ số hóa ma trận chính quy

Các phương pháp hệ số hóa ma trận chính quy (RMF) cải tiến MF bằng cách bổ sung các điều chuẩn đồ thị (Graph Regularization), giúp khai thác các mẫu phi tuyến, bậc thấp và thông tin cấu trúc giữa các đối tượng.

Zhang (2020) [48] đề xuất GRGMF (Graph Regularized Generalized Matrix Factorization), khai thác thông tin từ nút lân cận và xây dựng mô hình như một dạng phân tích ma trận tổng quát có điều chuẩn đồ thị. Mô hình áp dụng cơ chế học thích ứng để thu nhận biểu diễn tiềm ẩn của từng nút. Lian và cộng sự (2021) [49] sử dụng chính quy hóa đồ thị hai phần (Bipartite Graph Regularization) để biểu diễn thuốc và bệnh trong không gian chiều thấp, kết hợp Stream Learning với Sparse Learning để loại bỏ dữ liệu không cần thiết và nâng cao hiệu quả dự đoán. Wang và Wang (2021) [50] sử dụng tương đồng hóa học giữa các thuốc và tương đồng chuỗi gen giữa các bệnh, kết hợp Laplacian pairwise regularization để biểu diễn các vector nhân tố tiềm ẩn và xác định tương tác tiềm năng qua điểm số logistic.

RMF vừa khai thác đặc trưng tiềm ẩn, vừa tích hợp trực tiếp cấu trúc giữa các đối tượng, giúp cải thiện dự đoán trong môi trường dữ liệu phức tạp. Tuy nhiên, các mô hình này có chi phí tính toán cao và đòi hỏi kiểm soát nhiễu tốt.

Nhận xét chung và đánh giá

Tổng hợp lại, các phương pháp MC, LMF, CMF, MSBMF và RMF – cùng với các mô hình cụ thể như DRRS, OMC, DRIMC, SCMFDD, NRLMF, RLMD – đều nhằm khai thác cấu trúc tiềm ẩn trong ma trận tương tác thuốc–bệnh để suy luận các chỉ định mới. Ưu điểm chính của nhóm phương pháp này là:

- Cho phép dự đoán các tương tác tiềm năng mà không cần đầy đủ thông tin lý–hóa hoặc chú giải chức năng;
- Tích hợp được nhiều nguồn thông tin (đa ma trận tương đồng, đa mạng liên quan);
- Giảm bậc ma trận, xử lý hiệu quả dữ liệu thưa và giảm chi phí tính toán.

Tuy nhiên, các phương pháp này cũng tồn tại hạn chế: dễ rơi vào cực tiểu cục bộ (đặc biệt trong MC), nhạy cảm với việc lựa chọn độ đo tương đồng, khó tích hợp trực tiếp cấu trúc HIN đầy đủ và có độ phức tạp cao khi số lượng nguồn dữ liệu tăng.

1.4.3. Khai thác đồ thị

Trong những năm gần đây, khai thác đồ thị (graph mining) trên HIN đã trở thành một hướng nghiên cứu quan trọng trong dự đoán liên kết thuốc–bệnh. Phương pháp này tận dụng trực tiếp cấu trúc HIN, tích hợp thông tin thuốc, bệnh, protein, gen và các đặc trưng sinh học khác để phát hiện các tương tác tiềm năng. Các mô hình khai thác đồ thị có thể chia thành năm nhóm chính: (i) phân cụm trên đồ thị, (ii) meta-path/meta-graph, (iii) random walk/bi-random walk, (iv) lan truyền trên HIN, và (v) khuếch tán trên đồ thị.

Phân cụm trên đồ thị

Một trong những nghiên cứu sớm về dự đoán liên kết thuốc–bệnh dựa trên HIN là của Wu và cộng sự [51]. Nhóm tác giả xây dựng HIN gồm thuốc và bệnh, trong đó dữ liệu thuốc và bệnh được trích xuất từ KEGG Medicus [52], các liên kết thuốc–bệnh được đánh giá bằng hệ số Jaccard dựa trên gen chung, tương đồng thuốc–thuốc và bệnh–bệnh được đo từ các đặc trưng sinh học chia sẻ (quá trình sinh học, pathway, kiểu hình). Sau đó, phương pháp phân cụm với mở rộng lân cận chồng lấn (ClusterONE) [53] được áp dụng lên HIN có trọng số để phát hiện các cụm và suy luận các liên kết thuốc–bệnh mới. Gần đây,

Wang và cộng sự [54] sử dụng phân cụm mẫu (sample clustering) để xác định đặc trưng gen liên quan đến tái định vị thuốc điều trị ung thư, củng cố hiệu quả của cách tiếp cận phân cụm trong khai thác đồ thị y sinh.

Siêu đường dẫn

Wu và cộng sự [26] đề xuất mô hình EMP-SVD (ensemble meta-paths and singular value decomposition), một phương pháp dựa trên đường đi (path-based) trong HIN. Khác với nhiều phương pháp dựa trên mạng khác, EMP-SVD không sử dụng độ tương đồng thuốc–thuốc hoặc bệnh–bệnh, do các độ đo này có thể khác biệt đáng kể tùy đặc trưng. Thay vào đó, mô hình chỉ khai thác các quan hệ thuốc–bệnh, bệnh–protein và thuốc–protein để xây dựng HIN thống nhất. Từ mạng này, năm meta-path đặc trưng được định nghĩa, ví dụ:

- thuốc → bệnh
- thuốc → protein → bệnh
- thuốc → protein → thuốc → bệnh
- thuốc → bệnh → thuốc → bệnh
- thuốc → bệnh → protein → bệnh

Với mỗi meta-path, một ma trận "commuting" được xây dựng và các đặc trưng tiềm ẩn được trích xuất thông qua SVD. Các đặc trưng này sau đó được đưa vào bộ phân loại Random Forest, giúp giảm overfitting bằng việc kết hợp nhiều cây quyết định. Để khắc phục tình trạng chỉ có mẫu dương và các cặp chưa biết (không có mẫu âm rõ ràng), EMP-SVD lựa chọn mẫu âm từ các cặp ít khả năng có liên kết (không chia sẻ protein chung).

Nhiều phương pháp mở rộng khác dựa trên meta-path/meta-graph cũng được đề xuất. AMGDTI [55] sử dụng cơ chế Meta-Graph thích ứng nhằm tự động tìm kiếm meta-graph phù hợp cho bài toán dự đoán liên kết. MHTAN-DTI [56] tận dụng mô hình Transformer phân cấp kết hợp attention dựa trên

meta-path để học biểu diễn vector chiều thấp, cải thiện độ chính xác dự đoán liên kết thuốc–đích (DTI).

Bên cạnh đó, nhiều nghiên cứu gần đây tiếp tục khai thác meta-path kết hợp với các mô hình học sâu nhằm cải thiện khả năng biểu diễn của mạng dị thể. Ding và cộng sự [57] đề xuất MAPTrans, sử dụng kiến trúc Transformer với cơ chế mutual attention và chiến lược đánh giá tầm quan trọng của các meta-path để học tương tác giữa thuốc và bệnh. DRMGNE [58] kết hợp meta-path với Graph Convolutional Network (GCN) để khai thác cấu trúc topo của mạng sinh học, đồng thời áp dụng chiến lược adaptive negative enhancement nhằm nâng cao chất lượng mẫu âm. Huang và cộng sự [59] đề xuất MPHAM, sử dụng cơ chế hierarchical attention để tổng hợp thông tin từ nhiều meta-path trong HIN, qua đó cải thiện khả năng nắm bắt ngữ nghĩa giữa thuốc và bệnh. So với các phương pháp học biểu diễn sâu, luận án tập trung khai thác trực tiếp cấu trúc meta-path kết hợp với lan truyền thông tin trên mạng và phân rã ma trận, nhằm xây dựng mô hình có độ phức tạp tính toán thấp hơn và khả năng diễn giải tốt hơn.

Đường đi ngẫu nhiên trên đồ thị

Các phương pháp bước ngẫu nhiên (random walk) mô phỏng quá trình di chuyển trên mạng, trong đó xác suất chuyển đến các nút lân cận được xác định theo trọng số hoặc cấu trúc. Ý tưởng là các nút được "ghé thăm" thường xuyên hơn từ một nút nguồn sẽ có mối liên hệ chặt chẽ hơn với nút đó.

Luo và cộng sự [60] giới thiệu phương pháp MBiRW (Similarity measures and bi-random walk), trong đó độ tương đồng thuốc–thuốc và bệnh–bệnh được tính từ đặc trưng và liên kết thuốc–bệnh đã biết. Mô hình cập nhật mạng thuốc và mạng bệnh qua hai bước: (i) điều chỉnh độ tương đồng bằng hàm logistic để nhấn mạnh các cặp mang thông tin; (ii) phân cụm để tăng cường tương đồng trong cùng cụm. Sau đó, thuật toán bi-random walk được thực hiện trên mạng thuốc và bệnh theo hai phương trình hồi quy lặp, kết hợp kết quả để dự đoán liên kết mới. Liu và cộng sự [61] phát triển TP-NRWRH (Two-Pass Network-based

Random Walk with Restart on Heterogeneous network), mở rộng NRWRH [62] với hai giai đoạn random walk: hướng thuốc và hướng bệnh; kết quả được tích hợp bằng giá trị trung bình.

Khác với MBiRW sử dụng độ dài bước cố định, Wang và cộng sự [63] đề xuất DR-IBRW (drug repositioning based on individual bi-random walk), trong đó độ dài bước được xác định riêng cho từng nút nhằm phản ánh đóng góp khác nhau trong quá trình truyền tải thông tin. DR-IBRW khai thác dấu vân tay hóa học, triệu chứng bệnh, kernel GIP và chỉ số Jaccard sửa đổi để đo tương đồng, rồi áp dụng random walk hai chiều với khởi động lại để dự đoán liên kết.

Gần đây, một số mô hình lai (hybrid models) đã kết hợp random walk với các kỹ thuật học sâu nhằm cải thiện khả năng học biểu diễn trên mạng sinh học. Chẳng hạn, RWRGDR [64] kết hợp Random Walk with Restart (RWR) với GraphSAGE và cơ chế attention, trong đó RWR được sử dụng để tạo các đặc trưng ngữ cảnh ban đầu, còn mạng nơ-ron học các biểu diễn cuối cùng phục vụ dự đoán liên kết. Tương tự, MultiXVERSE [65] áp dụng RWR trên mạng đa lớp nhằm khám phá các quan hệ sinh học mang tính nhân quả, và các dự đoán của mô hình được kiểm chứng thông qua thí nghiệm sinh học.

Lan truyền trên đồ thị

Lan truyền trên đồ thị là nhóm các thuật toán trong đó thông tin (nhãn, điểm số, tài nguyên,...) được truyền từ một tập nút nguồn đã biết sang các nút khác trong mạng, qua đó suy ra mức độ liên quan của các nút chưa biết.

Wang và cộng sự [66] đề xuất TL_HGBI (triple-layer heterogeneous graph-based inference), kết hợp đồng thời dự đoán liên kết thuốc–bệnh và thuốc–đích. Mô hình xây dựng mạng ba lớp gồm thuốc, bệnh và gen, liên kết qua tương đồng thuốc–thuốc, bệnh–bệnh, gen–gen. Thuật toán lan truyền thông tin được áp dụng để cập nhật trọng số các cặp thuốc–bệnh, thuốc–gen và bệnh–gen, từ đó dự đoán liên kết tiềm năng. DrugNet [67] mở rộng bằng cách áp dụng lan truyền thông tin trên các mạng tương đồng thuốc, bệnh và protein, sử dụng ProphNet [68]. Quá trình lan truyền luân phiên trong nội mạng và liên mạng

cho đến khi hội tụ, điểm dự đoán cuối cùng được xác định dựa trên lượng thông tin tích lũy. Các mô hình như DRAGNN [69] khai thác thêm cơ chế tổng hợp thông tin láng giềng và truyền thông điệp, vốn là nguyên lý cốt lõi trong lan truyền trên đồ thị.

Zhao và cộng sự (2021) [70] đề xuất HINGRL – một mô hình dựa trên HIN để dự đoán chỉ định mới của thuốc. Phương pháp xây dựng HIN từ ba mạng thông tin (thuốc–bệnh, thuốc–protein, protein–bệnh), tích hợp kiến thức sinh học và dùng chiến lược học biểu diễn cùng mô hình rừng ngẫu nhiên để dự đoán các liên kết chưa biết. Yajie và cộng sự (2022) [71] đề xuất DRWBNCF, một mô hình lọc cộng tác thần kinh có trọng số, xây dựng ba mạng (thuốc–bệnh, tương đồng thuốc–thuốc, tương đồng bệnh–bệnh) và áp dụng tích chập song tuyến có trọng số (weighted bilinear graph convolution) để tích hợp thông tin; sau đó dùng MLP với α-balanced focal loss và graph regularization để dự đoán, đạt hiệu quả cao trên ba bộ dữ liệu.

Khuếch tán trên đồ thị

Tương tự lan truyền, các mô hình khuếch tán (diffusion) mô phỏng các quá trình vật lý như khuếch tán nhiệt, trong đó tài nguyên ban đầu tại một số nút sẽ lan ra toàn mạng đến khi đạt trạng thái cân bằng, phản ánh tầm ảnh hưởng toàn cục của các nút nguồn.

Xie và cộng sự [72] đề xuất BGMSDDA (Bipartite graph diffusion with multiple similarity integration) để dự đoán liên kết thuốc–bệnh, trong đó ma trận liên kết được tái xây dựng bằng thuật toán k-láng giềng có trọng số, tích hợp nhiều độ đo tương đồng (GIP kernel và linear neighborhood similarity), rồi thực hiện khuếch tán trên mạng hai phía (bipartite graph). Dựa trên tài nguyên lan truyền hai chiều thuốc–bệnh, điểm dự đoán cuối cùng được tính toán. Junkai và cộng sự (2023) [73] đề xuất AMDGT, một khung đa mô thức sử dụng dual-graph transformer để tích hợp đồng thời dữ liệu tương đồng và thông tin sinh–hóa phức tạp, kết hợp attention-aware modality interaction để học biểu diễn sâu hơn. Eslami và cộng sự [74] giới thiệu DTINet, một pipeline

mạnh cho dự đoán liên kết thuốc–đích (DTI) từ HIN, kết hợp random walk with restart với diffusion component analysis để tạo biểu diễn vector chiều thấp, sau đó suy luận các liên kết thuốc–đích tiềm năng. Các phương pháp random walk và diffusion đều dựa trên một ma trận chuyển tiếp Markov duy nhất được xây dựng từ ma trận kề của đồ thị. Ma trận này không phân biệt loại quan hệ và coi tất cả các cạnh có ý nghĩa như nhau. Trong HIN, mỗi loại quan hệ lại mang ngữ nghĩa riêng (ví dụ: "thuốc–protein" khác với "protein–bệnh"), do đó việc sử dụng một ma trận chuyển tiếp đồng nhất sẽ làm mất thông tin ngữ cảnh và không phản ánh được cấu trúc không đồng nhất của mạng.

Mạng tích chập đồ thị

GCN mở rộng CNN sang dữ liệu đồ thị, học biểu diễn cho các nút dựa trên cấu trúc mạng. Lijun và cộng sự (2021) [75] đề xuất DRHGCN, sử dụng tích chập đồ thị để trích xuất đặc trưng liên miền (inter-domain) và nội miền (intra-domain) từ các mạng thuốc–thuốc, bệnh–bệnh và thuốc–bệnh, sau đó hợp nhất song song để thu được biểu diễn toàn diện hơn. Yu và cộng sự (2021) [76] kết hợp GCN với cơ chế chú ý để học cách biểu diễn thuốc–bệnh từ HIN. Cai và cộng sự (2021) [77] phân tích đặc trưng liên miền/nội miền qua GCN để thu được biểu diễn đại diện hơn. Yu và cộng sự (2021) [78] phát triển LAGCN (layer attention graph convolutional network), áp dụng attention giữa các lớp GCN nhằm kết hợp hiệu quả biểu diễn từ nhiều tầng và giảm ảnh hưởng của dữ liệu không cân bằng.

GCN mạnh trong khai thác cấu trúc đồ thị và học biểu diễn nút, nhưng yêu cầu chi phí tính toán đáng kể trên HIN lớn và dễ gặp hiện tượng làm mịn quá mức (over-smoothing) khi tăng số lớp.

Suy luận Bayes

Suy luận Bayes cung cấp một khuôn khổ thống nhất cho việc tích hợp dữ liệu đa nguồn và mô hình hóa bất định trong dự đoán liên kết thuốc–bệnh.

Các nghiên cứu sử dụng phương pháp Bayesian trong định vị lại thuốc đã đạt được những kết quả đáng chú ý. [79] kết hợp thông tin về sự tương đồng hóa học và di gen của thuốc và protein mục tiêu bằng phương pháp Bayesian để ước tính xác suất tương tác thuốc-mục tiêu, cải thiện độ chính xác trong môi trường dữ liệu thưa thớt. [80] sử dụng phương pháp Bayesian với xấp xỉ biến phân để cải thiện tính toán và định lượng sự không chắc chắn khi kết hợp các nguồn dữ liệu phụ qua ma trận kernel. [12] áp dụng Hoàn thiện ma trận suy luận Bayes theo phương pháp suy luận quy nạp (DRIMC) để kết hợp dữ liệu sự tương đồng của thuốc và bệnh, từ đó hoàn thiện ma trận và tăng cường độ chính xác dự đoán các mối quan hệ thuốc-bệnh chưa biết.

Nhận xét chung và hạn chế

Các phương pháp khai thác đồ thị thể hiện hiệu quả vượt trội trong tích hợp dữ liệu đa nguồn và dự đoán các liên kết thuốc–bệnh chưa biết. Tuy nhiên, chúng còn tồn tại một số hạn chế:

1. Chi phí tính toán phụ thuộc mạnh vào kích thước mạng; với mạng quá lớn, các bước random walk, propagation hay SVD trở nên tốn kém về thời gian và bộ nhớ;

2. Nhiều mô hình (đặc biệt meta-path, meta-graph) yêu cầu thiết kế thủ công, phụ thuộc vào kinh nghiệm chuyên gia và có thể mang tính chủ quan;

3. Các phương pháp random walk và diffusion thường chưa khai thác đầy đủ sự không đồng nhất trong trọng số và bối cảnh của dữ liệu;

4. Việc xây dựng tập mẫu âm trong DTI/DDA thường không đầy đủ, dễ gây thiên lệch trong huấn luyện và dự đoán.

Nhìn chung, khai thác đồ thị là hướng đi giàu tiềm năng nhưng vẫn cần các phương pháp cải tiến để xử lý mạng quy mô lớn, tích hợp thông tin đa dạng và tự động hóa lựa chọn meta-path/meta-graph.

1.5. Phương pháp đánh giá

Mục này trình bày khung đánh giá thực nghiệm cho mô hình dự đoán liên kết thuốc–bệnh, bao gồm các phương pháp xác thực và các chỉ số đo lường hiệu suất.

Xác thực chéo

Xác thực chéo (cross-validation) là phương pháp phổ biến trong học máy nhằm đánh giá khả năng tổng quát hóa của mô hình trên dữ liệu chưa thấy. Dữ liệu được chia thành k phần bằng nhau; mô hình được huấn luyện k lần, mỗi lần dùng một phần làm tập kiểm tra và k–1 phần còn lại làm tập huấn luyện. Hiệu suất cuối cùng được tính bằng trung bình kết quả của k lần kiểm tra.

Trong tái định vị thuốc – nơi dữ liệu liên kết thuốc–bệnh thường thưa thớt – xác thực chéo đặc biệt hữu ích trong việc hạn chế hiện tượng overfitting và đánh giá độ ổn định của mô hình. Trong luận án này, tất cả các thí nghiệm sử dụng 5-fold cross-validation.

Ma trận nhầm lẫn và các chỉ số đánh giá hiệu suất

Trong bài toán phân loại, hiệu suất của mô hình thường được đánh giá dựa trên ma trận nhầm lẫn (confusion matrix), với bốn thành phần cơ bản:

- TP (True Positive): Số lượng các cặp thuốc–bệnh có liên kết thật sự và mô hình dự đoán đúng là có liên kết.
- TN (True Negative): Số lượng các cặp thuốc–bệnh không có liên kết và mô hình dự đoán đúng là không liên kết.
- FP (False Positive): Số lượng các cặp thuốc–bệnh không có liên kết, nhưng mô hình dự đoán sai là có liên kết.
- FN (False Negative): Số lượng các cặp thuốc–bệnh có liên kết thực sự, nhưng mô hình dự đoán sai là không liên kết.

Từ các giá trị này, nhiều chỉ số quan trọng được rút ra:

- Độ nhạy (Sensitivity, SE):

$$SE = \frac{TP}{TP + FN}, \tag{1.7}$$

- Độ đặc hiệu (Specificity, SP):

$$SP = \frac{TN}{FP + TN}, \tag{1.8}$$

- Trung bình hình học (G-Mean):

$$G\text{-}Mean = \sqrt{SP \times SE}, \tag{1.9}$$

- Độ chính xác tổng thể (Accuracy, ACC):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{1.10}$$

- Độ hồi tưởng (Recall, REC):

$$REC = \frac{TP}{TP + FN}, \tag{1.11}$$

- Độ chính xác (Precision, PRE):

$$PRE = \frac{TP}{TP + FP}, \tag{1.12}$$

- Chỉ số F1 (F1-score):

$$F1\text{-score} = \frac{2 \times REC \times PRE}{REC + PRE}, \tag{1.13}$$

- Hệ số tương quan Matthews (Matthews Correlation Coefficient – MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{1.14}$$

Đường cong đánh giá tổng quát

AUC (Area Under the ROC Curve) phản ánh khả năng phân biệt hai lớp của mô hình bằng cách xem xét quan hệ giữa TPR (True Positive Rate) và FPR (False Positive Rate) khi thay đổi ngưỡng phân loại. AUC tiến gần 1 cho thấy năng lực phân loại tốt, và ít bị ảnh hưởng bởi sự mất cân bằng dữ liệu.

AUPR (Area Under the Precision–Recall Curve) tập trung vào hiệu suất đối với lớp dương tính (lớp thiểu số)—đặc điểm quan trọng trong tái định vị

thuốc. Chỉ số này đo diện tích dưới đường cong Precision–Recall và thường nhạy hơn AUC khi tỷ lệ mẫu dương rất thấp.

Việc kết hợp các thước đo như AUC, AUPR, F1-score và MCC cho phép đánh giá mô hình một cách toàn diện, khách quan.

So sánh với các phương pháp khác

Để đánh giá hiệu quả của mô hình đề xuất, các kết quả thực nghiệm được so sánh với nhiều phương pháp dự đoán liên kết thuốc–bệnh đã được công bố trước đó. Việc so sánh này giúp đánh giá mức độ cải thiện của mô hình đề xuất so với các phương pháp hiện có.

Phân tích cắt bỏ

Phân tích cắt bỏ được sử dụng để đánh giá đóng góp của từng thành phần trong mô hình. Bằng cách lần lượt loại bỏ hoặc thay đổi từng thành phần (ví dụ meta-path, chiến lược chọn mẫu âm hoặc bước giảm chiều), có thể xác định mức độ ảnh hưởng của mỗi thành phần đến hiệu suất tổng thể của mô hình.

Kiểm định thống kê

Để đánh giá mức độ tin cậy của kết quả thực nghiệm, luận án sử dụng kiểm định thống kê nhằm xác định xem sự khác biệt giữa các phương pháp có ý nghĩa thống kê hay không. Các kiểm định này giúp đảm bảo rằng sự cải thiện hiệu suất của mô hình đề xuất không phải do ngẫu nhiên.

Các nghiên cứu điển hình (Case study)

Ngoài các đánh giá định lượng, luận án còn thực hiện các nghiên cứu trường hợp nhằm kiểm tra tính hợp lý sinh học của các dự đoán có xác suất cao. Các cặp thuốc–bệnh được dự đoán được đối chiếu với tài liệu y sinh hoặc cơ sở dữ liệu chuyên ngành để xác minh khả năng ứng dụng thực tiễn của mô hình.

1.6. Kết luận chương 1

Chương 1 đã trình bày tổng quan về bài toán dự đoán liên kết thuốc–bệnh, một nhiệm vụ trọng tâm hỗ trợ tái định vị thuốc. Các khái niệm nền tảng về mạng thông tin không đồng nhất (HIN), siêu đường dẫn (meta-path) và ma trận kết hợp đã được giới thiệu, tạo cơ sở lý thuyết cho các phương pháp tiếp theo. Phần tổng quan tình hình nghiên cứu đã hệ thống hóa và phân tích các hướng tiếp cận chính, đồng thời chỉ ra những thách thức còn tồn tại như dữ liệu thưa, mất cân bằng lớp và vấn đề âm tính giả.

Trên cơ sở đó, luận án xác định ba hướng tiếp cận chính nhằm nâng cao hiệu quả dự đoán:

- Khai thác siêu đường dẫn và ứng dụng suy luận Bayes để tăng cường khả năng phát hiện các mối quan hệ tiềm ẩn giữa thuốc và bệnh trên HIN.
- Phát triển các kỹ thuật xử lý dữ liệu để giải quyết vấn đề mất cân bằng lớp.
- Đề xuất phương pháp lọc và xây dựng tập mẫu nhằm giảm thiểu ảnh hưởng của âm tính giả.

Cuối cùng, chương này cũng đã thiết lập khung phương pháp luận, bao gồm môi trường thực nghiệm và các chỉ số đánh giá hiệu suất mô hình, làm cơ sở cho việc kiểm chứng các đề xuất trong các chương tiếp theo.

CHƯƠNG 2. KHAI THÁC SIÊU ĐƯỜNG DẪN TRONG DỰ ĐOÁN THUỐC–BỆNH

Kế thừa phân tích về thách thức dữ liệu thưa và các mối quan hệ phức tạp, ẩn giấu giữa thuốc, protein và bệnh từ Chương 1, Chương này đề xuất một hướng tiếp cận mới dựa trên HIN. Trọng tâm là khai thác siêu đường dẫn để mã hóa các quan hệ tiềm ẩn giữa thuốc và bệnh, từ đó bổ sung thông tin và nâng cao hiệu quả dự đoán liên kết thuốc–bệnh.

Để hiện thực hóa mục tiêu này, luận án tập trung vào hai giải pháp then chốt:

1. Khai thác quan hệ gián tiếp thông qua siêu đường dẫn cơ bản: Đề xuất xây dựng các siêu đường dẫn mới để mô hình hóa các tương tác sinh học gián tiếp. Các siêu đường dẫn này sẽ được chuyển đổi thành các đặc trưng có thể học và phân lớp với mô hình LightGBM, nhằm mở rộng không gian tìm kiếm và phát hiện các liên kết tiềm năng bị che khuất trong dữ liệu thưa thớt.

2. Tăng cường ngữ nghĩa thông qua tích hợp ma trận đồng nhất: Để bổ sung bối cảnh và phản ánh tương quan nội tại giữa các thực thể cùng loại, luận án tích hợp sáu ma trận đồng nhất (thuốc–thuốc, bệnh–bệnh, protein–protein). Từ đó, xây dựng các nhóm siêu đường dẫn mới kết hợp cả quan hệ trực tiếp lẫn gián tiếp, giúp mô hình tăng cường độ chính xác và khả năng khái quát hóa.

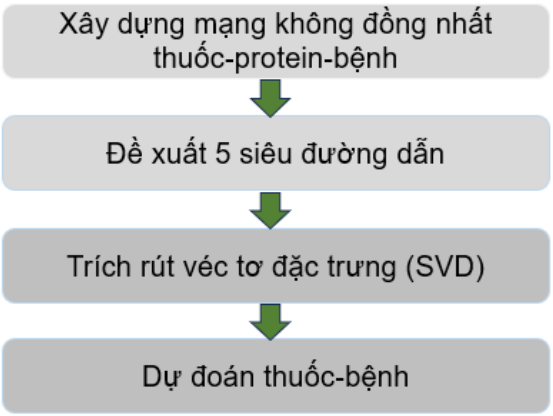
Dựa trên phân tích những hạn chế của EMP-SVD, luận án sẽ trình bày các phương pháp cải tiến nhằm khai thác hiệu quả hơn cấu trúc HIN.

2.1. Mô hình EMP-SVD

2.1.1. Giới thiệu EMP-SVD

Mô hình EMP-SVD [26] kết hợp các kỹ thuật meta-path, SVD và Random Forest, tích hợp dữ liệu tương tác giữa các thực thể sinh học như thuốc, protein

và bệnh để dự đoán mối liên kết thuốc và bệnh. Mô hình EMP-SVD được thực hiện qua các bước như Hình 2.1 và được mô tả như sau:



Hình 2.1: Sơ đồ quy trình mô hình EMP-SVD

Bước 1. Xây dựng mạng dữ liệu không đồng nhất

Mạng không đồng nhất được thiết lập dựa trên ba loại nút: thuốc, protein và bệnh. Các cạnh biểu diễn quan hệ thuốc–protein A_{dt}, bệnh–protein A_{st} và thuốc–bệnh A_{ds}. Bộ dữ liệu gồm:

- 1.186 thuốc, 1.147 protein, 449 bệnh;
- 4.642 tương tác thuốc–protein, 1.365 tương tác bệnh–protein, và 1.827 liên kết thuốc–bệnh đã biết.

Mạng thu được có cấu trúc thưa (sparse), phản ánh mối tương tác sinh học đa dạng.

Bước 2. Đề xuất 5 siêu đường dẫn và các ma trận kết hợp

Để khai thác mối quan hệ tiềm ẩn giữa thuốc và bệnh, năm meta-path (độ dài ≤ 3) được xác định, mỗi meta-path tương ứng với một ma trận và được gán nhãn là W_1, W_2, W_3, W_4 và W_5, được thể hiện trong các công thức (2.1) đến (2.5) như sau:

$$\text{Meta-path-1: Thuốc} \rightarrow \text{Bệnh} \quad W_1 = C_{ds}, \tag{2.1}$$

$$\text{Meta-path-2: Thuốc} \rightarrow \text{Protein} \rightarrow \text{Bệnh} \quad W_2 = C_{dt} \times {C_{st}}^T, \tag{2.2}$$

$$\text{Meta-path-3: Thuốc} \rightarrow \text{Protein} \rightarrow \text{Thuốc} \rightarrow \text{Bệnh} \quad W_3 = C_{dt} \times {C_{dt}}^T \times C_{ds}, \tag{2.3}$$

$$\text{Meta-path-4: Thuốc} \rightarrow \text{Bệnh} \rightarrow \text{Thuốc} \rightarrow \text{Bệnh} \quad W_4 = C_{ds} \times {C_{ds}}^T \times C_{ds}, \tag{2.4}$$

$$\text{Meta-path-5: Thuốc} \rightarrow \text{Bệnh} \rightarrow \text{Protein} \rightarrow \text{Bệnh} \quad W_5 = C_{ds} \times C_{st} \times {C_{st}}^T, \tag{2.5}$$

Phần tử $W(i, j)$ của ma trận kết hợp W biểu thị số lượng đường đi từ thuốc d_i đến bệnh s_j theo siêu đường dẫn tương ứng. Các meta-path này giúp phản ánh các cơ chế sinh học tiềm ẩn khác nhau giữa thuốc và bệnh.

Bước 3. Rút trích đặc trưng bằng SVD

Các ma trận meta-path thường có kích thước lớn và chứa nhiều thông tin dư thừa, khiến việc sử dụng trực tiếp làm đặc trưng dễ dẫn đến quá khớp. Để khắc phục điều này, EMP-SVD áp dụng phân rã giá trị kỳ dị (SVD) nhằm nén thông tin từ ma trận vào một không gian có cấu trúc gọn hơn. SVD được thực hiện theo công thức:

$$W = U \Sigma V^T. \tag{2.6}$$

Quá trình phân rã giúp trích xuất các đặc trưng tiềm ẩn đại diện cho thuốc và bệnh, hỗ trợ mô hình học hiệu quả hơn.

Bước 4. Xây dựng và tổng hợp các mô hình phân lớp

Mỗi *meta-path* sinh ra một bộ đặc trưng riêng và được huấn luyện bằng một bộ phân lớp Random Forest (RF) độc lập. Mỗi mô hình cơ sở trả về xác suất dự đoán liên kết thuốc–bệnh.

Giả sử:

- x là vector đặc trưng của một cặp thuốc–bệnh chưa gán nhãn.
- $h_i(x)$, với $i = 1, 2, \ldots, 5$, là xác suất dự đoán từ mô hình cơ sở thứ i.

Các xác suất từ năm mô hình cơ sở được kết hợp bằng phương pháp *voting* trung bình, cho xác suất cuối cùng của liên kết thuốc–bệnh như được thể hiện ở công thức (2.7) như sau:

$$H(x) = \frac{1}{5} \sum_{i=1}^{5} h_i(x) \tag{2.7}$$

Trong đó, $H(x)$ là xác suất dự đoán cuối cùng cho cặp thuốc–bệnh x.

2.1.2. Phân tích hạn chế

Qua phân tích và so sánh hiệu suất của các meta-path trong [26], có thể rút ra một số hạn chế sau:

Thứ nhất, các meta-path không đi qua protein (Meta-Path-1 và Meta-Path-4) cho kết quả dự đoán thấp hơn đáng kể. Nguyên nhân chính là do protein đóng vai trò trung gian quan trọng trong nhiều quá trình sinh học, việc bỏ qua thông tin này làm suy giảm khả năng khái quát hóa của mô hình.

Thứ hai, trong năm meta-path được đề xuất, số lần khai thác quan hệ thuốc–bệnh (6 lần) vượt trội so với quan hệ thuốc–protein và bệnh–protein (mỗi loại chỉ 3 lần). Đặc biệt, quan hệ thuốc–protein chỉ xuất hiện trong 2 trên 5 meta-path. Sự thiên lệch này dẫn đến việc chưa tận dụng đầy đủ các nguồn thông tin tiềm ẩn trong mạng dữ liệu.

Thứ ba, mô hình EMP-SVD chủ yếu tập trung vào quan hệ không đồng

nhất (thuốc – bệnh, thuốc – protein, bệnh – protein) mà chưa khai thác quan hệ đồng nhất như thuốc–thuốc, bệnh–bệnh hay protein–protein. Điều này hạn chế khả năng mô hình hóa đầy đủ cấu trúc và ngữ nghĩa phức tạp trong HIN.

Thứ tư, vấn đề âm tính giả tuy đã được đề cập với bộ lọc mẫu âm tính, nhưng cách tiếp cận này vẫn còn đơn giản, chưa khai thác sâu các đặc trưng sinh học để đảm bảo tính tin cậy.

Thứ năm, vấn đề mất cân bằng dữ liệu được xử lý theo cách truyền thống: chọn ngẫu nhiên một số lượng mẫu âm bằng số lượng mẫu dương. Cách làm này tuy thuận tiện nhưng có thể bỏ sót các mẫu tiềm năng quan trọng, ảnh hưởng đến chất lượng dự đoán.

Tóm lại, mô hình EMP-SVD được đánh giá cao nhờ khả năng tích hợp dữ liệu từ nhiều nguồn khác nhau và giảm chiều dữ liệu hiệu quả, song vẫn còn tồn tại những hạn chế trong việc khai thác ngữ nghĩa phức tạp và xử lý dữ liệu.

2.1.3. Định hướng phát triển trong luận án:

EMP-SVD được lựa chọn làm mô hình nền tảng của luận án vì hai lý do chính.

Thứ nhất, mô hình này thể hiện ưu điểm nổi bật trong việc khai thác cấu trúc của mạng thông tin không đồng nhất (HIN) thông qua hệ thống siêu đường dẫn (meta-path), đồng thời sử dụng SVD để giảm chiều hiệu quả, góp phần xử lý vấn đề dữ liệu thưa và nhiễu trong các ma trận tương tác sinh học. Nhờ đó, EMP-SVD có khả năng khái quát hóa tốt và dễ dàng mở rộng sang nhiều dạng dữ liệu sinh học khác nhau.

Thứ hai, mặc dù có nhiều điểm mạnh, EMP-SVD vẫn tồn tại những hạn chế quan trọng. Cụ thể, mô hình chưa khai thác đầy đủ vai trò trung gian của protein trong các cơ chế sinh học, dẫn đến việc một số meta-path không đạt hiệu quả mong muốn. Ngoài ra, mức độ sử dụng các loại quan hệ trong HIN còn mất cân bằng, và nhóm quan hệ đồng nhất (thuốc–thuốc, bệnh–bệnh, protein–protein) hoàn toàn chưa được xem xét. Bên cạnh đó, dù EMP-SVD đã

đề cập đến vấn đề âm tính giả, mô hình vẫn chưa xử lý triệt để bài toán mất cân bằng dữ liệu—vốn là đặc trưng cố hữu của bài toán dự đoán liên kết sinh học.

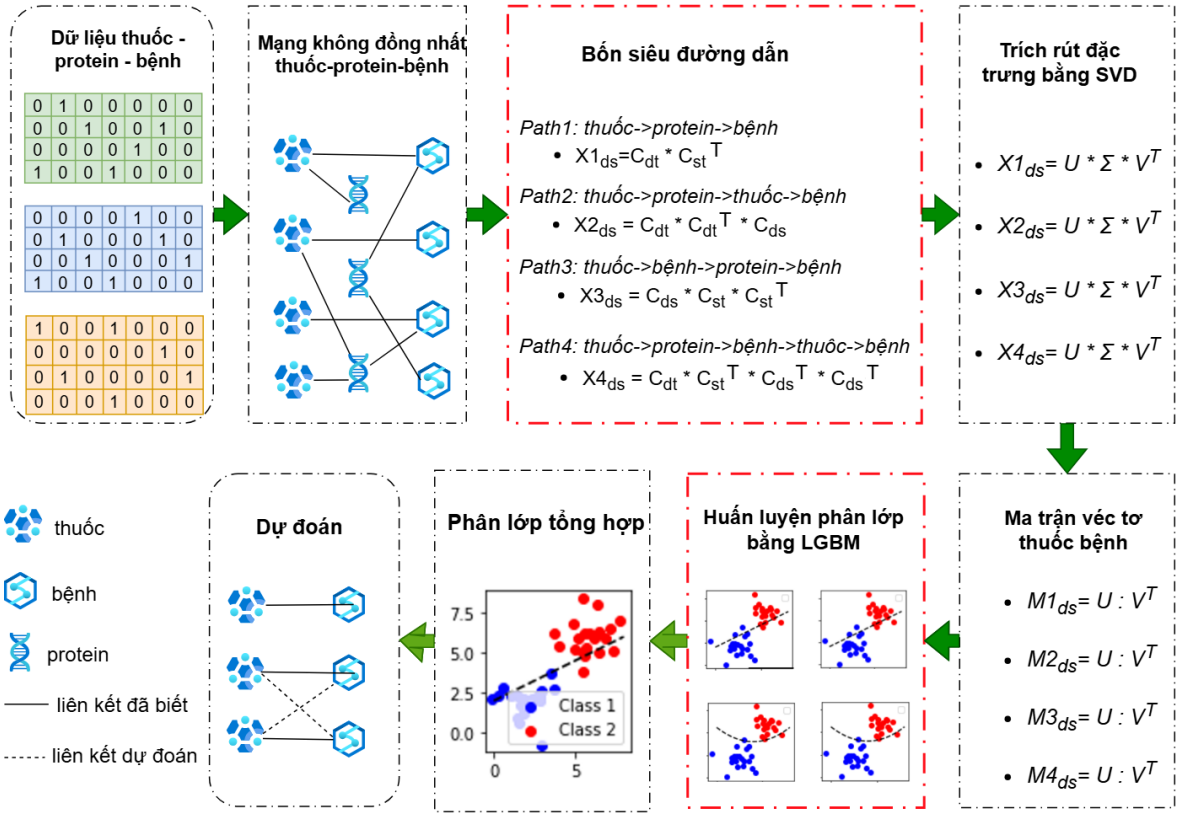
Những hạn chế trên cho thấy tồn tại không gian cải tiến rõ ràng và có cơ sở khoa học, đặc biệt trong việc mở rộng hệ thống siêu đường dẫn để tăng cường khả năng mô hình hóa ngữ nghĩa trong HIN và cải thiện hiệu quả dự đoán liên kết thuốc–bệnh. Các hướng cải tiến cụ thể dựa trên phân tích trên sẽ được trình bày trong các mục tiếp theo của chương này.

2.2. Mô hình DR-LGBM-MH: Siêu đường dẫn mới và LightGBM

2.2.1. Giới thiệu mô hình

Trên cơ sở kế thừa khung phương pháp EMP-SVD của Wu và cộng sự [26], luận án tiến hành mở rộng bước 2 giai đoạn xác định meta-path bằng cách khai thác bốn siêu đường dẫn, trong đó có một siêu đường dẫn mới do luận án đề xuất, với sự tham gia của protein trong vai trò nút trung gian. Việc lựa chọn các meta-path này được thực hiện một cách có định hướng và dựa trên ba nền tảng chính: (i) cơ sở sinh học, phản ánh vai trò trung tâm của protein trong cơ chế tác động giữa thuốc và bệnh; (ii) cơ sở toán học, đảm bảo độ dài đường dẫn không quá lớn nhằm hạn chế nhiễu và tránh gia tăng độ phức tạp không cần thiết; và (iii) cơ sở khoa học dữ liệu, tận dụng sự đa dạng về số lượng nút trung gian để khai thác các mức độ liên kết khác nhau trong mạng không đồng nhất.

Hình 2.2 mô tả toàn bộ luồng công việc của mô hình được đề xuất, trong đó bước 2 của EMP-SVD được cải tiến thông qua việc xây dựng hệ thống meta-path mới, và LightGBM được lựa chọn vì đặc trưng đầu vào sau khi trích xuất và giảm chiều bằng SVD có kích thước nhỏ, dạng bảng (tabular), phù hợp với các thuật toán boosting trên cây quyết định. Mô hình này xử lý hiệu quả dữ liệu thưa và mất cân bằng (ít mẫu dương, nhiều mẫu âm) nhờ cơ chế GOSS (Gradient-based One-Side Sampling), trong khi học sâu thường yêu cầu dữ liệu lớn để tránh quá khớp. Hơn nữa, LightGBM cho tốc độ huấn luyện nhanh, ổn



Hình 2.2: Sơ đồ luồng công việc của mô hình DR-LGBM-MH

định với tập mẫu nhỏ và cung cấp khả năng giải thích mô hình qua feature importance – một yếu cầu quan trọng trong nghiên cứu y sinh. Mô hình hoàn chỉnh được gọi tắt là DR-LGBM-MH (Drug Repositioning using LightGBM and Meta-Paths in HIN).

2.2.2. Quy trình thực hiện

Mô hình được triển khai theo năm bước chính sau đây:

Bước 1: Mạng không đồng nhất thuốc-protein-bệnh.

Khởi tạo ký hiệu và ma trận liên kết sinh học

Nhằm đơn giản hóa ký hiệu và tránh phân tán sự chú ý vào các chi tiết kỹ thuật không cần thiết, luận án sử dụng các tập hợp như sau:

- D: Tập hợp các thuốc, $D = \{d_i \mid i = 1, \ldots, n\}$

- T: Tập hợp các protein, $T = \{t_j \mid j = 1, \ldots, k\}$
- S: Tập hợp các bệnh, $S = \{s_i \mid i = 1, \ldots, m\}$

Hình 2.3 minh họa mối quan hệ thuốc–protein–bệnh. Ba loại mối liên kết sinh học chính được sử dụng bao gồm:

- Mối liên kết "liên kết với" giữa thuốc và protein.
- Mối liên kết "gây ra/gây ra bởi" giữa protein và bệnh.
- Mối liên kết "điều trị/được điều trị bởi" giữa thuốc và bệnh.

Các mối quan hệ này được mã hóa dưới dạng ma trận nhị phân như sau trong các công thức (2.8) đến (2.10):

$$C_{ds}[i,j] = \begin{cases} 1, & \text{nếu thuốc } d_i \text{ có quan hệ với bệnh } s_j, \\ 0, & \text{ngược lại.} \end{cases} \tag{2.8}$$

$$C_{dt}[i,j] = \begin{cases} 1, & \text{nếu thuốc } d_i \text{ có quan hệ với protein } t_j, \\ 0, & \text{ngược lại.} \end{cases} \tag{2.9}$$

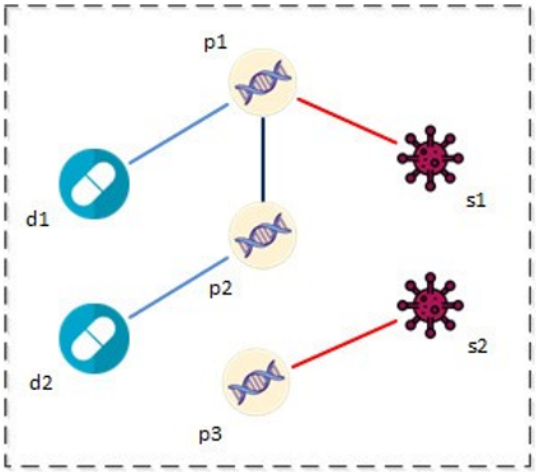
$$C_{st}[i,j] = \begin{cases} 1, & \text{nếu bệnh } s_i \text{ có quan hệ với protein } t_j, \\ 0, & \text{ngược lại.} \end{cases} \tag{2.10}$$

Việc mã hóa dữ liệu dưới dạng ma trận giúp chuẩn hóa biểu diễn mối quan hệ giữa các thực thể sinh học, tạo nền tảng cho các bước phân tích và xây dựng mô hình sau này.

Mạng không đồng nhất thuốc-protein-bệnh

Dựa trên các tập thực thể D, S và T đã được định nghĩa ở phần trên, mạng thuốc–protein–bệnh có thể được mô tả dưới dạng sơ đồ mạng:

$$N = (O, R),$$



Hình 2.3: Sơ đồ minh họa mối quan hệ thuốc–protein–bệnh.

trong đó $O = D \cup S \cup T$ là tập các thực thể và R là tập các loại quan hệ tương ứng giữa chúng.

Kích thước của tập thực thể được xác định bởi (2.11):

$$|O| = n + m + z. \tag{2.11}$$

Tập quan hệ R bao gồm ba loại quan hệ đã mô tả trước đó (thuốc–bệnh, thuốc–protein và bệnh–protein), với số lượng quan hệ tiềm năng như (2.12):

$$|R| = (n \times m) + (n \times z) + (m \times z). \tag{2.12}$$

Trong mạng thông tin thuốc–protein–bệnh, một *siêu đường dẫn* được định nghĩa là một chuỗi các thực thể và các loại quan hệ liên kết giữa chúng. Cụ thể, siêu đường dẫn có dạng:

$$M = o_1 \xrightarrow{r_1} o_2 \xrightarrow{r_2} \dots \xrightarrow{r_k} o_k,$$

trong đó mỗi thực thể $o_i \in O = D \cup S \cup T$ tương ứng với thuốc, bệnh hoặc protein, và mỗi quan hệ $r_i \in R$ thuộc một trong ba loại quan hệ của mạng: thuốc–bệnh, thuốc–protein hoặc bệnh–protein.

Bước 2: Đề xuất siêu đường dẫn mới

Từ những hạn chế đã phân tích ở mục 2.1.2, có thể thấy rằng mô hình EMP-SVD tuy hiệu quả trong việc tích hợp dữ liệu và đề xuất 5 meta-path,

nhưng chưa khai thác đầy đủ ngữ nghĩa phức tạp của HIN, đặc biệt là vai trò trung gian quan trọng của protein cũng như các quan hệ đồng nhất (thuốc–thuốc, bệnh–bệnh, protein–protein). Do vậy, trong phần này luận án giới thiệu 1 siêu đường dẫn mới kết hợp với 3 siêu đường dẫn được giới thiệu ở phương pháp EMP-SVD nhằm khai thác tối đa vai trò trung gian của protein cũng như cân bằng quan hệ giữa các đối tượng trong mạng.

Sự lựa chọn này hướng tới sự cân bằng tối ưu giữa độ sâu thông tin và hiệu quả tính toán của mô hình, đồng thời giảm thiểu nhiễu tiềm ẩn do việc sử dụng các meta-path quá dài. Cấu trúc chi tiết của các meta-path này sẽ được trình bày ở phần tiếp theo.

Meta-path-1: $d \to t \to s$ (Thuốc → Protein → Bệnh)

Đây là meta-path đơn giản nhất nhưng mang tính sinh học rõ ràng và trực tiếp. Nó mô tả một lộ trình trong đó một loại thuốc liên kết với một protein, và protein đó có liên quan đến (hoặc là nguyên nhân gây ra) một bệnh. Mối quan hệ này phản ánh cơ chế tác động của thuốc thông qua mục tiêu phân tử, cụ thể là protein – vốn là thành phần trung gian quan trọng trong nhiều quá trình bệnh lý.

Do đó, meta-path này cung cấp cơ sở đáng tin cậy để suy luận mối liên hệ tiềm năng giữa thuốc và bệnh dựa trên liên kết sinh học nền tảng. Với X_1 ký hiệu là ma trận kết hợp giữa thuốc và bệnh thông qua Meta-path-1, theo (1.1) X_1 được tính theo công thức (2.13) như sau:

$$X_1 = C_{dt} \times {C_{st}}^T, \tag{2.13}$$

Meta-path-2: $d \to t \to d \to s$ (Thuốc → Protein → Thuốc → Bệnh)

Meta-path này mở rộng lộ trình bằng cách thêm một loại thuốc trung gian. Cụ thể, một loại thuốc đầu tiên liên kết với một protein, protein này lại có liên kết với một loại thuốc khác, và thuốc thứ hai có liên quan đến một bệnh.

Con đường này khai thác mối tương quan giữa các thuốc có cùng mục tiêu protein hoặc tương tác sinh học, từ đó xác định khả năng hai thuốc điều trị các bệnh tương tự. Meta-path này đặc biệt hữu ích trong việc phát hiện các

cơ hội tái định vị thuốc, khi một loại thuốc có thể được gợi ý sử dụng cho các bệnh mới mà trước đây chưa được chỉ định.

Với X_2 ký hiệu là ma trận kết hợp giữa thuốc và bệnh thông qua Meta-path-2. Theo (1.1), X_2 tính theo công thức (2.14) như sau:

$$X_2 = C_{dt} \times {C_{dt}}^T \times C_{ds}, \tag{2.14}$$

Meta-path-3: $d \to s \to t \to s$ (Thuốc → Bệnh → Protein → Bệnh)

Meta-path này mô tả một lộ trình phức tạp hơn, trong đó một loại thuốc có liên quan đến một bệnh, bệnh này có liên hệ với một protein (có thể là nguyên nhân gây bệnh), và protein đó tiếp tục là yếu tố gây ra một bệnh khác.

Ở đây, bệnh trung gian đóng vai trò như một cầu nối giúp khám phá các mối quan hệ gián tiếp giữa thuốc và bệnh mới thông qua các cơ chế bệnh lý chung. Điều này hỗ trợ trong việc xác định các nhóm bệnh có cơ chế sinh học tương đồng, từ đó mở rộng phạm vi ứng dụng tiềm năng của thuốc.

Với X_3 ký hiệu là ma trận kết hợp giữa thuốc và bệnh thông qua Meta-path-3. Theo (1.1), X_3 được tính theo công thức (2.15) như sau:

$$X_3 = C_{ds} \times C_{st} \times {C_{st}}^T, \tag{2.15}$$

Meta-path-4: $d \to t \to s \to d \to s$ (Thuốc → Protein → Bệnh → Thuốc → Bệnh)

Đây là Meta-path-4 phức tạp nhất, kết hợp nhiều thực thể trung gian bao gồm protein, bệnh và thuốc khác. Theo lộ trình này, một loại thuốc đầu tiên liên kết với một protein, protein đó là nguyên nhân của một bệnh, bệnh này được điều trị bởi một loại thuốc khác, và thuốc thứ hai có liên quan đến một bệnh mục tiêu.

Chuỗi quan hệ nhiều tầng này cho phép mô hình khai thác các mối liên hệ gián tiếp sâu hơn giữa thuốc và bệnh, phản ánh các tương tác sinh học đa bước có thể không được quan sát trực tiếp nhưng mang ý nghĩa sinh học thực tiễn. Nhờ đó, meta-path-4 này hỗ trợ việc phát hiện các quan hệ tiềm ẩn trong mạng HIN và nâng cao hiệu quả dự đoán.

Với X_4 ký hiệu là ma trận kết hợp giữa thuốc và bệnh thông qua Meta-

path-4. Theo (1.1), X_4 được tính theo công thức (2.16) sau:

$$X_4 = C_{dt} \times {C_{st}}^T \times {C_{ds}}^T \times C_{ds}, \tag{2.16}$$

Bốn meta-path được đề xuất tạo thành một hệ thống hoàn chỉnh để khai thác các mối quan hệ thuốc-bệnh ở nhiều mức độ phức tạp khác nhau. Trong đó:

- Meta-path-1: Tập trung vào quan hệ trực tiếp thông qua protein
- Meta-path-2 & 3: Khai thác quan hệ gián tiếp thông qua thuốc, bệnh và protein trung gian
- Meta-path-4: Phản ánh các tương tác đa bước phức tạp trong mạng thông tin

Việc đề xuất bốn meta-path mới trong luận án này không phải là sự lựa chọn ngẫu nhiên trong vô số khả năng kết hợp có thể có, mà là sự lựa chọn có định hướng, dựa trên ba cơ sở khoa học, toán học và sinh học cụ thể. Theo nghiên cứu của Xiao [81], việc sử dụng các meta-path có nhiều nút trung gian không nhất thiết mang lại hiệu quả cao hơn trong việc biểu diễn quan hệ, thậm chí còn có thể làm giảm hiệu suất tổng thể của mô hình. Ngược lại, Tian [82] chỉ ra rằng hiệu suất của mô hình có thể được cải thiện khi kết hợp các meta-path với số lượng nút khác nhau, nhờ khả năng khai thác thông tin ở nhiều mức độ liên kết.

Như vậy, bốn meta-path được lựa chọn không chỉ phản ánh trực tiếp các cơ chế sinh học tiềm năng mà còn mở rộng không gian suy luận so với các meta-path trong EMP-SVD. Tuy nhiên, những meta-path này vẫn chưa tận dụng được các mối tương quan nội tại giữa các thực thể cùng loại như thuốc–thuốc, bệnh–bệnh và protein–protein. Việc tích hợp các mối quan hệ đồng nhất này sẽ được trình bày trong mục tiếp theo.

Bước 3: Phân tách giá trị kỳ dị và trích rút vector thuốc - bệnh

Sau khi xây dựng các ma trận meta-path thể hiện số lượng đường đi từ thuốc d_i đến bệnh s_j, yêu cầu đặt ra là cần rút trích đặc trưng từ các ma trận này sao cho vừa giữ được thông tin quan trọng, vừa tránh tình trạng số chiều quá lớn dẫn đến quá khớp. Mỗi hàng và mỗi cột của ma trận meta-path đều mang thông tin cấu trúc hữu ích, nhưng kích thước của chúng thường quá lớn so với số lượng mẫu dương, khiến các mô hình phân lớp khó đạt khả năng tổng quát hóa cao.

Để giải quyết vấn đề này, luận án sử dụng phân tách giá trị kỳ dị SVD nhằm nén ma trận commuting X vào một không gian tiềm ẩn có số chiều thấp hơn. SVD phân tách ma trận như sau:

$$X = U \Sigma V^T, \tag{2.17}$$

trong đó Σ chứa các giá trị kỳ dị sắp xếp theo thứ tự giảm dần. Chỉ giữ lại k giá trị kỳ dị dương lớn nhất ($\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$) cho phép tạo ra một xấp xỉ hạng thấp của ma trận gốc:

$$X \approx U_{m\times k}\, \Sigma_{k\times k}\, V^T_{k\times n}. \tag{2.18}$$

Trong đó, các hàng của U biểu diễn đặc trưng tiềm ẩn của thuốc, còn các hàng của V biểu diễn đặc trưng tiềm ẩn của bệnh. Cụ thể, ta có:

$$U = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{bmatrix}, \qquad V^T = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1k} \\ b_{21} & b_{22} & \dots & b_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mk} \end{bmatrix}.$$

Vector đặc trưng của thuốc thứ i và bệnh thứ j lần lượt như sau:

$$U_{i,:} = [a_{i1}, a_{i2}, \dots, a_{ik}],$$

$$(V^T)_{j,:} = [b_{j1}, b_{j2}, \ldots, b_{jk}].$$

Từ đó, đặc trưng của cặp thuốc–bệnh (i, j) được hình thành bằng phép nối hai vector tiềm ẩn:

$$x_{i,j} = U_{i,:} \,\|\, (V^T)_{j,:} = [a_{i1}, \ldots, a_{ik},\, b_{j1}, \ldots, b_{jk}]. \tag{2.19}$$

Để làm rõ hơn cách biểu diễn đặc trưng tiềm ẩn của cặp thuốc–bệnh được xây dựng từ phân tách giá trị kỳ dị (SVD), xét một ví dụ minh họa đơn giản như sau. Sau khi thực hiện phân tách SVD trên ma trận meta-path (ví dụ: Drug–Target–Disease), ta thu được hai ma trận U và V^T, trong đó các hàng của U biểu diễn đặc trưng tiềm ẩn của các thuốc, còn các cột của V^T (hay tương đương các hàng của V) biểu diễn đặc trưng tiềm ẩn của các bệnh trong không gian đặc trưng k-chiều.

Giả sử với số chiều tiềm ẩn $k = 3$, một thuốc D_i có vector đặc trưng:

$$u_{D_i} = [0.8,\ 0.1,\ 0.6]$$

và một bệnh B_j có vector đặc trưng:

$$v_{B_j} = [0.75,\ 0.2,\ 0.55]$$

Theo công thức 2.19, vector đặc trưng của cặp thuốc–bệnh (D_i, B_j) được xây dựng bằng phép nối hai vector này:

$$x_{D_i,B_j} = [0.8,\ 0.1,\ 0.6,\ 0.75,\ 0.2,\ 0.55]$$

Các thành phần trong các vector tiềm ẩn này không biểu diễn các thuộc tính sinh học cụ thể một cách tường minh, mà phản ánh các mẫu quan hệ ẩn được trích xuất từ cấu trúc của mạng meta-path. Chẳng hạn, các thành phần có giá trị lớn đồng thời ở cả thuốc và bệnh có thể tương ứng với các cơ chế sinh học chung (như pathway viêm hoặc tương tác protein tương tự), trong khi các thành phần có giá trị nhỏ thể hiện mức độ liên quan thấp hơn trong không gian đặc trưng.

Nhờ sử dụng SVD, số chiều đặc trưng được giảm xuống còn $2k$, nhỏ hơn rất nhiều so với kích thước ban đầu $m+n$, giúp mô hình phân lớp hoạt động ổn định hơn, giảm hiện tượng overfitting và giữ lại cấu trúc ngữ nghĩa quan trọng trong mạng không đồng nhất.

Bước 4: Huấn luyện mô hình LightGBM

Sau khi thu được các vector đặc trưng thuốc–bệnh thông qua phân rã SVD ở Bước 3, luận án tiến hành xây dựng các mô hình phân loại độc lập tương ứng với bốn meta-path đã đề xuất. Mỗi meta-path tạo ra một không gian đặc trưng khác nhau; do đó, để đảm bảo tính khách quan và khả năng so sánh, bốn mô hình LightGBM được huấn luyện riêng biệt nhưng sử dụng chung một cấu hình tham số.

Trong nghiên cứu này, Light Gradient Boosting Machine (LightGBM hoặc LGBM) được lựa chọn làm bộ phân loại cơ sở nhờ khả năng xử lý hiệu quả các ma trận đặc trưng lớn và thưa, vốn là đặc trưng của dữ liệu được sinh ra từ mạng HIN thuốc–protein–bệnh. LightGBM xây dựng mô hình dưới dạng một tập hợp cây quyết định theo cơ chế boosting, trong đó mỗi cây liên tiếp học từ phần sai số còn lại của cây trước. Thuật toán sử dụng hai kỹ thuật tối ưu quan trọng nhằm tăng tốc độ huấn luyện mà vẫn duy trì chất lượng mô hình:

Trong quá trình xây dựng mô hình, các tham số LightGBM được thiết lập thống nhất cho cả bốn mô hình như sau:

- `n_estimators = 256`: số lượng cây trong mô hình boosting; lựa chọn này bảo đảm mô hình có đủ năng lực học các quan hệ phi tuyến mà không gây quá tải tính toán.
- `learning_rate = 0.1`: tốc độ học trung bình, cân bằng giữa tốc độ hội tụ và khả năng tổng quát hóa.
- `max_depth = -1`: không giới hạn độ sâu của cây quyết định, cho phép LightGBM phát triển cây theo chiến lược leaf-wise nhằm mô hình hóa hiệu quả các quan hệ phức tạp trong dữ liệu.

- `random_state = 1`: bảo đảm tính tái lập của toàn bộ thí nghiệm.

Việc sử dụng cấu hình tham số đồng nhất cho cả bốn mô hình giúp đảm bảo tính công bằng trong việc đánh giá mức độ đóng góp của từng meta-path, đồng thời tránh hiện tượng sai lệch do lựa chọn tham số không đồng nhất. Sau khi huấn luyện, mỗi bộ phân loại cơ sở trả về một xác suất $h_i(x)$ biểu thị mức độ tin cậy rằng cặp thuốc–bệnh x có quan hệ điều trị. Các xác suất này sau đó được sử dụng trong bước tổ hợp mô hình nhằm cải thiện độ ổn định và độ chính xác của dự đoán cuối cùng.

Bước 5: Dự đoán và đánh giá

Sau khi huấn luyện bốn mô hình LightGBM tương ứng với bốn meta-path, bước tiếp theo là tiến hành dự đoán quan hệ thuốc–bệnh và đánh giá hiệu suất của mô hình. Như đã trình bày ở Bước 4, mỗi mô hình cơ sở trả về một xác suất $h_i(x)$ cho biết mức độ tin cậy rằng cặp thuốc–bệnh x thuộc lớp dương (có quan hệ điều trị). Để tận dụng thông tin từ nhiều nguồn đặc trưng khác nhau, luận án áp dụng cơ chế tổ hợp dựa trên trung bình cộng xác suất.

Tổ hợp dự đoán từ các mô hình cơ sở

Giả sử với mỗi cặp thuốc–bệnh x, bốn bộ phân loại tương ứng với bốn meta-path đưa ra các xác suất $h_1(x), h_2(x), h_3(x), h_4(x)$. Xác suất dự đoán tổng hợp $q(x)$ được tính theo công thức (2.20):

$$q(x) = \frac{1}{4} \sum_{i=1}^{4} h_i(x). \tag{2.20}$$

Cơ chế tổ hợp này giúp giảm phương sai của mô hình, tăng tính ổn định và khai thác được thông tin bổ sung từ các meta-path có độ dài và mức độ liên kết khác nhau. Các meta-path ngắn phản ánh quan hệ trực tiếp, trong khi meta-path dài cung cấp thông tin gián tiếp về cấu trúc liên kết sâu hơn trong mạng HIN.

Xác định ngưỡng phân loại tối ưu

Để đưa ra nhãn dự đoán cuối cùng, mô hình cần xác định một ngưỡng phân loại τ để chuyển đổi xác suất $q(x)$ thành nhãn nhị phân. Ngưỡng τ được lựa chọn sao cho cân bằng giữa độ chính xác (precision) và độ hồi tưởng (recall). Trong luận án, ngưỡng tối ưu τ^* được xác định bằng cách tối đa hóa chỉ số F1-score theo (2.21):

$$\tau^* = \arg\max_{\tau} \text{ F1-score}(\tau). \tag{2.21}$$

F1-score là thước đo hài hòa giữa precision và recall, đặc biệt hữu ích trong các bài toán có phân bố nhãn mất cân đối như dự đoán quan hệ thuốc–bệnh, nơi số lượng mẫu âm thường lớn hơn rất nhiều so với mẫu dương.

Nhãn dự đoán cuối cùng được xác định theo quy tắc như (2.22):

$$\hat{y}(x) = \begin{cases} 1, & \text{nếu } q(x) \geq \tau^*, \\ 0, & \text{ngược lại.} \end{cases} \tag{2.22}$$

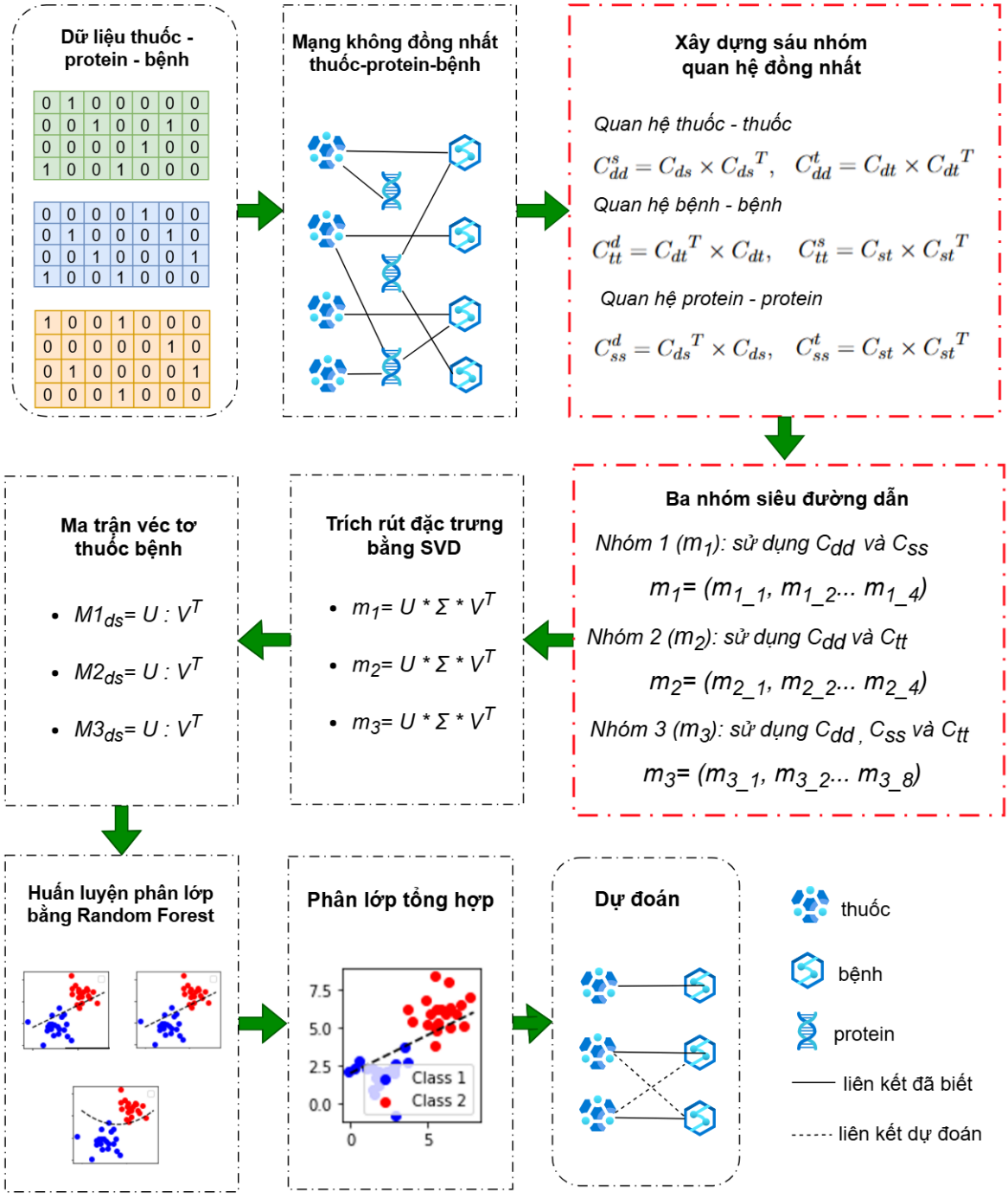
2.3. Mô hình HS-TMP: Quan hệ đồng nhất và siêu đường dẫn

2.3.1. Giới thiệu mô hình HS-TMP

Mặc dù bốn meta-path được đề xuất trong Mục 2.2.1 đã phản ánh được các mối liên hệ sinh học quan trọng giữa thuốc, protein và bệnh, nhưng chúng vẫn chưa khai thác toàn diện các quan hệ đồng nhất (Homogeneous Relations) giữa các thực thể cùng loại như thuốc–thuốc, bệnh–bệnh và protein–protein. Các quan hệ đồng nhất này lại đóng vai trò quan trọng trong việc mô tả sự tương đồng về cơ chế tác động, chức năng sinh học hoặc cấu trúc phân tử.

Do đó, trong mục này, luận án mở rộng bước 2 của mô hình EMP-SVD cũng như bước 2 của mô hình DR-LGBM-MH (giới thiệu tại Mục 2.2) bằng cách đề xuất mô hình HS-TMP (Homogeneous Similarities and Three Meta-Path) với sáu mối tương quan đồng nhất và ba nhóm siêu đường dẫn. Mục tiêu là làm phong phú thêm không gian suy luận trong dự đoán quan hệ thuốc-bệnh. Luồng

công việc được minh họa trong Hình 2.4.



Hình 2.4: Sơ đồ luồng công việc 3 nhóm meta-path

2.3.2. Xây dựng sáu mối tương quan đồng nhất

Trong việc thiết kế một mạng không đồng nhất, có sự lựa chọn về các loại cạnh. Mạng lưới được xây dựng có các cạnh cho các tương quan thuốc-

protein, bệnh-protein và thuốc-bệnh. Luận án đề xuất thêm ba loại cạnh mới là thuốc-thuốc, bệnh-bệnh và protein-protein.

Tương quan thuốc-thuốc. Chỉ định tương quan thuốc-thuốc là rất đáng chú ý và nó thực sự có thể được thực hiện bằng cách nghiên cứu tương quan thuốc-bệnh hoặc thuốc-protein. Tất nhiên, mối liên hệ giữa hai loại thuốc có thể được thiết lập nếu tồn tại một căn bệnh mà cả hai loại thuốc đều liên quan.

Tương quan thuốc-thuốc được tạo ra bởi bệnh làm trung gian, được đại diện bởi một ma trận C^s_{dd}. Tương tự, tương quan thuốc-thuốc có thể được quan sát bằng cách kiểm tra tương quan thuốc-protein. Nó phải đánh dấu mối liên quan của hai loại thuốc bất cứ khi nào tồn tại một protein có liên quan đến cả hai loại thuốc.

Tương quan thuốc-thuốc được tạo ra bởi protein như một trung gian được đại diện bởi ma trận C^t_{dd}. Hai ma trận C^s_{dd} và C^t_{dd} được tính theo công thức (2.23) và (2.24) như sau:

$$C^s_{dd} = C_{ds} \times {C_{ds}}^T, \tag{2.23}$$

$$C^t_{dd} = C_{dt} \times {C_{dt}}^T, \tag{2.24}$$

Tương quan protein-protein

Luận án đã xem xét các kỹ thuật để xác định tương quan này. Một cách để phát hiện liên kết protein–protein là tìm kiếm một loại thuốc làm trung gian cho liên kết protein–thuốc–protein để tính ma trận liên kết $C^d_{tt} \in \mathbb{R}^{z \times z}$.

Với sự sẵn có của tương quan protein-bệnh, tất nhiên luận án có thể đánh giá ngay lập tức mối quan hệ protein-bệnh-protein để có được mối liên hệ protein-protein. Điều này dẫn đến một ma trận $C^s_{tt} \in \mathbb{R}^{z \times z}$.

Hai ma trận C^d_{tt} và C^s_{tt} được tính theo công thức (2.25) và (2.26) như sau:

$$C^d_{tt} = {C_{dt}}^T \times C_{dt}, \tag{2.25}$$

$$C^s_{tt} = C_{st} \times {C_{st}}^T, \tag{2.26}$$

Tương quan bệnh–bệnh

Loại tương quan này có thể được xác định dựa trên khả năng hai bệnh

khác nhau cùng được điều trị hoặc điều trị bằng một hoặc nhiều loại thuốc giống nhau. Trong luận án, mối quan hệ bệnh-thuốc-bệnh được trích xuất bằng cách kiểm tra các liên kết giữa bệnh và thuốc.

Cụ thể, nếu hai bệnh cùng liên quan đến một loại thuốc, chúng được xem là có mối liên hệ và được đánh dấu trong ma trận tương quan $C_{ss}^{d} \in \mathbb{R}^{m \times m}$.

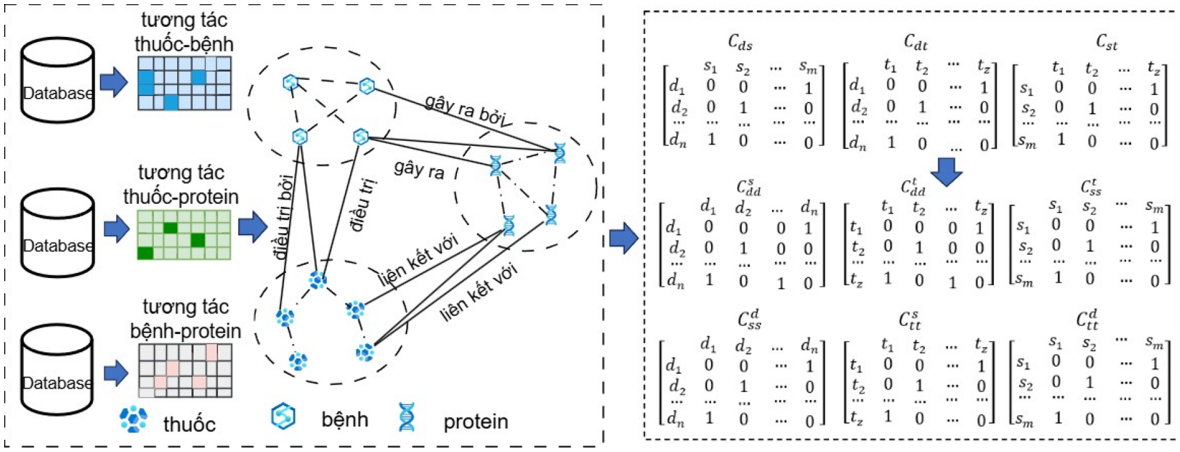
Tương tự, mỗi liên kết bệnh–protein trong dữ liệu quan hệ bệnh–protein cũng được sử dụng để suy luận mối quan hệ giữa các bệnh, thông qua việc kiểm tra xem liệu hai bệnh có cùng gây ra hoặc bị gây ra bởi một hoặc nhiều protein giống nhau. Kết quả thu được sẽ được biểu diễn trong ma trận $C_{ss}^{t} \in \mathbb{R}^{m \times m}$.

Hai ma trận C_{ss}^{d} và C_{ss}^{t} được tính theo công thức (2.27) và (2.28) như sau:

$$C_{ss}^{d} = {C_{ds}}^{T} \times C_{ds}, \tag{2.27}$$

$$C_{ss}^{t} = C_{st} \times {C_{st}}^{T}, \tag{2.28}$$

Việc kết hợp các quan hệ thông qua các thực thể trung gian như thuốc và protein cho phép luận án xây dựng đầy đủ sáu ma trận tương quan quan trọng (thuốc–thuốc, bệnh–bệnh, protein–protein), như được minh họa trong Hình 2.5



Hình 2.5: Mô hình HIN và các ma trận tương quan

2.3.3. Đề xuất ba nhóm siêu đường dẫn

Trong nghiên cứu trước đây, Wu và cộng sự [26] đã chứng minh 5 meta-path có hiệu quả trong việc ước tính các mối liên hệ thuốc-bệnh. Trong luận án này, luận án đề xuất 3 nhóm meta-path mới để dự đoán các loại thuốc tiềm năng cho bệnh. Các meta-path này được thiết kế dựa trên một logic thực tế và khả thi. Cụ thể, mỗi mẫu meta-path được xây dựng bằng cách kết hợp các thành phần liên kết, và mỗi cách kết hợp sẽ tạo ra một phiên bản meta-path con cụ thể.

Chẳng hạn, mẫu meta-path đầu tiên được thiết kế để bao gồm các quan hệ thuốc-thuốc, thuốc-bệnh và bệnh-bệnh, với dạng tổng quát: $C_{dd}C_{ds}C_{ss}$

Nhóm meta-path thứ nhất được thiết kế để bao gồm các quan hệ thuốc-thuốc, thuốc-bệnh và bệnh-bệnh, với dạng tổng quát: $m_1 = C_{dd} \times C_{ds} \times C_{ss}$. Nhóm meta-path này có hai lựa chọn liên kết thuốc-thuốc là C^s_{dd}, C^t_{dd} và hai lựa chọn liên quan đến bệnh-bệnh là C^d_{ss}, C^t_{ss}. Bằng cách kết hợp các tùy chọn này, nhóm meta-path đầu tiên chứa bốn meta-path con tương ứng với các công thức (2.29) đến (2.32):

$$m1\_1 = C^s_{dd} \times C_{ds} \times C^d_{ss} \tag{2.29}$$

$$m1\_2 = C^s_{dd} \times C_{ds} \times C^t_{ss} \tag{2.30}$$

$$m1\_3 = C^t_{dd} \times C_{ds} \times C^d_{ss} \tag{2.31}$$

$$m1\_4 = C^t_{dd} \times C_{ds} \times C^t_{ss} \tag{2.32}$$

Nhóm thứ nhất mô tả cơ chế suy luận dựa trên độ tương đồng giữa các thuốc và giữa các bệnh. Về mặt sinh học, cách tiếp cận này dựa trên giả định rằng các thuốc có đặc tính dược lý hoặc cơ chế tác động tương tự có xu hướng điều trị hiệu quả các bệnh có đặc điểm sinh học tương đồng, chẳng hạn như liên quan đến cùng gen, protein hoặc các con đường sinh học.

Nhóm meta-path thứ hai được xây dựng dựa trên việc nghiên cứu các mối liên kết thuốc-protein và bệnh-protein, bằng cách kết hợp các quan hệ

thuốc-thuốc và protein-protein với ma trận kết hợp $m_2 = C_{dd} \times C_{dt} \times C_{tt} \times {C_{st}}^T$. Các quan hệ thuốc-thuốc và protein-protein có các lựa chọn thay thế tương ứng bao gồm C^s_{dd}, C^t_{dd}, C^s_{tt} và C^d_{tt}. Tổng cộng có bốn meta-path con cho nhóm meta-path này, được mô tả trong các công thức (2.33) đến (2.36):

$$m2\_1 = C^s_{dd} \times C_{dt} \times C^s_{tt} \times {C_{st}}^T \tag{2.33}$$

$$m2\_2 = C^s_{dd} \times C_{dt} \times C^d_{tt} \times {C_{st}}^T \tag{2.34}$$

$$m2\_3 = C^t_{dd} \times C_{dt} \times C^s_{tt} \times {C_{st}}^T \tag{2.35}$$

$$m2\_4 = C^t_{dd} \times C_{dt} \times C^d_{tt} \times {C_{st}}^T \tag{2.36}$$

Về mặt sinh hóm thứ hai đi sâu hơn vào tầng chức năng phân tử khi lan truyền thông tin từ thuốc qua protein đích và mạng tương tác protein, từ đó suy ra mối liên hệ với bệnh. Nhóm này có ý nghĩa sinh học rõ ràng vì trong thực tế, tác dụng của thuốc không chỉ phụ thuộc vào một protein đích riêng lẻ mà còn chịu ảnh hưởng bởi các protein tương tác trong cùng con đường tín hiệu hoặc cùng quá trình sinh học.

Nhóm meta-path thứ ba được thiết kế để khai thác đồng thời ba mối quan hệ đồng nhất: thuốc-thuốc, protein-protein và bệnh-bệnh, cùng với các quan hệ không đồng nhất thuốc-protein và protein-bệnh. Nhóm meta-path này có ma trận kết hợp $m_3 = C_{dd} \times C_{dt} \times C_{tt} \times {C_{st}}^T \times C_{ss}$, với tám meta-path con tương ứng (2.37) đến (2.44):

$$m3\_1 = C^s_{dd} \times C_{dt} \times C^s_{tt} \times {C_{st}}^T \times C^d_{ss} \tag{2.37}$$

$$m3\_2 = C^s_{dd} \times C_{dt} \times C^s_{tt} \times {C_{st}}^T \times C^t_{ss} \tag{2.38}$$

$$m3\_3 = C^s_{dd} \times C_{dt} \times C^d_{tt} \times {C_{st}}^T \times C^d_{ss} \tag{2.39}$$

$$m3\_4 = C^s_{dd} \times C_{dt} \times C^d_{tt} \times {C_{st}}^T \times C^t_{ss} \tag{2.40}$$

$$m3\_5 = C^t_{dd} \times C_{dt} \times C^s_{tt} \times {C_{st}}^T \times C^d_{ss} \tag{2.41}$$

$$m3\_6 = C^t_{dd} \times C_{dt} \times C^s_{tt} \times {C_{st}}^T \times C^t_{ss} \tag{2.42}$$

$$m3\_7 = C^t_{dd} \times C_{dt} \times C^d_{tt} \times {C_{st}}^T \times C^d_{ss} \tag{2.43}$$

$$m3\_8 = C_{dd}^{t} \times C_{dt} \times C_{tt}^{d} \times {C_{st}}^{T} \times C_{ss}^{t} \tag{2.44}$$

Nhóm thứ 3 là nhóm có ngữ nghĩa phong phú nhất vì đồng thời tích hợp cả tương đồng thuốc, tương tác protein và tương đồng bệnh; do đó, nhóm này cho phép mô hình hóa các cơ chế tác động đa tầng, trong đó một thuốc có thể được suy luận là phù hợp với một bệnh không chỉ do đặc tính của chính thuốc đó mà còn do sự lan truyền ảnh hưởng qua mạng protein và sự tương tự về đặc điểm sinh học giữa các bệnh. Theo nghĩa này, ba nhóm siêu đường dẫn không chỉ làm tăng số lượng đặc trưng mà còn mở rộng chiều sâu sinh học của quá trình suy luận liên kết thuốc–bệnh.

Do đó, mỗi nhóm meta-path được đề xuất bao gồm một tập hợp các meta-path con, và chính các meta-path con này được sử dụng để xây dựng các tương quan thuốc-bệnh.

Ba nhóm siêu đường dẫn được đề xuất trong luận án bao gồm các cấu trúc meta-path có độ phức tạp khác nhau: nhóm thứ nhất có 4 meta-path con, nhóm thứ hai có 4 meta-path con và nhóm thứ ba có 8 meta-path con, tạo thành tổng cộng 128 meta-path khi kết hợp. Các meta-path này sẽ được xử lý theo các bước tiếp theo như đã trình bày trong Mục 2.2.1 và trong luồng công việc ở Hình 2.4. Khi lựa chọn một meta-path con từ mỗi nhóm để hình thành một tổ hợp meta-path hoàn chỉnh, số lượng tổ hợp có thể tạo ra được tính như sau: $4 \times 4 \times 8 = 128$. Mỗi tổ hợp meta-path đại diện cho một hướng suy luận khác nhau trong mạng không đồng nhất, phản ánh các mức độ liên kết sinh học và cấu trúc đa tầng trong quan hệ thuốc–bệnh. Dựa trên 128 tổ hợp này, luận án xây dựng tương ứng 128 bộ phân loại cơ sở; mỗi bộ phân loại được huấn luyện trên tập đặc trưng thu được từ một meta-path cụ thể nhằm dự đoán khả năng tồn tại các liên kết thuốc–bệnh mới. Trong 3 nhóm siêu đường dẫn trên, mỗi nhóm đều được xây dựng dựa trên việc tích hợp tối thiểu 2 loại quan hệ đồng nhất (thuốc-thuốc, bệnh-bệnh, protein-protein). Trong khi đó, 4 siêu đường dẫn được giới thiệu trong mô hình DR-LGBM-MH (Chương 2.2) chỉ sử dụng các quan hệ không đồng nhất cơ bản (thuốc–protein, protein–bệnh, thuốc–bệnh) và

không thuộc nhóm nào trong 3 nhóm trên. Sự khác biệt này thể hiện hai hướng tiếp cận bổ sung: DR-LGBM-MH tập trung khai thác cấu trúc gián tiếp thông qua vai trò trung gian của protein, còn HS-TMP làm giàu ngữ nghĩa mạng bằng cách tích hợp thông tin đồng nhất nội tại giữa các thực thể cùng loại.

2.4. Thực nghiệm và đánh giá

2.4.1. Dữ liệu

Để đánh giá mô hình DR-LGBM-MH và HS-TMP, luận án sử dụng ba bộ dữ liệu quan hệ sinh học được tích hợp từ các nguồn công khai và uy tín: DrugBank [83] (thông tin thuốc và chỉ định), OMIM [84] (mô tả bệnh và triệu chứng), và bộ dữ liệu của Gottlieb [85]. Các nguồn bổ trợ khác bao gồm DailyMed, SIDER (tác dụng phụ), HPO, và ClinicalTrials.gov. Hệ thống UMLS được sử dụng để chuẩn hóa tên gọi các thực thể.

Toàn bộ dữ liệu được chia thành 5 phần (5-fold) để tiến hành đánh giá chéo. Trong mỗi lần chạy, một phần được dùng làm tập kiểm tra và bốn phần còn lại dùng làm tập huấn luyện. Chi tiết các bộ dữ liệu thành phần như sau:

- **Thuốc-Bệnh**: Gồm 1.186 thuốc, 449 bệnh với 1.827 tương tác điều trị đã được xác nhận.
- **Thuốc-Protein**: Gồm 1.186 thuốc, liên kết với 4.642 protein mục tiêu.
- **Bệnh-Protein**: Gồm 449 bệnh, liên quan đến 1.467 protein với 4.642 tương tác.

2.4.2. Môi trường thực nghiệm

Tất cả các thí nghiệm được triển khai trong môi trường lập trình Python 3.x, sử dụng các thư viện chuyên dụng cho học máy và xử lý đồ thị như Scikit-learn, LightGBM, NetworkX... Cấu hình phần cứng sử dụng để đảm bảo tính nhất quán là với CPU Intel i7-12700, RAM 32GB, GPU RTX 3080. Như đã nêu ở mục trước, phương pháp đánh giá chéo 5-fold được áp dụng. Để đảm bảo

đánh giá khách quan và tránh rò rỉ thông tin, trong mỗi lần lặp của quá trình cross-validation, tất cả các cạnh (liên kết đã biết) thuộc tập kiểm tra đều được loại bỏ khỏi mạng HIN trước khi tiến hành xây dựng các ma trận kết hợp dựa trên meta-path. Quy trình này mô phỏng chính xác kịch bản thực tế khi dự đoán các liên kết thuốc-bệnh hoàn toàn mới.

2.4.3. Các chỉ số đánh giá

Trong đánh giá hiệu suất của các mô hình học máy, đặc biệt đối với các bộ dữ liệu y sinh trong bài toán dự đoán liên kết thuốc–bệnh, việc lựa chọn bộ chỉ số đo lường phù hợp giữ vai trò then chốt. Nhằm đảm bảo quá trình đánh giá toàn diện và khách quan, luận án sử dụng một hệ thống thước đo bao gồm:

Nhóm chỉ số cơ bản: Accuracy (ACC), Recall (REC), Precision (PRE).

Nhóm chỉ số tổng hợp: F1-score và hệ số tương quan Matthews (MCC), các thước đo nhấn mạnh mức độ cân bằng giữa các loại lỗi trong phân loại nhị phân.

Nhóm chỉ số dựa trên đường cong: Diện tích dưới đường cong ROC (AUC) và diện tích dưới đường cong Precision–Recall (AUPR), dùng để đánh giá khả năng phân biệt mô hình trong các ngưỡng phân loại khác nhau, đặc biệt hữu ích trong các bài toán mất cân bằng lớp.

Các công thức tính toán tương ứng của những chỉ số này được trình bày trong các biểu thức từ (1.7) đến (1.14).

2.4.4. Thiết lập thực nghiệm và Phương pháp so sánh

Kịch bản 1: Thực nghiệm mô hình DR-LGBM-MH

- Tính ma trận kết hợp từ các siêu đường dẫn
- Lựa chọn đặc trưng tiềm ẩn từ phân tách giá trị kỳ dị
- Huấn luyện mô hình LightGBM
- Đánh giá hiệu suất mô hình

Chứng minh hiệu quả từng meta-path và mô hình tổ hợp. Phân tích ma trận nhầm lẫn chỉ ra rằng các meta-path riêng lẻ có xu hướng tạo FP và FN cao hơn. Mô hình tổ hợp giúp giảm đáng kể hai loại lỗi này và tăng TP, TN, khẳng định lợi thế của ensemble learning.

- So sánh với các mô hình truyền thống. LightGBM cho kết quả vượt trội so với RF, XGB, SVM, KNN và AC trên toàn bộ các thước đo đánh giá.
- Kiểm chứng vai trò của SVD. Thử nghiệm ablation (có/không sử dụng SVD) cho thấy LightGBM kết hợp SVD cho hiệu suất cao nhất, chứng minh tính hiệu quả của giảm chiều.
- Kiểm định thống kê. Kiểm định t-test hai mẫu xác nhận sự khác biệt có ý nghĩa thống kê giữa mô hình DR-LGBM-MH và các mô hình so sánh (p-value < 0.05).
- So sánh với các nghiên cứu trước. Đặt cạnh các phương pháp cổ điển (PREDICT, TL-HGBI, LRSSL, SCMFDD, MBiRW) và hiện đại (EMP-SVD, AICI2023), mô hình DR-LGBM-MH đạt kết quả nổi bật, đặc biệt ở Recall, MCC và F1-score.
- Nghiên cứu điển hình

Kịch bản 2: Thực nghiệm mô hình HS-TMP

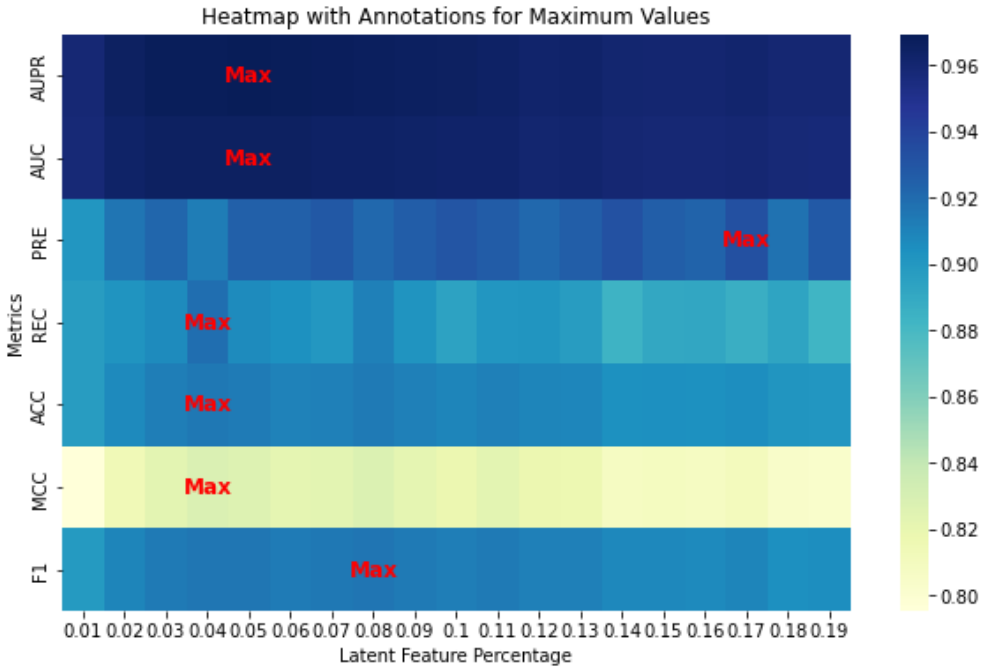
- Thiết lập sáu loại tương quan: thuốc–thuốc, bệnh–bệnh và protein–protein nhằm làm giàu thông tin trong mạng tri thức.
- Xây dựng 3 nhóm mô hình cơ sở và các ma trận tổng hợp tương ứng các công thức (2.23 - 2.44)
- Huấn luyện 128 mô hình cơ sở Random Forest
- Lựa chọn kết hợp được kết hợp từ 3 mô hình cơ sở, mỗi mô hình thuộc 1 nhóm (có tổng 128 cách kết)

- Đánh giá hiệu quả so với các nghiên cứu trước.
- Các nghiên cứu điển hình.

2.4.5. Kết quả và Thảo luận cho Mô hình DR-LGBM-MH

Lựa chọn giá trị tiềm ẩn cho SVD

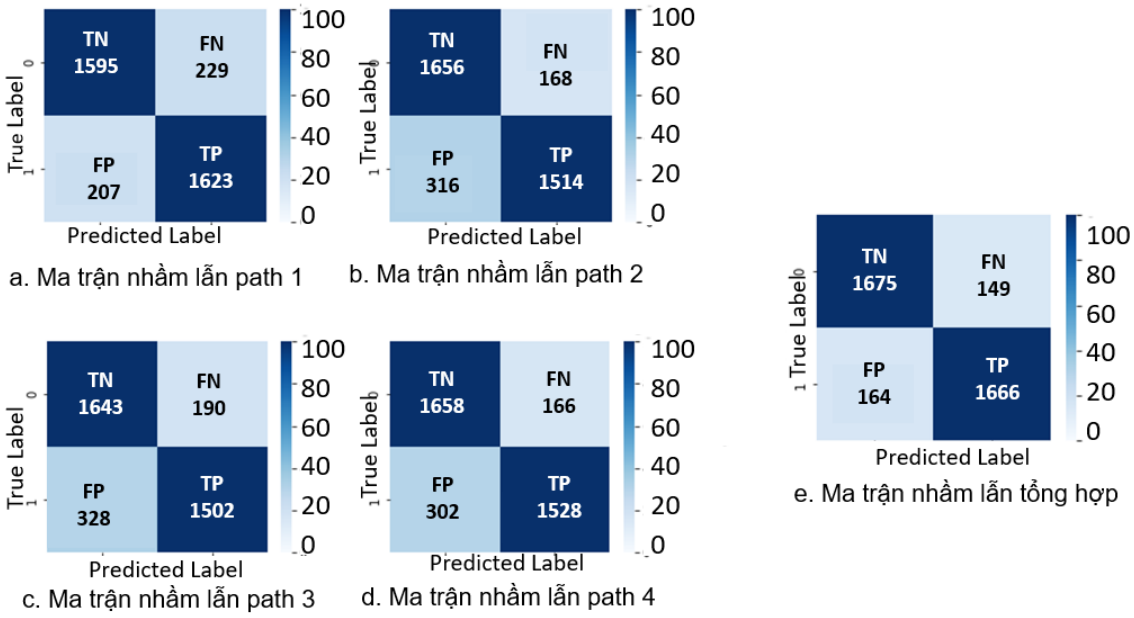
Việc lựa chọn số chiều đặc trưng tiềm ẩn k sau khi phân rã SVD có ảnh hưởng quan trọng đến hiệu suất và độ phức tạp của mô hình. Trong luận án, k được xác định thông qua một tham số tỷ lệ β, với công thức: $k = \min(n, m) \times \beta$ trong đó n và m lần lượt là số lượng thuốc và bệnh. Để tìm giá trị β tối ưu, một loạt thí nghiệm được thực hiện với β thay đổi từ 0.01 đến 0.20. Kết quả chi tiết



Hình 2.6: Hiệu suất của mô hình DR-LGBM-MH theo β
Các điểm đánh dấu (max) chỉ ra giá trị tối ưu của từng chỉ số.

được trình bày trong Hình 2.6. Quan sát cho thấy các chỉ số đánh giá khác nhau đạt giá trị cực đại ở những mức β khác nhau: Recall (0.919), Accuracy (0.914) và MCC (0.828) đạt cao nhất khi $\beta = 0.04$; AUPR (0.969) và AUC (0.966) đạt cao nhất khi $\beta = 0.05$; F1-score (0.916) đạt cao nhất khi $\beta = 0.08$, trong khi Precision (0.932) đỉnh điểm ở $\beta = 0.17$. Mặc dù $\beta = 0.05$ cho AUPR và AUC cao

nhất, sự chênh lệch so với giá trị tại $\beta = 0.04$ là không đáng kể (lần lượt là 0.968 so với 0.969 và 0.965 so với 0.966). Trong khi đó, tại $\beta = 0.04$, ba chỉ số quan trọng là Recall, Accuracy và MCC lại đồng thời đạt cực đại. Việc lựa chọn một giá trị β nhỏ hơn (0.04 thay vì 0.05, 0.08 hay 0.17) cũng giúp giảm thiểu nguy cơ quá khớp (overfitting) bằng cách loại bỏ nhiều thành phần nhiễu và tối ưu hóa khả năng khái quát hóa của mô hình. Do đó, dựa trên sự cân bằng tổng thể giữa các chỉ số và ưu tiên tính tổng quát, luận án lựa chọn $\beta = 0.04$ làm tham số mặc định cho tất cả các thí nghiệm tiếp theo.



Hình 2.7: Ma trận nhầm lẫn của các siêu đường dẫn.

Phân tích hiệu quả của học tổ hợp

Sau khi xác định được cấu hình SVD tối ưu, bước tiếp theo là đánh giá hiệu quả của chiến lược học tổ hợp. Hình 2.7 trình bày ma trận nhầm lẫn (confusion matrix) ứng với từng meta-path riêng lẻ và ma trận nhầm lẫn tổng hợp từ mô hình học tổ hợp.

Khi xem xét từng meta-path riêng lẻ, có thể nhận thấy hiệu suất dự đoán còn nhiều hạn chế. Cụ thể:

- Meta-path-1 cho thấy số lượng dương tính sai (FP = 207) và âm tính sai

(FN = 229) khá cân bằng. Điều này phản ánh rằng meta-path này duy trì được sự ổn định tương đối giữa việc phát hiện đúng mẫu dương tính và loại bỏ chính xác mẫu âm tính, mặc dù vẫn còn tồn tại tỷ lệ sai sót đáng kể.

- Ngược lại, Meta-path-2, Meta-path-3 và Meta-path-4 gặp khó khăn rõ rệt trong việc phân loại chính xác các trường hợp âm tính. Kết quả là tỷ lệ FP rất cao (lần lượt 316, 328 và 302). Điều này cho thấy các meta-path này có xu hướng dự đoán quá nhiều trường hợp dương tính so với thực tế, dẫn đến giảm độ đặc hiệu (Specificity) và làm suy giảm khả năng tổng quát hóa của mô hình nếu chỉ dựa vào một meta-path riêng lẻ.

Khi áp dụng phương pháp học tổ hợp (ensemble learning) để kết hợp dự đoán từ nhiều meta-path, hiệu suất được cải thiện đáng kể. Kết quả tổng hợp thể hiện qua ma trận nhầm lẫn cho thấy:

- FN giảm mạnh còn 149 (so với mức trên 200 ở các meta-path riêng lẻ), chứng tỏ mô hình tổ hợp khả năng phát hiện đúng các trường hợp dương tính được cải thiện rõ rệt.

- FP cũng giảm xuống 164, tức mô hình tổ hợp đã loại bỏ tốt hơn các trường hợp âm tính so với từng meta-path riêng biệt.

- Đồng thời, số lượng TN tăng lên 1.675 và TP đạt 1.666, phản ánh sự cân bằng và hiệu quả hơn trong cả hai chiều phân loại (dương tính và âm tính).

Những kết quả này khẳng định rằng việc kết hợp đa dạng các meta-path thông qua học tổ hợp là một chiến lược hiệu quả. Cơ chế này cho phép bù trừ sai số của từng thành phần riêng lẻ, tận dụng thế mạnh bổ sung của các quan hệ khác nhau trong mạng HIN, từ đó nâng cao độ chính xác và độ ổn định tổng thể của mô hình dự đoán

Để kiểm chứng tính ưu việt của phương pháp đề xuất, các thí nghiệm tiếp theo tiến hành so sánh với các mô hình phân loại truyền thống và hiện đại.

So sánh với các bộ phân loại truyền thống và hiện đại

Để đánh giá khách quan, luận án so sánh các mô hình đề xuất với hai nhóm phương pháp:

- Nhóm phương pháp cổ điển: PREDICT [85], TL-HGBI [86], LRSSL [87], SCMFDD [88], MBiRW [89].
- Nhóm phương pháp hiện đại dựa trên HIN/meta-path: EMP-SVD [26] (mô hình nền tảng), AICI2023 [**CT02**].
- So sánh thuật toán phân lớp: Để kiểm chứng việc lựa chọn LightGBM, các thí nghiệm cũng so sánh với Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) và Adaptive Classifier (AC) trên cùng một tập đặc trưng.

Kết quả thực nghiệm được trình bày trong Bảng 2.1 và được trực quan hóa trên Hình 2.8.

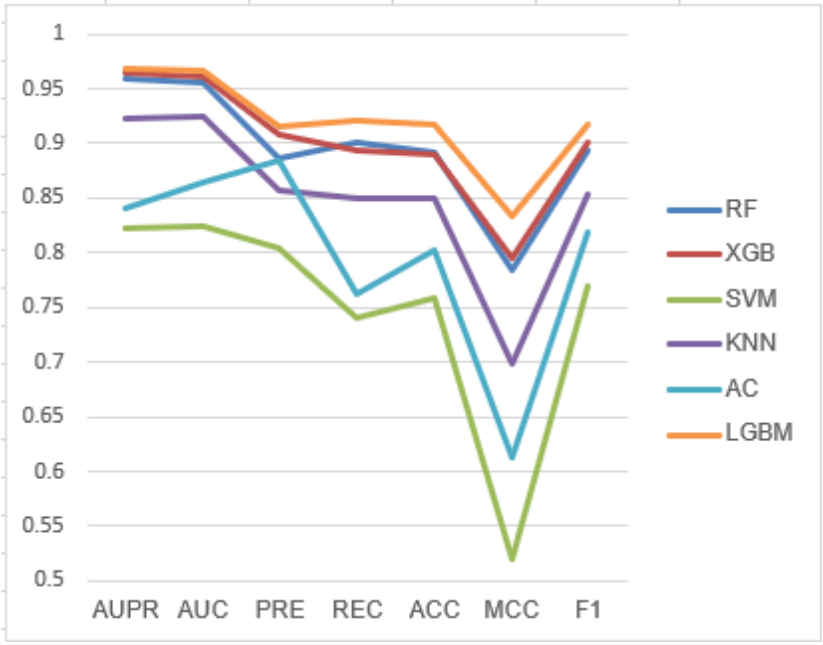
Bảng 2.1: Hiệu suất của các phương pháp liên quan trên tập dữ liệu

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
RF	0,959	0,953	0,888	0,891	0,887	0,776	0,889
XGB	0,964	0,961	0,909	0,893	0,891	0,796	0,900
SVM	0,822	0,824	0,805	0,740	0,759	0,520	0,770
KNN	0,923	0,925	0,857	0,850	0,849	0,699	0,853
AC	0,841	0,865	0,884	0,762	0,802	0,613	0,818
LightGBM	**0,969**	**0,966**	**0,915**	**0,921**	**0,917**	**0,834**	**0,918**

Các giá trị tốt nhất được in đậm.

Kết quả ở Bảng 2.1 cho thấy LightGBM đạt giá trị cao nhất trên tất cả các chỉ số, khẳng định tính vượt trội về khả năng dự đoán chính xác và khả năng tổng quát hóa. Đáng chú ý, ngay cả khi sử dụng Random Forest (RF) với 4 siêu đường dẫn mới (tập trung khai thác quan hệ qua protein trung gian) cũng cho kết quả tốt hơn trên mọi chỉ số so với mô hình EMP-SVD dùng 5 siêu đường

dẫn. Điều này chứng tỏ việc lựa chọn và thiết kế các siêu đường dẫn dựa trên vai trò trung gian của protein đã mang lại biểu diễn quan hệ tin cậy và hiệu quả hơn, tạo nền tảng vững chắc để các bộ phân lớp (đặc biệt là LightGBM) phát huy tối đa hiệu suất.



Hình 2.8: Hiệu suất các thuật toán

Nghiên cứu cắt bỏ về ảnh hưởng của SVD

Để kiểm chứng vai trò của kỹ thuật giảm chiều bằng SVD trong việc cải thiện hiệu suất, luận án tiến hành một nghiên cứu cắt bỏ (*ablation study*), trong đó các mô hình RF và LightGBM được huấn luyện với và không có SVD. Kết quả được trình bày trong Bảng 2.2.

Bảng 2.2: Kết quả nghiên cứu cắt bỏ về ảnh hưởng của SVD

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
RF no SVD	0,955	0,949	0,905	0,861	0,877	0,756	0,882
RF with SVD	0,959	0,953	0,888	0,891	0,887	0,776	0,889
LightGBM no SVD	0,960	0,955	0,898	0,893	0,894	0,787	0,895
LightGBM with SVD	**0,969**	**0,966**	**0,915**	**0,921**	**0,917**	**0,834**	**0,918**

Các giá trị tốt nhất được in đậm.

Kết quả cho thấy, LightGBM với SVD đạt hiệu suất cao nhất trên toàn bộ các thước đo, vượt trội so với cả RF và LightGBM không dùng SVD. Điều này khẳng định tác động tích cực của LightGBM kết hợp giảm chiều dựa trên SVD, giúp mô hình loại bỏ nhiễu và khai thác tốt hơn đặc trưng tiềm ẩn từ dữ liệu mạng không đồng nhất.

Kiểm định thống kê

Để xác nhận rằng sự cải thiện hiệu suất không phải ngẫu nhiên, luận án tiến hành kiểm định t-test hai mẫu cho các thước đo chính. Kết quả được trình bày trong Bảng 2.3.

Bảng 2.3: So sánh hiệu suất bằng kiểm định t-test hai mẫu

Phương pháp	**Phương pháp luận án(LGB có SVD)**						
	AUPR	AUC	PRE	REC	ACC	MMC	F1
RF without SVD	0,0128	0,00997	0,6240	0,0089	0,00236	0,0020	0,0017
RF with SVD	0,0055	0,0002	0,1180	0,0960	0,0044	0,0046	0,0024
LightGBM without SVD	0,0330	0,0150	0,1240	0,0793	0,0039	0,0038	0,0027

Kết quả kiểm định cho thấy, phương pháp đề xuất đạt cải thiện có ý nghĩa thống kê trên các chỉ số AUPR, AUC, ACC, MCC và F1-score (p-value $< 0,05$). Điều này đảm bảo rằng các cải thiện về hiệu suất không chỉ là kết quả ngẫu nhiên, mà phản ánh tác động thực sự của phương pháp. Dù Precision và Recall chưa đạt mức ý nghĩa thống kê, nhưng xu hướng tiến gần tới ngưỡng ý nghĩa gợi mở khả năng cải thiện thêm khi mở rộng dữ liệu hoặc tối ưu tham số.

So sánh với các phương pháp trong nghiên cứu trước

Cuối cùng, luận án so sánh mô hình DR-LGBM với các phương pháp được công bố trước đây. Nhóm phương pháp cổ điển bao gồm PREDICT, TL-HGBI,

LRSSL, SCMFDD, MBiRW; trong khi nhóm phương pháp gần đây có EMP-SVD, AICI2022, AICI2023 và Three meta-path. Kết quả được trình bày trong Bảng 2.4 và minh họa tại Hình 2.9.

Độ tin cậy của mô hình được chứng minh thông qua việc so sánh với các phương pháp được đề xuất trong các nghiên cứu trước đây, bao gồm: Dự đoán chỉ định thuốc (PREDICT) [85], Suy luận dựa trên đồ thị không đồng nhất ba lớp (TL-HGBI) [86], Học không gian con thưa có chuẩn hóa Laplacian (LRSSL) [87], Phân rã ma trận có ràng buộc độ tương đồng để dự đoán liên kết thuốc–bệnh (SCMFDD) [88] , và Phép duyệt ngẫu nhiên hai chiều với các thước đo tương đồng toàn diện (MBiRW) [89] . Các phương pháp này đều đã được giới thiệu cách đây hơn năm năm.

Cụ thể, PREDICT tận dụng độ tương đồng giữa thuốc và bệnh để xây dựng đặc trưng phân loại. TL-HGBI khai thác mối tương đồng giữa các thực thể như bệnh, thuốc và mục tiêu thuốc thông qua tích hợp thông tin đa chiều. LRSSL sử dụng ma trận Laplacian để trích xuất các đặc trưng quan trọng từ không gian thưa nhằm dự đoán mối liên hệ thuốc–bệnh. SCMFDD áp dụng phân rã ma trận có ràng buộc dựa trên độ tương đồng, trong khi MBiRW dựa vào các thước đo tương đồng toàn diện trong mạng không đồng nhất để xác định các liên kết tiềm năng.

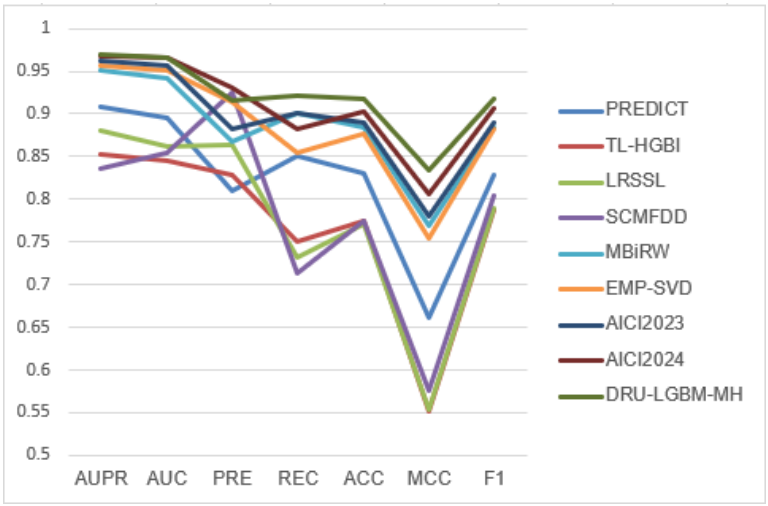
Trong vòng năm năm trở lại đây, một số phương pháp hiện đại đã được phát triển, bao gồm: Tập hợp meta-path kết hợp phân rã giá trị suy biến Wu [26] và Meta-path xử lý dữ liệu không đồng nhất (AICI2023) [CT02]. Các phương pháp này khai thác meta-paths trong mạng dữ liệu không đồng nhất để tính toán độ tương đồng giữa thuốc và bệnh. Đáng chú ý, tất cả các thuật toán trên đều được đánh giá trên cùng một bộ dữ liệu Wu [26].

Bảng 2.4 và Hình 2.9 tổng hợp so sánh toàn diện. DR-LGBM-MH đứng đầu về AUPR (0.969), AUC (0.966), Recall (0.921), ACC (0.917), MCC (0.834) và F1-score (0.918). Mặc dù Precision (0.915) thấp hơn một chút so với SCMFDD và AICI2023, nhưng Recall cao hơn đáng kể, cho thấy DR-LGBM-MH có khả năng phát hiện mẫu dương tốt hơn mà vẫn duy trì độ chính xác cao. So với mô

Bảng 2.4: Kết quả so sánh hiệu suất giữa các phương pháp

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
PREDICT	0,908	0,895	0,809	0,850	0,830	0,662	0,828
TL-HGBI	0,852	0,846	0,829	0,750	0,774	0,552	0,787
LRSSL	0,881	0,861	0,864	0,732	0,770	0,553	0,790
SCMFDD	0,836	0,854	0,926	0,713	0,774	0,575	0,805
MBiRW	0,952	0,942	0,867	0,901	0,884	0,769	0,884
EMP-SVD	0,956	0,951	0,913	0,854	0,876	0,755	0,882
AICI2023	0,968	0,966	**0,930**	0,882	0,903	0,806	0,906
DR-LGBM-MH	**0,969**	**0,966**	0,915	**0,921**	**0,917**	**0,834**	**0,918**

Các giá trị tốt nhất được in đậm.



Hình 2.9: So sánh hiệu năng giữa các mô hình dự đoán thuốc–bệnh khác nhau

hình nền tảng EMP-SVD, DR-LGBM-MH cải thiện vượt bậc, đặc biệt là Recall (từ 0.854 lên 0.921) và F1-score (từ 0.882 lên 0.918).

Phân tích tổng hợp và kết luận về DR-LGBM-MH

Nhìn chung, DR-LGBM-MH không chỉ đạt được giá trị cao ở từng chỉ số riêng lẻ mà còn duy trì được sự cân bằng giữa các tiêu chí. Điều này không chỉ xác nhận tính hiệu quả của DR-LGBM-MH trên dữ liệu hiện tại, mà còn chứng minh tính vượt trội khi đặt trong cùng bối cảnh với các phương pháp đã được cộng đồng nghiên cứu công bố rộng rãi.

Tóm lại, kết quả so sánh tổng thể cho thấy phương pháp DR-LGBM-MH

không chỉ vượt trội so với các phương pháp tiên tiến đã công bố mà còn cải thiện đáng kể so với chính mô hình nền tảng EMP-SVD. Sự vượt trội này xuất phát từ hai cải tiến then chốt: (1) Hệ thống siêu đường dẫn mới tập trung vào protein trung gian và (2) Việc áp dụng thuật toán LightGBM tối ưu.

Đặc biệt, kết quả từ Bảng 2.1 và Bảng 2.2 cho thấy, ngay cả khi sử dụng cùng thuật toán Random Forest (RF), 4 siêu đường dẫn mới của DR-LGBM-MH đã cho kết quả tốt hơn 5 siêu đường dẫn của EMP-SVD. Điều này chứng tỏ chất lượng và sự tập trung thông tin của siêu đường dẫn quan trọng hơn số lượng. Việc thiết kế các siêu đường dẫn mới (như $d \to t \to s \to d \to s$) nhấn mạnh vai trò trung gian của protein đã tạo ra biểu diễn quan hệ giàu ngữ nghĩa và hiệu quả hơn. Trên nền tảng siêu đường dẫn tối ưu đó, LightGBM – với khả năng xử lý dữ liệu lớn, thưa và học các quan hệ phi tuyến phức tạp – đã phát huy tối đa hiệu suất, dẫn đến sự cải thiện vượt bậc, đặc biệt là khả năng phát hiện (Recall) và chất lượng phân loại tổng thể (MCC, F1).

Các biểu thức (2.13)–(2.19) trong mô hình đề xuất chủ yếu bao gồm các phép nhân ma trận. Xét phép nhân hai ma trận đặc trưng $A \in \mathbb{R}^{m \times n}$ và $B \in \mathbb{R}^{n \times z}$, chi phí tính toán của phép nhân này là $O(mnz)$. Do đó, trong trường hợp coi số phép nhân ma trận trong mô hình DR-LGBM-MH là một hằng số và xét theo bậc lớn, độ phức tạp của phần xây dựng biểu diễn có thể được xấp xỉ là $O(mnz)$.

Như vậy, sự thành công của DR-LGBM-MH là minh chứng cho một chiến lược hiệu quả: kết hợp giữa việc thiết kế hệ thống siêu đường dẫn có định hướng sinh học và lựa chọn thuật toán học máy phù hợp với đặc thù dữ liệu, mở ra hướng tiếp cận mạnh mẽ cho bài toán dự đoán liên kết thuốc–bệnh trên mạng không đồng nhất.

Các nghiên cứu điển hình

Phần này của luận án tập trung phân tích các cặp thuốc – bệnh có giá trị dự đoán cao nhất, được kiểm chứng thủ công thông qua việc tra cứu kỹ lưỡng các công trình nghiên cứu đã công bố và nguồn dữ liệu khoa học trực tuyến.

Việc đối chiếu với tài liệu y học giúp đánh giá độ tin cậy sinh học của mô hình dự đoán, đồng thời xác định các ứng viên thuốc có tiềm năng tái định vị trong điều trị bệnh. Cụ thể, trong bộ dữ liệu, nhiều cặp thuốc – bệnh không chỉ đạt xác suất dự đoán cao (0,99) mà còn được chứng thực bởi bằng chứng y học thực nghiệm hoặc lâm sàng.

Bảng 2.5 trình bày 10 loại thuốc ứng viên hàng đầu cho nhiều loại bệnh khác nhau. Kết quả cho thấy, đối với bệnh "Renal failure, progressive, with hypertension; RFH1", ba loại thuốc Bisoprolol, Atenolol và Metipranolol được xác định là các ứng viên điều trị tiềm năng. Các thuốc này đều thuộc nhóm β-blocker, có tác dụng làm giảm nhịp tim, giảm sức co bóp của cơ tim và hạ huyết áp, do đó phù hợp về mặt cơ chế sinh học với bệnh cảnh suy thận tiến triển kèm tăng huyết áp. Sự phù hợp này được củng cố bởi các công trình nghiên cứu mang mã PMID: 18165208, 7408390 và 2870216, xác nhận tính hiệu quả của nhóm thuốc β-blocker trong việc giảm tiến triển tổn thương thận do tăng huyết áp mạn tính.

Tương tự, đối với bệnh "Neuropathy, hereditary sensory and autonomic, type I, with cough and gastroesophageal reflux", hai loại thuốc Nortriptyline và Imipramine được xác định là các ứng viên có tiềm năng điều trị cao. Đây là các thuốc thuộc nhóm tricyclic antidepressants (TCA), có cơ chế tác động lên hệ thần kinh trung ương, giúp giảm đau thần kinh và điều hòa hoạt động cảm giác – tự động. Điều này cho thấy mối tương quan hợp lý giữa đặc tính dược lý và cơ chế bệnh sinh, đã được xác nhận bởi các nghiên cứu có mã PMID: 25569864 và 16311270.

Bên cạnh đó, Bảng 2.6 trình bày danh sách 10 loại thuốc có xác suất dự đoán cao nhất cho bệnh Suy thận tiến triển kèm theo tăng huyết áp "Renal failure, progressive, with hypertension". Các kết quả này tiếp tục khẳng định tính ổn định của mô hình dự đoán, khi các thuốc như Bisoprolol, Atenolol và Metipranolol vẫn giữ vị trí hàng đầu. Ngoài ra, Epinephrine và Ergotamine cũng được xác minh qua các công trình nghiên cứu có mã PMID: 4012388 và 24029265, cho thấy mối liên hệ sinh lý học giữa các hoạt chất tác động trên hệ tim

Bảng 2.5: Top 10 thuốc ứng viên tiềm năng cho các bệnh khác nhau

Xếp hạng	**Tên thuốc**	**Tên bệnh**	**Xác suất**	**Tài liệu xác thực**
1	Dextromethorphan	Insensitivity to pain with hyperplastic myelinopathy	0,99	NA
2	Nortriptyline	Neuropathy, hereditary sensory and autonomic, type I, with cough and gastroesophageal reflux	0,99	PMID:25569864 [90]
3	Doxazosin	Neuropathy, hereditary sensory and autonomic, type I, with cough and gastroesophageal reflux	0,99	NA
4	Bisoprolol	Renal failure, progressive, with hypertension; RFH1	0,99	PMID:18165208 [91]
5	Terazosin	Neuropathy, hereditary sensory and autonomic, type I, with cough and gastroesophageal reflux	0,99	NA
6	Atenolol	Renal failure, progressive, with hypertension; RFH1	0,99	PMID:7408390 [92]
7	Metipranolol	Renal failure, progressive, with hypertension; RFH1	0,99	PMID:2870216 [93]
8	Imipramine	Neuropathy, hereditary sensory and autonomic, type I, with cough and gastroesophageal reflux	0,99	PMID:16311270 [94]
9	Alfuzosin	Neuropathy, hereditary sensory and autonomic, type I, with cough and gastroesophageal reflux	0,99	NA
10	Prazosin	Neuropathy, hereditary sensory and autonomic, type I, with cough and gastroesophageal reflux	0,99	NA

mạch và chức năng thận. Tuy nhiên, các thuốc Flavoxate, Florinef, Midodrine và Phenylephrine – mặc dù có xác suất dự đoán cao (0,99) – hiện chưa có công bố khoa học xác thực. Điều này gợi ý rằng các cặp thuốc – bệnh này có thể đại diện cho các hướng khám phá mới, đặc biệt nếu được kiểm chứng qua thí nghiệm in vitro, in vivo hoặc nghiên cứu lâm sàng giai đoạn sớm. Mặc dù một số cặp thuốc – bệnh trong Bảng 2.5 và Bảng 2.6 chưa được xác nhận bởi tài liệu hiện có, song những dự đoán này vẫn mang giá trị khoa học và thực tiễn cao. Xác suất dự đoán lớn (0.99) phản ánh mức độ tương đồng cao trong không gian đặc trưng sinh học giữa thuốc và bệnh, cho thấy mô hình đã nắm bắt được mối liên hệ ngầm giữa cấu trúc phân tử, đặc tính dược lý và biểu hiện bệnh học. Do đó, các cặp này xứng đáng được xem xét trong các nghiên cứu tiền lâm sàng hoặc

Bảng 2.6: Top 10 dự đoán thuốc cho bệnh suy thận tiến triển kèm theo tăng huyết áp

Xếp hạng	Tên thuốc	Xác suất	Tài liệu xác thực
1	Bisoprolol	0,99	PMID:18165208 [91]
2	Atenolol	0,99	PMID:7408390 [92]
3	Metipranolol	0,99	PMID:2870216 [93]
4	Epinephrine	0,99	PMID: 4012388 [95]
5	Flavoxate	0,99	NA
6	Florinef	0,99	NA
7	Ergotamine	0,99	PMID:24029265 [96]
8	Amiodarone	0,99	PMID:19640392 [97]
9	Midodrine	0,99	NA
10	Phenylephrine	0,99	NA

thử nghiệm lâm sàng có kiểm soát, nhằm xác định khả năng ứng dụng thực tiễn trong tái định vị thuốc và phát triển liệu pháp điều trị mới.

2.4.6. Kết quả và thảo luận cho mô hình HS-TMP

Mô hình và tham số

Mô hình có sử dụng 6 ma trận tương quan đồng nhất và 3 nhóm siêu đường dẫn.

Tham số. Để thể hiện tính rõ ràng cho phần thực nghiệm, các tham số được cấu hình như sau: Đối với meta-path 1, tham số thuốc $d \in [0, 1]$ có hai tham số để tính toán liên kết thuốc-thuốc DD, tham số bệnh $s \in [0, 1]$ có hai tham số để tính toán liên kết thuốc-thuốc SS, cụ thể:

- Tham số đầu tiên $(d) = \mathbf{0}$ sử dụng C_{dd}^{s} luận án đã đánh dấu `m1d0`
- Tham số đầu tiên $(d) = \mathbf{1}$ Sử dụng C_{dd}^{t} luận án đánh dấu là `m1d1`
- Tham số đầu tiên $(s) = \mathbf{0}$ sử dụng C_{ss}^{d} luận án đã đánh dấu `m1s0`
- Tham số đầu tiên $(s) = \mathbf{1}$ Sử dụng C_{ss}^{t} luận án đánh dấu là `m1s1`

Tương tự như vậy, siêu đường dẫn thứ hai và đường dẫn thứ ba chứa các tham số. Có 2^2 tùy chọn, là sự kết hợp của các tham số đầu vào d, s cho siêu đường

dẫn 1, 2^2 tùy chọn cho d, p trong siêu đường dẫn 2 và 2^3 tùy chọn cho d, p, s trong siêu đường dẫn 3 theo phương pháp được mô tả ở phần 3. Như vậy, có tổng cộng $2^7 = 128$ lựa chọn học. Trong quá trình huấn luyện cho từng tùy chọn học, ba siêu đường dẫn đã được thực hiện và sau đó một phương pháp tổng hợp được thực hiện từ các đường dẫn để có được phương pháp học tổng hợp.

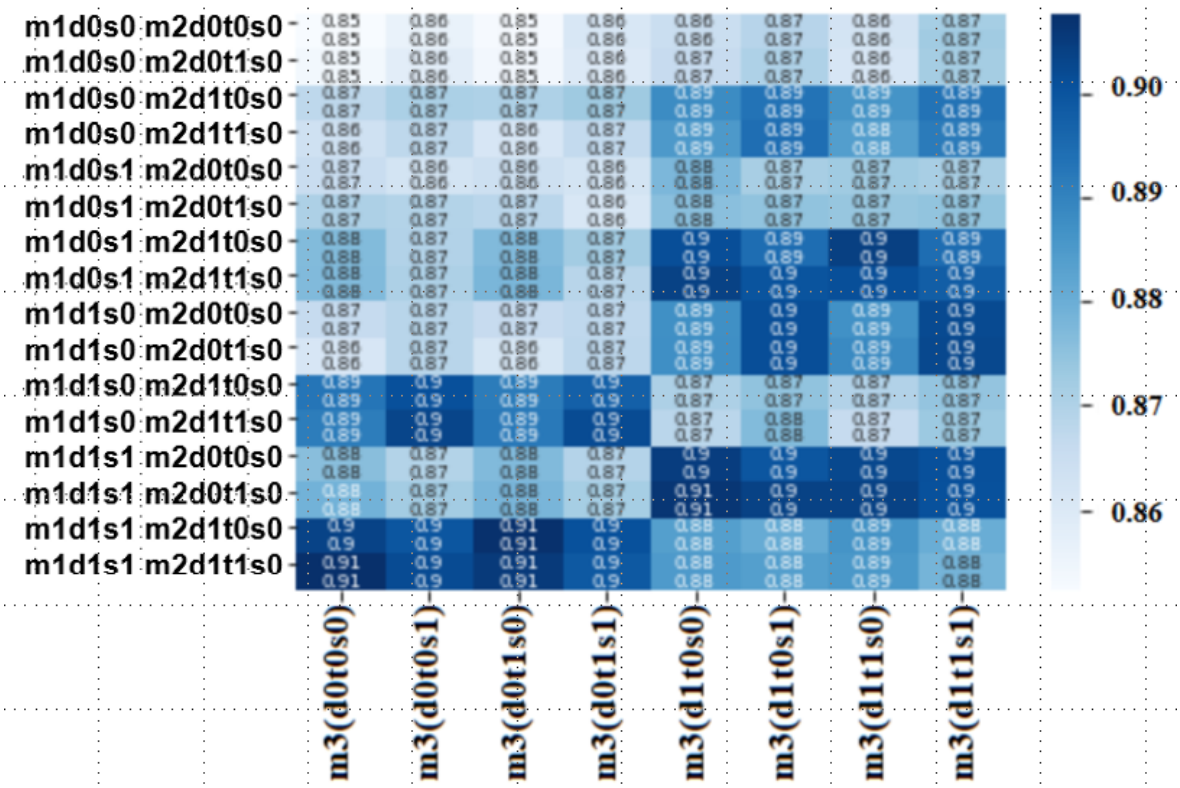
Kết quả đánh giá của tất cả 128 cấu hình được tổng hợp trong Hình 2.10. Có thể thấy, độ chính xác (Accuracy) phân bổ trong một dải khá rộng, với nhiều cấu hình đạt hiệu suất trên 0.9. Điều này cho thấy việc tích hợp thông tin đồng nhất có tiềm năng lớn, nhưng hiệu quả cụ thể phụ thuộc nhiều vào cách kết hợp các mối quan hệ.

Kết quả thực nghiệm của ba meta-path

Luận án có thể minh họa các tham số đã chọn tương ứng với giá trị chính xác có thể. Bản đồ trong Hình 2.10 khám phá sự tương ứng bằng cách sử dụng trục dọc cho các tham số của siêu đường dẫn 1 và siêu đường dẫn 2, trong khi trục ngang được bao phủ bởi các tham số của siêu đường dẫn 3. Đặc biệt, độ chính xác cao hơn 0,9 đã được nhìn thấy ở một số nơi trên bản đồ.

Để báo cáo kết quả tốt nhất của việc thực hiện ba đường dẫn bằng 2^7 tùy chọn, Bảng 2.7 sử dụng một ghi chú cụ thể cho một tùy chọn bao gồm các tham số. Tại đường dẫn 1, lưu ý của $m1(ds)$ là hiển thị thông số của thuốc (d) cho độ chính xác thay đổi từ 0,906 đến 0,908. Trên thực tế, chỉ có một lưu ý về con đường 1 yêu cầu cập nhật trước của thuốc bằng tương tác thuốc-protein và trước khi bệnh bằng tương tác bệnh-protein. Tuy nhiên, con đường thứ ba cần cập nhật trước bệnh bằng tương tác thuốc-bệnh vì số 0 có thể được nhìn thấy ở cuối tất cả các ghi chú trong cột thứ ba. Và bệnh có thể là 0 hoặc 1. Vì vậy, nút 1-1 trong trường hợp này có nghĩa là $d = 1, s = 1$.

Bảng 2.7 cho thấy tám lựa chọn tốt nhất cho độ chính xác thay đổi từ 0,906 đến 0,908. Trên thực tế, chỉ có một lưu ý 1-1 cho đường 1 $m1(ds)$, yêu cầu cập nhật giá trị trước của thuốc bằng tương tác thuốc-protein và giá trị trước của bệnh bằng tương tác bệnh-protein. Tuy nhiên, đường dẫn thứ ba cần cập



Hình 2.10: Bản đồ độ chính xác của các tham số meta-path

nhật trước bệnh bằng tương tác thuốc-bệnh vì số 0 có thể được nhìn thấy ở cuối tất cả các ghi chú trong cột thứ ba.

So sánh với các nghiên cứu trước

Để đánh giá hiệu quả của phương pháp đề xuất, luận án tiếp tục so sánh với một số mô hình tiên tiến trước đó, bao gồm EMP-SVD, LRSSL, MBiRW, MPG-DDA, PREDICT, SCMFDD và TL-HGBI. Kết quả so sánh được trình bày trong Bảng 2.8 theo các chỉ số AUPR, AUC, PRE, REC, ACC, MCC và F1-score.

Kết quả trong Bảng 2.8 cho thấy phương pháp của luận án đạt hiệu suất cao nhất trên hầu hết các chỉ số. Cụ thể, AUPR đạt 0,968 và AUC đạt 0,963 — đều vượt trội hơn các phương pháp so sánh khác, phản ánh khả năng phân loại và dự đoán liên kết thuốc–bệnh vượt trội.

Mặc dù các mô hình như EMP-SVD và MBiRW cũng thể hiện hiệu năng tốt (AUPR lần lượt 0,956 và 0,952), phương pháp của luận án vẫn vượt trội hơn

Bảng 2.7: 8 cấu hình meta-path cho kết quả tốt nhất

$m1(ds)$	$m2(ds)$	$m3(ds)$	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
1-1	1-1-0	0-0-0	0.968	0.963	0.895	0.922	0.908	0.816	0.908
1-1	1-1-1	0-0-0	0.968	0.963	0.895	0.922	0.908	0.816	0.908
1-1	1-0-0	0-1-0	0.968	0.963	0.903	0.914	0.907	0.815	0.909
1-1	1-0-1	0-1-0	0.968	0.963	0.903	0.914	0.907	0.815	0.909
1-1	0-1-0	1-0-0	0.966	0.960	0.906	0.910	0.906	0.813	0.908
1-1	0-1-1	1-0-0	0.966	0.960	0.906	0.910	0.906	0.813	0.908
1-1	1-1-0	0-1-0	0.967	0.963	0.897	0.915	0.906	0.811	0.906
1-1	1-1-1	0-1-0	0.967	0.963	0.897	0.915	0.906	0.811	0.906

Bảng 2.8: Hiệu suất của các phương pháp liên quan

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
EMP-SVD	0,956	0,951	0,913	0,854	0,876	0,755	0,882
LRSSL	0,881	0,861	0,864	0,732	0,770	0,553	0,790
MBiRW	0,952	0,942	0,867	0,901	0,884	0,769	0,884
MPG-DDA	0,944	0,930	0,886	0,842	0,867	NA	0,863
PREDICT	0,908	0,895	0,809	0,850	0,830	0,662	0,828
SCMFDD	0,836	0,854	**0,926**	0,713	0,774	0,575	0,805
TL-HGBI	0,852	0,846	0,829	0,750	0,774	0,552	0,787
Our method	**0,968**	**0,963**	0,895	**0,922**	**0,908**	**0,816**	**0,908**

Các giá trị tốt nhất được in đậm.

ở cả độ nhạy (REC = 0,922) và độ chính xác tổng thể (ACC = 0,908). Trong khi SCMFDD đạt giá trị PRE cao nhất (0,926), DR-LGBM-MH vẫn đạt PRE = 0,895, cho thấy khả năng dự đoán dương tính chính xác và duy trì sự cân bằng tốt giữa các chỉ số đánh giá.

Đặc biệt, chỉ số MCC = 0,816 và F1 = 0,908 chứng minh mô hình duy trì được sự cân bằng giữa dự đoán đúng và sai, đồng thời đảm bảo tính ổn định của kết quả. Những cải thiện này cho thấy mô hình đề xuất không chỉ nâng cao độ chính xác mà còn tăng tính tin cậy và khả năng khái quát hóa trong bài toán dự đoán tương tác thuốc–bệnh.

Kết quả này cho thấy mô hình HS-TMP, dựa trên việc khai thác có hệ thống các quan hệ đồng nhất, đạt được hiệu suất cạnh tranh và trong đó chỉ số

như Recall còn vượt trội so với các phương pháp hiện đại.

Từ góc độ ứng dụng, các nhóm siêu đường dẫn này có giá trị thực tiễn trong bài toán tái định vị thuốc ở giai đoạn tiền lâm sàng. Cụ thể, nhóm thứ nhất phù hợp với việc sàng lọc nhanh các ứng viên thuốc cho các bệnh có tính tương đồng cao; nhóm thứ hai hỗ trợ phát hiện các ứng viên gắn với cơ chế phân tử và mạng chức năng protein; còn nhóm thứ ba đặc biệt hữu ích cho việc phát hiện các chỉ định mới phức tạp, nơi quan hệ thuốc–bệnh không thể được giải thích bằng một liên kết trực tiếp đơn lẻ. Vì vậy, ý nghĩa ứng dụng của ba nhóm siêu đường dẫn nằm ở khả năng ưu tiên hóa các cặp thuốc–bệnh tiềm năng để phục vụ các bước xác thực tiếp theo như phân tích tài liệu y sinh, thí nghiệm in vitro, in vivo hoặc nghiên cứu tiền lâm sàng có kiểm soát. Kết quả thực nghiệm tốt của mô hình HS-TMP cho thấy các siêu đường dẫn được đề xuất không chỉ hợp lý về mặt cấu trúc toán học mà còn có giá trị trong việc biểu diễn các quan hệ sinh học tiềm ẩn phục vụ khám phá thuốc.

Nghiên cứu cắt bỏ SVD và số lượng meta-path

Để làm rõ vai trò của kỹ thuật phân rã giá trị kỳ dị (SVD) trong việc trích xuất đặc trưng, luận án tiến hành một thí nghiệm bổ sung. Trong thí nghiệm này, SVD không được sử dụng và toàn bộ đặc trưng được đưa trực tiếp vào quá trình huấn luyện. Đồng thời, luận án cũng khảo sát hiệu quả của phương pháp đề xuất khi chỉ kết hợp hai meta-path (m2_1, m2_2, m2_3, m2_4) theo các cách kết hợp khác nhau. Kết quả so sánh được trình bày trong Bảng 2.9.

Cụ thể, phương pháp không sử dụng SVD với ba meta-path đạt AUPR = 0,923 và AUC = 0,920, thấp hơn đáng kể so với khi tích hợp SVD (AUPR = 0,968, AUC = 0,963). Điều này cho thấy việc giảm chiều dữ liệu thông qua SVD có vai trò quan trọng trong việc loại bỏ nhiễu và giữ lại thông tin đặc trưng có ý nghĩa cho mô hình. Trong khi đó, phương pháp chỉ sử dụng hai meta-path cho kết quả kém hơn rõ rệt (AUPR = 0,845, AUC = 0,833), chứng minh rằng số lượng và sự đa dạng của meta-path có ảnh hưởng trực tiếp đến hiệu quả dự đoán.

Bảng 2.9: So sánh hiệu suất theo số lượng meta-path và SVD

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
Our method with 3 paths no SVD	0,923	0,920	0,867	0,838	0,848	0,696	0,852
Our method with 2 paths	0,845	0,833	0,822	0,731	0,756	0,518	0,773
Our method	**0,968**	**0,963**	**0,895**	**0,922**	**0,908**	**0,816**	**0,908**

Các giá trị tốt nhất được in đậm.

Kết quả trong Bảng 2.9 cho thấy rằng cả hai yếu tố — số lượng meta-path và việc áp dụng SVD — đều có ảnh hưởng đáng kể đến hiệu suất của mô hình. Khi chỉ sử dụng hai meta-path, các chỉ số hiệu suất giảm đáng kể (AUPR = 0,845; AUC = 0,833; F1 = 0,773), phản ánh việc thiếu thông tin quan hệ trong mạng không đồng nhất.

Trong khi đó, việc mở rộng lên ba meta-path giúp cải thiện rõ rệt các chỉ số (AUPR = 0,923; AUC = 0,920). Tuy nhiên, khi kết hợp thêm phép phân rã SVD để rút trích đặc trưng ẩn và giảm nhiễu trong không gian đặc trưng, hiệu suất tăng mạnh trên toàn bộ các tiêu chí, đặc biệt với AUPR = 0,968 và F1 = 0,908.

Ngoài ra, để chứng minh tính hiệu quả của các phương pháp mà luận án đề xuất trong việc khai thác thông tin từ meta-path trong mạng không đồng nhất, luận án đã tiến hành đối sánh với các hướng tiếp cận gần đây dựa trên học biểu diễn đồ thị và cơ chế khai thác meta-path như metapath2vec, các mô hình attention phân cấp, cũng như các kiến trúc transformer có tích hợp meta-path. Các nghiên cứu tiêu biểu có thể kể đến như DRMGNE [58] (2025), phương pháp tích hợp meta-path với lấy mẫu âm thích ứng, phương pháp tổng hợp meta-path dựa trên cơ chế attention phân cấp, hay các mô hình transformer với chiến lược pruning meta-path động.

Trong phạm vi thực nghiệm của luận án, phương pháp DRMGNE được lựa chọn làm đại diện so sánh trực tiếp do đây là một trong những mô hình tiêu biểu và gần đây nhất, kết hợp hiệu quả giữa hướng dẫn meta-path và chiến lược lấy mẫu âm thích ứng trong bài toán dự đoán liên kết thuốc–bệnh. Kết quả

so sánh được thể hiện trong Bảng 2.10. Kết quả trong Bảng 2.10 cho thấy các

Bảng 2.10: So sánh với các phương pháp sử dụng meta-path gần đây

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
DRMGNE	0.953	0.962	0.912	**0.924**	**0.917**	**0.836**	**0.918**
DR-LGBM-MH	**0,969**	**0,966**	**0,915**	0,921	**0,917**	0,834	**0,918**
Our method	0,968	0,963	0,895	0,922	0,908	0,816	0,908

Các giá trị tốt nhất được in đậm.

phương pháp đề xuất trong luận án đạt hiệu năng cạnh tranh và nhìn chung vượt trội so với phương pháp DRMGNE (2025) trong việc khai thác thông tin từ các meta-path. Cụ thể, mô hình DR-LGBM-MH đạt giá trị AUPR cao nhất (0.969) và AUC cao nhất (0.966), cho thấy khả năng phân biệt giữa các cặp thuốc–bệnh có liên kết và không có liên kết là tốt hơn so với phương pháp đối sánh. Đồng thời, chỉ số Recall của các phương pháp đề xuất (0.921 và 0.922) cũng duy trì ở mức cao, phản ánh khả năng phát hiện đầy đủ các liên kết tiềm năng.

Tuy nhiên, có thể nhận thấy một số chỉ số như Precision và MCC của phương pháp đề xuất chưa vượt trội hoàn toàn so với DRMGNE. Điều này cho thấy mặc dù mô hình có khả năng phát hiện nhiều liên kết hơn (Recall cao), nhưng vẫn tồn tại một số dự đoán sai dương tính. Nguyên nhân có thể đến từ việc các đặc trưng tiềm ẩn được học từ SVD mang tính khái quát cao, dẫn đến sự chồng lấn trong không gian đặc trưng giữa các cặp có và không có liên kết.

Điều này khẳng định rằng việc khai thác cấu trúc siêu đường dẫn cùng với kỹ thuật SVD không chỉ nâng cao khả năng học biểu diễn mà còn giúp mô hình đạt được sự cân bằng tối ưu giữa độ chính xác và độ bao phủ trong dự đoán liên kết thuốc–bệnh.

Luận án đề xuất phân tích các mối liên hệ bệnh-thuốc bằng cách trình bày các mối liên hệ thông qua ba con đường meta mới. Qua các thí nghiệm, luận án cũng đã chứng minh đây là điểm mới và đóng góp chính của luận án, thể hiện vai trò của ba meta-path mới này. Tuy nhiên, một hạn chế của luận án là chưa được giải thích một cách khoa học bằng cách sử dụng các cơ sở y sinh

để thấy được ý nghĩa thực tiễn. Nếu thực hiện được, đây sẽ là một đóng góp đột phá. Hiện nay, nghiên cứu chủ yếu chứng minh điều đó thông qua các thí nghiệm và đo lường để đánh giá.

Tương tự như phân tích độ phức tạp của mô hình DR-LGBM-MH, trong mô hình HS-TMP luận án đề xuất sử dụng các công thức (2.23) đến (2.44) Độ phức tạp trong mô hình này là $O(mnz)$.

Các nghiên cứu điển hình

Ở bước tiếp theo, mô hình được mở rộng bằng cách sử dụng 6 ma trận quan hệ kết hợp với 3 nhóm siêu đường dẫn, nhằm khai thác tốt hơn mối quan hệ ẩn giữa các thực thể trong mạng không đồng nhất.

Thông qua phương pháp chuyển nhãn (label transfer), các mối quan hệ thuốc – bệnh mới được suy diễn từ các mối quan hệ đã biết. Đáng chú ý, một số mối liên hệ mới được phát hiện không xuất hiện trong tập dữ liệu ban đầu, cũng như không trùng lặp với các kết quả từ mô hình Wu [26].

Mối liên hệ giữa fludrocortisone và bệnh hạ huyết áp tư thế đứng được phát hiện. Tài liệu của Veazie và cộng sự [98] chỉ ra rằng Fludrocortisone – một loại thuốc mineralocorticoid – có tác dụng tăng thể tích máu và huyết áp, được xem là một trong các liệu pháp điều trị đầu tay cho bệnh lý này.

Một trường hợp khác là mối liên hệ giữa oseltamivir (Tamiflu) [81] và bệnh não. Dù oseltamivir thường dùng để điều trị cúm, nghiên cứu của Yen và cộng sự đã ghi nhận một trường hợp thực tế trong đó bệnh nhân mắc bệnh não đã hồi phục sau khi được điều trị bằng oseltamivir, cho thấy tiềm năng điều trị ngoài chỉ định của loại thuốc này.

Mặc dù không phải tất cả các liên kết mới đều có tài liệu xác nhận, kết quả này cho thấy khả năng của mô hình trong việc đề xuất các mối quan hệ thuốc – bệnh có tiềm năng, hỗ trợ hiệu quả cho công tác tái định vị thuốc và mở rộng hiểu biết y sinh.

2.5. Kết luận chương 2

Chương 2 đã trình bày và đánh giá chi tiết hai mô hình dự đoán liên kết thuốc-bệnh dựa trên việc khai thác nâng cao cấu trúc mạng thông tin không đồng nhất (HIN).

Mô hình DR-LGBM-MH tập trung vào việc đề xuất và kết hợp các siêu đường dẫn (meta-path) mới có chứa protein, kết hợp với bộ phân lớp LightGBM và kỹ thuật giảm chiều SVD. Kết quả thực nghiệm cho thấy mô hình này đạt hiệu suất vượt trội (AUPR=0.969, F1=0.918) so với các phương pháp so sánh, đặc biệt trong việc cải thiện độ nhạy (Recall) mà vẫn duy trì độ chính xác tổng thể cao.

Mô hình HS-TMP đi sâu vào việc tích hợp các quan hệ đồng nhất (thuốc-thuốc, bệnh-bệnh, protein-protein) thông qua ba nhóm siêu đường dẫn. Mặc dù độ phức tạp cao với 128 cấu hình con, mô hình vẫn đạt được hiệu suất rất cạnh tranh (AUPR=0.968, Recall=0.922), chứng minh giá trị của thông tin đồng nhất trong việc làm giàu ngữ nghĩa cho HIN.

Cả hai mô hình đều được xác thực bằng các nghiên cứu điển hình, cho thấy khả năng dự đoán các liên kết thuốc-bệnh tiềm năng có ý nghĩa sinh học. Tuy nhiên, các thách thức như âm tính giả và sự mất cân bằng dữ liệu vẫn tồn tại, ảnh hưởng đến độ tin cậy của một số dự đoán. Để giải quyết triệt để những vấn đề này, Chương 3 sẽ đề xuất một hướng tiếp cận mới dựa trên Suy luận Bayes, nhằm mô hình hóa trực tiếp sự không chắc chắn và cải thiện độ ổn định của mô hình dự đoán. Các đóng góp ở chương 2 đã được công bố trong các công trình nghiên cứu [**CT01**], [**CT02**], [**CT03**], và [**CT04**] .

CHƯƠNG 3. SUY LUẬN BAYES TRONG DỰ ĐOÁN LIÊN KẾT THUỐC - BỆNH

Chương trước đã phân tích những hạn chế cốt lõi của các mô hình dựa trên meta-path (như EMP-SVD và các mở rộng của nó) trong bài toán dự đoán thuốc-bệnh, đặc biệt là các vấn đề như dữ liệu thưa, mất cân bằng, nguy cơ âm tính giả và khả năng giải thích hạn chế. Mặc dù Chương 2 đã có những cải tiến đáng kể bằng cách mở rộng hệ thống siêu đường dẫn và tích hợp thông tin đồng nhất, nhưng các thách thức nền tảng liên quan đến chất lượng dữ liệu huấn luyện và tính bất định trong kiến thức sinh học vẫn chưa được giải quyết một cách nguyên lý.

Do đó, Chương này đề xuất một hướng tiếp cận mới, đột phá hơn, dựa trên nền tảng của Suy luận Bayes. Thay vì tiếp tục mở rộng các đặc trưng thông qua meta-path, luận án này chuyển sang một mô hình xác suất nguyên lý, nhằm trực tiếp mô hình hóa các cơ chế sinh học, tích hợp sự không chắc chắn, và cải thiện triệt để chất lượng dữ liệu. Phương pháp luận mới này hứa hẹn mang lại tính minh bạch cao hơn, khả năng giải thích tốt hơn, và độ tin cậy được nâng cao cho các dự đoán, được thể hiện qua mô hình DDA-BNS (Drug-Disease Association prediction using Bayesian inference with Enhanced Negative Sampling and Data Balancing).

3.1. Mô hình DDA-BNS

3.1.1. Giới thiệu về DDA-BNS

Ở chương 1 luận án đã giới thiệu về mô hình EMP-SVD cũng như mô tả chi tiết khung mô hình DR-LGBM-MH trong đó các bước cơ ban gồm: Bước 1: mạng không đồng nhất thuốc -protein-bệnh; Bước 2: Đề xuất siêu đường dẫn; Bước 3: Phân tách giá trị trị kỳ dị và vector thuốc bệnh; Bước 4: Huấn luyện mô

hình; Bước 5 dự đoán và đánh giá. Trong mô hình này luận án kế thừa khung mô hình DR-LGBM-MH với các đóng góp mới như (1) Suy luận Bayes, (2) Lọc mẫu âm chất lượng cao, và (3) Cân bằng dữ liệu. Tuy nhiên, thay vì tiếp tục mở rộng hệ thống siêu đường dẫn (meta-path), DDA-BNS tập trung vào ba cải tiến mang tính đột phá về phương pháp luận, được tích hợp vào các bước then chốt của khung này:

1. **Thay thế siêu đường dẫn bằng suy luận Bayes:** sử dụng xác suất hậu nghiệm thay cho meta-path để mô hình hóa mối quan hệ thuốc–bệnh.
2. **Đề xuất cơ chế chọn mẫu âm chất lượng cao:** nhận diện và loại bỏ các mẫu âm tính giả dựa trên thông tin sinh học.
3. **Áp dụng chiến lược cân bằng dữ liệu hiệu quả hơn:** thay thế chọn mẫu âm ngẫu nhiên bằng các kỹ thuật cân bằng dữ liệu phù hợp với mạng thông tin đa nguồn.

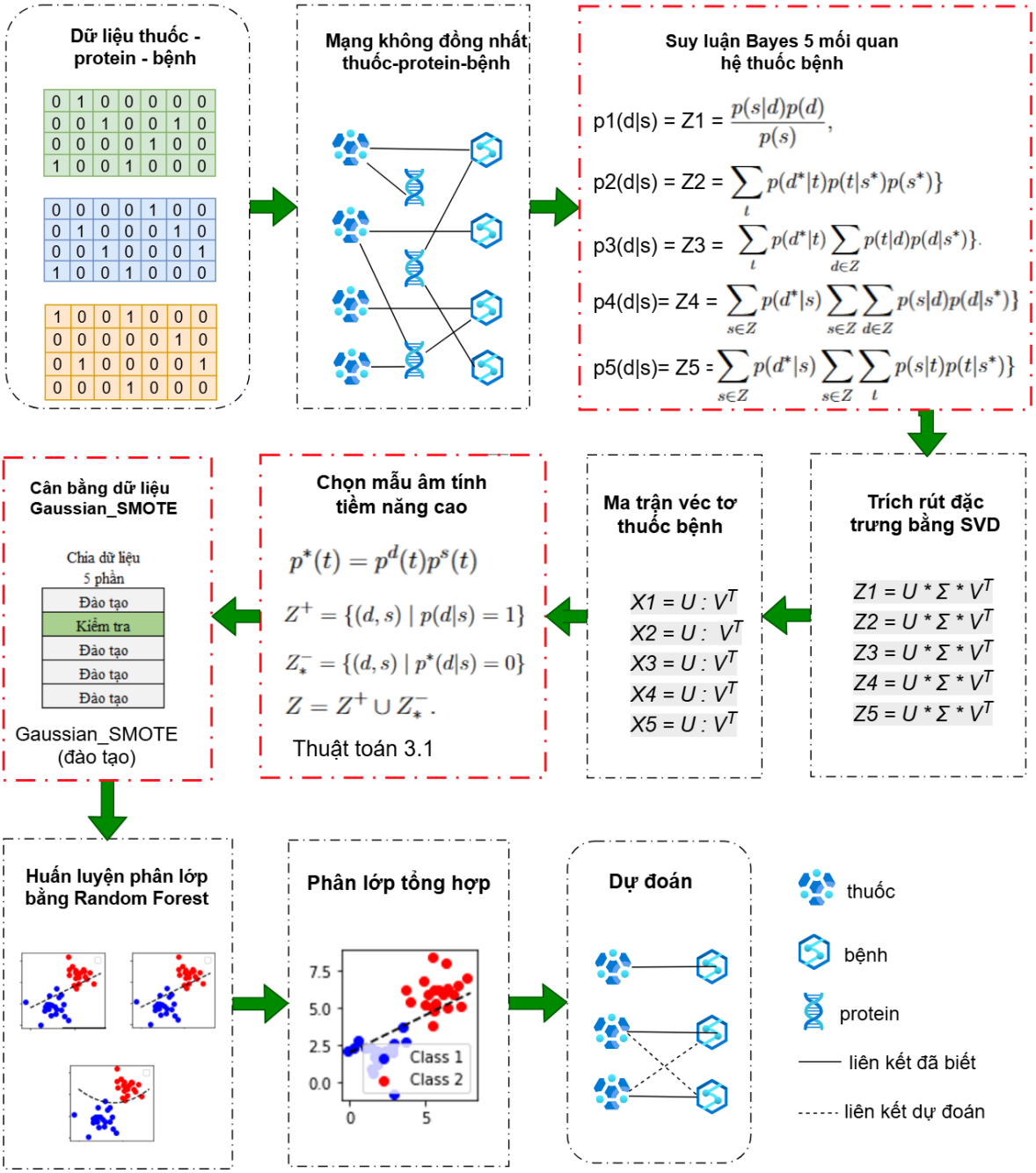
Ba cải tiến này tạo thành một hệ thống tích hợp, trong đó mô hình Bayes cung cấp nền tảng để đánh giá chất lượng dữ liệu, và dữ liệu chất lượng cao lại củng cố độ tin cậy cho các suy luận. Hình 3.1 minh họa khung phương pháp DDA-BNS với ba đóng góp chính này.

Những đóng góp mới này sẽ được trình bày chi tiết từng thành phần của khung phương pháp này.

3.1.2. Suy luận Bayes với năm mối quan hệ thuốc bệnh

Kết quả của dữ liệu huấn luyện được thực hiện như một phần của số lượng mẫu cân bằng cho các lớp liên quan đến thuốc và bệnh $p(d|s)$, bao gồm một loạt nghiên cứu về protein. Suy luận Bayes có thể mở rộng theo (1.6) cho phép dự đoán mối quan hệ tương tác thuốc-bệnh mới. Ở đây, xác suất tiên nghiệm của bệnh $p(s^*)$ được ước tính từ dữ liệu huấn luyện cho một bệnh cụ thể s^*:

$$p(s^*) = \sum_{(d,s)\in Z} p(s^*|d)p(d). \tag{3.1}$$



Hình 3.1: Khung phương pháp DDA-BNS

Do đó, suy luận Bayes (1.6) cho thấy một dự đoán về tương tác giữa thuốc d và bệnh s không được trình bày trong dữ liệu huấn luyện Z. Dự đoán cuối cùng được tạo ra bằng cách tổng hợp kết quả từ năm dự đoán khác nhau. Dự đoán đầu tiên sử dụng dữ liệu huấn luyện được chuẩn bị từ tập dữ liệu $Z_1 = Z$ (3.16) với nhãn cho từng cặp thuốc và bệnh, và áp dụng quy tắc Bayes được thể hiện như công thức (3.2):

$$p_1(d|s) = \frac{p(s|d)p(d)}{p(s)}, \quad Z_1 = Z. \tag{3.2}$$

Vì một cặp thuốc–bệnh có thể liên quan đến nhiều protein khác nhau, suy luận Bayes yêu cầu tổng hợp thông tin từ toàn bộ các protein. Do đó, công thức (3.3) chính là sự phân rã xác suất qua biến trung gian protein t, phản ánh cơ chế lan truyền giữa thuốc d^* và bệnh s^* thông qua protein. Bộ dữ liệu thứ hai Z_2 và xác suất p_2 được biểu diễn như công thức (3.3):

$$p_2(d|s), Z_2 = \{(d^*, s^*) \mid p(d^*|s^*) = \sum_t p(d^*|t)p(t|s^*)p(s^*)\}. \tag{3.3}$$

Bộ dữ liệu thứ hai chỉ xem xét lan truyền trực tiếp qua protein. Tuy nhiên, trong thực tế sinh học, protein có thể liên quan tới bệnh thông qua nhiều thuốc khác nhau. Vì vậy, dự đoán thứ ba mở rộng cơ chế lan truyền bằng cách bổ sung thêm tầng thông tin từ các thuốc liên quan.

Trong mô hình này, thuốc d^* tác động lên protein t với xác suất $p(d^*|t)$, còn vai trò của protein đối với bệnh s^* được ước lượng gián tiếp thông qua tập các thuốc d có khả năng liên quan đến bệnh. Kết hợp hai bước lan truyền qua protein và các thuốc liên quan, ta thu được dự đoán thứ ba như công thức (3.4):

$$p_3(d|s), Z_3 = \{(d^*, s^*) \mid p(d^*|s^*) = \sum_t p(d^*|t) \sum_{d \in Z} p(t|d)p(d|s^*)\}. \tag{3.4}$$

Bên cạnh các cơ chế lan truyền qua protein và các thuốc liên quan, mạng thuốc–bệnh còn cho phép suy luận thông qua chuỗi quan hệ lặp lại giữa thuốc và bệnh. Trong trường hợp này, thông tin được lan truyền từ thuốc sang bệnh thông qua $p(d^*|s)$, và sau đó phản hồi trở lại qua các quan hệ bệnh–thuốc bằng phân bố $p(s|d)$. Sự kết hợp hai hướng lan truyền này tạo thành dữ liệu huấn luyện thứ tư như công thức (3.5).

$$p_4(d|s), Z_4 = \{(d^*, s^*) \mid p(d^*|s^*) = \sum_{s \in Z} p(d^*|s) \sum_{s \in Z} \sum_{d \in Z} p(s|d)p(d|s^*)\}. \tag{3.5}$$

Bên cạnh các cơ chế lan truyền dựa trên quan hệ thuốc–bệnh và protein–thuốc, mô hình còn cần xem xét trực tiếp mối liên hệ hai chiều giữa bệnh và protein. Thực tế sinh học cho thấy một bệnh s có thể liên quan đến nhiều protein t, và ngược lại, mỗi protein cũng có thể tham gia vào nhiều cơ chế bệnh lý khác nhau. Do đó, việc kết hợp cả hai phân bố $p(s|t)$ và $p(t|s^*)$ cho phép mở rộng không gian suy luận và hình thành dữ liệu huấn luyện thứ năm.

Trong cơ chế này, xác suất thuốc d^* liên quan đến bệnh trung gian s được kết hợp với xác suất bệnh s liên quan đến protein t và protein t quay lại tác động lên bệnh mục tiêu s^*. Quá trình lan truyền hai chiều này dẫn đến dữ liệu huấn luyện thứ 5 như công thức (3.6):

$$p_5(d|s), Z_5 = \{(d^*, s^*) \mid p(d^*|s^*) = \sum_{s \in Z} p(d^*|s) \sum_{s \in Z} \sum_{t} p(s|t)p(t|s^*)\}. \tag{3.6}$$

Tóm lại, năm mô hình suy luận Bayes được xây dựng trong bước này phản ánh năm cơ chế lan truyền thông tin khác nhau trong mạng thuốc–protein–bệnh. Mỗi cơ chế tương ứng với một dạng phân rã xác suất qua các biến trung gian, cho phép mô hình khai thác tối đa cấu trúc dị thể của dữ liệu sinh học. Các dự đoán thu được từ năm quan hệ này cung cấp những góc nhìn bổ sung và không trùng lặp, từ các quan hệ trực tiếp thuốc–bệnh cho tới các chuỗi lan truyền phức tạp thông qua protein, thuốc liên quan hoặc các bệnh trung gian.

Nguồn dữ liệu suy luận thu được từ năm mô hình trên sẽ tiếp tục được sử dụng làm đầu vào cho bước tiếp theo, trong đó luận án áp dụng phân tách giá trị kỳ dị để xây dựng các vector biểu diễn thuốc–bệnh phục vụ cho quá trình dự đoán.

3.1.3. Suy luận Bayes trích rút mẫu âm tính chất lượng cao

Thật vậy, quá trình học để định vị lại thuốc đòi hỏi dữ liệu huấn luyện bao gồm các cặp thuốc d và bệnh s và nhãn của chúng cho biết tương tác âm tính hay dương tính dựa trên các tài liệu đã công bố cho từng cặp. Đối với một dữ liệu huấn luyện nhất định Z, ghi chú Z^+ được cấp cho một tập hợp các mẫu dương tính và ghi chú Z^- được tạo cho một tập hợp các mẫu âm tính. Z được

tính theo công thức

$$Z = Z^+ \cup Z^-. \tag{3.7}$$

Việc xác định mối quan hệ thuốc-bệnh đã biết được thực hiện bằng cách kiểm tra các loại thuốc đã được các công ty dược phẩm chấp thuận và được tính theo công thức

$$Z^+ = \{(d, s) \mid p(d|s) = 1\}. \tag{3.8}$$

Mối quan hệ thuốc-bệnh chưa biết liên quan đến những cặp thuốc-bệnh mà thông tin về chúng không đầy đủ. Z^- được tính theo công thức

$$Z^- = \{(d, s) \mid p(d|s) = 0\}. \tag{3.9}$$

Có sự mất cân bằng rõ rệt trên dữ liệu huấn luyện. Đối với bất kỳ loại thuốc nào d, chỉ một số ít bệnh s sẵn sàng cho các mẫu dương tính $p(d|s) = 1$ trong khi các bệnh khác được thiết lập trong các mẫu âm tính, công thức sau thể hiện sự mất cân bằng:

$$\text{count}(Z^+) \ll \text{count}(Z^-). \tag{3.10}$$

Phân loại mà tập dữ liệu có tỷ lệ lớp lệch được gọi là mất cân bằng. Các lớp có tỷ lệ lớn trong tập dữ liệu được gọi là lớp đa số. Các lớp chiếm tỷ lệ nhỏ hơn là lớp thiểu số. Sự mất cân bằng có thể gây ra các vấn đề về học. Nếu số lượng mẫu dương quá nhỏ so với mẫu âm, quá trình huấn luyện sẽ dành phần lớn thời gian cho các mẫu âm và các mẫu dương không được học đủ. Vấn đề trên dẫn luận án đến việc áp dụng một giải pháp như sau.

Phát hiện rằng xác suất tiên nghiệm của protein $p(t)$ có thể được phục hồi từ việc kiểm tra chéo mối liên kết protein-thuốc và tìm kiếm trong quá trình huấn luyện, mang lại suy luận cho protein $p^d(t)$ với dấu hiệu d. $p^d(t)$ được thể hiện theo công thức

$$p^d(t) = \sum_d p(t|d)p(d|t). \tag{3.11}$$

Tương tự như vậy, có thể có một cách khác để xác định xác suất tiên nghiệm của protein $p(t)$ thông qua mối liên hệ giữa protein và bệnh, luận án đánh dấu nó bằng $p^s(t)$ và được thể hiện theo công thức

$$p^s(t) = \sum_s p(t|s)p(s|t). \tag{3.12}$$

Do đó, các xác suất tiên nghiệm được kết hợp trong một cái nhìn chung từ triển vọng thuốc cũng như từ triển vọng bệnh tật. Điều này tạo ra xác suất cuối cùng cho protein $p(t)$ ký hiệu là $p^*(t)$ và được thể hiện theo

$$p^*(t) = p^d(t)p^s(t) = \sum_d p(t|d)p(d|t) \sum_s p(t|s)p(s|t). \tag{3.13}$$

Để cải thiện chất lượng học, các mẫu âm tính để huấn luyện có thể được chọn từ tập hợp các mẫu có xác suất của chúng là $p(d|s) = 0$. Lưu ý rằng, phép tính của $p^*(d|s)$ được đề xuất sử dụng phương trình (1.6) với protein trước $p^*(t)$ được xác định trước ở trên bởi (3.13):

$$p^*(d|s) = \sum_t \frac{p(d|t)p^*(t)p(t|s)}{p(s)}. \tag{3.14}$$

Tập dữ liệu huấn luyện các mẫu âm tính từ (3.14) có chất lượng tốt hơn so với tập dữ liệu gốc do cách chọn mẫu đã được mô tả như công thức

$$Z_*^- = \{(d, s) \mid p^*(d|s) = 0\}. \tag{3.15}$$

Do đó, luận án đang mô tả rằng việc xây dựng dữ liệu huấn luyện có chọn lọc với dữ liệu mẫu âm tính có tiềm năng cao là có thể thông qua việc đánh giá xác suất tiên nghiệm đối với protein, hoàn thành nhiệm vụ chuẩn bị dữ liệu với tập hợp mẫu để huấn luyện được biểu diễn như công thức

$$Z = Z^{+} \cup Z_{*}^{-}. \tag{3.16}$$

Trong luận án này, luận án tiến hành chọn các mẫu âm tính có xác suất cao bằng cách sử dụng các công thức đã phân tích trước đó (3.15). Theo các phương trình (3.11)–(3.14), luận án trích xuất các mẫu âm tính chất lượng cao (High Negative Samples – HNS) bằng Thuật toán 3.1. Cuối cùng, một tập dữ liệu mới được tạo ra bằng cách kết hợp các mẫu dương tính với các mẫu âm tính chất lượng cao này theo phương trình.

Tập dữ liệu mới này được gọi là **HNdataset**. Ngoài ra, luận án cũng xây dựng một tập dữ liệu khác dựa trên phương pháp lọc mẫu âm tính (Filtered Negative Samples – FNS) do Wu [26] giới thiệu, có tên là **FNdataset**. Dựa trên các xác suất tiên nghiệm của protein, luận án đề xuất một thuật toán lựa chọn các mẫu âm tính chất lượng cao (High Negative Samples – HNS), nhằm giảm thiểu ảnh hưởng của âm tính giả và cải thiện chất lượng dữ liệu huấn luyện.

Thuật toán 3.1 Chọn mẫu âm tính tiềm năng cao

Input : $A_{dt}[n \times k]$ (ma trận thuốc-protein)

$A_{st}[m \times k]$ (ma trận bệnh-protein-)

$A_{tt}[k \times k]$ (ma trận protein-protein), xem công thức (3.11, 3.12)

Output: Z^- (tập mẫu âm tính)

Begin Algorithm

$A_{ds} \leftarrow A_{dt} \cdot A_{tt} \cdot A_{st}^T$, xem công thức (3.13)

$Z^- \leftarrow \emptyset$

$i \leftarrow 1$

while $i \leq m$ **do**

  $j \leftarrow 1$

  while $j \leq n$ **do**

    if $A_{ds}(i,j) = 0$ **then**

      $Z^- \leftarrow Z^- \cup \{(i,j)\}$

    end

    $j \leftarrow j + 1$

  end

  $i \leftarrow i + 1$

end

return Z^-

Kết thúc thuật toán

Mục đích của việc học là tối đa hóa niềm tin vào lý luận và do đó đưa hậu nghiệm gần đúng càng gần với hậu nghiệm thực càng tốt. Để có được dự đoán cuối cùng cho dữ liệu thử nghiệm với các cặp (d, s), luận án áp dụng cơ chế tổng hợp cho năm dự đoán:

$$p(d|s) = \max_{i=1,\ldots,5} p_i(d|s). \tag{3.17}$$

Trích rút và xây dựng đặc trưng

Như đã phân tích trong Chương 2, mỗi ma trận tổng hợp Z (Z_1, Z_2, Z_3, Z_4, Z_5) được xây dựng theo các công thức (3.16–3.20), thể hiện mối quan hệ tiềm ẩn giữa thuốc và bệnh. Ma trận có kích thước 1.186×449, trong đó 1.186 loại thuốc được biểu diễn theo hàng và 449 loại bệnh được biểu diễn theo cột.

Mỗi véc-tơ thuốc d_i và bệnh s_j được hình thành từ sự kết hợp của 449 đặc trưng thuốc với 1.186 đặc trưng bệnh, dẫn đến số chiều rất lớn. Dữ liệu có tính chất thưa và nhiều đặc trưng không mang nhiều ý nghĩa. Việc giữ nguyên toàn bộ đặc trưng sẽ làm tăng chi phí tính toán, gây nguy cơ quá khớp và khó diễn giải.

Để giảm chiều dữ liệu và bảo toàn các mối quan hệ cốt lõi, SVD được áp dụng. Ma trận Z được phân rã thành ba ma trận thành phần:

$$Z \approx U_{n\times r}\, \Sigma_{r\times r}\, V^T_{r\times m}, \tag{3.22}$$

trong đó:

- $U_{n\times r}$: ánh xạ thuốc vào không gian đặc trưng tiềm ẩn với số chiều r.
- $\Sigma_{r\times r}$: ma trận đường chéo chứa các giá trị kỳ dị, đại diện cho độ quan trọng của từng thành phần tiềm ẩn.
- $V_{r\times m}$: ánh xạ bệnh vào cùng không gian đặc trưng tiềm ẩn.

Bằng cách giữ lại các giá trị kỳ dị lớn, SVD giúp giảm chiều dữ liệu, lọc nhiễu và giữ lại các đặc trưng quan trọng, từ đó cải thiện khả năng dự đoán liên kết thuốc – bệnh.

3.1.4. Cân bằng dữ liệu

Để đánh giá và so sánh hiệu quả, luận án phân loại các phương pháp thành ba nhóm chính. Nhóm đầu tiên bao gồm các thuật toán cân bằng dữ liệu lấy mẫu dưới mức. Đó là SPY [99], NearMiss [100], TomekLinks [101], RandomUnderSampler, OneSidedSelection [102] và NeighbourhoodCleaningRule [103].

Nhóm thứ hai bao gồm các phương pháp cân bằng dữ liệu lấy mẫu quá mức. Đó là SMOTE [104], Borderline-SMOTE [105], CURE-SMOTE [106], SMOTE-TomekLinks [107], AND-SMOTE [108], SMOTE-D [109], Random-SMOTE [110], Kmean-SMOTE [111], Gaussian-SMOTE [112] và SMOTE- WB [113]. Nhóm thứ ba bao gồm các kỹ thuật từ nghiên cứu trước đây để cân bằng dữ liệu trước khi học máy. Luận án đã triển khai các phương pháp này trên hai tập dữ liệu chuẩn hóa, HNdataset và FNdataset. Tất cả các thí nghiệm đều được tiến hành trong điều kiện giống hệt nhau để đảm bảo so sánh công bằng.

Kỹ thuật lấy mẫu quá mức thiểu số tổng hợp (SMOTE), được công nhận rộng rãi vì khả năng tạo mẫu tổng hợp cho lớp thiểu số bằng cách tạo dữ liệu mới dọc theo đường kết nối một thể hiện lớp thiểu số và một số lượng nhất định các láng giềng cùng lớp của nó, đã cho thấy tiềm năng đáng kể trong việc giảm thiểu các vấn đề mất cân bằng dữ liệu. Tiến xa hơn nữa, kỹ thuật lấy mẫu quá mức thiểu số tổng hợp với phát hiện tăng cường và nhiễu (SMOTEWB) đại diện cho một phương pháp tiếp cận kết hợp SMOTE và lấy mẫu quá mức ngẫu nhiên (ROS), kết hợp các kỹ thuật tăng cường và phát hiện nhiễu bổ sung. Mục tiêu của sự kết hợp này là nâng cao hiệu quả của việc tạo mẫu tổng hợp, do đó cải thiện độ chính xác của các mô hình phân loại. Kỹ thuật tăng cường khuếch đại các tính năng của mẫu dữ liệu tổng hợp bằng cách tập trung vào các trường hợp khó phân loại, trong khi phát hiện nhiễu giảm thiểu tác động của dữ liệu nhiễu lên quá trình huấn luyện, đảm bảo chất lượng cao của các mẫu tổng hợp.

3.1.4.1. Độ phức tạp tính toán và khả năng mở rộng

Giả sử n, m và z lần lượt là số lượng thuốc, bệnh và protein. Dựa trên các công thức (3.11)–(3.15), độ phức tạp của khung DDA-BN được phân tích như sau: tính $p_d(t)$ theo (3.11) có độ phức tạp $O(nz)$; tính $p_s(t)$ theo (3.12) có độ phức tạp $O(mz)$; tính $p^*(t)$ theo (3.13) có độ phức tạp $O(z)$; tính $p^*(d|s)$ theo (3.14) cho toàn bộ các cặp thuốc–bệnh có độ phức tạp $O(nmz)$; và bước chọn mẫu âm theo (3.15) có độ phức tạp $O(nm)$. Do đó, độ phức tạp tổng thể của

phần suy luận Bayes và chọn mẫu âm là

$$O(nz + mz + z + nmz + nm) = O(nmz).$$

Như vậy, bước chi phối của mô hình là công thức (3.14). Khi số lượng protein tăng lớn, chi phí tính toán tăng tuyến tính theo số lượng protein z, cho thấy mô hình vẫn có khả năng mở rộng, nhưng cần tối ưu thêm về lưu trữ ma trận thưa và tính toán song song đối với các bộ dữ liệu rất lớn.

3.2. Thực nghiệm và đánh giá

3.2.1. Tập dữ liệu thử nghiệm

Dữ liệu thử nghiệm A_dataset

Trong luận án này, việc dự đoán liên kết giữa thuốc và bệnh đòi hỏi dữ liệu đáng tin cậy về các tương tác: thuốc–bệnh, thuốc–protein và bệnh–protein. Các dữ liệu này được thu thập từ nhiều nguồn khác nhau và đã được Wu et al. [26] lựa chọn, bao gồm: OMIM [84], bộ dữ liệu của Gottlieb [85] và DrugBank [83], như trình bày trong Bảng 3.1.

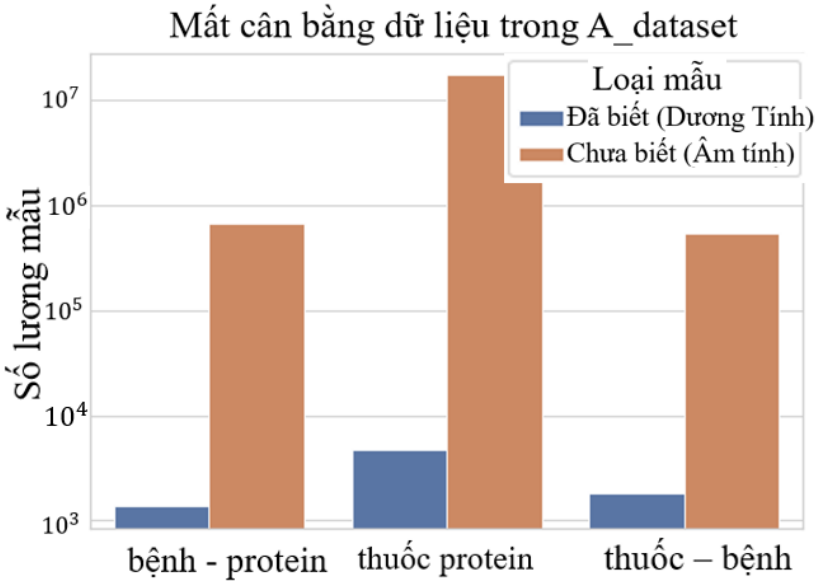
Do dữ liệu đến từ các nguồn khác nhau, chúng có định dạng và kiểu dữ liệu đa dạng. Cụ thể:

- **Bệnh–protein**: trích xuất từ OMIM, gồm 1.365 tương tác giữa 449 bệnh và 1.147 protein.
- **Thuốc–bệnh**: thu thập từ bộ dữ liệu của Gottlieb, gồm 1,827 tương tác giữa 551 thuốc và 302 bệnh.
- **Thuốc–protein**: từ DrugBank, gồm 4.642 tương tác giữa 1.186 thuốc và 1.147 protein.

Điểm đáng chú ý là tập dữ liệu mang tính không đồng nhất do được tổng hợp từ các nguồn khác nhau. Để nghiên cứu mối quan hệ giữa một loại thuốc và một bệnh cụ thể, cần xác định xem có tồn tại con đường kết nối giữa thuốc và bệnh thông qua mạng các protein, bệnh và thuốc hay không.

Bảng 3.1: Tổng quan về A_dataset

Mối quan hệ	**Số lượng tương tác dương**	**Số lượng tương tác chưa xác minh**	**Tỷ lệ % dương**
Bệnh–protein (449 x 1.147)	1.365	65.318	0,207
Thuốc–protein (1.186 x 1.147)	4.642	1.352.200	0,267
Thuốc–bệnh(551 x 302)	1.827	530.687	0,344



Hình 3.2: Mô tả mất cân bằng dữ liệu A-dataset

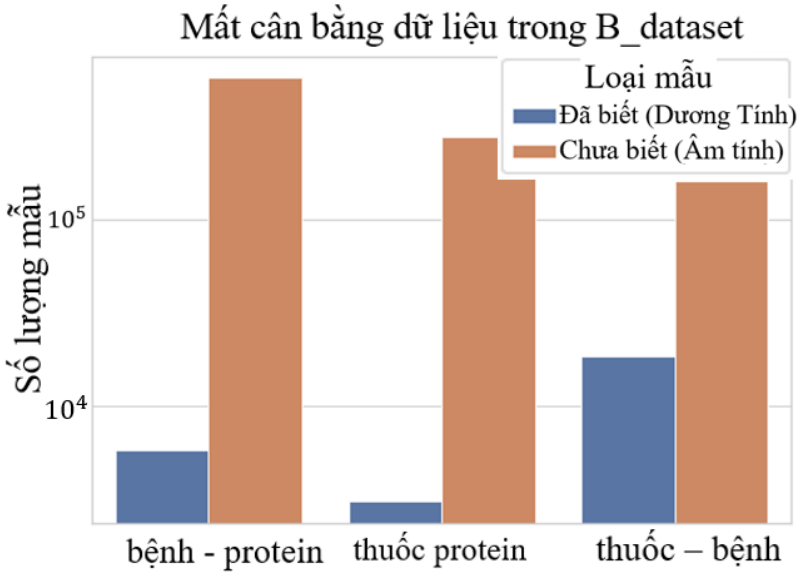
Dữ liệu thử nghiệm B_dataset

Tập dữ liệu do W. Zhang [88] và BW. Zhao [114] xây dựng được gọi là B_dataset. Tập dữ liệu này bao gồm 269 loại thuốc, 598 bệnh và 1.021 protein, với số lượng liên kết dương như sau: 18.416 cho thuốc–bệnh, 3.110 cho thuốc–protein và 5.898 cho bệnh–protein (Bảng 3.2).

Trong đó, n là số lượng thuốc, m là số lượng bệnh, và k là số lượng protein. Các mẫu dương tính trong tập dữ liệu thử nghiệm được xác định theo công thức (3.8), trong khi các mẫu còn lại được coi là chưa xác định và có thể là dương tính hoặc âm tính.

Bảng 3.2: Mô tả tập thử nghiệm B_dataset

Mối quan hệ	**Số lượng tương tác**	**Số lượng không có tương tác**	**Tỷ lệ %**
Protein–bệnh (598 x 1021)	5.898	569.718	1,04
Protein–thuốc (269 x 1021)	3.110	274.649	1,13
Bệnh–thuốc (269 x 598)	18.416	160.862	11,45



Hình 3.3: Mô tả mất cân bằng dữ liệu B-dataset

Thiết lập thực nghiệm

Các bộ dữ liệu A_dataset và B_dataset được biểu diễn trên mạng không đồng nhất thuốc–protein–bệnh. Mẫu âm tính được trích rút với độ tin cậy cao theo hai cách: phương pháp của Wu [26] và thuật toán đề xuất (Thuật toán 1).

- Với A_dataset: tạo ra A_FNdataset (theo Wu) và A_HNdataset (theo thuật toán 1).

- Với B_dataset: tạo ra B_FNdataset và B_HNdataset tương ứng.

Tất cả dữ liệu được chia thành bộ huấn luyện và bộ kiểm tra theo tỷ lệ 80:20.

3.2.2. Môi trường thực nghiệm

Các thí nghiệm được thực hiện trên máy tính cấu hình: CPU Intel Core i5-12400, RAM 16GB DDR4, hệ điều hành Windows 11 Pro. Các thí nghiệm sử dụng Python 3.11.5 và Scikit-learn 1.3.0 để phát triển và đánh giá mô hình học máy. Thư viện smote_variants [98] được sử dụng để triển khai nhiều kỹ thuật lấy mẫu quá mức khác nhau để xử lý các tập dữ liệu mất cân bằng. Tất cả các phương pháp lấy mẫu quá mức đều được áp dụng bằng cách sử dụng các thiết lập tham số mặc định của thư viện, vì các cấu hình này được ghi chép đầy đủ và đã được xác thực trong các nghiên cứu trước đây. Lựa chọn này đảm bảo tính nhất quán và khả năng tái tạo trong các thí nghiệm trong khi tập trung vào việc đánh giá phương pháp được đề xuất.

3.2.3. Tham số đánh giá

Trong quá trình đánh giá các mô hình học máy, nhiều tham số khác nhau được đề xuất để đánh giá hiệu suất một cách chính xác. Việc lựa chọn các tham số phù hợp với từng mô hình và đặc điểm của tập dữ liệu là rất quan trọng. Trong các tình huống liên quan đến các tập dữ liệu mất cân bằng nghiêm trọng, độ hồi tưởng (SE) và độ đặc hiệu (SP) thường là các số liệu được sử dụng, xem (1.7, 1.8). Kubat và cộng sự đã giới thiệu giá trị trung bình hình học (G-Mean) [92] để đánh giá các mô hình học máy trên dữ liệu mất cân bằng (1.9). Ngoài ra, các số liệu như độ chính xác (ACC) (1.10), độ hồi tưởng (REC) (1.11), độ chính xác (PRE) (1.12), điểm F1 (1.13), diện tích dưới đường cong AUPR, hệ số tương quan Matthews (MCC) (1.14), diện tích dưới đường cong (AUC) và diện tích dưới đường cong PR-AUC được sử dụng để đánh giá và so sánh hiệu quả của các phương pháp luận với nghiên cứu gần đây.

3.2.4. Mục tiêu và quy trình thử nghiệm

Mục tiêu của các thí nghiệm bao gồm:

1. Đánh giá hiệu quả của phương pháp đề xuất trong việc phát hiện các liên

kết thuốc–bệnh mới.

2. So sánh với các phương pháp hiện tại sử dụng các bộ dữ liệu chuẩn và các chỉ số đánh giá tiêu chuẩn.

3. Sử dụng phương pháp thống kê và kiểm định để đánh giá mô hình.

Quy trình thực nghiệm được thực hiện như sau:

- Biểu diễn A_dataset và B_dataset trên mạng không đồng nhất thuốc–protein–bệnh.
- Chọn các mẫu âm tính có độ tin cậy cao theo phương pháp của Wu [26] và theo thuật toán 1. Các tập dữ liệu mới được tạo ra ký hiệu là A_FNdataset, A_HNdataset, B_FNdataset và B_HNdataset.
- Chia dữ liệu thành bộ huấn luyện và bộ kiểm tra theo tỷ lệ 80:20.

Kịch bản 1: Thực nghiệm trên dữ liệu A_FNdataset

- Áp dụng các kỹ thuật cân bằng dữ liệu: SPY, OS, NCR, Borderline-SMOTE, SMOTETomek, AND_SMOTE, KMeans_SMOTE, Gaussian_SMOTE.
- Sử dụng phương pháp trích rút mẫu âm tính Wu [26].
- Chỉ số đánh giá: F1, G-mean, PR_AUC.
- Mô hình tốt nhất (Gaussian_SMOTE) được so sánh với các nghiên cứu trước.
- Lấy 20 dự đoán có xác suất cao nhất để kiểm tra điển hình.

Kịch bản 2: Thực nghiệm trên A_FNdataset và A_HNdataset

- Áp dụng các kỹ thuật cân bằng dữ liệu dưới mức: SPY, NearMiss, Tomek-Links, RandomUnderSampler, OneSidedSelection, NeighbourhoodCleaningRule và các kỹ thuật cân bằng dữ liệu quá mức: SMOTE, Borderline-SMOTE, SMOTETomek, AND_SMOTE, KMeans_SMOTE, Gaussian_SMOTE, SMOTE_D, Random-SMOTE, SMOTEWB lần lượt lên A_FNdataset và A_HNdataset

- Chỉ số đánh giá: F1, G-mean, PR_AUC.
- Mô hình tốt nhất (Gaussian_SMOTE) được chọn cho các thí nghiệm tiếp theo.
- Sử dụng kỹ thuật cắt bỏ cắt bỏ so sánh mô hình có và không sử dụng kỹ thuật cân bằng dữ liệu kết hợp với chọn mẫu âm tính
- Sử dụng kỹ thuật thống kê chứng minh ý nghĩa thống kê khi kết hợp cân bằng dữ liệu và chọn mẫu âm tính
- Lấy 10 dự đoán có xác suất cao nhất làm kiểm tra điển hình.

Kịch bản 3: Thực nghiệm trên B_dataset

- Thực nghiệm trên dữ liệu B_dataset.
- So sánh với các mô hình trước đó

3.2.5. Kết quả thực nghiệm và đánh giá

Kết quả theo kịch bản 1

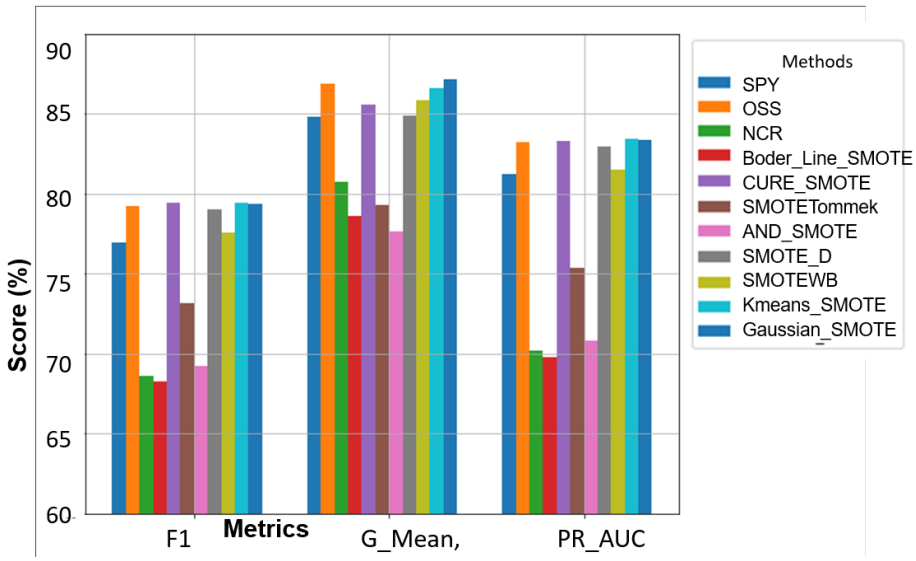
Trước tiên, các thí nghiệm được tiến hành trên nhiều biến thể khác nhau của phương pháp SMOTE, bao gồm: SPY, OSS, NCR, Borderline_SMOTE, CURE_SMOTE, SMOTE_Tomek, AND_SMOTE, SMOTE_D, SMOTEWB, KMeans_SMOTE và Gaussian_SMOTE. Với mỗi phương pháp, luận án áp dụng kỹ thuật kiểm thử chéo 5 lần (5-fold cross-validation). Kết quả trung bình về hiệu suất được trình bày trong Bảng 3.3.

Kết quả so sánh các kỹ thuật cân bằng dữ liệu trong Bảng 3.3 cho thấy Gaussian-SMOTE đạt hiệu suất tổng thể tốt nhất, với G-mean = 0.872. Mặc dù KMeans-SMOTE đạt F1 cao hơn (0.794 so với 0.793 của Gaussian-SMOTE) và PR-AUC cao hơn (0.834 so với 0.833), sự chênh lệch là rất nhỏ ($<$0.002). Gaussian-SMOTE được lựa chọn do tính ổn định và khả năng cân bằng giữa các lớp tốt hơn, thể hiện qua chỉ số G-mean cao nhất. Phương pháp này tạo mẫu

Bảng 3.3: Hiệu suất của các phương pháp cân bằng dữ liệu

Phương pháp	**F1**	**G_mean**	**PR_AUC**
SPY	0.769	0.848	0.812
OSS	0.792	0.868	0.832
NCR	0.686	0.807	0.701
Borderline_SMOTE	0.682	0.786	0.698
CURE_SMOTE	0.794	0.855	0.833
SMOTE_Tomek	0.731	0.792	0.754
AND_SMOTE	0.692	0.776	0.708
SMOTE_D	0.790	0.848	0.829
SMOTEWB	0.776	0.858	0.815
KMeans_SMOTE	**0.794**	0.866	**0.834**
Gaussian_SMOTE	0.793	**0.872**	0.833

tổng hợp dựa trên phân phối Gaussian quanh các mẫu thiểu số, tạo dữ liệu mới tự nhiên và đa dạng hơn so với SMOTE thông thường, giúp mô hình học được biên phân lớp chính xác mà không gây quá khớp, phù hợp với đặc thù dữ liệu mất cân bằng nghiêm trọng trong bài toán. Kết quả này được minh họa trực quan trong Hình 3.4.

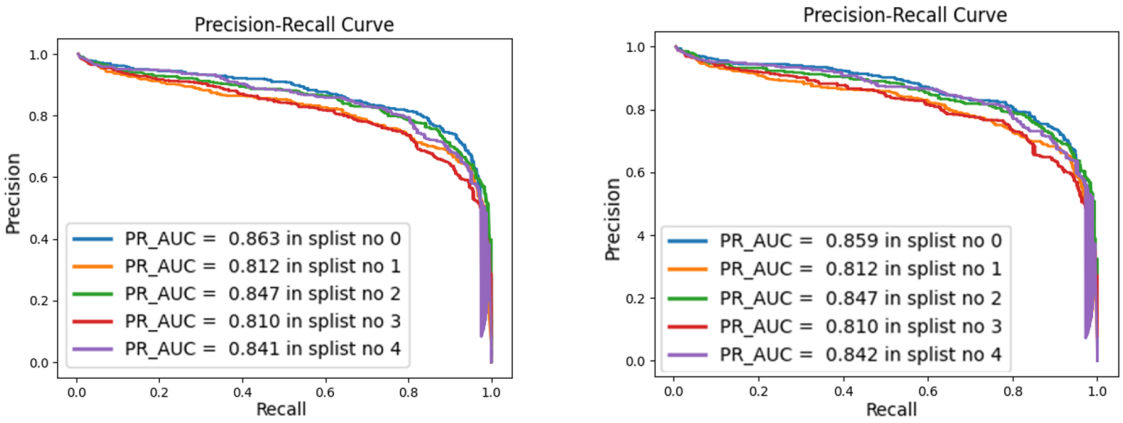


Hình 3.4: So sánh hiệu suất các kỹ thuật cân bằng dữ liệu

Như đã phân tích trong Chương 1, đối với các bài toán dữ liệu mất cân bằng, chỉ số G_mean là thước đo quan trọng, vì nó phản ánh mức cân bằng giữa độ nhạy (sensitivity) và độ đặc hiệu (specificity). Do đó, luận án lựa chọn

Gaussian_SMOTE làm kỹ thuật chính, nhờ hiệu suất G_mean vượt trội và khả năng nâng cao nhận diện đúng cả hai lớp thiểu số và đa số.

PR_AUC là chỉ số đặc biệt hữu ích trong các bài toán có tỷ lệ lớp không cân bằng, thể hiện năng lực phát hiện mẫu dương ở nhiều ngưỡng khác nhau. Trong các biến thể SMOTE, KMeans_SMOTE và Gaussian_SMOTE có G_mean lần lượt là 0.866 và 0.872. Đường cong PR_AUC của hai phương pháp này được minh họa trong Hình 3.5a và Hình 3.5b. Cả hai đều cho kết quả cao, trong đó Gaussian_SMOTE đạt mức ấn tượng hơn.



(a) KMeans_SMOTE_PR_AUC (b) Gaussian_SMOTE_PR_AUC

Hình 3.5: PR_AUC: So sánh KMeans_SMOTE và Gaussian_SMOTE

Như đã trình bày trong Chương 1 (Các nghiên cứu liên quan), nhiều nghiên cứu gần đây về tái định vị thuốc đã tập trung giải quyết vấn đề mất cân bằng dữ liệu trong học máy. Trong luận án này, nhóm tác giả cũng áp dụng các kỹ thuật từ [115] và [116] để so sánh hiệu quả trên cùng tập dữ liệu. Kết quả tổng hợp được thể hiện trong Bảng 3.4.

Bảng 3.4: So sánh hiệu suất với các phương pháp trong nghiên cứu trước

Kỹ thuật	**F1**	**G_mean**	**PR_AUC**
Seyedeh et al. [116]	0.641	0.7697	0.629
Korkmaz et al. [115]	0.764	0.856	0.804
Our method	**0.793**	**0.872**	**0.833**

Bảng 3.4 khẳng định phương pháp đề xuất vượt trội hơn các nghiên cứu

trước ở cả ba chỉ số F1, G_mean và PR_AUC. Đặc biệt, G_mean đạt 0.872 thể hiện khả năng cân bằng tốt giữa hai lớp, giúp mô hình hoạt động ổn định hơn trên dữ liệu mất cân bằng. So sánh với các phương pháp nghiên cứu trước, cho thấy DDA-BNS đạt AUPR=0.932, vượt trội so với DRIMC (0.912) và SCMFDD (0.905). Nguyên nhân: Các phương pháp trước chủ yếu dựa trên phân rã ma trận và chưa giải quyết triệt để vấn đề âm tính giả và mất cân bằng dữ liệu, trong khi DDA-BNS tích hợp suy luận Bayes với quy trình tiền xử lý dữ liệu chặt chẽ.

Các nghiên cứu điển hình

Phần này trình bày nghiên cứu điển hình trên bệnh nhân mắc Bạch cầu tủy cấp tính (AML, ID 601626). Tất cả các mối tương quan giữa AML và các thuốc khác trong tập dữ liệu được gán giá trị 0 trong giai đoạn huấn luyện để loại bỏ ảnh hưởng học trước. Mô hình sau khi huấn luyện được sử dụng để dự đoán các liên kết thuốc–bệnh mới; 10 thuốc có xác suất cao nhất được trình bày trong Bảng 3.5.

Bảng 3.5: Mười thuốc có xác suất dự đoán cao nhất cho AML

Mã thuốc	**Tên thuốc**	**Quan hệ đã biết***	**Bằng chứng xác thực**
DB00541	Vincristine	1	PMID: 5279331 [117]
DB00997	Doxorubicin	1	PMID: 16549995 [118]
DB00290	Bleomycin	0	NA
DB00694	Daunorubicin	1	PMID: 1449119 [119]
DB01204	Mitoxantrone	1	PMID: 9352324 [120]
DB01177	Idarubicin	1	PMID: 8290969 [121]
DB01033	Mercaptopurine	0	NA
DB00515	Cisplatin	0	PMID: 27127664 [122]
DB01073	Fludarabine	0	PMID: 11426551[123]
DB00444	Teniposide	0	NA

* Quan hệ đã biết trong tập dữ liệu ký hiệu là 1, ngược lại là 0.

Kết quả đối soát cho thấy: (i) 5/10 cặp thuốc–bệnh đã tồn tại trong tập dữ liệu gốc, phản ánh khả năng tái hiện tri thức đã biết của mô hình; (ii) 5 cặp còn lại là các dự đoán mới. Độ tin cậy của các dự đoán được kiểm chứng thông qua y văn quốc tế, trong đó các liên kết chưa có bằng chứng được ký hiệu là "NA". Đáng chú ý, các thuốc được mô hình dự đoán đều có cơ sở lâm sàng hợp lý.

Cụ thể, Vincristine là 1 trong 3 thuốc chính trong phác đồ chuẩn(Daunorubicin + Vincristine + Cytosine arabinoside). Ngoài ra Vincristine còn được sử dụng trong điều trị duy trì cùng với Cyclophosphamide và Cytosine arabinoside [117]. Tương tự, Doxorubicin và Daunorubicin là thuốc nền tảng trong phác đồ điều trị chuẩn AML [118], [119], trong khi Idarubicin có hiệu lực mạnh cho tỷ lệ lui bệnh hoàn toàn 84.2% [121]. Mitoxantrone thường được sử dụng trong các trường hợp kháng trị hoặc không dung nạp anthracycline, với ưu điểm giảm độc tính trên tim [121].

Đối với 5 liên kết mới, luận án tiến hành tra cứu và tìm thấy 2 liên kết có bằng chứng hỗ trợ, tương ứng với hai thuốc *Cisplatin* và *Fludarabine* trong điều trị AML; Cụ thể, Mody et al. [122] mô tả một trường hợp lâm sàng trong đó bệnh nhân được chẩn đoán đồng thời AML và ung thư biểu mô tế bào vảy thanh quản đã đạt thuyên giảm hoàn toàn (complete remission) của AML sau khi được điều trị bằng hóa xạ trị dựa trên Cisplatin, cho thấy khả năng Cisplatin có thể liên quan đến đáp ứng điều trị AML trong một số điều kiện lâm sàng nhất định. Trong khi đó, Carella et al. [123] báo cáo kết quả nghiên cứu trên 41 bệnh nhân AML tiên lượng xấu được điều trị bằng phác đồ FLAG (Fludarabine, Cytarabine và G-CSF), trong đó 23 bệnh nhân (56%) đạt thuyên giảm hoàn toàn, cho thấy hiệu quả đáng kể của Fludarabine trong các phác đồ điều trị AML.

Phương pháp đề xuất trong luận án này kết hợp mô hình dựa trên xử lý mất cân bằng dữ liệu với kỹ thuật Gaussian_SMOTE, nhằm cải thiện độ chính xác trong dự đoán liên kết thuốc–bệnh. Không chỉ chứng minh hiệu quả qua các chỉ số định lượng, mô hình còn dự đoán chính xác các thuốc có tiềm năng thực tiễn, mở ra hướng tiếp cận mới cho bài toán tái định vị thuốc (drug repositioning) trong lĩnh vực dược phẩm, giúp rút ngắn thời gian và chi phí phát triển thuốc.

Kết quả kịch bản 2

Kết quả thực nghiệm với các kỹ thuật cân bằng dưới mức trên A_FNdataset và A_HNdataset

Đề xuất 1 đã cho thấy Gaussian_SMOTE là cách tiếp cận hiệu quả để xử lý mất cân bằng dữ liệu. Trong Đề xuất 2, luận án kiểm chứng hiệu quả khi kết hợp Gaussian_SMOTE với chọn mẫu âm tính chất lượng cao (HNS), đồng thời mở rộng đánh giá cả nhóm kỹ thuật lấy mẫu dưới mức (undersampling) và lấy mẫu quá mức trên hai tập A_FNdataset và A_HNdataset.

Bảng 3.6: Hiệu suất mô hình theo các kỹ thuật lấy mẫu dưới mức

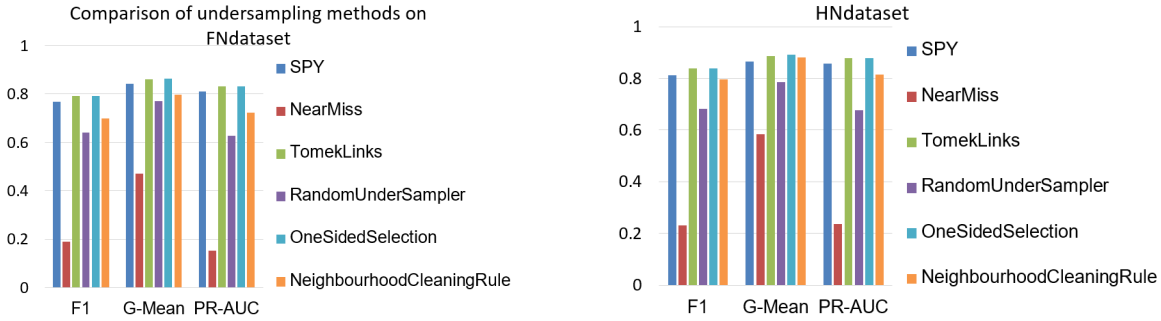
Phương pháp	A_FNdataset (mẫu dưới mức)			A_HNdataset (mẫu dưới mức)		
	F1	**G-Mean**	**PR-AUC**	**F1**	**G-Mean**	**PR-AUC**
SPY	0.769	0.842	0.811	0.813	0.865	0.856
NearMiss	0.191	0.472	0.153	0.231	0.585	0.236
TomekLinks	0.791	0.860	**0.832**	**0.839**	0.887	**0.8790**
RandomUnderSampler	0.641	0.770	0.628	0.683	0.786	0.677
OneSidedSelection	**0.792**	**0.863**	**0.832**	**0.839**	**0.891**	0.8789
NeighbourhoodCleaningRule	0.698	0.796	0.722	0.797	0.880	0.814

*Giá trị cao nhất được in đậm.

Ba thước đo được sử dụng là F1, G-Mean và PR-AUC. Kết quả chi tiết thể hiện ở Bảng 3.6 và minh họa trong Hình 3.6 - 3.7. Nhìn chung, A_HNdataset cho kết quả vượt trội: TomekLinks trên A_HNdataset đạt F1 = 0.839 và PR-AUC = 0.879, trong khi OneSidedSelection đạt G-Mean = 0.891.

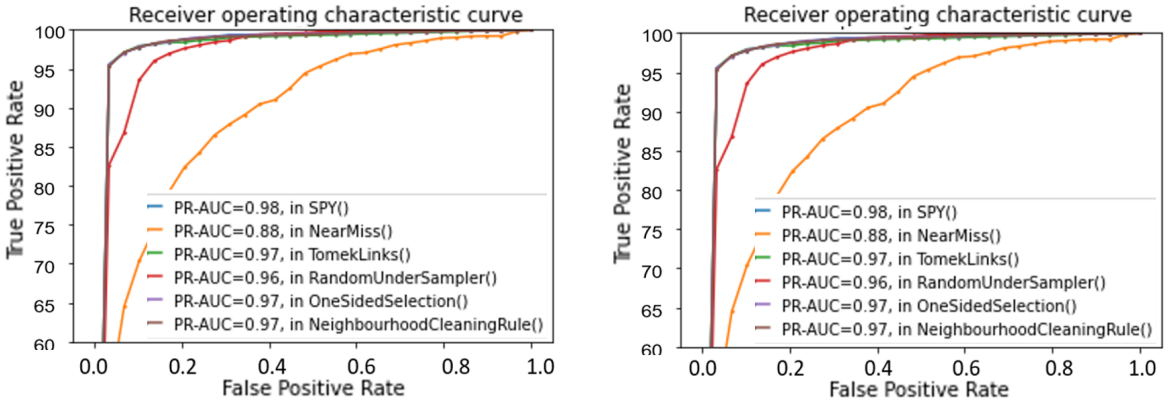
Kết quả thực nghiệm với các kỹ thuật cân bằng quá mức trên A_FNdataset và A_HNdataset

Các thí nghiệm lấy mẫu quá mức (SMOTE và biến thể) trên A_FNdataset và A_HNdataset được tổng hợp trong Bảng 3.7 và minh họa ở Hình 3.8–3.9. Kết



(a) A_FNdataset (undersampling) (b) A_HNdataset (undersampling)

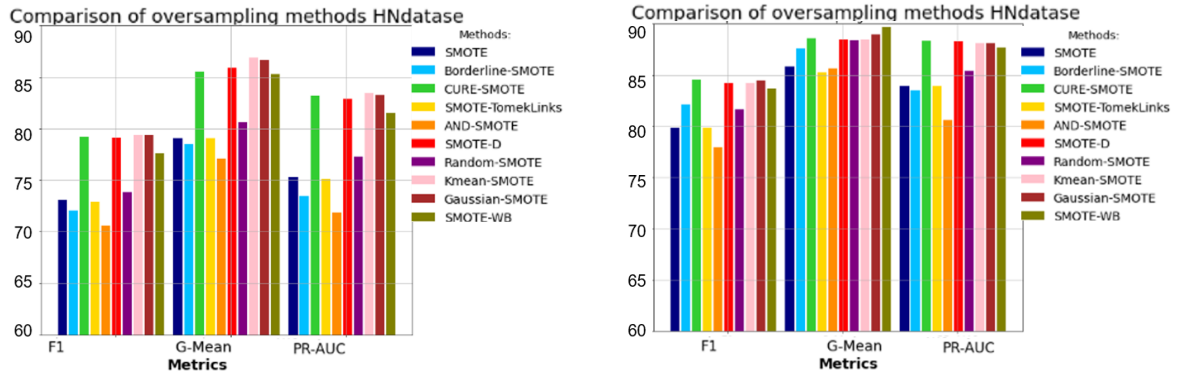
Hình 3.6: Hiệu suất mô hình với các kỹ thuật lấy mẫu dưới mức.



(a) PR curve cho A_FNdataset (b) PR curve cho A_HNdataset

Hình 3.7: PR-AUC cho A_FNdataset và A_HNdataset lấy mẫu dưới mức.

quả cho thấy HNdataset luôn vượt trội FNdataset trên mọi thước đo. Đáng chú ý, trên A_HNdataset, Gaussian_SMOTE đạt F1 = 0.8509, G-Mean = 0.8980, và PR-AUC = 0.8839; CURE_SMOTE cũng cạnh tranh với F1 = 0.8453, G-Mean = 0,8858 và PR-AUC = 0.8836.



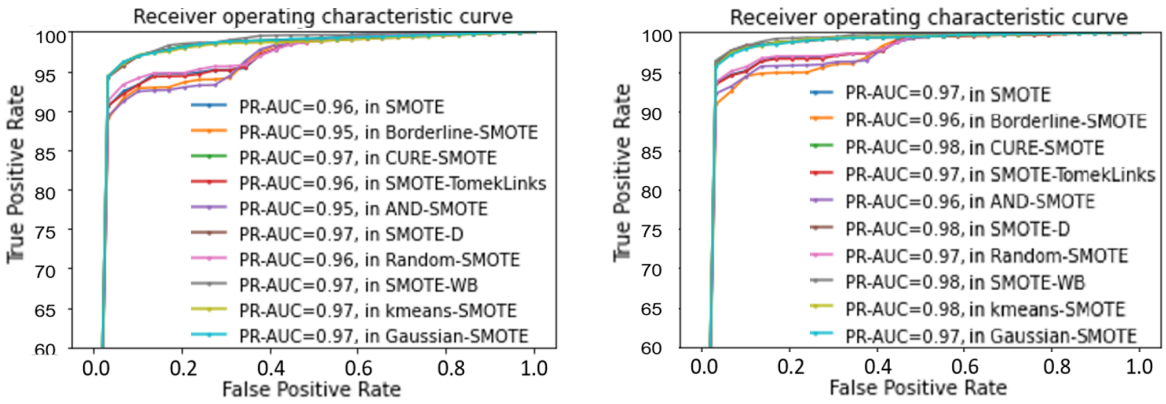
(a) A_FNdataset (oversampling) (b) A_HNdataset (oversampling)

Hình 3.8: Hiệu suất mô hình với các kỹ thuật lấy mẫu quá mức.

Bảng 3.7: Hiệu suất mô hình theo các kỹ thuật lấy mẫu quá mức

Phương pháp	A_FNdataset (mẫu quá mức)			A_HNdataset (mẫu quá mức)		
	F1	**G-Mean**	**PR-AUC**	**F1**	**G-Mean**	**PR-AUC**
SMOTE	0.7308	0.7909	0.7531	0.7985	0.7586	0.8395
Borderline-SMOTE	0.7206	0.7854	0.7345	0.8217	0.8761	0.8351
CURE-SMOTE	0.7920	0.8556	0.8320	0.8453	0.8858	0.8836
SMOTE-TomekLinks	0.7292	0.7907	0.7516	0.7988	0.8528	0.8393
AND-SMOTE	0.7057	0.7708	0.7181	0.7794	0.8565	0.8063
SMOTE-D	0.7911	0.8589	0.8288	0.8424	0.8848	0.8831
Random-SMOTE	0.7384	0.8061	0.7728	0.8168	0.8842	0.8541
KMeans-SMOTE	**0.7940**	**0.8686**	0.8045	0.8426	0.8851	0.8809
SMOTEWB	0.7757	0.8527	0.8152	0.8371	0.8970	0.8767
Gaussian_SMOTE	0.7936	0.8667	**0.8327**	**0.8509**	**0.8980**	**0.8839**

*Giá trị cao nhất được in đậm.



(a) PR curve cho A_FNdataset (b) PR curve cho A_HNdataset

Hình 3.9: PR-AUC cho A_FNdataset và A_HNdataset lấy mẫu quá mức.

Kết luận kết quả thực nghiệm với các kỹ thuật cân bằng quá mức trên A_FNdataset và A_HNdataset (1) Trên cả hai chiến lược chọn âm (FN vs HN), A_HNdataset cho kết quả hơn nhất quán về F1, G-Mean, PR-AUC;

(2) Trong nhóm oversampling, Gaussian_SMOTE trên A_HNdataset đạt

F1 = 0.8509, G-Mean = 0.8980, PR-AUC = 0.8839, vượt qua các biến thể khác;

(3) Kết hợp HNS + Gaussian_SMOTE tạo ra tập huấn luyện giàu thông tin, giúp ranh giới quyết định cân bằng và ổn định hơn trên dữ liệu mất cân bằng.

Mặc dù kết quả F1 tốt nhất trong Chương 3 (Bảng 3.7) là 0.8509, thấp hơn so với kết quả F1=0.918 của mô hình DR-LGBM-MH trong Chương 2 (Bảng 2.4), sự chênh lệch này không phản ánh chất lượng kém hơn của phương pháp suy luận Bayes. Nguyên nhân chính là do sự khác biệt căn bản trong thiết kế thực nghiệm giữa hai chương:

- Chương 2 (Mô hình dựa trên meta-path): Dữ liệu được cân bằng trước khi chia thành tập huấn luyện và tập kiểm tra, tạo điều kiện đánh giá tối ưu khả năng học mẫu hình từ cấu trúc đồ thị trong môi trường lý tưởng.
- Chương 3 (Mô hình suy luận Bayes): Dữ liệu giữ nguyên trạng thái mất cân bằng nghiêm trọng khi chia tập. Các kỹ thuật cân bằng chỉ được áp dụng trên tập huấn luyện, trong khi tập kiểm tra vẫn giữ tỷ lệ mất cân bằng tự nhiên (lớp dương chiếm tỷ lệ rất thấp). Thiết kế này nhằm mô phỏng sát điều kiện thực tế, nơi các liên kết mới cần được dự đoán trên nền dữ liệu không cân bằng.

Do đó, việc đạt được F1 $\approx$ 90% trên tập kiểm tra mất cân bằng cho thấy tính hiệu quả, ổn định và bền vững của phương pháp suy luận Bayes trong điều kiện khó khăn hơn. Hai chương có mục tiêu và ngữ cảnh đánh giá khác nhau, thể hiện các đóng góp bổ sung: Chương 2 tối ưu hóa việc khai thác cấu trúc đồ thị, trong khi Chương 3 tập trung xử lý các thách thức về chất lượng dữ liệu và mất cân bằng trong điều kiện thực tế.

Phương pháp cắt bỏ trên A_HNdataset

Để kiểm chứng hiệu quả, luận án so sánh bốn cấu hình chính kết quả như Bảng 3.8:

(1) Original: dùng tập dữ liệu gốc, không áp dụng High Negative (HN) hay Gaussian_SMOTE;

(2) Gaussian_SMOTE: chỉ áp dụng Gaussian_SMOTE, không dùng HN;

(3) High Negative: chỉ áp dụng HN, không kết hợp Gaussian_SMOTE; (4) Đề xuất: kết hợp đồng thời HN và Gaussian_SMOTE. Mỗi cấu hình được đánh giá bằng 5-fold cross-validation để bảo đảm tính khách quan và ổn định.

Bảng 3.8: So sánh F1, G-Mean và AUPR giữa các cấu hình

Phương pháp	**F1**	**G-Mean**	**AUPR**
Original	0.5264	0.6976	0.5067
High Negative	0.8339	0.8779	0.8746
Gaussian_SMOTE	0.5413	0.6933	0.5248
Our method	**0.8509**	**0.8980**	**0.8839**

Kết quả ở Bảng 3.8 cho thấy phương pháp kết hợp đạt hiệu suất cao nhất trên cả ba thước đo: F1 = 0.8509, G-Mean = 0.8980 và AUPR = 0.8839.

Phương pháp thống kê trên A_HNdataset

Để đánh giá ý nghĩa thống kê, luận án thực hiện kiểm định *t-test* hai mẫu, hai đuôi trên các điểm số của 5 fold (giả định phương sai bằng nhau). Kết quả p-value được trình bày trong Bảng 3.9.

Bảng 3.9: Kiểm định *t*-test hai mẫu.

So sánh	**F1**	**G-Mean**	**AUPR**
Original vs Gaussian_SMOTE	0,000010	0,000120	0,000006
Original vs High Negative	0,370000	0,780000	0,066000
Original vs Our method	0,000005	0,000051	0,000006
Gaussian_SMOTE vs High Negative	0,000005	0,000370	0,000005
Gaussian_SMOTE vs Our method	0,000002	0,000190	0,000004
High Negative vs Our method	0,000310	0,000440	0,000180

(i) Kết hợp HN + Gaussian_SMOTE cải thiện đáng kể so với từng thành phần đơn lẻ (p-value $< 0{,}001$ ở hầu hết phép so sánh);

(ii) So với High Negative đơn thuần, phương pháp kết hợp vẫn đạt cải thiện có ý nghĩa trên F1 và AUPR (p<0,05) và trên G-Mean (p=0,000313);

(iii) Cơ chế bổ trợ là: HN cung cấp âm tính "giàu thông tin", trong khi Gaussian_SMOTE tăng cường lớp thiểu số theo phân bố mượt, giúp ranh giới quyết định cân bằng hơn và nâng cao khả năng khái quát hoá.

Kết quả kịch bản 3

Trong phần này, mô hình sử dụng Gaussian-SMOTE kết hợp với tập dữ liệu âm tính chất lượng cao (B_HNdataset) được xây dựng từ Thuật toán 1. Kết quả hiệu suất của phương pháp này được so sánh với bảy nghiên cứu nổi bật trước đây, trong đó các mô hình đều hướng tới dự đoán mối liên hệ thuốc–bệnh dựa trên mạng không đồng nhất.

- **deepDR** [124]: Sử dụng khuôn khổ học sâu dựa trên mạng để tái sử dụng thuốc in silico, tích hợp 10 mạng liên quan và dùng bộ mã hóa tự động biến thiên để chuyển các đặc trưng thuốc–bệnh về không gian tiềm ẩn có chiều thấp.

- **DDAGDL** [114]: Áp dụng học sâu hình học trên mạng thông tin không đồng nhất, kết hợp cơ chế chú ý để xử lý cấu trúc phi Euclid trong mạng y sinh học.

- **HINGRL** [70]: Tận dụng mạng thông tin không đồng nhất kết hợp học biểu diễn đồ thị và bộ phân loại Rừng ngẫu nhiên nhằm nâng cao hiệu quả định vị lại thuốc.

- **HNet-DNN** [125]: Xây dựng mạng nơ-ron sâu trên đồ thị thuốc–bệnh, tích hợp các mạng tương đồng thuốc–thuốc và bệnh–bệnh cùng đặc trưng cấu trúc.

- **DRWBNCF** [71]: Dựa trên lọc cộng tác thần kinh song tuyến tính có trọng số, kết hợp tích chập đồ thị song tuyến tính và perceptron đa lớp để dự đoán mối liên hệ thuốc–bệnh mới.

- **DRHGCN** [75]: Sử dụng mạng tích chập đồ thị và cơ chế chú ý lớp để hợp nhất thông tin từ mạng thuốc–thuốc, bệnh–bệnh và thuốc–bệnh.

- **AMDDT** [73]: Dựa trên bộ chuyển đổi đồ thị kép, tận dụng mạng nơ-ron đồ thị tiên tiến để dự đoán các mối liên kết thuốc–bệnh.

Bảng 3.10: So sánh hiệu suất với các nghiên cứu gần đây trên **Bdataset**

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
deepDR	0.804	0.820	0.883	0.233	0.601	0.299	0.369
DDAGDL	0.831	0.842	0.761	0.770	0.764	0.529	0.765
HINGRL	0.877	0.884	0.800	0.808	0.803	0.607	0.804
HNet-DNN	0.891	0.892	0.782	0.828	0.810	0.621	0.804
DRWBNCF	0.901	0.900	0.981	0.202	0.599	0.326	0.335
DRHGCN	0.910	0.909	0.867	0.771	0.826	0.658	0.816
AMDDT	0.930	0.933	0.861	0.865	0.862	0.725	0.863
Model with FNS	0.892	0.959	0.835	0.843	0.932	0.793	0.835
Our method	**0.915**	**0.966**	**0.862**	**0.851**	**0.938**	**0.817**	**0.856**

* Giá trị cao nhất được in đậm.

Trong luận án này, kiểm định chéo 5 lần được áp dụng nhất quán cho mọi thí nghiệm, bao gồm cả trường hợp kết hợp hai kỹ thuật lấy mẫu âm: HNS (High Negative Sampling) và FNS (Full Negative Sampling). Hai hàng cuối của Bảng 3.10 thể hiện so sánh trực tiếp giữa HNS và FNS.

Kết quả cho thấy chiến lược HNS vượt trội rõ rệt so với FNS trên tất cả các chỉ số. Cụ thể: AUPR tăng từ 0.892 → 0.915; AUC từ 0.959 → 0.966; PRE từ 0.835 → 0.862; REC từ 0.843 → 0.851; ACC từ 0.932 → 0.938; MCC từ 0.793 → 0.817; và F1 từ 0.835 → 0.856. Những cải thiện này khẳng định hiệu quả của HNS trong việc tận dụng các mẫu âm giàu thông tin để nâng cao chất lượng huấn luyện.

Hơn nữa, so với các phương pháp tiên tiến trước đó, mô hình của luận án đạt các chỉ số nổi bật: AUPR = 0.980, AUC = 0.983, REC = 0.940, ACC = 0.946, MCC = 0.890, F1 = 0.935. Các giá trị này vượt trội so với toàn bộ kết quả trong Bảng 3.10, cho thấy năng lực dự đoán tối ưu và độ tổng quát cao của mô hình khi áp dụng chiến lược HNS trong bài toán dự đoán tương tác thuốc–bệnh.

Phân tích tổng hợp và kết luận về DDA-BNS

Sự vượt trội của phương pháp đề xuất không chỉ thể hiện qua các con số cụ thể trong từng bảng, mà còn nằm ở cách tiếp cận toàn diện từ xử lý dữ liệu đến mô hình hóa. Đây là phương pháp đầu tiên kết hợp thành công suy luận Bayes với quy trình xử lý dữ liệu chuyên sâu cho bài toán dự đoán liên kết thuốc-bệnh, giải quyết được những hạn chế cố hữu của các phương pháp trước đó.

Các phương pháp so sánh trong các bảng đa số chỉ tập trung vào cải tiến thuật toán học máy hoặc kiến trúc mô hình, mà chưa chú trọng xử lý các vấn đề nền tảng về chất lượng dữ liệu. Trong khi đó, phương pháp đề xuất coi việc xử lý dữ liệu là bước then chốt, tạo nền tảng vững chắc để mọi mô hình sau đó có thể phát huy tối đa hiệu suất. Đây chính là điểm khác biệt cốt lõi dẫn đến kết quả vượt trội.

Lý do phương pháp đề xuất đạt kết quả cao:

- Mô hình hóa xác suất chặt chẽ: DDA-BNS sử dụng suy luận Bayes để mô hình hóa trực tiếp quan hệ thuốc-protein-bệnh thông qua xác suất có điều kiện, cung cấp khung giải thích rõ ràng và tích hợp thông tin đa nguồn một cách tự nhiên.

- Xử lý triệt để vấn đề dữ liệu: Phương pháp đề xuất giải quyết đồng thời hai thách thức lớn: âm tính giả (bằng HNS) và mất cân bằng nghiêm trọng (bằng Gaussian-SMOTE). Đây là điểm mà nhiều nghiên cứu trước chưa làm được một cách hệ thống.

- Quy trình tối ưu hóa từ dữ liệu đến mô hình: Luận án không chỉ đề xuất mô hình mới mà còn xây dựng một quy trình xử lý dữ liệu có hệ thống (lọc nhiễu → cân bằng → huấn luyện), đảm bảo đầu vào chất lượng cao cho mô hình.

3.2.6. Các nghiên cứu điển hình

Mối quan hệ giữa thuốc và bệnh được trích xuất từ DrugBank đã được xác định bằng 1.187 mẫu dương tính và được bổ sung bằng 530,687 cặp thuốc-bệnh chưa biết, được phân loại là mẫu âm tính cao. Để cân bằng tập dữ liệu cho việc huấn luyện mô hình, kỹ thuật Gaussian-SMOTE đã được sử dụng. Sau đó, mô hình được huấn luyện trên các cặp chưa biết này để dự đoán mối quan hệ thuốc-bệnh tiềm ẩn. Xác thực nghiêm ngặt các kết quả đã được tiến hành thông qua việc xem xét các tài liệu y sinh đáng tin cậy.

Bảng 3.11 trình bày 20 cặp thuốc–bệnh chưa được ghi nhận trong tập dữ liệu huấn luyện nhưng có xác suất dự đoán cao nhất theo mô hình đề xuất. Đối với các cặp dự đoán đã được xác nhận trong các nghiên cứu y sinh đáng tin cậy, luận án cung cấp liên kết tài liệu tham khảo tương ứng; những cặp chưa tìm thấy bằng chứng trong tài liệu được ký hiệu là "NA". Kết quả cho thấy trong 20 dự đoán hàng đầu có 10 cặp thuốc–bệnh đã được ghi nhận trong các nghiên cứu y sinh, cho thấy khả năng phát hiện các liên kết hợp lý của mô hình.

Ví dụ, Levobunolol đã được báo cáo có hiệu quả trong điều trị glaucoma góc mở nguyên phát (Glaucoma, Primary Open Angle – POAG; ID: 137760) và Glaucoma 1, Primary Open Angle, C (GLC1C; ID: 601682). Các nghiên cứu lâm sàng cho thấy levobunolol, một thuốc chẹn β-adrenergic dùng tại chỗ trong nhãn khoa, có hiệu quả tương đương timolol trong việc giảm áp lực nội nhãn, yếu tố quan trọng trong kiểm soát bệnh glaucoma [126].

Gliclazide, một thuốc thuộc nhóm sulfonylurea, cũng được ghi nhận có hiệu quả đối với một số dạng Maturity-Onset Diabetes of the Young (MODY), bao gồm MODY3 (ID: 600496), MODY1 (ID: 125850) và MODY2 (ID: 125851). Một nghiên cứu lâm sàng cho thấy điều trị bằng 20 mg gliclazide mỗi ngày giúp

giảm HbA1c từ 7,2% xuống 6,5% sau 3 tháng ở bệnh nhân MODY3, đồng thời các nghiên cứu khác cho thấy thuốc có thể khôi phục cơ chế ức chế glucagon do glucose kích thích trong nghiệm pháp dung nạp glucose đường uống [127, 128].

Triamcinolone acetonide (TCA) đã được nghiên cứu trong điều trị Multiple Sclerosis (MS; ID: 126200), đặc biệt ở các trường hợp MS tiến triển với triệu chứng tủy sống. Một nghiên cứu trên 31 bệnh nhân MS cho thấy việc tiêm nội tủy 40 mg TCA giúp cải thiện đáng kể Expanded Disability Status Scale (EDSS) và khoảng cách đi bộ, cho thấy tiềm năng điều trị của thuốc trong các dạng MS tiến triển [129, 130].

Prednisone/Prednisolone cũng liên quan đến Otitis Media, Susceptibility To (OMS; ID: 166760). Trong thử nghiệm lâm sàng OPAL, prednisolone đường uống được đánh giá trong điều trị viêm tai giữa cấp ở trẻ em và cho thấy khả năng giảm phản ứng viêm và cải thiện triệu chứng lâm sàng, gợi ý vai trò tiềm năng của corticosteroid này trong kiểm soát bệnh [131].

Đối với Estradiol, các nghiên cứu về liệu pháp estradiol qua da trong điều trị ung thư tuyến tiền liệt tiến triển cho thấy thuốc có thể ức chế sản xuất androgen và kiểm soát sự tiến triển của khối u, qua đó được xem là một lựa chọn trong điều trị nội tiết của Prostate Cancer, Hereditary 1 (HPC1; ID: 601518) [132].

Các ví dụ khác cũng cung cấp bằng chứng hỗ trợ cho các dự đoán của mô hình. Chẳng hạn, Hydrocortisone, một corticosteroid chống viêm, đã được báo cáo có hiệu quả trong điều trị sarcoidosis (Sarcoidosis, Susceptibility To, 1; SS1; ID: 181000) thông qua việc cải thiện các tổn thương viêm [133]. Cimetidine được chứng minh có khả năng giảm các cytokine viêm trong mô dạ dày ở mô hình nhiễm Helicobacter pylori, từ đó cải thiện tình trạng viêm dạ dày liên quan đến Helicobacter pylori infection (ID: 600263) [134]. Bên cạnh đó, Daunorubicin, một thuốc hóa trị nhóm anthracycline, đã cho thấy đáp ứng hoàn toàn ở 1 bệnh nhân và đáp ứng một phần ở 2 bệnh nhân trong nghiên cứu trên 19 bệnh nhân lymphoma tái phát hoặc kháng trị, chứng minh hoạt tính điều trị của thuốc đối với Lymphoma, Hodgkin, Classic (CHL; ID: 236000) [135].

Những kết quả này cho thấy mô hình không chỉ dự đoán được các liên kết thuốc–bệnh chưa có trong dữ liệu huấn luyện mà còn phù hợp với nhiều bằng chứng trong tài liệu y sinh hiện có. Mặc dù một số dự đoán chưa có bằng chứng trực tiếp trong các công bố hiện tại, các kết quả này mở ra những hướng nghiên cứu và thử nghiệm lâm sàng tiềm năng, góp phần rút ngắn thời gian và chi phí phát triển thuốc, đồng thời hỗ trợ các chiến lược y học cá thể hóa với độ chính xác và hiệu quả cao hơn.

Bảng 3.11: 20 cặp thuốc-bệnh có xác suất dự đoán cao nhất

Thứ hạng	**Mã số thuốc**	**Tên thuốc**	**Mã số bệnh**	**Tên bệnh**	**Xác suất dự đoán**	**Tài liệu xác thực**
1	DB01120	Gliclazide	600496	Maturity-Onset Diabetes Of The Young, Type 3; Mody3	0,982	[127], [128]
2	DB00783	Estradiol	192000	Uterine Anomalies	0,973	NA
3	DB01120	Gliclazide	125850	Maturity-Onset Diabetes Of The Young, Type 1; Mody1	0,967	[127], [128]
4	DB01120	Gliclazide	125851	Maturity-Onset Diabetes Of The Young, Type 2; Mody2	0,966	NA
5	DB01210	Levobunolol	137760	Glaucoma, Primary Open Angle; Poag	0,948	[126]
6	DB01210	Levobunolol	601682	Glaucoma 1, Primary Open Angle, C; Glc1C	0,946	[126]
7	DB00232	Methyclothiazide	600351	Enteropathy, Familial, With Villous Edema And Immunoglobulin G2 Deficiency	0,944	NA
8	DB00620	Triamcinolone	126200	Multiple Sclerosis, Susceptibility To; Ms	0,943	[129], [130]
9	DB00712	Flurbiprofen	133690	Exostoses With Anetodermia And Brachydactyly, Type E	0,941	NA

(còn tiếp)

Thứ hạng	Mã số thuốc	Tên thuốc	Mã số bệnh	Tên bệnh	Xác suất dự đoán	Tài liệu xác thực
10	DB00590	Doxazosin	157950	Permanent Molars, Secondary Retention Of	0,940	NA
11	DB01070	Dihydrotachysterol	259660	Malignant Hyperthermia, Susceptibility To, 3	0,930	NA
12	DB00635	Prednisone	166760	Otitis Media, Susceptibility To; Oms	0,934	[131]
13	DB00783	Estradiol	601518	Prostate Cancer, Hereditary, 1; Hpc1	0,926	[132]
14	DB00443	Betamethasone	188030	Immune Thrombocytopenia	0,926	NA
15	DB00741	Hydrocortisone	181000	Sarcoidosis, Susceptibility To, 1; Ss1	0,925	[133]
16	DB00481	Raloxifene	215470	Boucher-Neuhauser Syndrome; Bnhs	0,925	NA
17	DB01013	Clobetasol propionate	233810	Growth Retardation, Small And Puffy Hands And Feet, And Eczema	0,925	NA
18	DB00501	Cimetidine	600263	Helicobacter Pylori Infection, Susceptibility To	0,919	[134]
19	DB00694	Daunorubicin	236000	Lymphoma, Hodgkin, Classic; Chl	0,917	[135]
20	DB00136	Calcitriol	241519	Hypophosphatemia, Renal, with Intracerebral Calcifications	0,913	NA

3.3. Kết luận chương 3

Chương 3 của luận án tập trung vào việc ứng dụng suy luận Bayes để ước lượng xác suất của các mối quan hệ thuốc–bệnh, một phương pháp tiên tiến

nhằm khai thác sâu hơn các tương tác gián tiếp giữa thuốc và bệnh. Bằng cách kết hợp thông tin từ các quan hệ thuốc–protein và bệnh–protein, phương pháp này không chỉ giúp suy luận các tương tác tiềm năng mà còn cung cấp các xác suất quan trọng, góp phần tăng độ chính xác và độ tin cậy của mô hình dự đoán.

Chương này cũng trình bày chi tiết về việc phát triển chiến lược lựa chọn mẫu âm tính chất lượng cao (High-Confidence Negative Sampling). Đây là một giải pháp hiệu quả để loại bỏ các mẫu âm tính giả, vốn là yếu tố làm giảm chất lượng của dữ liệu huấn luyện và gây sai lệch trong mô hình phân lớp. Việc áp dụng chiến lược này đã giúp cải thiện chất lượng dữ liệu, từ đó nâng cao độ chính xác của mô hình.

Ngoài ra, luận án đã đề xuất kỹ thuật Gaussian Sampling để giải quyết vấn đề mất cân bằng dữ liệu, một vấn đề phổ biến trong các bộ dữ liệu sinh học. Kỹ thuật này giúp tái cân bằng phân bố mẫu, giảm thiểu nhiễu và góp phần tối ưu hóa hiệu suất của mô hình. Khi kết hợp với chiến lược lựa chọn mẫu âm tính chất lượng cao, phương pháp này đã chứng tỏ hiệu quả trong việc cải thiện tổng thể độ chính xác của mô hình dự đoán.

Như vậy, việc kết hợp suy luận Bayes với các kỹ thuật lấy mẫu âm tính chất lượng cao và Gaussian Sampling không chỉ nâng cao độ tin cậy của mô hình mà còn giúp tăng cường khả năng phát hiện các tương tác thuốc–bệnh trong bối cảnh dữ liệu thiếu hụt và mất cân bằng, đóng góp quan trọng vào việc cải thiện kết quả dự đoán.

Các đóng góp ở chương 3 đã được công bố trong các công trình nghiên cứu [**CT05**], [**CT06**], và [**CT07**] .

KẾT LUẬN VÀ HƯỚNG NGHIÊN CỨU

Luận án đã tập trung giải quyết bài toán dự đoán liên kết thuốc–bệnh trong bối cảnh tái định vị thuốc, một lĩnh vực đang thu hút sự quan tâm ngày càng lớn trong y sinh học tính toán. Bài toán này gặp phải nhiều thách thức, bao gồm đặc điểm dữ liệu không đồng nhất, mất cân bằng nghiêm trọng, sự tồn tại của nhiều mẫu âm tính giả và mức độ thưa thớt trong các mối quan hệ sinh học. Những vấn đề này ảnh hưởng trực tiếp đến hiệu quả của các mô hình học máy truyền thống, làm giảm độ chính xác và độ tin cậy của dự đoán.

Để khắc phục các thách thức này, luận án đã đề xuất áp dụng các phương pháp học máy tiên tiến kết hợp với siêu đường dẫn và suy luận Bayes. Các kỹ thuật này không chỉ giúp giảm tác động của dữ liệu nhiễu mà còn khai thác hiệu quả các đặc trưng của mạng thông tin không đồng nhất và cải thiện độ cân bằng dữ liệu. Những phương pháp này đặc biệt hữu ích trong việc phát hiện các liên kết thuốc–bệnh hiếm gặp hoặc ít được ghi nhận trong thực nghiệm, nơi các phương pháp truyền thống có thể gặp khó khăn trong việc phát hiện.

Đóng góp cụ thể của luận án

Luận án bao gồm hai đóng góp chính, mỗi đóng góp tương ứng với một hướng tiếp cận nhằm nâng cao hiệu quả dự đoán liên kết thuốc–bệnh, đặc biệt trong các dữ liệu phức tạp và mất cân bằng.

Đóng góp chính 1: Khai thác siêu đường dẫn trong mạng thông tin không đồng nhất

1. **Đề xuất một siêu đường dẫn mới và xây dựng mô hình tổng hợp từ bốn mô hình phân lớp cơ sở.** Luận án đã đề xuất một siêu đường dẫn (meta-path) mới, nhằm phản ánh đầy đủ hơn các mối quan hệ tiềm ẩn

giữa thuốc và bệnh. Siêu đường dẫn này được kết hợp với ba siêu đường dẫn do Wu và cộng sự phát triển, tạo thành một tập hợp bốn siêu đường dẫn. Dựa trên bốn siêu đường dẫn này, bốn mô hình phân lớp cơ sở riêng biệt đã được xây dựng. Từ đó, luận án phát triển một mô hình phân lớp tổng hợp sử dụng cơ chế biểu quyết từ bốn mô hình cơ sở này. Kết quả thực nghiệm cho thấy mô hình tổng hợp có hiệu quả dự đoán ổn định và vượt trội so với các mô hình đơn lẻ, khẳng định lợi ích của việc kết hợp đa siêu đường dẫn để tăng cường khả năng biểu diễn đặc trưng và nâng cao hiệu quả dự đoán.

2. **Đề xuất ba nhóm siêu đường dẫn dựa trên các quan hệ đồng nhất và phát triển mô hình tổng hợp thứ hai.** Luận án tiếp tục mở rộng khả năng biểu diễn mạng thông tin không đồng nhất bằng cách xây dựng ba loại quan hệ đồng nhất: thuốc–thuốc, bệnh–bệnh và protein–protein. Từ ba loại quan hệ này, ba nhóm siêu đường dẫn mới đã được thiết lập, phản ánh các dạng tương đồng và mối liên hệ sinh học giữa các thực thể. Trên mỗi nhóm siêu đường dẫn, một mô hình phân lớp cơ sở được phát triển và sau đó hợp nhất thành một mô hình phân lớp tổng hợp thứ hai, thông qua cơ chế biểu quyết từ ba mô hình cơ sở. Kết quả thực nghiệm cho thấy mô hình tổng hợp này giúp mở rộng không gian biểu diễn và nâng cao khả năng phát hiện các tương tác thuốc–bệnh tiềm năng, đặc biệt trong trường hợp các tương tác ít được ghi nhận trong dữ liệu.

Đóng góp chính 2: Ứng dụng suy luận Bayes và phát triển kỹ thuật lấy mẫu âm tính chất lượng cao

1. **Áp dụng suy luận Bayes để ước lượng xác suất của năm mối quan hệ thuốc–bệnh.** Luận án đã áp dụng phương pháp suy luận Bayes nhằm xây dựng và ước lượng xác suất tồn tại của năm loại quan hệ thuốc–bệnh. Đây là một hướng tiếp cận mới trong dự đoán liên kết thuốc–bệnh, cho phép khai thác sâu hơn thông tin từ các mối quan hệ thuốc–protein và

bệnh–protein, từ đó suy luận các tương tác gián tiếp giữa thuốc và bệnh. Kết quả suy luận này cung cấp các xác suất quan trọng, giúp đánh giá độ tin cậy của các cặp thuốc–bệnh trong mạng không đồng nhất thuốc-protein-bệnh, từ đó cải thiện độ chính xác của mô hình dự đoán.

2. **Xây dựng chiến lược lựa chọn mẫu âm tính chất lượng cao.** Dựa trên các xác suất suy luận Bayes, luận án đã phát triển chiến lược lựa chọn mẫu âm tính chất lượng cao. Chiến lược này giúp loại bỏ hiệu quả các mẫu âm tính giả, vốn là một trong những yếu tố chính gây sai lệch mô hình, từ đó cải thiện chất lượng dữ liệu huấn luyện và nâng cao độ chính xác của mô hình phân lớp.

3. **Đề xuất sử dụng Gaussian Sampling để cân bằng dữ liệu.** Để giải quyết vấn đề mất cân bằng dữ liệu trong các bộ dữ liệu sinh học, luận án đã đề xuất sử dụng Gaussian Sampling nhằm tái cân bằng phân bố mẫu. Khi kết hợp với kỹ thuật lựa chọn mẫu âm tính chất lượng cao, phương pháp này giúp tạo ra một tập dữ liệu huấn luyện cân bằng hơn, giảm thiểu nhiễu và góp phần nâng cao hiệu năng tổng thể của mô hình dự đoán.

Các kết quả của luận án không chỉ được chứng minh thông qua các chỉ số đánh giá mô hình mà còn được kiểm chứng bằng phân tích tài liệu từ các bài báo uy tín trong lĩnh vực y sinh. Điều này góp phần khẳng định tính thực tiễn và độ tin cậy của các kết quả nghiên cứu.

Ngoài ra, gần đây NCS cũng đã công bố thêm được 01 công trình trên tạp chí quốc tế [CT08] (Q3, Scopus, ESCI), 01 công trình trên hội nghị quốc tế (Scopus) và 01 đề tài nghiên cứu khoa học cấp trường [CT09] cả 3 công trình này đều nằm trong phạm vi của luận án. Tổng cộng là 10 công trình nghiên cứu.

Hướng nghiên cứu tiếp theo

Mặc dù luận án đã đạt được những kết quả khả quan trong việc dự đoán liên kết thuốc–bệnh, vẫn còn những khía cạnh tiềm năng cần được tiếp tục khai

thác và mở rộng. Trong thời gian tới, nghiên cứu sẽ tập trung vào ba hướng chính nhằm tối ưu hóa hiệu quả của mô hình:

- Thứ nhất, nghiên cứu sẽ hướng tới việc tự động hóa quy trình lựa chọn và kết hợp các siêu đường dẫn (meta-path). Thay vì xác định thủ công dựa trên tri thức chuyên gia, việc áp dụng các thuật toán học máy để tự động trích xuất các meta-path phù hợp với đặc thù của từng bộ dữ liệu cụ thể sẽ giúp mô hình trở nên linh hoạt và khách quan hơn.

- Thứ hai, nghiên cứu sẽ phát triển cơ chế dán trọng số động để đánh giá chính xác mức độ đóng góp của từng meta-path trong quá trình suy luận. Việc định lượng tầm quan trọng này không chỉ giúp tăng cường độ chính xác cho dự đoán mà còn cho phép loại bỏ các meta-path dư thừa hoặc gây nhiễu, từ đó tối ưu hóa chi phí tính toán và nâng cao tính diễn giải của mô hình.

- Cuối cùng, hướng phát triển quan trọng là mở rộng khả năng tích hợp đa dữ liệu. Bên cạnh cấu trúc mạng hiện tại, nghiên cứu sẽ khai thác thêm các dạng dữ liệu thô như văn bản (text), chuỗi gen, protein (sequence), đường chuyển hóa (pathway), cấu trúc phân tử (2D/3D). Việc kết hợp đa nguồn dữ liệu hứa hẹn sẽ cung cấp cái nhìn toàn diện hơn về cơ chế tương tác giữa thuốc và bệnh.

DANH MỤC CÔNG TRÌNH CÔNG BỐ

[**CT01**] Anh Dao, N, Le MH, Tho Dang, X. (2024). "Label Transfer for Drug Disease Association in Three Meta-Paths", Evol Bioinform Online. 2024 Sep 13;20:11769343241272414. doi: 10.1177/11769343241272414. PMID: 39279816; PMCID: PMC11401013.

[**CT02**] Tho Dang, X., Hung Le, M., Anh Dao, N. (2023). "Drug Repositioning for Drug Disease Association in Meta-paths", In: Phuong, N.H., Kreinovich, V. (eds) Deep Learning and Other Soft Computing Techniques. Studies in Computational Intelligence, vol 1097. Springer, Cham. https://doi.org/10.1007/978-3-031-29447-1_4

[**CT03**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2025). "Drug Repositioning by XGBoost for Meta-Paths in Heterogeneous Networks", In: Hoang Phuong, N., Huyen Chau, N.T., Vladik Kreinovich, (editors), Explainable AI and Other Soft Computing Techniques: Biomedical and Related Applications, Springer, (To appear in 2026) (indexed in Scopus)

[**CT04**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2025). "Drug Repositioning by LightGBM for Meta-Paths in Heterogeneous Networks", The National Conference on Fundamental and Applied IT Research, 2025.

[**CT05**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2025). "Bayes Inference for Drug Discovery by High Negative Samples and Oversampling", Bioinformatics and Biology Insights. 2025;19. doi:10.1177/11779322251328269

[**CT06**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2024). "Enhancing Drug Discovery Through A Meta-Path Based Oversampling Approach For Imbalanced Data", Journal of Science and Technique-Section on Information and Communication Technology 13.01 (2024).

[**CT07**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2024). "High Potential Negative Sampling for Drug Disease Association Prediction", In: Hoang Phuong, N., Huyen Chau, N.T., Kreinovich, V. (eds) Machine Learning and Other Soft Computing Techniques: Biomedical and Related Applications. Studies in Systems, Decision and Control, vol 543. Springer, Cham. https://doi.org/10.1007/978-3-031-63929-6_7

[**CT08**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2026). Drug repositioning by belief networks and ensemble method. Biomed Phys Eng Express. 2026 Feb 19;12(2).

doi: 10.1088/2057-1976/ae43f0. PMID: 41666480.

[**CT09**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2026). Drug Repositioning by Multilayer Perceptron with KernelPCA in Heterogeneous Networks, In: Hoang Phuong, N., Huyen Chau, N.T., Vladik Kreinovich, (editors), Explainable AI and Other Soft Computing Techniques: Biomedical and Related Applications, Springer, (To appear in 2026), (indexed in Scopus).

TÀI LIỆU THAM KHẢO

[1] Y. Hua, X. Dai, Y. Xu, et al. "Drug repositioning: Progress and challenges in drug discovery for various diseases". In: *European Journal of Medicinal Chemistry* 234 (2022), p. 114239.

[2] N. Novac. "Challenges and opportunities of drug repositioning". In: *Trends in Pharmacological Sciences* 34.5 (2013), pp. 267–272.

[3] T. T. Ashburn and K. B. Thor. "Drug repositioning: identifying and developing new uses for existing drugs". In: *Nature Reviews Drug Discovery* 3.8 (2004), pp. 673–683.

[4] T. U. Singh, S. Parida, M. C. Lingaraju, et al. "Drug repurposing approach to fight COVID-19". In: *Pharmacological Reports* 72.6 (2020), pp. 1479–1508.

[5] F. Cheng, R. J. Desai, D. E. Handy, et al. "Network-based approach to prediction and population-based validation of *in silico* drug repurposing". In: *Nature Communications* 9 (2018), p. 2691.

[6] Y. Zhou, Y. Hou, J. Shen, et al. "Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2". In: *Cell Discovery* 6 (2020), p. 14.

[7] P. Sibilio, S. Bini, G. Fiscon, et al. "*In silico* drug repurposing in COVID-19: a network-based analysis". In: *Biomedicine & Pharmacotherapy* 9 (2021). Xác minh volume/số nếu cần, p. 111954.

[8] Z. He, J. Zhang, X.-H. Shi, et al. "Predicting drug–target interaction networks based on functional groups and biological features". In: *PLoS ONE* 5.3 (2010), e9603.

[9] L. T. Kohn, J. Corrigan, M. S. Donaldson, et al. *To Err is Human: Building a Safer Health System*. Vol. 6. Washington, DC: National Academy Press, 2000.

[10] Y. Ma and Y. Ma. "Hypergraph-based logistic matrix factorization for metabolite–disease interaction prediction". In: *Bioinformatics* 38.2 (2022), pp. 435–443.

[11] Y. Ma, Y. Zhao, and Y. Ma. "Kernel Bayesian nonlinear matrix factorization based on variational inference for human–virus protein–protein interaction prediction". In: *Scientific Reports* 14.1 (2024), p. 5693.

[12] W. Zhang, H. Xu, X. Li, et al. "DRIMC: an improved drug repositioning approach using Bayesian inductive matrix completion". In: *Bioinformatics* 36.9 (2020), pp. 2839–2847.

[13] M. Yang, H. Luo, Y. Li, et al. "Overlap matrix completion for predicting drug-associated indications". In: *PLoS Computational Biology* 15.12 (2019), e1007541.

[14] F. Wang, X. Lei, B. Liao, et al. "Predicting drug-drug interactions by graph convolutional network with multi-kernel". In: *Briefings in Bioinformatics* 23.1 (2022), bbab511.

[15] H. Wang, C. Dai, Y. Wen, et al. "GADRP: graph convolutional networks and autoencoders for cancer drug response prediction". In: *Briefings in Bioinformatics* 24.1 (2023), bbac501.

[16] X. Liu, C. Song, F. Huang, et al. "GraphCDR: a graph neural network method with contrastive learning for cancer drug response prediction". In: *Briefings in Bioinformatics* 23.1 (2022), bbab457.

[17] Y. Meng, Y. Wang, J. Xu, et al. "Drug repositioning based on weighted local information augmented graph neural network". In: *Briefings in Bioinformatics* 25.1 (2023), bbad431.

[18] Y. Gu, S. Zheng, Q. Yin, et al. "REDDA: Integrating multiple biological relations to heterogeneous graph neural network for drug-disease association prediction". In: *Computers in Biology and Medicine* 150 (2022), p. 106127.

[19] G. Li, P. Bai, C. Liang, et al. "Node-adaptive graph Transformer with structural encoding for accurate and robust lncRNA-disease association prediction". In: *BMC Genomics* 25.1 (2024), p. 73.

[20] J. Liu, S. Guan, Q. Zou, et al. "AMDGT: Attention aware multi-modal fusion using a dual graph transformer for drug–disease associations prediction". In: *Knowledge-Based Systems* 284 (2024), p. 111329.

[21] Z. Huang, Z. Xiao, C. Ao, et al. "Computational approaches for predicting drug-disease associations: a comprehensive review". In: *Frontiers in Computer Science* 19.5 (2025), p. 195909.

[22] J. Li, S. Zheng, B. Chen, et al. "A survey of current trends in computational drug repositioning". In: *Briefings in Bioinformatics* 17.1 (2016), pp. 2–12.

[23] H. Askr, E. Elgeldawi, H. Aboul Ella, et al. "Deep learning in drug discovery: an integrative review and future challenges". In: *Artificial Intelligence Review* 56.7 (2023), pp. 5975–6037.

[24] Ke-Jia Chen et al. "On Link Formation in Heterogeneous Information Networks: A View Based on Multi-Label Learning". In: *Association for Computing Machinery* 20.4 (2017), pp. 50–53.

[25] Chuan Shi et al. "A Survey of Heterogeneous Information Network Analysis". In: *CoRR* abs/1511.04854.5 (2015), pp. 1–45.

[26] G. Wu, J. Liu, and X. Yue. "Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition". In: *BMC Bioinformatics* 20.3 (2019), p. 134.

[27] F. Huang, Y. Qiu, Q. Li, et al. “Predicting Drug-Disease Associations via Multi-Task Learning Based on Collective Matrix Factorization”. In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), p. 218.

[28] Z. Gao, H. Ma, X. Zhang, et al. “Similarity measures-based graph co-contrastive learning for drug-disease association prediction”. In: *Bioinformatics* 39.6 (2023), btad357.

[29] J. Liu, Z. Zuo, and G. Wu. “Link Prediction Only With Interaction Data and its Application on Drug Repositioning”. In: *IEEE Transactions on Nanobioscience* 19.3 (2020), pp. 547–555.

[30] D. Barber. *Bayesian Reasoning and Machine Learning*. `https://www.cambridge.org/highereducation/books/bayesian-reasoning-and-machine-learning/37DAFA214EEE41064543384033D2ECF0`, accessed: 2025-07-21. Cambridge University Press, 2012.

[31] M. W. Gonzalez and M. G. Kann. “Protein interactions and disease”. In: *PLoS Computational Biology* 8.12 (2012), e1002819.

[32] M. Bagherian, E. Sabeti, K. Wang, et al. “Machine learning approaches and databases for prediction of drug-target interaction: a survey paper”. In: *Briefings in Bioinformatics* 22.1 (2021), pp. 247–269.

[33] Y. Kim, Y.-S. Jung, J.-H. Park, et al. “Drug-Disease Association Prediction Using Heterogeneous Networks for Computational Drug Repositioning”. In: *Biomolecules* 12.10 (2022), p. 1497.

[34] L. Cai, J. Chu, J. Xu, et al. “Machine learning for drug repositioning: Recent advances and challenges”. In: *Current Research in Chemical Biology* 3 (2023), p. 100042.

[35] Ran Zhang et al. “MHTAN-DTI: Metapath-based hierarchical transformer and attention network for drug–target interaction prediction”. In: *Briefings in Bioinformatics* 24.2 (2023), bbad079.

[36] H. Chen, H. Zhang, Z. Zhang, et al. “Network-based inference methods for drug repositioning”. In: *Computational and Mathematical Methods in Medicine* 2015 (2015), p. 130620.

[37] T. Zhou, Z. Kuscsik, J. G. Liu, et al. “Solving the apparent diversity–accuracy dilemma of recommender systems”. In: *Proceedings of the National Academy of Sciences of the USA* 107 (2010), pp. 4511–4515.

[38] L. Qi, L. Yu, et al. “A heterogeneous network embedding framework for predicting similarity-based drug–target interactions”. In: *Briefings in Bioinformatics* 22.6 (2021), bbab275.

[39] Wei Wang, Yongqing Wang, et al. “PPDTS: predicting potential drug–target interactions based on network similarity”. In: *IET Systems Biology* 16.1 (2022), pp. 18–27.

[40] M. Yang et al. “Overlap matrix completion for predicting drug-associated indications”. In: *PLoS Computational Biology* 15.7 (2019), e1007541.

[41] Y. Yan, M. Yang, H. Zhao, et al. “Drug repositioning based on multi-view learning with matrix completion”. In: *Briefings in Bioinformatics* (2022). Bổ sung volume/issue/pages nếu cần.

[42] W. Zhang et al. “DRIMC: An improved drug repositioning approach using Bayesian inductive matrix completion”. In: *Bioinformatics* 36 (2020), pp. 2839–2847.

[43] Wei Zhang et al. “Predicting drug-disease associations by using similarity constrained matrix factorization”. In: *BMC Bioinformatics* 19.1 (2018), p. 233. DOI: 10.1186/s12859-018-2220-4. URL: https://doi.org/10.1186/s12859-018-2220-4.

[44] T. Ban, M. Ohue, and Y. Akiyama. “NRLMFβ: beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction”. In: *Biochemical and Biophysical Reports* 18 (2019), p. 100615.

[45] X. Wang and R. Yan. “DDAPRED: a computational method for predicting drug repositioning using regularized logistic matrix factorization”. In: *Journal of Molecular Modeling* 26 (2020), p. 60.

[46] M. Yang et al. “Computational drug repositioning based on multi-similarities bilinear matrix factorization”. In: *Briefings in Bioinformatics* 22 (2021), bbaa267.

[47] Liang-Yong Xia et al. “Improved prediction of drug–target interactions using self-paced learning with collaborative matrix factorization”. In: *Journal of Chemical Information and Modeling* 59.7 (2019), pp. 3340–3351.

[48] Z.-C. Zhang, X.-F. Zhang, M. Wu, et al. “A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks”. In: *Bioinformatics* 36 (2020). Bổ sung pages/issue nếu cần.

[49] M. Lian, W. Du, X. Wang, et al. “Drug–target interaction prediction based on multi-similarity fusion and sparse dual-graph regularized matrix factorization”. In: *IEEE Access* 9 (2021), pp. 99718–99730.

[50] Aizhen Wang and Minhui Wang. “Drug–target interaction prediction via dual Laplacian graph regularized logistic matrix factorization”. In: *BioMed Research International* 2021 (2021), p. 5599263.

[51] C. Wu et al. “Computational drug repositioning through heterogeneous network clustering”. In: *BMC Systems Biology* 7 (2013), S6.

[52] M. Kanehisa et al. “KEGG: New perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Research* 45 (2017), pp. D353–D361.

[53] T. Nepusz, H. Yu, and A. Paccanaro. “Detecting overlapping protein complexes in protein–protein interaction networks”. In: *Nature Methods* 9 (2012), pp. 471–472.

[54] F. Wang, Y. Ding, X. Lei, et al. "Identifying gene signatures for cancer drug repositioning based on sample clustering". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2020). doi:10.1109/TCBB.2020.3019781.

[55] Yansen Su, Zhiyang Hu, Fei Wang, et al. "AMGDTI: drug–target interaction prediction based on adaptive meta-graph learning in heterogeneous network". In: *Briefings in Bioinformatics* 25.1 (2024), bbad474.

[56] M. Li et al. "Metapath-aggregated heterogeneous graph neural network for drug–target interaction prediction". In: *Briefings in Bioinformatics* 24.1 (2023), bbac578.

[57] Shanyang Ding et al. "MAPTrans: mutual attention transformer with dynamic meta-path pruning for drug repositioning". In: *Briefings in Bioinformatics* 26.4 (2025), bbaf382. DOI: 10.1093/bib/bbaf382.

[58] Y. Zhou et al. "Drug repositioning with metapath guidance and adaptive negative sampling enhancement". In: *Journal of Biomedical Informatics* 171 (2025), p. 104916.

[59] Y. Huang et al. "A Novel Drug Repositioning Method Using Meta-Path Aggregating via Hierarchical Attention Mechanism". In: *IEEE Transactions on Computational Biology and Bioinformatics* 23.1 (2026), pp. 259–270.

[60] H. Luo et al. "Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm". In: *Bioinformatics* 32 (2016), pp. 2664–2671.

[61] H. Liu et al. "Inferring new indications for approved drugs via random walk on drug–disease heterogeneous networks". In: *BMC Bioinformatics* 17 (2016), p. 539.

[62] X. Chen, M.-X. Liu, and G.-Y. Yan. "Drug–target interaction prediction by random walk on the heterogeneous network". In: *Molecular BioSystems* 8 (2012), pp. 1970–1978.

[63] Y. Wang et al. “Drug repositioning based on individual bi-random walks on a heterogeneous network”. In: *BMC Bioinformatics* 20 (2019), p. 547.

[64] B. M. Momanyi et al. “RWRGDR: Random Walk and GraphSAGE-based Framework for Enhanced Drug Repositioning”. In: *Current Drug Targets* (2026).

[65] L. Pio-Lopez, A. Valdeolivas, L. Tichit, et al. “MultiVERSE: a multiplex and multiplex-heterogeneous network embedding approach”. In: *Scientific Reports* 11 (2021), p. 8794.

[66] W. Wang et al. “Drug repositioning by integrating target information through a heterogeneous network model”. In: *Bioinformatics* 30 (2014), pp. 2923–2930.

[67] V. Martinez et al. “DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data”. In: *Artificial Intelligence in Medicine* 63 (2015), pp. 41–49.

[68] V. Martinez, C. Cano, and A. Blanco. “ProphNet: A generic prioritization method through propagation of information”. In: *BMC Bioinformatics* 15 (2014), S5.

[69] Y. Meng et al. “Drug repositioning based on weighted local information augmented graph neural network”. In: *Briefings in Bioinformatics* 25.1 (2023), bbad431.

[70] B.-W. Zhao, L. Hu, Z.-H. You, et al. “HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks”. In: *Brief Bioinform* 23.1 (2022), bbab515.

[71] Y. Meng, C. Lu, M. Jin, et al. “A weighted bilinear neural collaborative filtering approach for drug repositioning”. In: *Brief Bioinform* 23.2 (2022), bbab581.

[72] G. Xie et al. “BGMSDDA: A bipartite graph diffusion algorithm with multiple similarity integration for drug–disease association prediction”. In: *Molecular Omics* 17 (2021), pp. 997–1011.

[73] J. Liu et al. “AMDGT: attention-aware multi-modal fusion using a dual graph transformer for drug-disease associations prediction”. In: *Briefings in Bioinformatics* 23 (2022), bbab515. DOI: 10.1016/j.knosys.2023.111329.

[74] Yunan Luo et al. “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information”. In: *Nature Communications* 8.1 (2017), p. 573. DOI: 10.1038/s41467-017-00680-8. URL: https://doi.org/10.1038/s41467-017-00680-8.

[75] L. Cai, C. Lu, J. Xu, et al. “Drug repositioning based on the heterogeneous information fusion graph convolutional network”. In: *Brief Bioinform* 22.6 (2021), bbab319.

[76] Z. Yu, F. Huang, X. Zhao, et al. “Predicting drug–disease associations through layer attention graph convolutional network”. In: *Briefings in Bioinformatics* 22 (2021), bbaa243.

[77] L. Cai, C. Lu, J. Xu, et al. “Drug repositioning based on the heterogeneous information fusion graph convolutional network”. In: *Briefings in Bioinformatics* 22 (2021). Bổ sung pages/issue nếu cần.

[78] Z. Yu et al. “Predicting drug–disease associations through layer attention graph convolutional network”. In: *Briefings in Bioinformatics* 22 (2021), bbaa243.

[79] M. Gönen. “Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization”. In: *Bioinformatics* 28.18 (2012), pp. 2304–2310.

[80] Mehmet Gönen and Samuel Kaski. "Kernelized Bayesian Matrix Factorization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.10 (2014), pp. 2047–2060.

[81] Y. Xiao, J. Zhang, and L. Deng. "Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks". In: *Scientific Reports* 7.1 (2017), p. 3664.

[82] Z. Tian, Z. Teng, S. Cheng, et al. "Computational drug repositioning using meta-path-based semantic network analysis". In: *BMC Systems Biology* 12.9 (2018), p. 134.

[83] D. Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of Chemical Information and Computer Sciences* 28.1 (1988), pp. 31–36.

[84] N. M. O'Boyle, M. Banck, C. A. James, et al. "Open Babel: An open chemical toolbox". In: *Journal of Cheminformatics* 3.1 (2011), p. 33.

[85] A. Gottlieb, G. Y. Stein, E. Ruppin, et al. "PREDICT: a method for inferring novel drug indications with application to personalized medicine". In: *Molecular Systems Biology* 7 (2011), p. 496.

[86] W. Wang, S. Yang, X. Zhang, et al. "Drug repositioning by integrating target information through a heterogeneous network model". In: *Bioinformatics* 30.20 (2014), pp. 2923–2930.

[87] X. Liang, P. Zhang, L. Yan, et al. "LRSSL: predict and interpret drug-disease associations based on data integration using sparse subspace learning". In: *Bioinformatics* 33.8 (2017), pp. 1187–1196.

[88] W. Zhang, X. Yue, W. Lin, et al. "Predicting drug-disease associations by using similarity constrained matrix factorization". In: *BMC Bioinformatics* 19.1 (2018), p. 233.

[89] H. Luo, J. Wang, M. Li, et al. "Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm". In: *Bioinformatics* 32.17 (2016), pp. 2664–2671.

[90] S. Derry et al. "Nortriptyline for neuropathic pain in adults". In: *The Cochrane Database of Systematic Reviews* 1.1 (2015), p. CD011209.

[91] S. Paterna, G. Parrinello, P. Di Pasquale, et al. "Medium-term effects of bisoprolol administration on renal hemodynamics and function in mild to moderate essential hypertension". In: *Advances in Therapy* 24.6 (2007), pp. 1260–1270.

[92] J. McAinsh et al. "Atenolol kinetics in renal failure". In: *Clinical Pharmacology and Therapeutics* 28.3 (1980), pp. 302–309.

[93] O. Müller and H. R. Knobel. "[Effectiveness and tolerance of metipranolol–results of a multi-center long-term study in Switzerland]". In: *Klinische Monatsblätter für Augenheilkunde* 188.1 (1986), pp. 62–63.

[94] P. J. Spring et al. "Autosomal dominant hereditary sensory neuropathy with chronic cough and gastro-oesophageal reflux: clinical features in two families linked to chromosome 3p22-p24". In: *Brain: A Journal of Neurology* 128.12 (2005), pp. 2797–2810.

[95] D. H. Levine et al. "Renal failure and other serious sequelae of epinephrine toxicity in neonates". In: *Southern Medical Journal* 78.7 (1985), pp. 874–877.

[96] M. Pakfetrat et al. "Ergotamine-induced acute tubulo-interstitial nephritis". In: *Saudi Journal of Kidney Diseases and Transplantation: An Official Publication of the Saudi Center for Organ Transplantation* 24.5 (2013), pp. 981–983.

[97] R. Luciani et al. "Acute renal failure due to amiodarone-induced hypothyroidism". In: *Clinical Nephrology* 72.1 (2009), pp. 79–80.

[98] S. Veazie. *Fludrocortisone for orthostatic hypotension*. `https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012868.pub2/full`. accessed: 2024-03-27. 2021.

[99] X. T. Dang, D. H. Tran, O. Hirose, et al. "SPY: A Novel Resampling Method for Improving Classification Performance in Imbalanced Data". In: *Seventh International Conference on Knowledge and Systems Engineering (KSE)*. 2015, pp. 280–285.

[100] I. Mani and I. Zhang. "kNN approach to unbalanced data distributions: a case study involving information extraction". In: *Proceedings of Workshop on Learning from Imbalanced Datasets, ICML United States*. 2003, pp. 1–7.

[101] "Two Modifications of CNN". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6.11 (1976), pp. 769–772.

[102] M. Kubat and S. Matwin. "Addressing the Curse of Imbalanced Training Sets One-Sided Selection". In: *International Conference on Machine Learning*. Vol. 14. 1997, pp. 179–186.

[103] J. Laurikkala. "Improving Identification of Difficult Small Classes by Balancing Class Distribution". In: *Artificial Intelligence in Medicine, Berlin, Heidelberg, Springer*. 2001, pp. 63–66.

[104] N. V. Chawla, K. W. Bowyer, L. O. Hall, et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[105] H. Han, W.-Y. Wang, and B.-H. Mao. "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning". In: *Advances in Intelligent Computing, Springer Berlin Heidelberg*. 2005, pp. 878–887.

[106] L. Ma and S. Fan. "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests". In: *BMC Bioinformatics* 18.1 (2017), p. 169.

[107] G. E. Batista, A. L. Bazzan, and M. C. Monard. "Balancing training data for automated annotation of keywords: a case study". In: *WOB* 3 (2003), pp. 10–18.

[108] J. Yun, J. Ha, and J. S. Lee. "Automatic determination of neighborhood size in SMOTE". In: *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. Da Nang, Vietnam, 2016. DOI: 10.1145/2857546.2857648.

[109] F. R. Torres, J. A. Carrasco-Ochoa, and J. F. Martinez-Trinidad. "SMOTE-D: a deterministic version of SMOTE". In: *Pattern Recognition*. Ed. by J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and V. Ayala-Ramirez. Vol. 9703. Lecture Notes in Computer Science. Springer, 2016, pp. 134–143. DOI: 10.1007/978-3-319-39393-3_18.

[110] Y. Dong and X. A. Wang. "New over-sampling approach: randomSMOTE for learning from imbalanced data sets". In: *Knowledge Science, Engineering and Management*. Ed. by L. Wang. Vol. 7091. Lecture Notes in Computer Science. Springer, 2011, pp. 471–482. DOI: 10.1007/978-3-642-25975-3_30.

[111] G. Douzas, F. Bacao, and F. Last. "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE". In: *Information Sciences* 467 (2018), pp. 1–20. DOI: 10.1016/j.ins.2018.07.040.

[112] H. Lee, J. Kim, and S. Kim. "Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions". In: *International Journal of Fuzzy Logic and Intelligent Systems* 17.4 (2017), pp. 229–234.

[113] F. Saglam and M. A. Cengiz. "A novel SMOTE-based resampling technique through noise detection and the boosting procedure". In: *Expert Systems with Applications* 192 (2022), p. 117023. DOI: 10.1016/j.eswa.2022.117023.

[114] B.-W. Zhao, X.-R. Su, P.-W. Hu, et al. "A geometric deep learning framework for drug repositioning over heterogeneous information networks". In: *Brief Bioinform* 23.6 (2022), bbac384.

[115] S. Korkmaz. "Deep learning-based imbalanced data classification for drug discovery". In: *Journal of Chemical Information and Modeling* 60.9 (2020), pp. 4180–4190. DOI: 10.1021/acs.jcim.9b01162.

[116] S. S. Seyedeh and R. K. Mohammad. "RUSDR: Class imbalance-aware ensemble learning for drug repurposing". In: *Proceedings of the 10th International Conference on Information and Knowledge Technology (IKT)*. 2019.

[117] MB Seegars et al. "A Pilot Phase II Study of the Feasibility and Efficacy of Vincristine Sulfate Liposome Injection in Patients With Relapsed or Refractory Acute Myeloid Leukemia". In: *J Hematol* 10.1 (2021). Epub 2021 Feb 6. PMID: 33643502; PMCID: PMC7891907, pp. 1–7. DOI: 10.14740/jh771.

[118] J Palle et al. "Doxorubicin pharmacokinetics is correlated to the effect of induction therapy in children with acute myeloid leukemia". In: *Anticancer Drugs* 17.4 (2006). PMID: 16549995, pp. 385–92. DOI: 10.1097/01.cad.0000198911.98442.16.

[119] WL Hwang et al. "DAE (daunorubicin, Ara-C, and etoposide) and intermediate dose Ara-C for remission induction and consolidation treatment of adult patients with acute myeloid leukemia". In: *Am J Clin Oncol* 15.6 (1992). PMID: 1449119, pp. 531–4. DOI: 10.1097/00000421-199212000-00014.

[120] X Thomas and E Archimbaud. "Mitoxantrone in the treatment of acute myelogenous leukemia: a review". In: *Hematol Cell Ther* 39.4 (1997). PMID: 9352324, pp. 63–74. DOI: 10.1007/s00282-997-0163-8.

[121] G Lambertenghi Deliliers et al. "Idarubicin in the therapy of acute myeloid leukemia: final analysis in 57 previously untreated patients". In: *Semin Oncol* 20.6 Suppl 8 (1993). PMID: 8290969, pp. 27–33.

[122] M. D. Mody et al. "Complete remission of acute myeloid leukemia following cisplatin based concurrent therapy with radiation for squamous

cell laryngeal cancer". In: *Case Reports in Hematology* 2016 (2016), p. 8581421. DOI: 10.1155/2016/8581421.

[123] A. M. Carella et al. "Treatment of "poor risk" Acute Myeloid Leukemia with Fludarabine, Cytarabine and G-CSF (FLAG regimen): A single center study". In: *Leukemia & Lymphoma* 40.3-4 (2001), pp. 295–303. DOI: 10.3109/10428190109057928.

[124] X. Zeng, S. Zhu, X. Liu, et al. "deepDR: a network-based deep learning approach to in silico drug repositioning". In: *Bioinformatics* 35.24 (2019), pp. 5191–5198.

[125] H. Liu, W. Zhang, Y. Song, et al. "HNet-DNN: Inferring New Drug-Disease Associations with Deep Neural Network Based on Heterogeneous Network Features". In: *Journal of Chemical Information and Modeling* 60.4 (2020), pp. 2367–2376.

[126] S. J. Sorensen and S. R. Abel. "Comparison of the ocular beta-blockers". In: *Annals of Pharmacotherapy* 30 (1996), pp. 43–54. DOI: 10.1177/106002809603000109.

[127] A. M. Habeb et al. "Response to oral gliclazide in a pre-pubertal child with hepatic nuclear factor-1 alpha maturity onset diabetes of the young". In: *Annals of Saudi Medicine* 31 (2011), pp. 190–193. DOI: 10.4103/0256-4947.75590.

[128] I. I. Spiliotis et al. "Gliclazide restores appropriate glucagon suppression during OGTT in MODY3 & MODY1 patients: results of the "Glucagon in MODY" study". In: *Journal of the Endocrine Society* 7 (2023), bvad1141045. DOI: 10.1210/jendso/bvad114.1045.

[129] C. Lukas, B. Bellenberg, H. K. Hahn, et al. "Benefit of repetitive intrathecal triamcinolone acetonide therapy in predominantly spinal multiple sclerosis: prediction by upper spinal cord atrophy". In: *Therapeutic Advances in Neurological Disorders* 2 (2009), pp. 42–49. DOI: 10.1177/1756285609343480.

[130] M. Abu-Mugheisib, R. Benecke, and U. K. Zettl. "Repeated intrathecal triamcinolone acetonide administration in progressive multiple sclerosis: a review". In: *Multiple Sclerosis International* 2011 (2011), p. 219049. DOI: 10.1155/2011/219049.

[131] R. W. Ranakusuma, A. R. McCullough, E. D. Safitri, et al. "Oral prednisolone for acute otitis media in children: a pilot, pragmatic, randomised, open-label, controlled study (OPAL study)". In: *Pilot and Feasibility Studies* 6 (2020), p. 121. DOI: 10.1186/s40814-020-00671-5.

[132] J. L. Ockrim et al. "Transdermal estradiol therapy for advanced prostate cancer–forward to the past?" In: *Journal of Urology* 169 (2003), pp. 1735–1737. DOI: 10.1097/01.ju.0000061024.75334.40.

[133] R. D. Sullivan, R. L. Mayock, and R. J. Jones. "Local injection of hydrocortisone and cortisone into skin lesions of sarcoidosis". In: *JAMA* 152 (1953), pp. 308–312. DOI: 10.1001/jama.1953.03690040012005.

[134] K. Higuchi, T. Tanigawa, M. Hamaguchi, et al. "Comparison of the effects of rebamipide with those of cimetidine on chronic gastritis associated with Helicobacter pylori in Mongolian gerbils". In: *Alimentary Pharmacology & Therapeutics* 18 (2003), pp. 1–7. DOI: 10.1046/j.1365-2036.2003.018s101.x.

[135] D. S. Richardson et al. "Early evaluation of liposomal daunorubicin (DaunoXome, Nexstar) in the treatment of relapsed and refractory lymphoma". In: *Investigational New Drugs* 15 (1997), pp. 247–253. DOI: 10.1023/a:1005879219554.